Supervised Machine Learning Models Evaluation to Forecast Preventive Maintenance Costs in a Thermal Power Plant

by Javier Rodriguez

B.S. in Electronic Engineering, Feb 2007, University of Santo Tomas
M.B.A in Business Administration, Feb 2011, Temple University

A Praxis submitted to

The Faculty of
The School of Engineering and Applied Science
of The George Washington University
in partial fulfillment of the requirements
for the degree of Doctor of Engineering

May 18, 2025

Praxis directed by

Rebecca M. Yassan
Professorial Lecturer in Engineering and Applied Science

The School of Engineering and Applied Science of The George Washington University certifies that Javier Rodriguez has passed the Final Examination for the degree of Doctor of Engineering as of May 18, 2025. This is the final and approved form of the Praxis.

Supervised Machine Learning Models Evaluation to Forecast Preventive Maintenance Costs in a Thermal Power Plant

Javier Rodriguez

Praxis Research Committee:

Rebecca M. Yassan, Professorial Lecturer in Engineering and Applied Science, Praxis Director

Joseph P. Blackford, Associate Professor of Engineering and Applied Science, Committee Chair

Cory Fagan, Professorial Lecturer of Engineering and Applied Science, Committee Member

© Copyright 2025 by Javier Rodriguez
All rights reserved

Dedication

This work is dedicated to my wife, Jeaneth, whose unwavering support and patience have been my greatest strength, especially during the times when my commitment to my work and this project meant less time with our family. Your love and encouragement have been the bedrock of my journey.

To my mother, Carmen, for her constant spiritual support and belief in me, reminding me that faith and perseverance can carry us through any challenge.

To my children, Sofia and Neithan, who are my inspiration to strive for better, to show that dedication and hard work can turn dreams into reality. You both are the light that guides me forward.

To my sisters and brother, who have always been by my side, and to my nieces and nephews, may this work serve as a reminder that any goal is achievable with effort and determination.

I also dedicate this to my mentors and colleagues, who have been instrumental in shaping the professional I am today. Each of you, through your guidance and high expectations, have challenged me to push beyond my limits. Your belief in my potential has been a driving force throughout this journey.

Finally, to the teams who have worked under my guidance, I hope this work inspires you to pursue your own objectives with tenacity and passion.

Acknowledgements

I am deeply grateful to the individuals and institutions that provided support throughout this journey, making this research possible.

First and foremost, I wish to express my heartfelt thanks to David Walton and Jesus Garcia for their unwavering emotional support and supply of essential resources. Their trust in my capabilities and their encouragement were particularly important during the process and helped me to stay focused and motivated.

I am especially indebted to Rebecca M. Yassan for her academic guidance, which played a key role in structuring my research effectively and ensuring that each milestone was met on time. Her insights and expertise have been invaluable throughout this process. Additionally, Kamen Damov was instrumental in the development and testing of machine learning algorithms, sharing his deep knowledge and providing essential support during the coding phases of this work.

I would like to extend my gratitude to George Washington University for their support and for believing in my potential. The university's commitment to excellence, as well as their organizational structure and professionalism throughout the D. Eng. program, has greatly enriched my academic experience.

I am also grateful to the D17 cohort of D. Eng. students, who have become a community of support. Our shared exchanges of professional information, contacts, experience, and words of encouragement were a source of strength throughout this journey.

To everyone who contributed to the completion of this research, both directly and indirectly, I offer my sincere thanks.

Abstract of Praxis

Supervised Machine Learning Models Evaluation to Forecast Preventive Maintenance Costs in a Thermal Power Plant

This praxis explores the application of supervised machine learning models to predict maintenance costs in a thermal power plant. It addresses the challenges faced by plant managers in achieving accurate cost predictions, mainly because traditional methods are typically limited in adaptability, and non effective for short-term forecasting, which leads to frequent budget overruns. Historical data about plant performance and maintenance activities is used to evaluate the predictive accuracy of four machine learning techniques: Time Series (ARIMA, SARIMA), Linear Regression (ElasticNet), Decision Trees (XGBoost), and Long Short-Term Memory (LSTM) networks through mean absolute percentage error (MAPE), and other performance metrics such as mean absolute error (MAE), and root mean squared error (RMSE) using a holdout data set. Results are aligned with ASTM E2516-11 cost estimate accuracy standards. According to the results, XGBoost achieved robust performance with a MAPE up to 20% followed by SARIMA at 28%. The most significant contributors to short-term (12-month) cost prediction are time-based factors (e.g., monthly cyclic trends and historical cost rolling statistics, which account for 90 to 96% of the weight). This study contributes to the maintenance management practices in the energy sector, offering a data-driven framework adaptable to similar industrial contexts. Future research may focus on model refinement using hybrid ensemble models, integration with other predictor factors, and testing cross-plant applicability and industry-wide adoption in asset-intensive plants.

Table of Contents

Dedication ..... **iv**

Acknowledgements ..... **v**

Abstract of Praxis ..... **vi**

List of Figures ..... **xii**

List of Tables ..... **xvi**

List of Symbols ..... **xviii**

List of Acronyms ..... **xix**

Chapter 1—Introduction **1**

1.1 Background 1

1.2 Research Motivation 2

1.3 Problem Statement 3

1.4 Thesis Statement 3

1.5 Research Objectives 4

1.6 Research Questions and Hypotheses 4

1.7 Scope of Research 5

1.8 Research Limitations 6

1.9 Organization of Praxis 7

Chapter 2—Literature Review **8**

2.1 Introduction 8

2.2 Thermal Power Plants Outlook 9

2.3 Factors Changing Operational Regimes on Thermal Power Plants 10
2.4 Power Plant Maintenance Costs Structures 12
2.5 Budget Estimation Methodologies 14
2.5.1 Deterministic Estimation Methods 14
2.5.2 Stochastic Estimation Methods 15
2.5.3 Mixed Estimation Methods 15
2.6 Annual Budget Updating 16
2.7 Maintenance Budget Line of Sight 16
2.8 Maintenance Budget Overruns 17
2.9 Cost Estimation Classification 18
2.10 Comparison of Traditional vs. Machine Learning Approaches 19
2.11 Machine Learning Prediction Techniques 20
2.11.1 Time Series 20
2.11.2 Linear Regression (Elastic Net) 21
2.11.3 Decision Trees 22
2.11.4 LSTM 23
2.12 Summary and Conclusion 24
Chapter 3—Methodology 26
3.1 Introduction 26
3.2. Praxis Maintenance Cost Estimation Method 26
3.3. Preventive Maintenance Cost Allocation 27
3.4 Data Collection 28
3.4.1 Cost Transactions and Work Orders – Raw Data 29

3.4.2 Reference Data Tables – Maintenance Type 31
3.4.3 Support Data Tables 32
3.4.4 Features (Predictors) Data - Plant Performance 32
3.5 Data Pre-Processing 33
3.5.1 Software Resources 33
3.5.2 Data Cleaning 34
3.5.3 Data Filtering 34
3.5.4 Missing Values 35
3.5.5 Data Assembly and Merging 35
3.6. Data Exploratory Analysis 36
3.7 Developing Models 38
3.7.1 Data Split Sets 40
3.7.2 Single Target Regression Models 40
3.7.3 Time Series Models (ARIMA and SARIMA) 40
3.7.4 Linear Regression Model (ElasticNet) 42
3.7.5 XGBoost Regression Model 43
3.7.6 LSTM Regression Model 44
3.8 Models Evaluation and Ranking 46
Chapter 4—Results 49
4.1 Introduction 49
4.2 Data Pre-Processing Results and Observations 50
4.2.1 Output Target Data 50
4.2.2 Features (Predictors) Data 52

4.3 Data Exploratory Analysis 53
4.3.1 Output Target Data Exploratory Results 53
4.3.2 Features (Predictors) Data Exploratory Analysis 55
4.3.3 Understanding Time Series 58
4.4 Analysis of Preventive Maintenance Costs Models 63
4.4.1 Hypothesis 1 and Results 64
4.4.2 Hypothesis 2 and Results 68
4.4.3 Hypothesis 3 and Results 72
4.4.4 Hypothesis 4 and Results 75
4.4.5 Hypothesis 5 and Results 77
4.5 Models Results Comparison 79
4.6 Selection of the Best Performance Model 82
Chapter 5—Discussion and Conclusions 84
5.1 Introduction 84
5.2 Discussion 84
5.2.1 Discussion of Hypothesis 1 84
5.2.2 Discussion of Hypothesis 2 85
5.2.3 Discussion of Hypothesis 3 87
5.2.4 Discussion of Hypothesis 4 88
5.2.5 Discussion of Hypothesis 5 90
5.3 Conclusions 91
5.4. Lessons Learned 93
5.5 Contributions to Body of Knowledge 94

5.6 Recommendations for Future Research 95

References 97

Appendix A 108

Appendix B 110

List of Figures

Figure 2-1. Hourly U.S. Electricity Generation and Load by Fuel for Selected Cases and Representative Years (billion kWh). Source: (EIA, 2023). Note: In the public domain 11

Figure 2-2. Life Cycle EPOF vs Age for Conventional Plants Operating in Cycling Regime and Major Life Periods. Adapted from Electric Power Research Institute (EPRI), 2013. 17

Figure 3-1. Praxis Estimation Method Workflow. 27

Figure 3-2. Power Plant Maintenance Type Category Cost Structure. 28

Figure 3-3. Output Target Data Flow Chart. 29

Figure 3-4. Preventive Maintenance Costs (Output Target) Data Workflow Map.. 30

Figure 3-5. Features Data Flow Chart. 33

Figure 3-6. Box Plot Example. 36

Figure 3-7. Sample Distribution Analysis Charts. 37

Figure 3-8. Sample Correlation Matrix. 37

Figure 3-9. Sample ADF and ACF Analysis. 38

Figure 3-10. Datasets distribution 40

Figure 3-11. Time Series Model Flow Chart. 41

Figure 3-12. Linear Regression Model Flow Chart. 42

Figure 3-13. XGBoost Regression Model Flow Chart. 43

Figure 3-14. LSTM Forecasting Model- A Flow Chart. 45

Figure 3-15. LSTM Forecasting Model-B Flow Chart. 45

Figure 3-16. Residual Analysis and Actual vs. Predicted Comparison Plots. 46

Figure 3-17. SHAP Values Plot example..... 47

Figure 3-18. Feature Importance plot example..... 47

Figure 3-19. Spider Plot Examples. 48

Figure 4-1. Original Data for Maintenance Type Costs Categories..... 51

Figure 4-2. Original data for Preventive Maintenance Costs 51

Figure 4-3. Preventive Maintenance Cost After Completing Data Preprocessing. .. 52

Figure 4-4. Features (Predictors) Data. 52

Figure 4-5. Preventive Maintenance Box Plot. 53

Figure 4-6.a. Preventive Maintenance Cost Distribution. 54

Figure 4-6.b Preventive Maintenance Quartile Analysis. 54

Figure 4-7. Preventive Maintenance Log Transformation Distribution..... 54

Figure 4-8. Preventive Maintenance Drift Analysis..... 55

Figure 4-9. Preventive Maintenance Correlation with Features. 55

Figure 4-10. Distribution of Predictors. 56

Figure 4-11.a. Initial Features Correlation..... 56

Figure 4-11.b. Revised Features Correlation 56

Figure 4-12. Predictors Testing and Holdout Distribution Analysis..... 57

Figure 4-13. Preventive Maintenance Costs ADF Test and ACF + PACF plots..... 60

Figure 4-14. Differenced Preventive Maintenance Costs ADF Test and ACF + PACF plots. 61

Figure 4-15. Log-transformed Preventive Maintenance Costs ADF Test and ACF + PACF plots..... 62

Figure 4-16. Dif+Log-transformed Preventive Maintenance Costs ADF Test and ACF + PACF plots. 63

Figure 4-17.a. Top 10 Important Features – ElasticNet Time Series Validation..... 64

Figure 4-17.b. Top 10 Important Features – ElasticNet Random Series (K-fold validation) 65

Figure 4-17.c. Top 10 Important Features – ElasticNet Random Series (Walk Forward Validation) 65

Figure 4-18.a. SHAP Values Plots for Elastic Net with Time Series Validation..... 66

Figure 4-18.b. SHAP Values Plots for Elastic Net with KFold Validation..... 66

Figure 4-18.c. SHAP Values Plots for Elastic Net with Walk Forward Validation.. 66

Figure 4-19.a. Important Features – XGBoost Model - GridSearchCV..... 67

Figure 4-19.b. Important Features – XGBoost Model - Optuna 67

Figure 4-20. ARIMA model without transformation results. 68

Figure 4-21. ARIMA model with log transformation results. 68

Figure 4-22.a. ARIMA model without log transformation error plots..... 69

Figure 4-22.b. ARIMA model with log transformation error plots. 69

Figure 4-23. SARIMA Model without transformations Results..... 70

Figure 4-24. SARIMA Model with log transformation Results..... 70

Figure 4-25.a. SARIMA model without log transformation error plots..... 71

Figure 4-25.b. SARIMA model with log transformation error plots. 71

Figure 4-26.a. Elastic Net with Time Split-Validation..... 72

Figure 4-26.b. Elastic Net with K-Fold Cross Validation..... 73

Figure 4-26.c. Elastic Net with Walk-Forward Validation. 73

Figure 4-27.a. ElasticNet Time Series Validation model error plots 74

Figure 4-27.b. ElasticNet k-Fold Validation error plots. 74

Figure 4-27.c. ElasticNet Walk-Forward Validation error plots 74

Figure 4-28. XGBoost Grid Search CV Model..... 75

Figure 4-29. XGBoost Optuna Model. 75

Figure 4-30.a. XGBoost with Grid Search error plots. 76

Figure 4-30.b. XGBoost Optuna error plots..... 77

Figure 4-31. LSTM Model Test One-Step Forecasting Results. 77

Figure 4-32. LSTM Model with Multi-Step Forecasting Results. 78

Figure 4-33.a. LSTM Train/Test split and Single-Step error plots. 79

Figure 4-33.b. LSTM Train/Validation/Test split and One-Step forecasting error plots 79

Figure 4-34. Preventive Maintenance - Models Performance – (Test Set). 80

Figure 4-35. Preventive Maintenance - Models Performance – (Holdout Set). 81

Figure 4-36. Best Models Actual (unseen) vs Predicted data – (Holdout Set). 83

Figure A-1. Power Plant Maintenance Cost Structures. 109

List of Tables

Table 2-1. Summary of Cost Estimation Classification (Adapted from ASTM E2516-11) 18

Table 3-1. Data Characterization for Cost Transactions and Work Orders 31

Table 3-2. Data Characterization for Maintenance Cost Classes..... 31

Table 3-3. Data Characterization for Support Tables..... 32

Table 3-4. Data Characterization for Plant Performance..... 33

Table 3-5. ARIMA and SARIMA Proposed Tuning Parameters..... 41

Table 3-6. Elastic Net Models Proposed Tuning Parameters 43

Table 3-7. XGBoost Model Proposed Tuning Parameters..... 44

Table 3-8. LSTM Models' Main Differences 45

Table 3-9. LSTM Model Proposed Tuning Parameters..... 46

Table 3-10. Sample Ranking Matrix for RMSE 48

Table 4-1. Output Target Data Transformation Steps..... 50

Table 4-2. Predictors Drift Analysis..... 57

Table 4.3. Statistically significant lags. 59

Table 4-4. ARIMA Models Performance Results 69

Table 4-5. SARIMA Model Test Set Performance Results. 71

Table 4-6. Elastic Net Models Performance Results – Test Set. 73

Table 4-7. XGBoost Model Performance Results – Test Set..... 76

Table 4-8. LSTM Optimized Model Performance Results – Test Set. 78

Table 4-9. Preventive Maintenance - Models Performance – (Test Sets). 80

Table 4-10. Preventive Maintenance - Models Performance – (Holdout Sets). 81

Table 4-11. Best Models Actual (Unseen) vs Predicted Data – (Holdout Set). 82

Table B-1. Cost Transactions Fields 110

Table B-2. Work Order Report Used Fields 111

List of Symbols

$\boldsymbol{P'}$	Feature data array
m	Represents each month on the study period from m_1 to m_j.
p_i	Refers to each feature attribute such as Plant Blend, Plant Dispatch Factor, Load Cycles, etc., from p_1 to p_r.
p_{ij}	The value of the feature i-th for the j-th month
$\boldsymbol{C'}$	Target output array
C_j	Represents the preventive maintenance cost for each month from C_1 to C_j

List of Acronyms

ACF	Autocorrelation Function
ADF	Augmented Dickey-Fuller
AEO	Annual Energy Outlook
AI	Artificial Intelligence
ANN	Artificial Neural Networks
ASTM	American Society for Testing and Materials
ARIMA	Autoregressive Integrated Moving Average
CAPEX	Capital Expenditures
CFBC	Circulating Fluidized-Bed Combustion
CFR	Code of Federal Regulations
CMMS	Computerized Maintenance Management System
EIA	U.S. Energy Information Administration
EFOF	Equivalent Planned Outage Factor
EPOF	Equivalent Forced Outage Factor
EPRI	Electric Power Research Institute
ERP	Enterprise Resource Planning
IEA	International Energy Agency
ISO	International Organization for Standardization
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error

ML	Machine Learning
MSE	Mean Squared Error
MW	Mega-watt
MWh	Mega-watt Hour
NaN	Not a Number
NERC	North American Electric Reliability Corporation
NFVC	Non-Fuel Variable Cost
NREL	National Renewable Energy Laboratory
OPEX	Operational Expenditures
O&M	Operation and Maintenance
PACF	Partial Autocorrelation Function
PCA	Principal Component Analysis
Plant_DF	Plant Dispatch Factor
PM	Preventive Maintenance
PV	Photovoltaic
R2	Coefficient of Determination
RMSE	Root Mean Squared Error
SARIMA	Seasonal Autoregressive Integrated Moving Average
SHAP	SHapley Additive exPlanations
U.S.	United States
USD	United States Dollars
WO	Work Order
XGBoost	Extreme Gradient Boost

Chapter 1—Introduction

1.1 Background

Thermal power plants still play an essential role in meeting worldwide energy demands, providing a significant portion of the global electricity supply (International Energy Agency [IEA], 2023). However, the reliability of conventional generation is significantly challenged by more frequent extreme weather, high-demand conditions, and a changing resource mix, resulting in higher overall outage rates (North American Electric Reliability Corporation [NERC], 2023). Therefore, the efficient operation of these plants requires careful management of equipment maintenance to ensure reliability, safety, and cost-effectiveness. Maintenance activities in thermal power plants encompass a wide range of tasks, including routine inspections, preventive maintenance, and corrective repairs, all aimed at preserving the integrity and functionality of critical assets.

One of the challenges faced by power plant managers is accurately forecasting maintenance costs. A contributing factor is that no clear standard exists for maintenance cost quantification (Lemes & Hvam, 2019).

Traditional maintenance budgeting approaches often rely on extensive resources and time-consuming deterministic methods requiring individual asset data to build up the estimations or methods using historical data to estimate budgets based on mathematical models (Bahr & Lennerts, 2010), which are complex to implement.

The budget is then adapted to either fixed schedules, reactive strategies, or condition monitoring (Galar et al., 2017). These strategies, however, may not be well suited for the power generation mix applications that are dynamic and complex.

Advances in machine learning (ML) and data analytics are generating opportunities to enhance maintenance cost forecasting in thermal power plants. Historical plant operation and maintenance data can be used in ML models to reveal patterns, identify predictive features, and provide insights for a more proactive maintenance resource allocation planning approach. Supervised learning techniques have also emerged as an effective approach for processing labeled data available in asset-intensive industrial settings. This addresses situations where organizing and analyzing can be challenging and expensive.

1.2 Research Motivation

For the last three years, the reliability of conventional power plants has been on a downtrend (NERC, 2023). This downtrend is due to the increasing complexity and operational demands of thermal power generation. This research focuses on better asset reliability and flexibility improvement through better maintenance cost forecasting methodologies. Traditional approaches that are time and resource intensive may struggle to capture equipment performance, operation, and maintenance dynamics, which then leads to budget deviations and operational inefficiencies. The use of predictive machine learning models in this research aims to bridge the gap between historical data insights and budgeting for preventive maintenance. When stakeholders are provided with accurate forecasting tools, they can better optimize resource allocation, minimize downtime, and bolster overall operational control and resilience. This work also seeks to offer practical solutions for thermal power plants and other similar asset intensive industries through sustainable and cost-effective maintenance budgeting strategies.

1.3 Problem Statement

Power Plant managers are unable to forecast preventive maintenance costs that adapt to the historical plant performance and reflect accurate monthly expenditures, both of which contribute to significant budget deviations.

This particular problem is made worse by a lack of adaptability to historical equipment performance, which in turn means current forecasting methods may not effectively implement past data. Traditional methods estimate maintenance budgets using either a value of 2-5% of asset value (Reliability Web, 2023), a resource and time consuming deterministic model like the one proposed by Galar et al. (2017), or a model adjusted with scalation factors between 2-5% (GHD, 2022).

1.4 Thesis Statement

A predictive Supervised Machine Learning model based on historical plant operation and maintenance data is necessary for accurately forecasting thermal power plant maintenance costs.

Historically, most deterministic (Galar et al., 2017; Abdenour, 2021) and stochastic budget estimation methods (Ferreira et al., 2013; Gupta et al., 2014, 2017) require time and resource efforts to obtain accurate estimation as demonstrated by the American Society for Testing and Materials standard ASTM E2516-11 (ASTM International, 2011), where the higher the accuracy, the higher the effort required. Additionally, little or no development has been made to standardize maintenance costs (Lemes & Hvam, 2019), providing room for development using machine learning techniques to predict maintenance costs with minimum time and effort.

1.5 Research Objectives

The objectives of this research are as follows:

- Develop preventive maintenance costs forecasting models leveraging historical plant operation and maintenance data using machine learning techniques.
- Evaluate the performance of various machine learning algorithms, including Time Series, Linear Regression, Decision Trees (XGBoost), and LSTMs in predicting preventive maintenance costs.
- Identify trends and factors that influence preventive maintenance costs by analyzing past performance and maintenance patterns.
- Assess the effectiveness of supervised machine learning techniques in enhancing the predictive accuracy and robustness of preventive maintenance cost forecasts.
- Validate the developed predictive models through rigorous testing and comparison against actual preventive maintenance cost data, using metrics such as mean absolute error (MAE), root mean squared error (RMSE), and Mean Absolute Percentage Error (MAPE).

1.6 Research Questions and Hypotheses

The following research questions and hypotheses are developed to support the evaluation and validation of the machine learning models.

RQ1: What are the significant factors in forecasting the preventive maintenance cost in a thermal power plant?

RQ2: Which ML models can most accurately predict the preventive maintenance cost of a thermal power plant?

RQ3: Which model provides the lowest mean absolute percentage error (MAPE) among Time Series, Linear Regression, XGBoost, or LSTM in forecasting the preventive maintenance costs of a thermal power plant?

H1: The Main Factors (Independent Variables) identified in this praxis significantly influence the maintenance costs of a thermal power plant.

H2: The Time Series models (ARIMA/SARIMA with and without log transformation) will achieve a mean absolute percentage error (MAPE) below 30%, demonstrating their predictive effectiveness in estimating preventive maintenance costs.

H3: The linear regression models (ElasticNet with time series validation, k-fold validation, and walk-forward validation) will achieve a mean absolute percentage error (MAPE) below 30%, demonstrating their predictive effectiveness in estimating preventive maintenance costs.

H4: The XGBoost models (with GridSearchCV or with Optuna) will achieve a mean absolute percentage error (MAPE) below 30%, demonstrating their predictive effectiveness in estimating preventive maintenance costs.

H5: The LSTM models (one with training/test split and Recursive One-Step Forecasting and other with training/validation/test split and Recursive Multi-Step Forecasting) will achieve a mean absolute percentage error (MAPE) below 30%, demonstrating their predictive effectiveness in estimating preventive maintenance costs.

1.7 Scope of Research

This research focuses on the three things: the development, validation, and testing of preventive maintenance cost forecasting models by using machine learning techniques with historical data sourced from a specific thermal power plant. The research covers the

collection, preprocessing, and analysis of historical plant operation and maintenance data to identify significant factors impacting maintenance costs. The findings will support power plant managers in accurately forecasting the next year of preventive maintenance costs (in a monthly basis) with minimum effort.

1.8 Research Limitations

Limitations that may affect the validity and/or generalizability of the models are the availability, completeness, and quality of the data used for this analysis.

The findings of this research are specific to the thermal power plant whose data was collected. This means that differences in equipment, operating conditions, maintenance practices, and organizational contexts may limit how applicable the results are to other power plants or industrial settings.

Assumptions made (e.g., linearity and independence of variables) may not fully capture the complexities of the processes for real-world maintenance cost forecasting. Such assumptions could introduce uncertainties or inaccuracies in the model predictions. For this research, assumptions are specifically described for each model.

This research is limited in scope to specific aspects of preventive maintenance cost forecasting in thermal power plants. Other maintenance categories or lower granularities are not considered which may restrict the breadth of the analysis and generalization.

The expertise and resources available for conducting the research, including access to specialized knowledge, technical skills, and computational resources, may constrain the depth and sophistication of the analysis and model development efforts.

The analysis in this praxis applied selected machine learning techniques, including ARIMA/SARIMA time series, Elastic Net Linear Regression, XGBoost, and a Long Short-Term Memory (LSTM) neural network. Other methods within the broader category of machine learning, such as k-Nearest Neighbors, Random Forest, and Support Vector Regression, were excluded due to time constraints and their limited applicability for the specific forecasting objectives in this study.

For the results of this research to be repeatable for other similar applications and industries, similar preprocessing, training, and testing datasets would need to be used to obtain a customized model.

1.9 Organization of Praxis

This praxis is organized as follows:

Chapter 2 contains the literature review, while Chapter 3 outlines the methods, techniques, and procedures used in this research. Chapter 4 presents the results with a comparative analysis of the model´s performance that aids in answering each hypothesis and supports determining the most effective model(s) based on performance criteria (e.g., MAE, RMSE, and MAPE). Chapter 5 discusses the findings, summarizes conclusions, and provides suggestions for future research and development.

Chapter 2—Literature Review

2.1 Introduction

This literature review explores key aspects of maintenance cost structures in thermal power plants amid an evolving energy landscape marked by increasing integration of renewable sources. It examines how operational challenges, fuel flexibility, and regulatory changes can impact maintenance costs and deals with the need for more sophisticated cost forecasting models.

Section 2.2 provides an overview of the current outlook for thermal power plants worldwide and the stance of fossil fuel generation fleets among the energy matrix., It reviews industry perspectives on power plant cost segregation and outlines how various international organizations and benchmarking institutions have been addressing maintenance cost estimation.

Section 2.3 looks at the factors that are driving changes in the operational regimes of thermal power plants and the impact these factors have on maintenance costs and equipment reliability, providing context for the study.

Section 2.4 focuses on power plant maintenance cost structures in order to identify the gaps in current methodologies and understand the need for more detailed cost organization and standardization.

Section 2.5 introduces budget estimation methodologies commonly applicable in industry setups and highlights the gaps encountered when applied in power plant maintenance.

Sections 2.6 to **2.9** address challenges, including annual budget updates, budget line of sight, maintenance cost overruns, and cost estimation classifications, underscoring the need for more adaptive predictive models.

Sections 2.10 and **2.11** discuss traditional and machine learning-based techniques for predicting maintenance costs. Approaches like time series models, linear regression, decision trees, and advanced neural networks such as Long Short-Term Memory (LSTM), are reviewed. Their effectiveness in handling complex datasets and their limitations are critically evaluated, showcasing the potential for implementing these models in preventive maintenance cost forecasting. **Section 2.12** provides a summary of the literature review.

This literature review lays the foundation for understanding the main drivers of preventive maintenance and provides a framework for developing machine-learning models to improve maintenance cost estimation in thermal power plants.

2.2 Thermal Power Plants Outlook

With the increasing adoption of renewable energy sources, the demand for coal-based generators is declining (U.S. Energy Information Administration [EIA], 2023). By 2030, the level of coal-fired generation capacity is predicted to be 50% of what it was in 2024 and is expected to gradually reduce to close to zero around 2050 (EIA, 2023). These numbers are based on supply and demand projections for the US energy sector, considering current policy along with geopolitical and technological advances.

In the last three years there has been an increase in unavailability of conventional generation plants, from 7.1% in 2020 to 8.5% in 2022, and coal has been the highest contributor to this unavailability (NERC, 2023). Unavailability was attributed in part to a

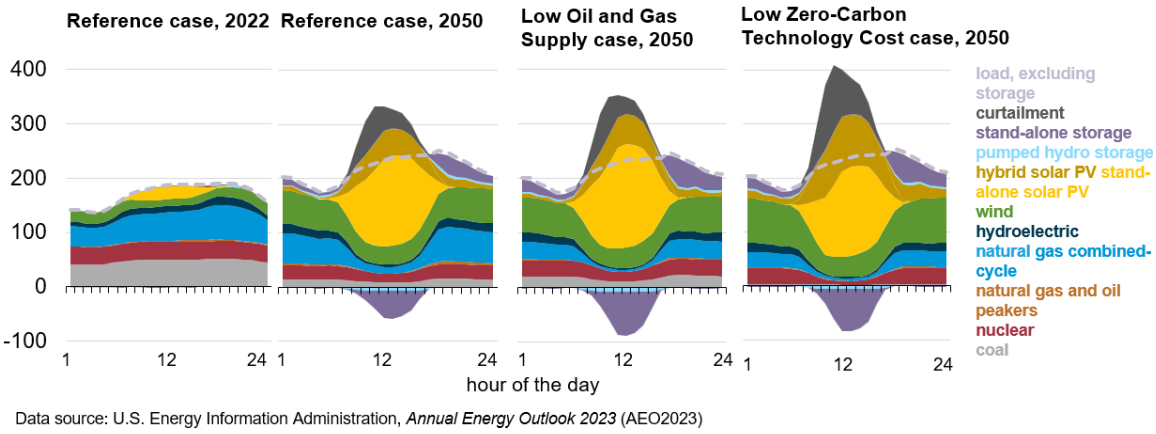
changing resource mix but primarily to extreme weather events that are affecting renewable infrastructure and leading to higher overall outage rates as well as increased operational demands for conventional power plants. In addition, new policies are leading to a reduction in coal-fired and other conventional generation methods without sufficiently reliable technology and operational processes in place (NERC, 2023). This points to increased risks posed by both new technologies and more frequent extreme weather. Water availability for power generation is also a concern because of limitations in various regions (Electric Power Research Institute [EPRI], 2018). The above mentioned constraints are adding challenges to the maintenance costs and performance of thermal power plants.

While the outlook for thermal power plants is shaped by declining reliance on fossil fuels and increased operational constraints, understanding the factors driving shifts in operational regimes, such as the integration of renewables and evolving fuel flexibility, provides critical insights into how these changes influence maintenance costs, as explored in the following section.

2.3 Factors Changing Operational Regimes on Thermal Power Plants

The growing integration of renewable energy sources has placed increasing pressure on conventional thermal power plants, particularly during periods of high renewable energy generation (EIA, 2023). The reliance on fossil fuel power generation is expected to decline during hours of high renewable generation (EIA, 2023). **Figure 2-1** illustrates the impact of renewable generation on a typical 24-hour operational cycle and provides a forecast for its continued influence on power generation through 2050. This shift results in more frequent cyclic operations for conventional power plants (which

mean either ramp-up and ramp-down operations between minimum and maximum load, or additional units startup and shutdown operations), as opposed to continuous base-load operations.



Data source: U.S. Energy Information Administration, *Annual Energy Outlook 2023* (AEO2023)

Figure 2-1. Hourly U.S. Electricity Generation and Load by Fuel for Selected Cases and Representative Years (billion kWh). Source: (EIA, 2023).

Note: In the public domain

Therefore, the transition to cyclic operation poses several challenges for conventional power plants, particularly in terms of increased maintenance costs and outages. The impact of cycling operations on the Equivalent Planned Outage Factor (EPOF) and Equivalent Forced Outage Factor (EFOF) was estimated by the EPRI using empirical analysis and polynomial regression modes based on the central limit theorem to be approximately 2% to 6% higher in cycling units than base-load operation units (EPRI, 2013). The impact of renewable energy integration on the cycling operations of fossil-fuel-based power plants has been also analyzed using regression analysis combining deterministic top-down cycling cost analysis with bottom-up cost estimations of primary unit, revealing that small subcritical units experienced an increase in cycling costs ranging from 0.2 to 0.7 $/MW (based on 25th to 75th percentile) (Kumar et al., 2012). An

estimated 100MW of additional solar PV installed capacity leads to an increment of marginal Non-Fuel Variable Cost (NFVC) rate between 0.05 and 0.11 USD/MWh (Bassi et al., 2018). Such cost increases are attributed to the additional strain on plants operating in a cyclic regime compared to those with continuous base-load operations. These conclusions serve as a valuable benchmark for estimating the variable operation and maintenance (O&M) costs.

Another factor that influences the operational regime of thermal power plants is fuel flexibility. An economic evaluation of circulating fluidized-bed combustion (CFBC) power plants indicates that it has advantages related to fuel flexibility and market responsiveness. CFBC power plants also benefit from lower fixed O&M costs (between 6.6% and 8.2%) and lower variable O&M costs (~10%) when compared to conventional pulverized fuel technology (Wheeldon & Thimsen, 2013). When burning different fuel blends (e.g., coal and petroleum coke with high sulfur content) the control equipment for compliance with environmental regulations requires additional capital expenditures (CAPEX). Such CAPEX also leads to additional costs for the O&M of the extra equipment. This research investigates the impact of fuel blend on maintenance costs to determine its significance as a contributing factor.

2.4 Power Plant Maintenance Costs Structures

The available literature on power plant maintenance costs has traditionally focused on two things: 1) establishing frameworks for categorizing non-fuel O&M expenses and 2) segmenting them into broad classifications of fixed and variable costs with yearly estimations (Meyers & Fuller, 1979; Kumar et al., 2012; Wheeldon & Thimsen, 2013; Badouard, 2018; Bassi et al., 2018). These studies generate average cost

factors expressed in terms of dollars per megawatt (MW) installed or megawatt-hour (MWh) generated. Attempts have been made to provide simplified asset system segregation using annualized data (Code of Federal Regulations CFR, 2024; Kumar et al., 2012), but unfortunately, they are neither comparable nor traceable. Although the annualized data is effective for long-term financial planning and benchmarking, these segregation models lack the granular detail necessary for accurate short-term budgeting and forecasting at a more operational level by maintenance managers (Lemes & Hvam, 2019).

While significant advancements have been made in developing cost estimation models for power plants under different scenarios, there is no standardized framework for maintenance cost structure. Although standards such as the International Organization for Standardization (ISO) series emphasize strategic asset management and encourage organizations to customize maintenance cost structures, they do not provide specific guidelines for maintenance cost segregation, leaving this to the discretion of individual organizations (ISO 14224, 2016; ISO 55000, 2014; ISO/TS 55010, 2019).

A limitation of most studies and industry standards is that they often overlook the nuanced needs of different stakeholders. For example, financial reporting focuses on broader categories of O&M costs, while operational and maintenance teams require detailed breakdowns by expense class, such as expense type, maintenance type, specialty, or asset system, for effective budgeting (Galar et al., 2017). This disparity highlights the opportunity to develop more detailed short-term variable cost models to assist mid-level managers in optimizing planning activities and operational decisions. This work will test

the modeling of preventive maintenance costs, and if found applicable, the approach can be extended to other expense sub-classes.

2.5 Budget Estimation Methodologies

This section reviews budget estimation methodologies relevant to power plant maintenance. While the literature does not mention specific methodologies for power plants, parallels can be drawn from other industries.

There are two main budget estimation methods used in construction projects: Stochastic and Deterministic, both of which are also applicable to power plant maintenance (Abdenour, 2021). In facility management, there are four maintenance budgeting methods, namely, key figure-oriented budgeting, value-oriented budgeting, analytical calculation, and budgeting by condition description (Bahr & Lennerts, 2010). The first two methods, which rely on historical data and fixed percentages, are widely used in power plants. However, they provide only average values and do not account for cyclical or seasonal fluctuations. The third method, which uses technical characteristics of assets, aligns well with this research, focusing on feature-based forecasting. The fourth method uses detailed asset condition-based data and requires extensive time and resources, as explained in the next section.

2.5.1 Deterministic Estimation Methods

There exists a long list of considerations and aspects required to prepare a maintenance budget, and a numerous of forms to collect individual asset/component data to build up the estimations (Galar et al., 2017). Yet, deterministic estimation methods that involve budgeting by condition combined with unplanned maintenance costs appear to be the most accurate approach (Bahr & Lennerts, 2010).

Deterministic methods calculate required resources (spare parts, consumables, manpower, etc.) for each maintenance activity and multiply them by unit costs (Abdenour, 2021) or use expert opinions (Ogunbayo et al., 2022). The activities are defined in maintenance plans, which derive from strategic asset management plans aligned with operational objectives (Galar et al., 2017). When strictly developed, this approach could provide a "precise" estimate but is time and resource-consuming and still depends on multiple assumptions that may not reflect actual conditions.

2.5.2 Stochastic Estimation Methods

Stochastic methods include: (i) key figure-oriented budgeting, (ii) value-oriented budgeting, and (iii) analytical calculation (Bahr & Lennerts, 2010). Stochastic methods typically use historical data and statistical analysis to account for uncertainty. Attempts have been made to generate mathematical models using plant operation and maintenance to estimate the annual maintenance budget based on asset replacement value (Gupta et al., 2014, 2017), but present complexity in implementation for large asset systems requiring large amounts of data entry. Stochastic methods have also been applied to define fixed and variable O&M costs as a percentage of Capital Expenditures - CAPEX (Badouard et al., 2018), which adapts method (iii) by using plant history and performance data as key features for cost estimation.

2.5.3 Mixed Estimation Methods

An example of mixed estimation methods, which combines reliability theory with maintenance budgeting, is modeling preventive maintenance as deterministic and corrective maintenance as stochastic (Ferreira et al., 2013). While effective for smaller systems, this approach becomes complex as the number of components increases,

limiting its application in larger power plants. Innovative machine learning decision tree classification analysis is also being used to identify the attributes that contribute most to maintenance decisions and then assign costs that can be used for forecasting (Amir & Henry, 2023).

This work proposes a novel model to support department managers in reviewing and preparing next year's budget by applying the stochastic method using a machine learning technique to forecast preventive maintenance costs. It will also explore the use of a decision tree (XGBoost) for prediction.

2.6 Annual Budget Updating

Annual budget reviews require department managers to review the performance of the previous year and develop the plan for next year with similar considerations as for the initial deterministic model but adding backlogs and additional requirements (Galar et al., 2017). This process is time and resource-consuming, leading managers to increase according to some ratios (Galar et al., 2017) or take shortcuts to simplify the exercise by making more generalized assumptions.

The scope of this work includes the forecast for the next twelve months of preventive maintenance costs.

2.7 Maintenance Budget Line of Sight

The normal design operational life for power plant equipment is between 25 and 35 years (EPRI, 2013). The failure rate follows a bathtub shaped curve across its lifecycle **(**Ben-Daya et al., 2016); **Figure 2-2** shows the bath tub curve represented as the Equivalent Planed Outage Factor (EPOF) and divided into three main periods: the commissioning period (5 years), the useful life period (from 5 to 20 years), and the wear

out period (>20 years) (EPRI, 2013). This provides a line of sight on how maintenance costs may behave for the plant being analyzed in this research, which is between years 3 and 9 (green shadow area).

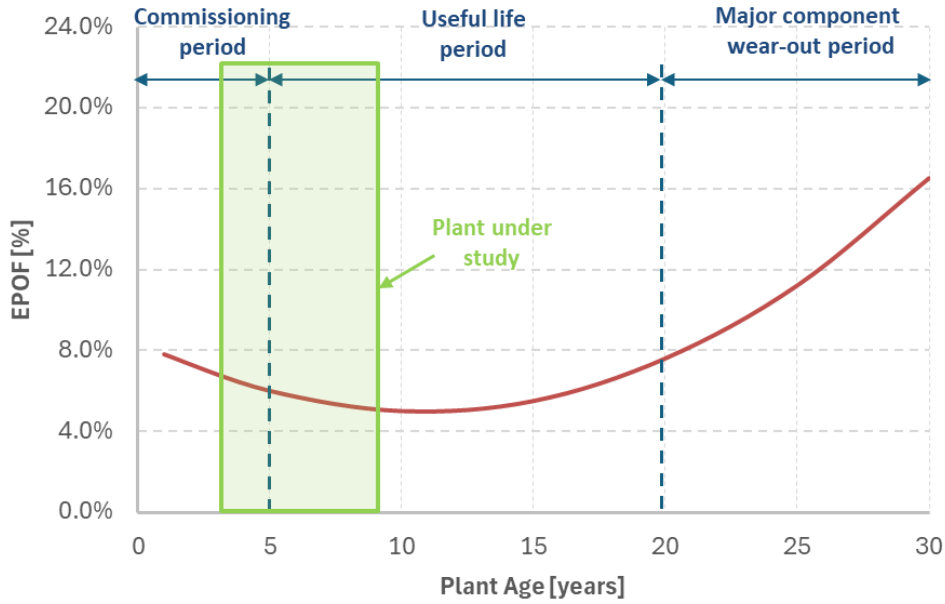


Figure 2-2. Life Cycle EPOF vs Age for Conventional Plants Operating in Cycling Regime and Major Life Periods. Adapted from Electric Power Research Institute (EPRI), 2013.

The power plant under analysis is a relatively young plant that has moved out of the commissioning period and has rolled over its initial years of useful life. It is expected that preventive maintenance costs will increase until stabilization, while corrective maintenance will be reduced to its minimum.

2.8 Maintenance Budget Overruns

Overruns are inevitable when reviewing a budget (Sweis, 2013; Cheng, 2014; Heravi & Mohammadian, 2021, as cited in Coffie & Cudjoe, 2023). Cost overruns in construction projects can come from technical issues such as scope change (differences

between as-built and as-designed) and inaccurate forecasting, as well as political, economic, technological, and psychological biases (Abdenour, 2021). This can similarly be applied to power plants, assuming maintenance costs as a project.

2.9 Cost Estimation Classification

ASTM International (2011) introduced the *Standard Classification for Cost Estimate Classification System* (ASTM E2516-11), which provides a generic classification matrix for categorizing cost estimates across various industries. This matrix uses the degree of project definition as the primary characteristic to classify estimates into five classes, ranging from Class 5 (least defined) to Class 1 (most defined). **Table 2-1** summarizes the classification system, which aims to improve communication and accuracy by aligning estimating methodologies—stochastic or deterministic—with project maturity and preparation effort.

Table 2-1. Summary of Cost Estimation Classification (Adapted from ASTM E2516-11)

Estimated Class	Usage	Methodology	Expected Accuracy	Preparation Effort
Class 5	Feasibility	Stochastic or judgment	-50% to +100%	1
Class 4	Concept study	Primarily stochastic	-30% to +100%	2 to 4
Class 3	Budget authorization	Mixed but primarily Stochastic	-20% to +50%	3 to 10
Class 2	Budget control	Primarily deterministic	-15% to +20%	5 to 20
Class 1	Execution comparison	Deterministic	-10% to +15%	10 to 100

The estimated classes 3, 2, and 1 are used for "budget authorization or control," which applies to this work. Class 3 typically expects accuracy between -20% to +50%,

while class 1 expects accuracy between -10 to +15%. While this standard serves as a benchmark to compare the accuracy of machine learning estimation techniques, this research hypothesis aims to reach an accuracy higher than 70%.

2.10 Comparison of Traditional vs. Machine Learning Approaches

As discussed in section 2.5, traditional cost estimation methods for maintenance in typically rely on deterministic or stochastic techniques, such as historical cost analysis, regression models, or expert judgment. The deterministic approaches provide precise calculations based on predefined parameters, but can be inflexible and time-consuming, and often fail to account for unexpected operational changes or variability in the data (Galar et al., 2017). Stochastic methods, on the other hand, introduce statistical models that can incorporate uncertainty and randomness, but are typically limited in their usefulness by the quality and scope of historical data available for use (Bahr & Lennerts, 2010). Traditional methods are effective for long-term budgeting but can struggle with short-term forecasting because they rely on fixed parameters and limited adaptability. Machine learning (ML) models, in contrast, offer dynamic and data-driven approaches for maintenance cost forecasting with minimum effort.

Decision trees, ensemble methods, and neural network techniques can all handle large, complex datasets while identifying critical patterns or relationships that more traditional methods may overlook. XGBoost, for example, has shown superior performance in cost prediction tasks because it can handle noisy data, missing values, and complex feature interactions (Elmousalami, 2021). Similarly, LSTM models excel in capturing temporal dependencies and nonlinear trends. This makes them ideal for time-series forecasting of maintenance costs.

The key advantage of ML lies in its ability to learn from data and continuously improves its predictive accuracy as new information becomes available. These advantages lead to a more flexible and accurate forecasting process for preventative maintenance in particular, where operational conditions and cost drivers can vary significantly over time. When compared to traditional approaches, integrating ML models into maintenance cost estimation allows power plant managers to obtain more reliable forecasts, better optimize resource allocation, and eventually reduce budget overruns.

2.11 Machine Learning Prediction Techniques

2.11.1 Time Series

Power plant cost data and plant operational features are both considered time series with historical maintenance costs and plant performance values that can, therefore, be used for forecasting.

The available literature on time series forecasting emphasizes that traditional models like ARIMA and SARIMA are used for handling time-dependent data, particularly in scenarios that involve seasonality and outliers (Aghelpour et al., 2019; Ng et al., 2023; Bunnag, 2024). These models have been widely adopted across various fields because they are robust in managing sudden variations and effective in linear time series scenarios. ARIMA and SARIMA can capture seasonal effects and accommodate multivariate inputs; they could additionally be enhanced with feature optimization techniques such as Principal Component Analysis (PCA) to improve forecasting accuracy (Tong et al., 2023; Draz et al., 2024), however they are limited when working with large data sets, high number of features, or complex, nonlinear relationships (Peixeiro, 2022).

One of the notable limitations reported in the reviewed literature is the absence of specific applications for industrial maintenance cost forecasting. Such a gap in published research provides an opportunity to adapt modern methodologies for preventive maintenance cost forecasting in thermal power plants, where time series data can exhibit seasonality and unexpected variations due to operational disruptions. The effectiveness of these models depends heavily on data preprocessing, quality, and careful model configuration (Peixeiro, 2022).

2.11.2 Linear Regression (Elastic Net)

Numerous studies on linear regression emphasize its capacity to predict an output from single or multiple feature relationships using different variations and optimizations (Alhamzawi & Ali, 2018; Tay et al., 2023; Drachal & Pawłowski, 2024; Zhang et al., 2017). Linear regression has been used to model many applications for forecasting and comparison with other models (Mirzaei et al., 2024; Drachal & Pawłowski, 2024; Abbas et al., 2020; Winalytra et al., 2018; Zournatzidou et al., 2024), most concluding that the linear regression technique underperformed, especially against decision trees and support vector machines, suggesting that non-linear models may be more suitable for capturing complex relationships in time-series data (Tran et al., 2022). In addition, embedding-based fuzzy approaches are more effective for non-stationary time series prediction than linear models, although they require custom implementations that are not directly available in standard machine-learning libraries (Bitencourt et al., 2023).

Despite the breadth of research in linear regression, there is a gap in the literature regarding its application to maintenance cost forecasting. However, many of the reviewed methods can be adapted to estimate them, where capturing the inter-dependencies among

maintenance costs, plant operation, and performance factors could be crucial for accurate forecasting.

2.11.3 Decision Trees

Traditional decision trees have provided a stratified way to make decisions similar to how the human brain works, but are not competitive with the best supervised learning approaches (e.g., linear or polynomial regressions). Because of this, random forest and tree-boosting methods have emerged to improve accuracy and have become widely used in multiple sectors of machine learning (Chen & Guestrin, 2016; James et al., 2017). One of the most notable advancements in decision tree techniques is XGBoost, which has already proven to be both reliable and efficient as a machine learning tool and is consistently ranked among the top performers in Kaggle competitions (Bentejac et al., 2019; Chen & Guestrin, 2016). XGBoost is very effective when working with small datasets and nonlinear patterns. This ability makes it useful for cost forecasting, where noisy data is a challenge (Elmousalami, 2021). Additionally, it is specifically engineered to lessen issues that can arise from overfitting, thus ensuring robust model performance while still maintaining computational efficiency (Coffie & Cudjoe, 2023).

Furthermore, XGBoost also excels at dealing with multiple potentially interrelated variables, which is an essential feature when analyzing influencing factors (Alashari et al., 2023). The ability of XGBoost to handle time series data—capturing seasonality, trends, and correlations with historical features—highlights its ability to effectively track fluctuations (Bezzar et al., 2022), thus making XGBoost a strong candidate for short-term forecasting of preventive maintenance costs.

2.11.4 LSTM

Long Short-Term Memory (LSTM) models have emerged as a powerful tool for time series forecasting, particularly when traditional models like ARIMA and SARIMA face challenges in handling large datasets, high-dimensional data, or nonlinear relationships (Peixeiro, 2022; Mani & Volety, 2021). LSTM excels in capturing complex temporal dependencies and mitigating issues like vanishing gradients, which commonly affect conventional neural networks.

A key strength of LSTM is its ability to retain long-term dependencies, making it highly effective in forecasting applications across various fields. For example, forecasting project costs, LSTM has demonstrated higher prediction accuracy compared to Artificial Neural Networks (ANN) (Le et al., 2020). Similarly, its capability to filter redundant data through feature optimization significantly improves fault diagnosis for industrial applications, which highlights the importance of data preprocessing (Feng et al., 2024). Additionally, LSTM models usually perform well in multi-output forecasting scenarios by considering interdependencies between outputs, which enhances predictive accuracy even on multiple locations (Crivellari & Beinat, 2020).

However, LSTM models still have limitations. One of the most significant is the high computational cost that impacts real-time applications and large deployments by making them heavily resource-intensive (Gafurov et al., 2023). Training these networks normally requires substantial computational power and memory. This can be a problematic barrier to implementation in environments where available resources are constrained. LSTM models are sensitive to data quality and preprocessing, which means

that careful handling of missing values, noise, and feature selection are key to ensure optimal performance.

2.12 Summary and Conclusion

This literature review provided insight into the complexities of maintenance cost estimation not only for thermal power plants, but also for other industries, highlighting the energy landscape driven towards renewable energy integration, including regulatory changes, and leading to operational challenges that adds to the maintenance requirements. This review highlighted key factors influencing maintenance costs, including challenges such as increased cyclic operations due to renewable energy penetration, fuel flexibility requirements, an aging infrastructure and planed and forced maintenance factors. Such factors result in new challenges in cost forecasting and management, emphasizing the need for predictive models that are more sophisticated.

For modern power plants, existing maintenance cost estimation methodologies (both deterministic and stochastic) often prove insufficient for capturing the dynamic and complex operations. Traditional approaches usually rely on fixed and variable cost structures but do not have the granularity needed for short-term budgeting and forecasting, nor capturing seasonality, which makes it difficult for mid-level managers to effectively optimize maintenance planning. The literature also shows that current practices for maintenance cost segregation are highly inconsistent across the industry, resulting in a substantial gap in standardized frameworks.

This research contributes to existing literature by addressing above mentioned gaps by testing the application of ML techniques, specifically models such as ARIMA/SARIMA, ElasticNet, XGBoost, and Long Short-Term Memory (LSTM)

networks. These models have demonstrated their ability to handle linear and nonlinear time series, capture seasonality and trends, and provide more accurate short-term forecasts when they are compared with the traditional approaches discussed. The use ML techniques introduces a data-driven approach to preventive maintenance cost estimation.

The literature analysis, along with some practical considerations, indicate that XGBoost and LSTM models may offer superior forecasting capabilities, but their implementation still needs a balance of feature engineering, accuracy, and computational efficiency.

Furthermore, this work aims to bridge the gap between high-level financial planning and operational-level maintenance management through a detailed cost forecasting framework incorporating both historical plant operation and maintenance performance data to provide useful insights for both budgeting and decision-making.

Chapter 3—Methodology

3.1 Introduction

This chapter introduces the methodological framework used to forecast preventive maintenance costs in thermal power plants by applying ML techniques, detailing the procedures and strategies used to develop accurate predictive models.

Section 3.2 is an overview of the Praxis Preventive Maintenance Cost Estimation. The method in **Section 3.3** covers the typical maintenance cost allocation structures and the allocation of preventive maintenance. **Section 3.4** describes the data collection process. **Section 3.5** covers data preprocessing, and **Section 3.6** explains the data exploratory analysis. Model development contained in **Section 3.7** describes the selection and configuration of the 11 models that cover different combinations of the four ML techniques. **Section 3.8** covers testing and validation techniques, while **Section 3.9** addresses comparative evaluations and ranking, followed by model selection criteria.

3.2. Praxis Maintenance Cost Estimation Method

This research proposes an estimation approach utilizing a stochastic method to predict power plant maintenance costs based on a set of historical plant features. The method integrates historical maintenance records, cost transactions related to the maintenance records, and plant process data typically available in industrial asset-intensive plants. This work trains machine learning models and generates preventive maintenance cost forecasts for a 12-month period. **Figure 3-1** summarizes the proposed cost estimation method workflow. Data is collected from different technological sources available in the power plant, such as enterprise resource planning (ERP), computerized

maintenance management system (CMMS), distributed control system (DCS), process control system, and energy meters. It is stored in Excel files and loaded into a Python file for filtering, cleaning, combining, and processing to prepare inputs to the different models. Each model performs its feature engineering and applies model hyperparameter configurations to forecast the next 12 months of preventive maintenance.

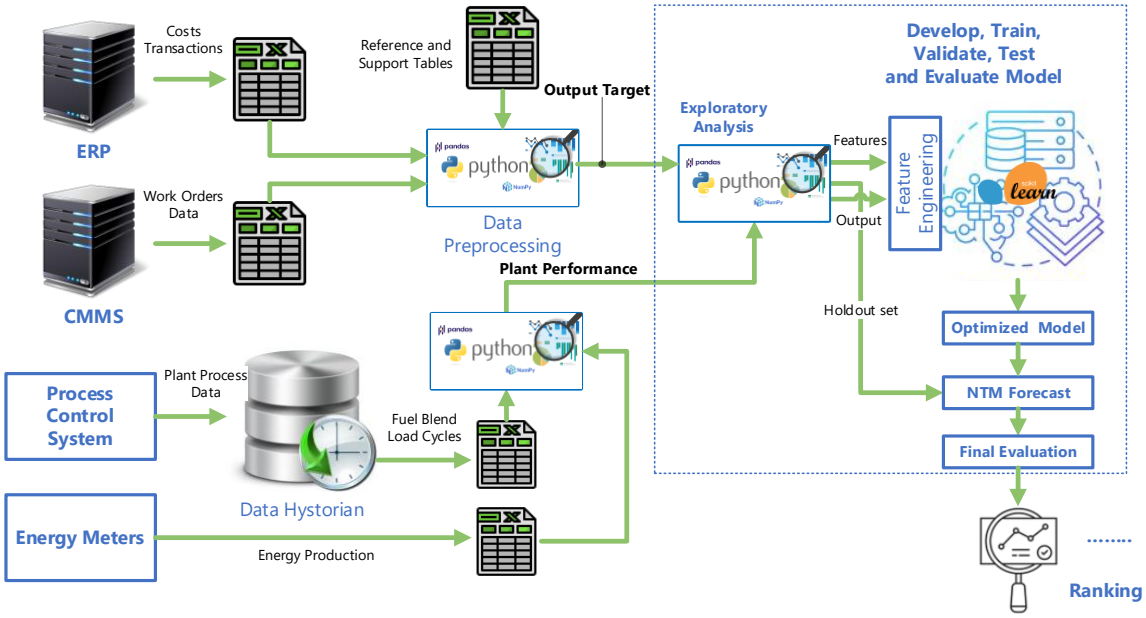


Figure 3-1. Praxis Estimation Method Workflow.

3.3. Preventive Maintenance Cost Allocation

Maintenance costs can be organized into different groups, as suggested by Galar et al. (2017). This research has located the preventive maintenance cost category within the Maintenance Type Class, as shown in **Figure 3-2**.

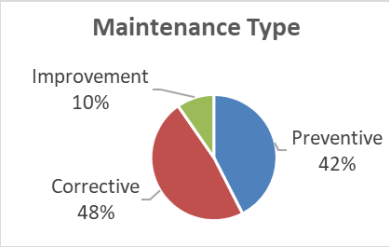


Figure 3-2. Power Plant Maintenance Type Category Cost Structure.

The transactions used for this structure exclude costs due to production losses, image damage, CAPEX, and major overhauls (or one-time major repair events) to minimize outliers and noise. This arrangement supports middle management in identifying cost patterns, planning future expenses effectively and focuses on the forecasting of the preventive maintenance category under the maintenance type class.

3.4 Data Collection

The data used in this research was collected from a comprehensive set of internal systems used in the thermal power plant under study, including the ERP system, a CMMS, and a Process Data Historian System. Here are the data sources:

1. **ERP System:** The ERP system supplied data on cost transactions, providing granular details on individual accounting records related to maintenance expenditures.
2. **CMMS System:** The CMMS system provided maintenance work order records, capturing information on maintenance activities, asset hierarchy, and cost allocations.
3. **Process Data Historian System:** This system delivered plant performance data, which were integrated as features in the predictive models.

The overall data flow process to prepare the Output Target data file for the models is shown in **Figure 3-3**, and a detailed Output Target data tables map is given in **Figure 3-4**.

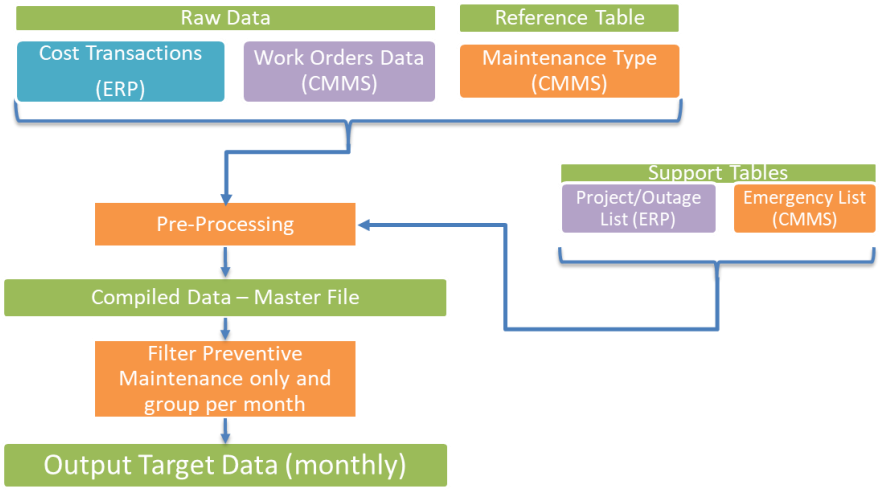


Figure 3-3. Output Target Data Flow Chart.

Each data source was characterized to identify the relevant fields and records. A detailed characterization of the data sources, fields used, and the rationale for their inclusion is presented in **Tables 3-1** through **3-4**.

3.4.1 Cost Transactions and Work Orders – Raw Data

Cost Transactions and Work Orders (WO) reports are automated on-request from the ERP and CMMS, respectively, to generate Excel lists containing the individual accounting records of each cost to the company and records of each maintenance work order executed at the plant, along with other related activities. The Cost Transactions and Work Orders data characterizations are summarized in **Table 3-1.** The numbers shown on the Used Records refer to the number of rows after filtering out non-related cost centers and outside the time of study. **Appendix A, Table A-1** displays the details of columns used for Cost Transactions, while Appendix A, Table A-2 displays the columns used for the Work Orders.

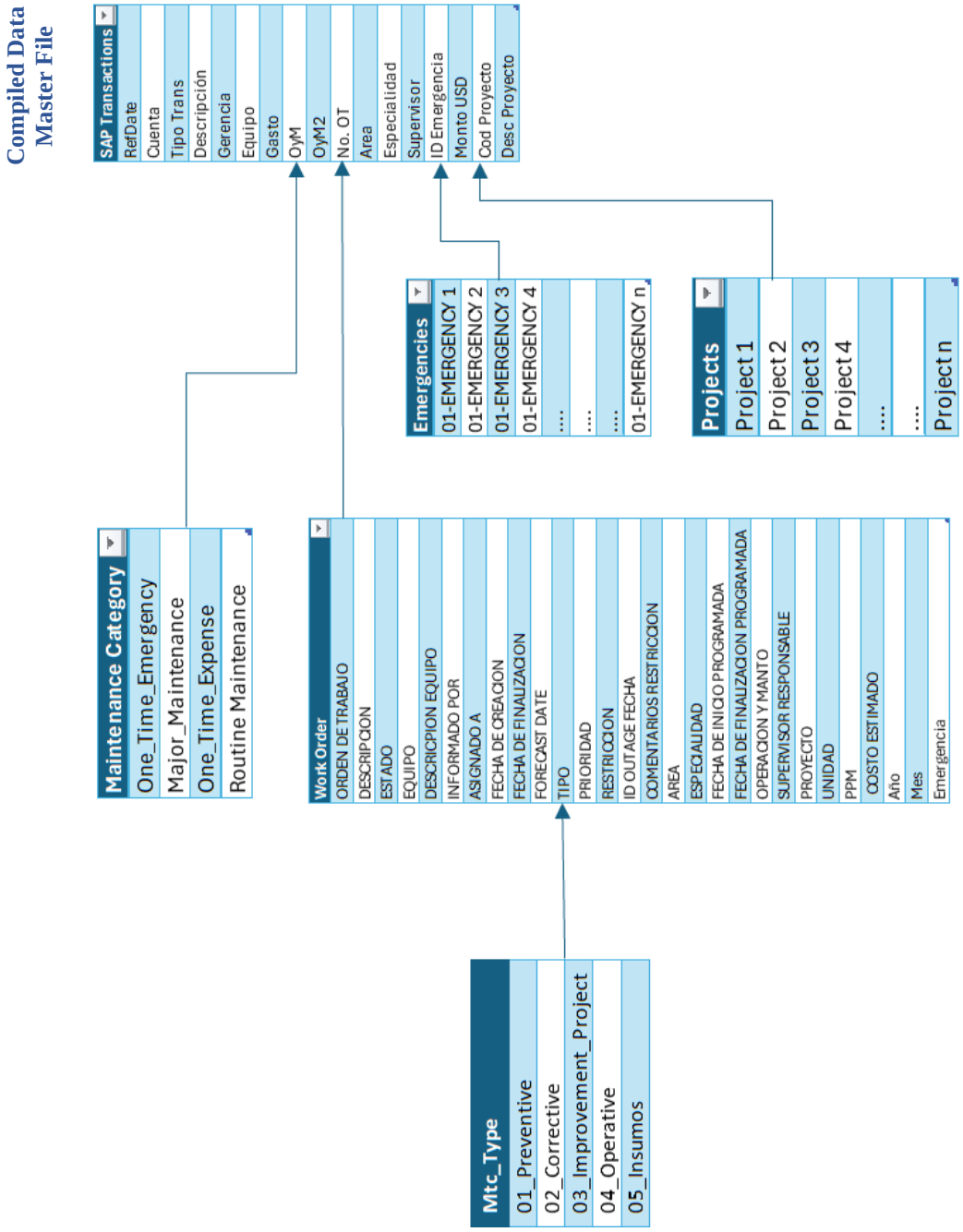


Figure 3-4. Preventive Maintenance Costs (Output Target) Data Workflow Map.

Table 3-1. Data Characterization for Cost Transactions and Work Orders

Id	df1	df2
Name	Cost Transactions	Work Orders
Source	ERP	CMMS
Brief Description	The transactions report contains all historical transactions IN/OUT registered as expenses.	The Work Order (WO) list contains all maintenance history information related to each asset in the plant.
Format	Excel File	Excel File
Total Columns	17	28
Used Columns	6	2
Total Time Span	Aug 2016 - Aug 2024	Aug 2016 - Aug 2024
Used Time Span	Mar 2018 - Jul 2024	Mar 2018 - Jul 2024
Total Records	170,256	106,594
Used Records	88,086	106,954

3.4.2 Reference Data Tables – Maintenance Type

Maintenance Type data characterization is summarized in **Table 3-2.** This table has been used as a key to filter out only preventive maintenance cost transactions.

Table 3-2. Data Characterization for Maintenance Cost Classes

Name	Maintenance Type
Source	CMMS
Brief Description	Preventive, Corrective, Improvement. Support for filtering preventive maintenance cost transactions
lFormat	Excel Table
Total Columns	2
Used Columns	2
Total Time Span	Up to date
Used Time Span	Up to date
Total Records	15
Used Records	15

3.4.3 Support Data Tables

Additional data tables are used to support filtering out data that represent outliers and one-time maintenance costs, as discussed in Section 3.3. Table 3-3 summarizes the characterization and use of support data tables.

Table 3-3. Data Characterization for Support Tables

Name	Projects List	Emergency List
Source	ERP	CMMS
Brief Description	List of CAPEX and OPEX projects. Support to filter out one-time expenses and major maintenance expenses.	List of declared emergencies attended. Support to filter out one-time emergencies.
Format	Excel Table	Excel Table
Total Columns	21	9
Used Columns	4	2
Total Time Span	Up to date	Up to date
Used Time Span	Up to date	Up to date
Total Records	81	109
Used Records	81	109

3.4.4 Features (Predictors) Data - Plant Performance

The plant performance is extracted from two sources: the plant data historian system via an hourly averages Excel file and the energy meters that are interrogated and downloaded to an Excel file. The files are merged into a single data frame and processed to generate daily and monthly performance indicators. Then, a processed Excel file with features columns is returned. This work has used the performance metrics described in the hypothesis as follows: Plant dispatch, Load Cycling coefficient, Plant fuel blend, Equivalent Planned Outage Factor (EPOF), and Equivalent Forced Outage Factor (EFOF)

factors. The overall data flow process for preparing the features data file is illustrated in **Figure 3-5,** and the data characterization is summarized in **Table 3-4.**

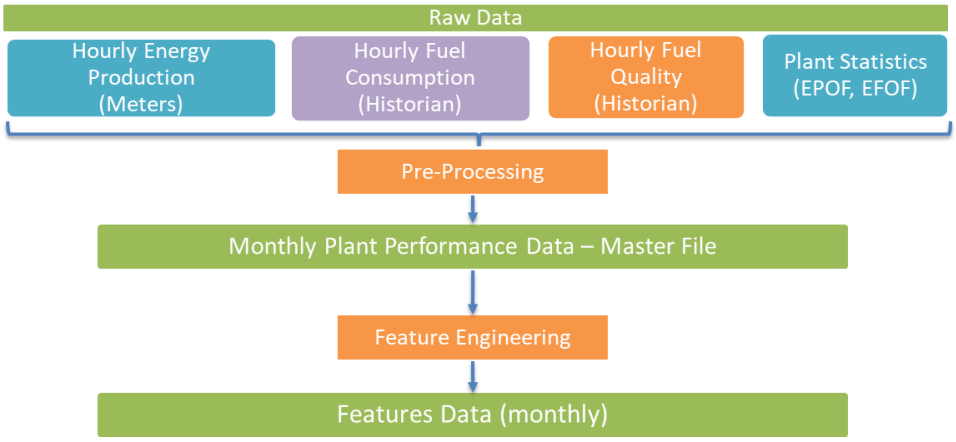


Figure 3-5. Features Data Flow Chart.

Table 3-4. Data Characterization for Plant Performance

Name	Monthly Plant Performance
Source	Plant Data Historian System and Energy Meters
Brief Description	Table with monthly data of features like Fuel Blend, Plant Load Factor, Plant Load Cycles, EPOF, EFOF, Month, Year.
Format	Excel File
Total Columns	5
Used Columns	5
Total Time Span	Mar 2018 - Dec 2023
Used Time Span	Mar 2018 - Dec 2023
Total Records	77
Used Records	77

3.5 Data Pre-Processing

3.5.1 Software Resources

Data extraction from ERP, CMMS, and Plant Data Historian systems used Excel (Microsoft, 2024) with preprocessing and feature engineering conducted using Python

(Python Software Foundation, 2023) through the Jupyter Notebook interface (Anaconda, Inc., 2023). Key packages leveraged included Numpy (Harris et al., 2020), Pandas (Pandas Development Team, 2020), and Seaborn (Waskom, 2021).

The model's hyperparameters tuning for performance optimization is executed using the Optuna Framework (Optuna Development Team, 2019) or GridSearchCV (Pedregosa et al., 2011). Whenever SearchCV is used, a comparative will be done with Optuna results.

3.5.2 Data Cleaning

The data cleaning process ensures that cost transactions and work orders fields used in the analysis were relevant and accurately represented within the study period. The initial step involved filtering out rows that fell outside the defined study period, from March 1, 2018, to July 31, 2024. Non-related cost centers were excluded, and unnecessary columns were removed to streamline the dataset. Data types for each column were configured to ensure compatibility and consistency for further analysis.

Work orders were refined by removing those outside the study period. Categories unrelated to maintenance activities, such as purely operational, logistics, and administrative, were also eliminated.

3.5.3 Data Filtering

Data filtering minimizes the impact of outliers that could skew results. Entries such as one-time emergencies, significant capital or operational expenditures (CAPEX and OPEX projects), and major maintenance activities, which could introduce variability uncharacteristic of routine maintenance costs, were excluded. The maintenance type was

also narrowed, filtering out non-preventive maintenance work orders such as corrective, improvement, etc., so only preventive maintenance costs were included.

3.5.4 Missing Values

To address missing data, missing cost values labeled as “NaN” (Not a Number) were replaced with zeros. Missing maintenance types were reconstructed by applying data mining techniques and cross-referencing with other fields, such as Supervisor Name or System assignments linked to the work orders. This approach helped to enrich the dataset and preserve data integrity.

An additional preprocessing step was required for transactions before Jan 2021, where the Work Order Number was extracted from cost transaction descriptions using string functions in Pandas. This was done because the system did not have a specific field to refer to the Work Order prior to the mentioned date; therefore, the number was manually embedded in the cost transaction description.

3.5.5 Data Assembly and Merging

The process of reassembly and merging was conducted to create master data files for target and feature variables. To create the master data file for target variables, the cost transactions and work orders data were merged using the unique Work Order Number as the reference field. The merged data was then aggregated into monthly segments by generating a pivot table with year-month as the index and the required preventive maintenance cost as the column. The feature variables were compiled from various sources and integrated into a unified data file, which involved consolidating hourly or daily data into monthly segments to align with the periodicity of the target variables and ensure consistency in representing plant operations.

3.6. Data Exploratory Analysis

Before entering the data into the models, an exploratory data analysis was carried out to uncover key insights, detect issues, and guide model configuration. This includes but is not limited to the following:

Box plots are used to visualize outliers, as shown in the example in **Figures 3-6.** If required, the model development could consider performing a quantile selection to prevent extreme outliers from affecting the model.

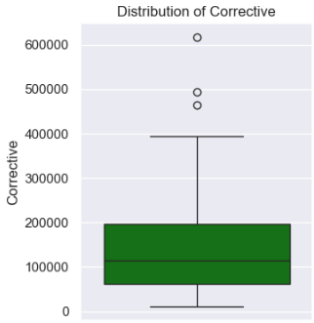


Figure 3-6. Box Plot Example.

Distribution plots help to understand the distribution of output targets and predictors. This includes the use of log transformation, quantile, and drift analysis between training and holdout sets as examples shown in **Figure 3-7**. Correlation matrixes (see **Figure 3-8**) will be used to identify highly correlated predictors and support feature selection. When necessary, multicollinearity analysis may be used.

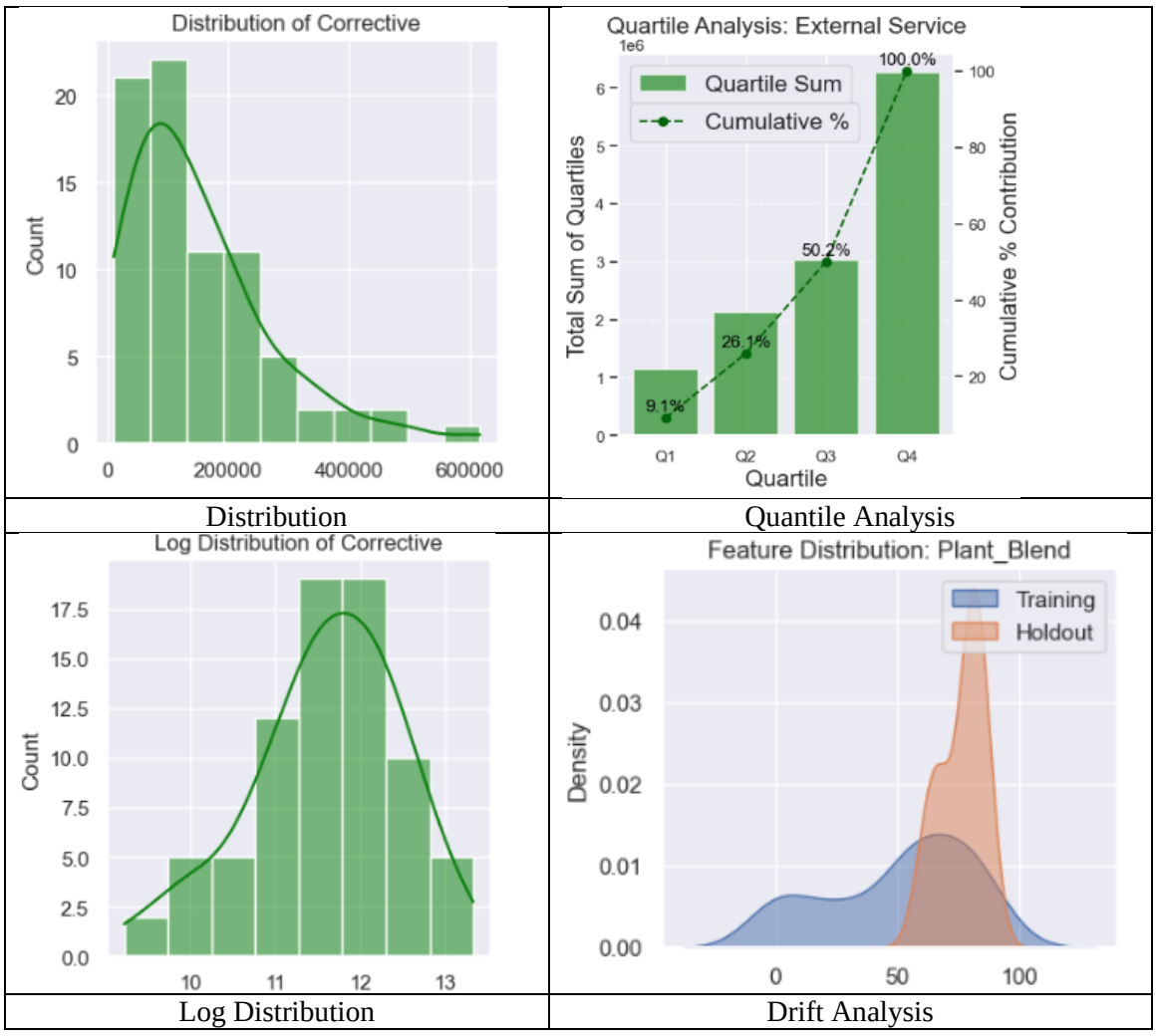


Figure 3-7. Sample Distribution Analysis Charts.

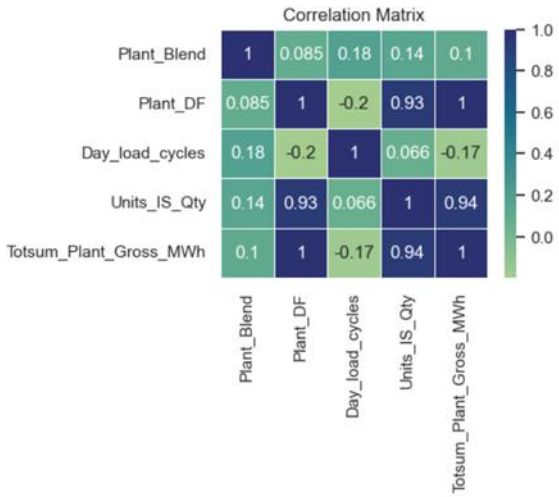


Figure 3-8. Sample Correlation Matrix.

For models sensitive to seasonality and stationarity, Augmented Dickey-Fuller (ADF) and Auto Correlation Function (ACF) tests are applied and summarized as shown in **Figure 3-9.** This analysis supports the decision of model hyper-parameters or the application of additional feature engineering steps.

Figure 3-9. Sample ADF and ACF Analysis.

3.7 Developing Models

The model development was structured to align with the requirements of the ML algorithm using feature and target data arrays defined by specific notations, as outlined below. The **Feature Data** array ***P′*** represents the attributes influencing maintenance

costs. The array consists of multiple features p_i and corresponding values p_{ij}. The feature data matrix is structured as follows:

$$P' = \begin{bmatrix} m_1 & p_{11} & p_{21} & \cdots & p_{i1} \\ m_2 & p_{12} & p_{22} & \cdots & p_{i2} \\ m_3 & p_{13} & p_{23} & \cdots & p_{i3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ m_j & p_{1j} & p_{2j} & \cdots & p_{ij} \end{bmatrix}$$

Where m represents each month on the study period from m_1 to m_j. It is used for indexing purposes. Variable p_i refers to each feature attribute, such as Plant Blend, Plant Dispatch Factor, Load Cycles, etc., from p_1 to p_r. The number of features depends on the feature engineering generated for each model. Variable p_{ij} represents the value of the i-th feature for the j-th month

The **Target Cost** array $\boldsymbol{C'}$ is the output preventive maintenance cost variable. The array consists of outputs C_1 for month 1 till C_j. The target data matrix is structured as follows:

$$C' = \begin{bmatrix} m_1 & C_1 \\ m_2 & C_2 \\ m_3 & C_3 \\ \vdots & \vdots \\ m_j & C_j \end{bmatrix}$$

Where m represents each month of the study period from m_1 to m_j which matches the month of the features array and is used for indexing purposes only. C_j represents the preventive maintenance cost for each month from C_1 to C_j.

For the Long Short-Term Memory (LSTM) model, only the Target Cost Array C' was used to train and generate the model because LSTM models are particularly effective in capturing temporal dependencies and trends within sequential data.

3.7.1 Data Split Sets

This study collected historical data from March 2018 until July 2024, splitting it into two sets: Test and Holdout. The test sets are additionally split according to the requirements of each ML model, while the holdout set comprises the last 12 months and acts as unseen data for all the models; this allows a common set for all models' performance evaluation. **Figure 3-10.** depicts the data set distribution of the case study.

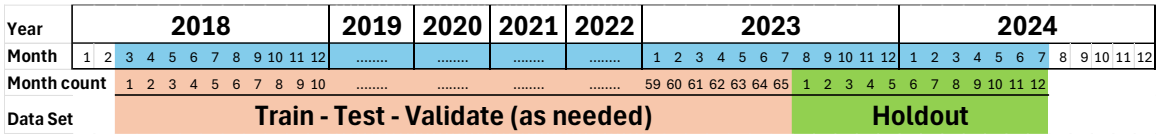


Figure 3-10. Datasets distribution

3.7.2 Single Target Regression Models

Due to the nature of the data, this research evaluated four machine learning techniques related to Time Series (ARIMA and SARIMA, including variations with and without log transformation), linear regression (ElasticNet with time series validation, k-fold cross-validation, and walk-forward validation), Decision Tree algorithms (XGBoost, one using GridSearchCV for hyperparameter tuning and another using Optuna), and deep learning methods (LSTM model-A training/test split with Recursive One-Step Forecasting and model-B with training/validation/test split with Multi-Step Forecasting), for a total of 11 models.

3.7.3 Time Series Models (ARIMA and SARIMA)

These models were chosen for their robustness in handling linear relationships and seasonal patterns in time-series data. The ARIMA and SARIMA time series model is shown in **Figure 3-11**.

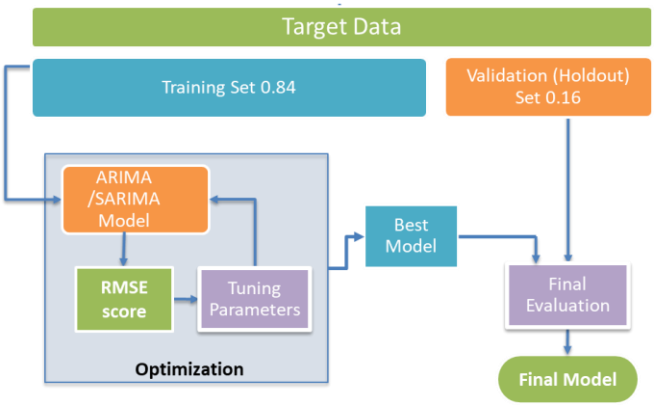


Figure 3-11. Time Series Model Flow Chart.

The data split ensures the training set (84% or 65-month data) has enough historical data to capture seasonal trends (5+ years). The validation (holdout) set reflects the most recent 1-year period (16%), testing the model's ability to generalize to future time steps. The tuning parameters range used for optimization of the models are listed in **Table 3-5**.

Table 3-5. ARIMA and SARIMA Proposed Tuning Parameters

Parameter	Range for ARIMA	Range for SARIMA
p (AutoRegression Order)	0 … 5	0 … 3
d (Differencing Order)	0 … 1	0 … 1
q (Moving Average Order)	0 … 5	0 … 3
P (Seasonal Autoregressive order)	NA	0 … 2
D (Seasonal Differencing order)	NA	0 … 1
Q (Seasonal Moving Average order)	NA	0 … 2
m (Number of periods per season)	NA	12

Both ARIMA and SARIMA models are tested in two versions each, with log transformation and without log transformation.

3.7.4 Linear Regression Model (ElasticNet)

The linear model passes through selection and regularization using the Shrinkage method (James et al., 2017), applying the Elastic-Net model function from SKlearn to obtain the L1 and L2-norm regularization of the coefficients. The overall linear regression model is shown in **Figure 3-12**.

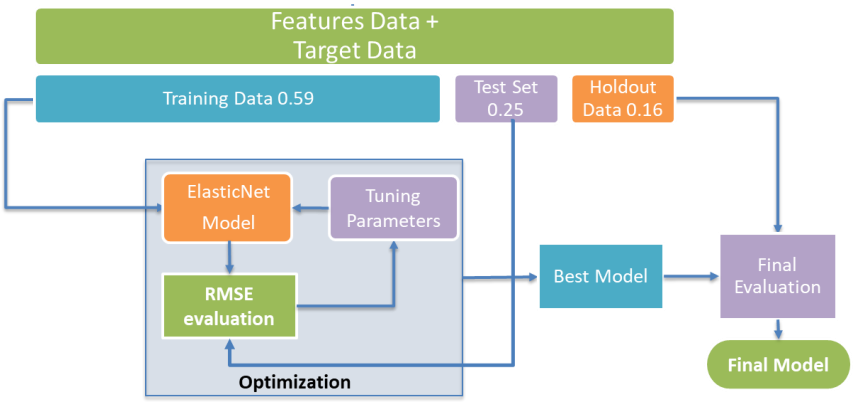


Figure 3-12. Linear Regression Model Flow Chart.

This model was selected for its ability to handle multicollinearity and perform feature selection through regularization. It was tested in three versions: time series, cross-validation, and walk-forward validation. The model configuration explored the Polynomial Features function to capture possible correlations between input values. The data split was selected so the holdout set would reflect the most recent 1-year period (16%), testing the model's ability to generalize and facilitate comparison to other ML models. The remaining 84% of was split into a 70/30 training/test split (equivalent to 59/25% of the total data set). The tuning parameter ranges used for all three models are listed in **Table 3-6**.

Table 3-6. Elastic Net Models Proposed Tuning Parameters

Parameter	Range
degree (Polynomial Order)	1 … 4
alpha (amount of regularization)	1e-4 … 1e2
l1_ratio (balance L1 (Lasso) and L2 (Ridge) regularization)	0.0 … 1.0
max_iter (maximum number of iterations)	1000 … 10000

3.7.5 XGBoost Regression Model

The non-linear regression tree implemented is XGBoost, which has demonstrated superiority over other decision trees and is recognized for high efficiency and flexibility. The overall XGBoost model process is shown in **Figure 3-13**. The data split keeps the holdout set reflecting the most recent 1-year period (16%), testing the model's ability to generalize and facilitate comparison to other models. The remaining 84% has been split into an 82/18 training/test split (equivalent to 68/16% of the total data set) with the aim of providing maximum training data.

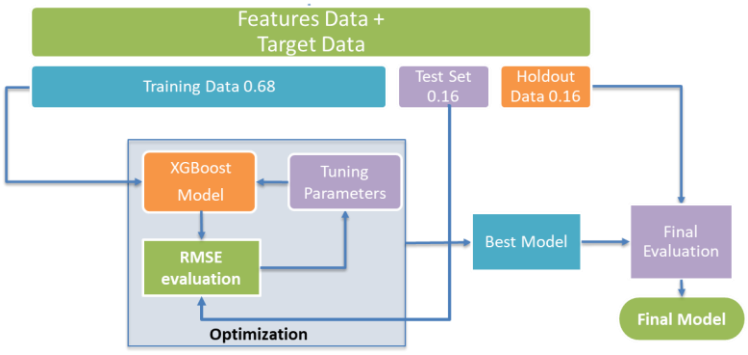


Figure 3-13. XGBoost Regression Model Flow Chart.

The model configuration explored two versions: one using grid search cross-validation (GridSearchCV) and the second using Optuna. Both models used reduced features according to the variance inflation factor (VIF) and data augmentation to improve generalization. Additionally, since the output variable is used within the feature

engineering, the last 12 months of output variable training set is used as input for forecasting. The hyperparameters used for tuning are listed in **Table 3-7**.

Table 3-7. XGBoost Model Proposed Tuning Parameters

Parameter	Range GridCV	Range Optuna
n_samples (Resample size) for data augmentation.	300	
n_estimators (number of boosting trees)	100, 300, 500	3000, 2000
learning_rate (how much the model adjusts per step)	0.01, 0.03, 0.05	0.0001, 0.1, log=True
max_depth (maximum number of iterations)	2, 5, 7	5, 10
reg_alpha (L1 regularization)	0.05, 0.1, 0.5	0.01, 1, log=True
reg_lambda (L2 regularization)	0.5, 1, 5	0.001, 10, log=True

3.7.6 LSTM Regression Model

The deep learning forecasting applying neural networks using LSTM has been selected due to its ability to handle nonlinearities, weak seasonality and non-stationarity. In this case, exogenous features are not used due to the nature of the time series, as defined by Peixeiro (2022).

The model configuration explored two versions, capturing the capabilities of LSTM aiming performance. The main differences between the models is shown in **Table 3.8**.

Table 3-8. LSTM Models' Main Differences

Description	LSTM Model-A	LSTM Model-B
Split	Training 0.68, Test 0.16, Holdout 0.16	Training 0.50, Validation 17, Test 0.17, Holdout 0.16
Validation	Not applicable	Time series cross-validation
Features	year, month sin/cos, rolling average for 3,6,12 months, Lags for 3,6,12 months, event_spike	year, month sin/cos, Rolling mean for 3,6, Rolling standard for 3, 6, Rolling exponential-median- average, months, Lags for 1, 3 months
Transformation	Log transformation	Log transformation Clip to prevent extreme low or high values
Normalization	Standard scaler	Standard scaler
Forecasting	Multi-Step Forecasting	Recursive One-Step Forecasting

The LSTM model-A main structure is shown in **Figure 3-14**.

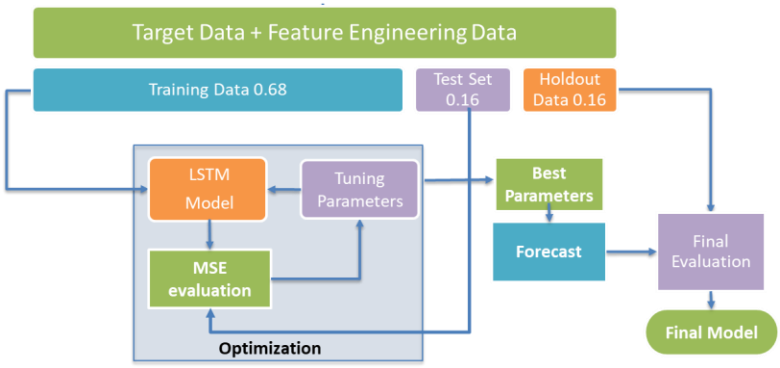


Figure 3-14. LSTM Forecasting Model- A Flow Chart.

The LSTM model-B main structure is shown in **Figure 3-15**.

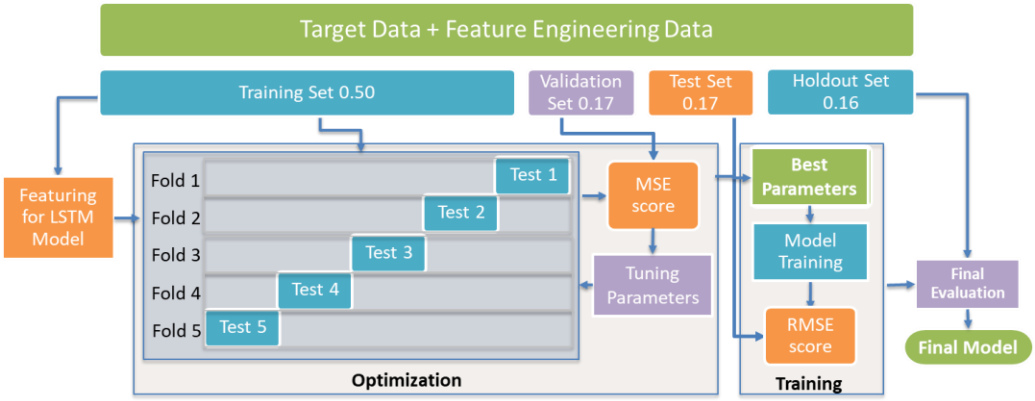


Figure 3-15. LSTM Forecasting Model-B Flow Chart.

The ranges of tuning parameters used for model optimization are in **Table 3-9**:

Table 3-9. LSTM Model Proposed Tuning Parameters

Parameter	Ranges Model-A	Ranges Model-B
n_units (neurons per layer)	32, 128, step=32	10, 50
n_layers (depth of LSTM network)	1 … 3	
dropout (random drop to prevent overfitting)	0.1 … 0.5	0.2, 0.4
lr (learning rate)	1e-4, 1e-2, log=True	5e-2, 2e-1
n_trials (L1 regularization)	20 … 100	20
epochs (L2 regularization)	50	60
batch_size (samples per training step)	8, 16, 32	32, 96

3.8 Models Evaluation and Ranking

This research implements an evaluation using a combination of mean absolute error (MAE), root mean squared error (RMSE), and mean absolute percentage error (MAPE). The coefficient R2 is not considered because the skewness of the target data made this parameter unreliable. Additionally, spider plots, as suggested by Korneva (2020), support and ensure a holistic performance assessment of each model.

The evaluation is complemented by the analysis of error plots and predicted vs. actual plots, which are useful for detecting bias, heteroscedasticity, systematic errors, generalization capacity and comparing results against a perfect fit. **Figure 3-16** shows an example of model error plots.

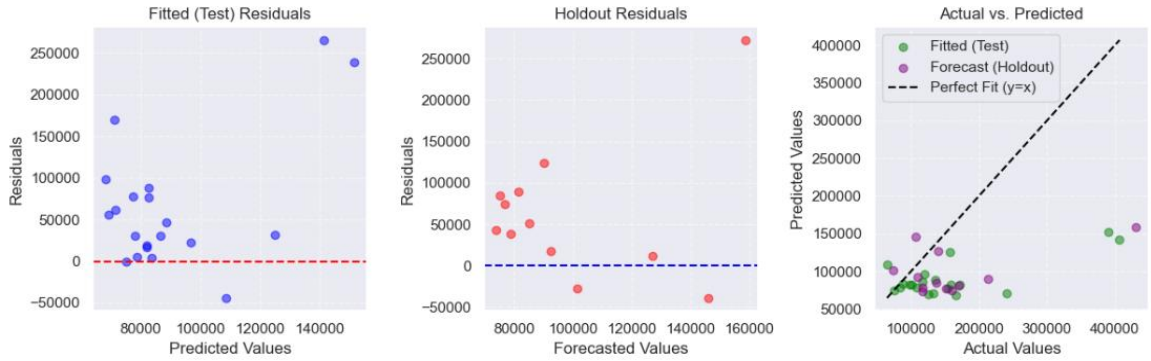


Figure 3-16. Residual Analysis and Actual vs. Predicted Comparison Plots.

The models that use the multiple input data sets (plant performance values as well as featured values) will find the most significant influencers on preventive maintenance costs using the SHapley Additive exPlanations (SHAP) value plots, as shown **in Figure 3-17**, or the Feature Importance plot, as shown in **Figure 3-18**.

Figure 3-17. SHAP Values Plot example.

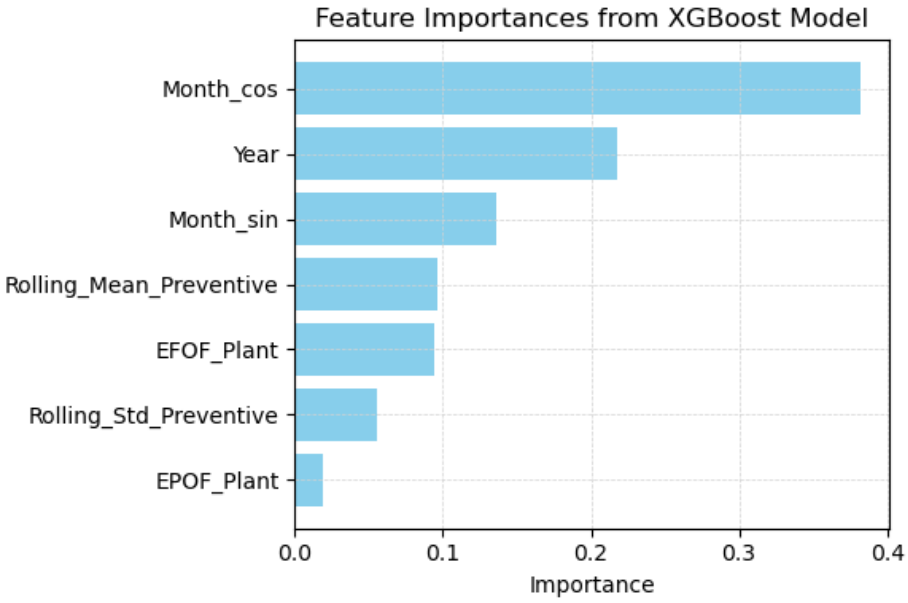


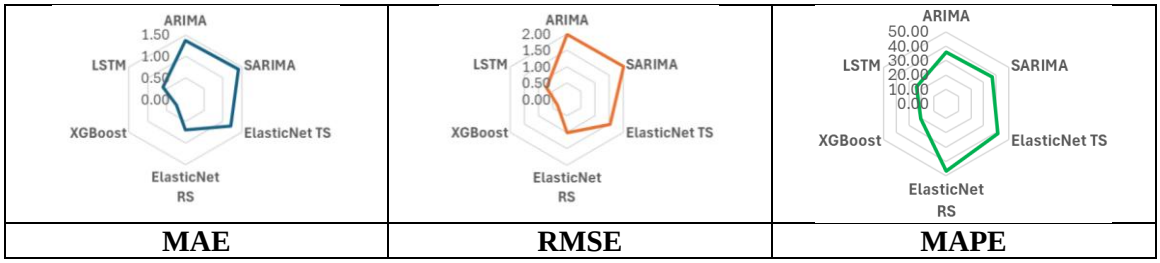
Figure 3-18. Feature Importance plot example.

Model size and computing time are not included in the evaluation criteria, as the primary focus is on prediction accuracy and model reliability.

The model selected will achieve the best overall rank based on quantitative performance measures MAE, RMSE and MAPE after compounding the rank obtained from Test and Holdout sets evaluation. This shall ensure selecting the model(s) that ensure forecast consistency. To facilitate comparison, a model selection matrix and ranking, as well as spider plots, will be generated, as per examples in **Table 3-10** and **Figure 3-19**.

Table 3-10. Sample Ranking Matrix for RMSE

Model / Parameter	MAE	RMSE	MAPE	RANK
ARIMA	0.59	0.72	23.70	4
SARIMA				3
ElasticNet TS				6
ElasticNet RS				5
XGBoost				1
LSTM				2



MAE | **RMSE** | **MAPE**

Figure 3-19. Spider Plot Examples.

Chapter 4—Results

4.1 Introduction

This chapter presents the results derived from the application of various machine learning models to predict preventive maintenance costs in thermal power plants, as developed through the methodologies outlined in the previous chapter. The findings are required for understanding the effectiveness and practical implications of different predictive models within the context of thermal power plant maintenance management. In total, eleven models were generated in this study from the four different machine-learning techniques that were selected.

Section 4.2 provides an overview of data pre-processing results, detailing the steps taken to enhance data quality, quantity, structure, and noise reduction. **Section 4.3** presents an exploratory analysis of the dataset, examining the characteristics of both the target variable and the predictors used in the models. **Sections 4.4** address the hypotheses set forth in the research study, presenting empirical evidence to confirm or refute each hypothesis.

Section 4.5 presents a comparative analysis of the model´s results that support determining the most effective model(s) based on a set of performance criteria. Towards the end, **Section 4.6** concludes with the identification of the best performance model.

4.2 Data Pre-Processing Results and Observations

This section presents the results of the pre-processing steps and data analysis aimed at obtaining meaningful insights from the historical preventive maintenance costs.

4.2.1 Output Target Data

Table 4-1 shows the transformation of the number of records through data cleaning, filtering, and merging.

Table 4-1. Output Target Data Transformation Steps

Step	Table Name	Cost Transactions	Work Orders
	Used Time Span	Mar 2018 - Jul 2024	Mar 2018 - Jul 2024
0	**Initial Records**	**170,256**	**106,594**
1	After filtering out non-related cost centers and outside time span	88,086	105,398
2	After Merging with Work Orders and Removing Operational, Logistics, and Administrative Work Orders	85,407	
3	After Removing same Negative and Positive Transactions	73,287	
4	After Removing One-Time Emergencies, One-Time Expenses, Major Maintenance, and Transactions without Work Order	71,510	
5	Records After Filtering by Preventive Maintenance (**Compiled Data Master File**)	**26,109**	
6	Records After Grouping by Month **Output Target Data (monthly)**	**77**	

To visualize the output target data transformation (preventive maintenance costs), **Figure 4-1** shows the monthly maintenance cost history for the categories within the maintenance type class (Preventive, Corrective, Improvement, and Others) without any data filtering (data from Step 2), which illustrates the contribution of preventive

maintenance costs; **Figure 4-2** shows the preventive maintenance data alone. Note that negative values occur, normally on External Services, due to the accrual of finance transactions made at different times. **Figure 4-3** shows the filtered preventive maintenance data after completing step 6.

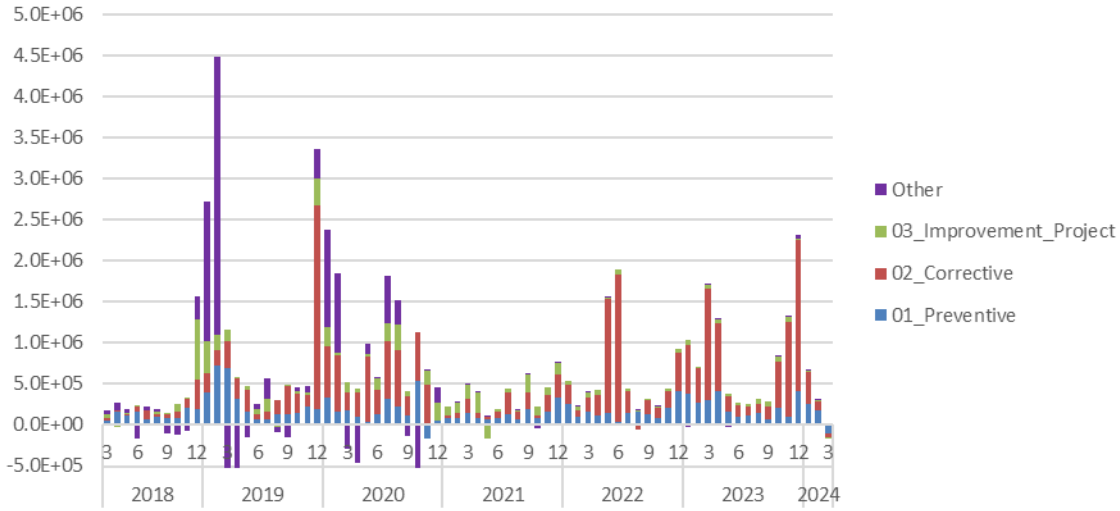


Figure 4-1. Original Data for Maintenance Type Costs Categories

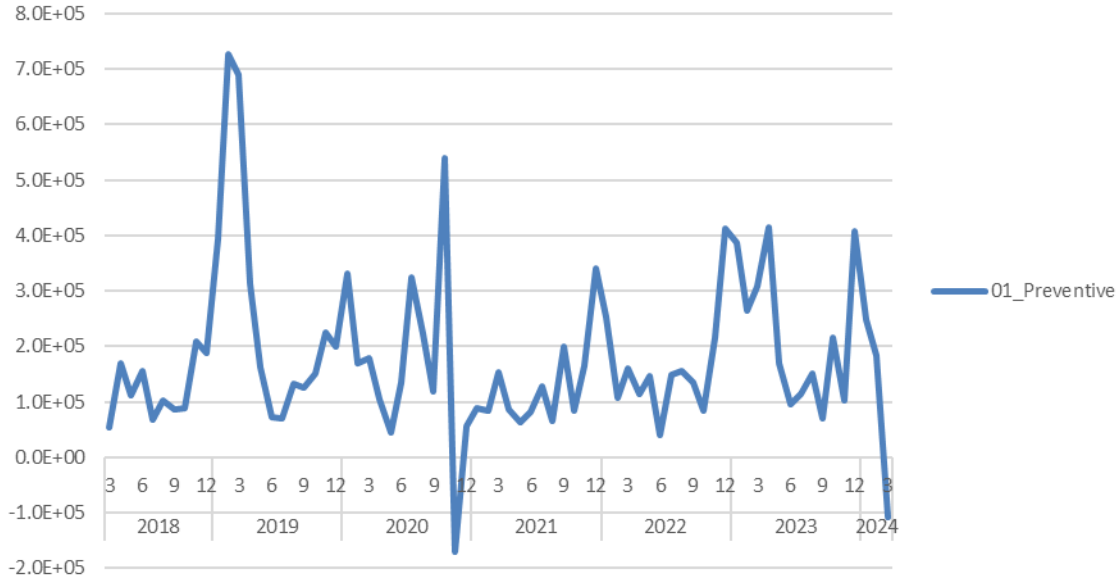


Figure 4-2. Original data for Preventive Maintenance Costs

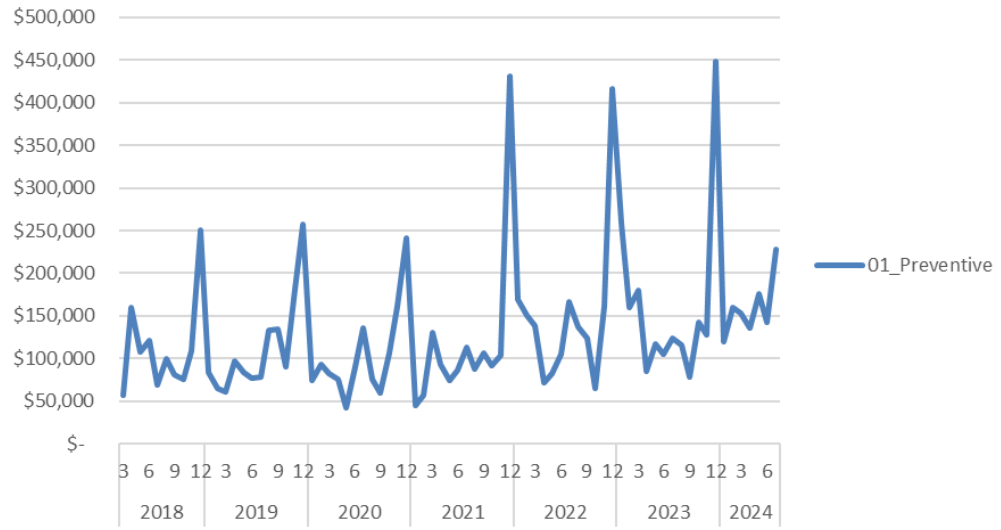


Figure 4-3. Preventive Maintenance Cost After Completing Data Preprocessing.

4.2.2 Features (Predictors) Data

According to the literature discussed in section 2, the predictors selected to be used as inputs of the ML models for this work are Plant Blend, Plant Dispatch Factor (Plant_DF), Equivalent Planned Outage Factor (EPOF), Equivalent Forced Outage Factor (EFOF) and day load cycles (Day_load_cycles). as major contributors, which are shown in **Figure 4-4**.

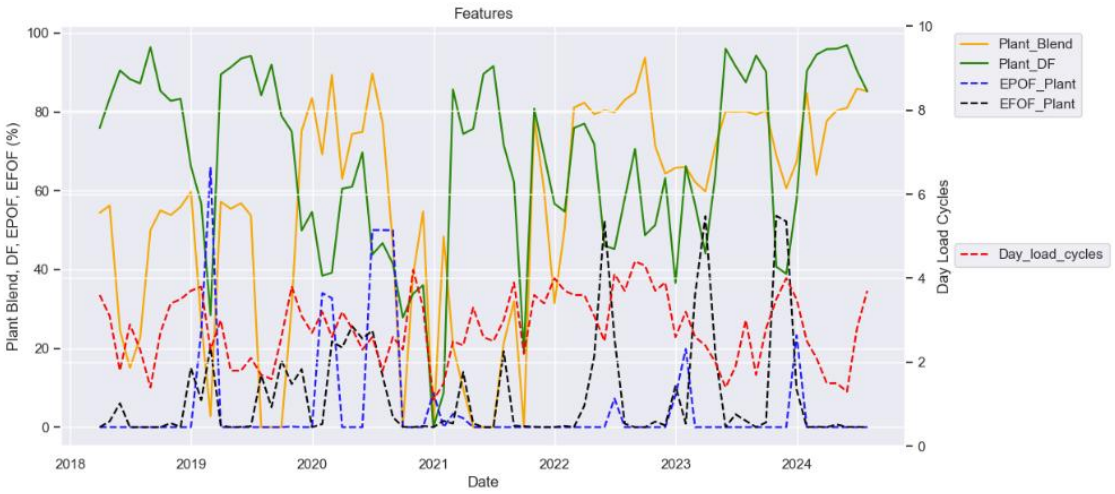


Figure 4-4. Features (Predictors) Data.

4.3 Data Exploratory Analysis

This section presents the results of data analysis aimed at obtaining meaningful insights from the historical preventive maintenance costs (Output Target Data) and the candidate features (Predictors).

4.3.1 Output Target Data Exploratory Results

A distribution analysis confirmed that preventive maintenance is a highly skewed distribution, as shown in **Figures 4-5** and **4-6.a,** due to the nature of maintenance costs. There are a significant number of transactions related to preventive maintenance costs on the low-cost side and fewer on the high-cost side.

A quartile contribution in **Figure 4-6.b** shows that quartiles 1 and 2 (where most of the cost transactions are allocated) contribute only around 30% of the total cost of preventive maintenance, while quartile four accounts for approximately 45% of the total cost. Therefore, those who appear as outliers are purposely not removed because they provide meaningful information about cost expenditures. The handling of this highly skewed data is done using a log transformation as shown in **Figure 4-7**.

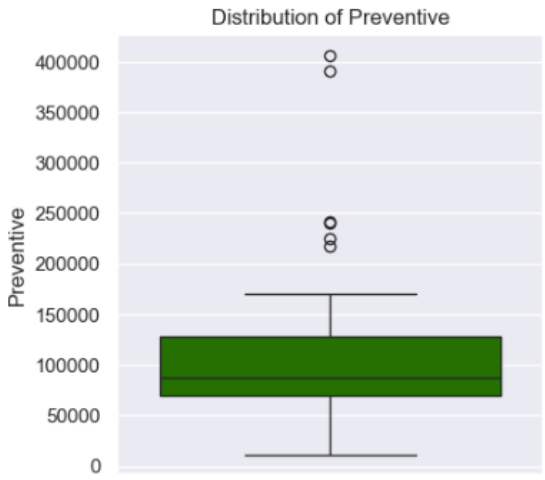


Figure 4-5. Preventive Maintenance Box Plot.

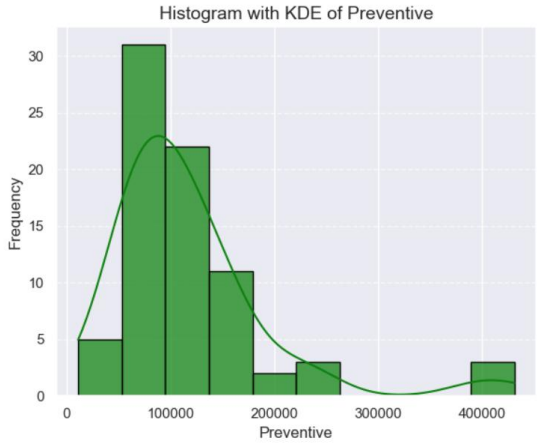


Figure 4-6.a. Preventive Maintenance Cost Distribution.

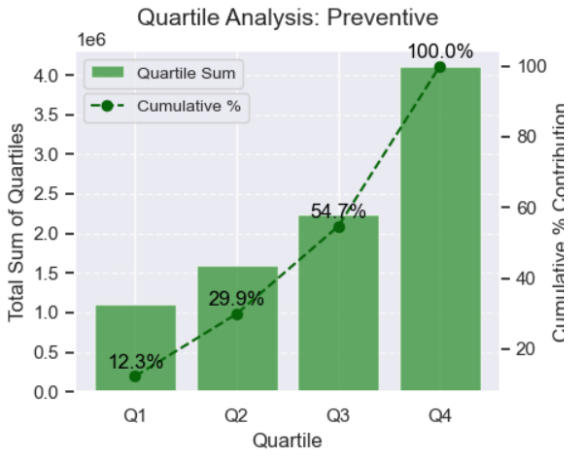


Figure 4-6.b Preventive Maintenance Quartile Analysis.

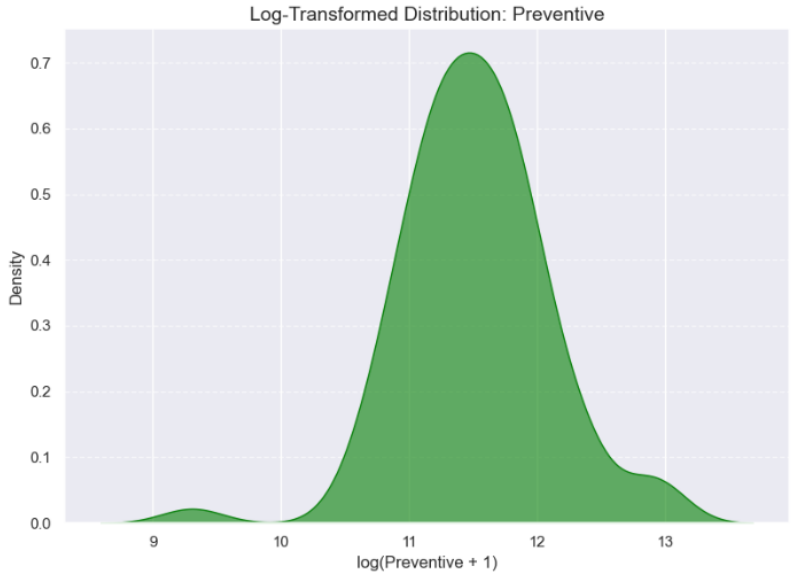
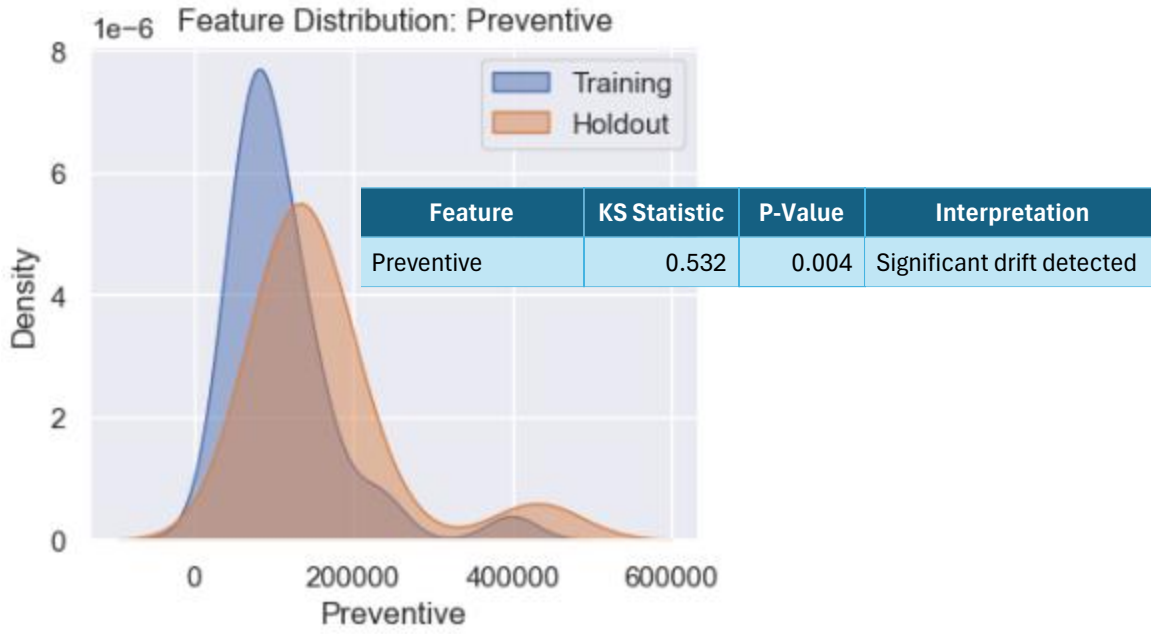


Figure 4-7. Preventive Maintenance Log Transformation Distribution.

The drift analysis reveals significant drifts between training and holdout sets on some features, as shown in **Figure 4-8.**



Feature	KS Statistic	P-Value	Interpretation
Preventive	0.532	0.004	Significant drift detected

Figure 4-8. Preventive Maintenance Drift Analysis.

Further correlation analysis of **Figure 4-9**. of preventive maintenance with each of the features indicate that there is not a strong correlation with any of the expected predictors.

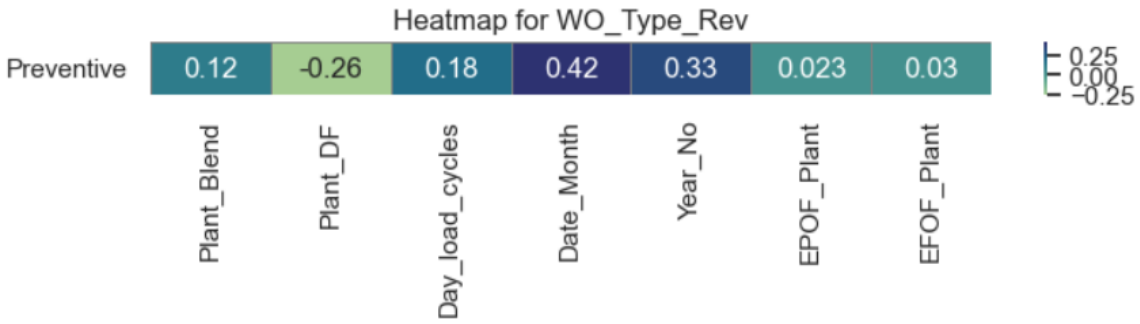


Figure 4-9. Preventive Maintenance Correlation with Features.

4.3.2 Features (Predictors) Data Exploratory Analysis

Figure 4-10 shows the distribution of the predictors. **Figure 4-11.a** shows the correlation heat map for the potential predictors indicating the high direct correlation

between Plant load Factor (Plant_DF) with the total generation output (Totsum_Plant_Gross_MWh) and the quantity of units in service (Units_IS_Qty).

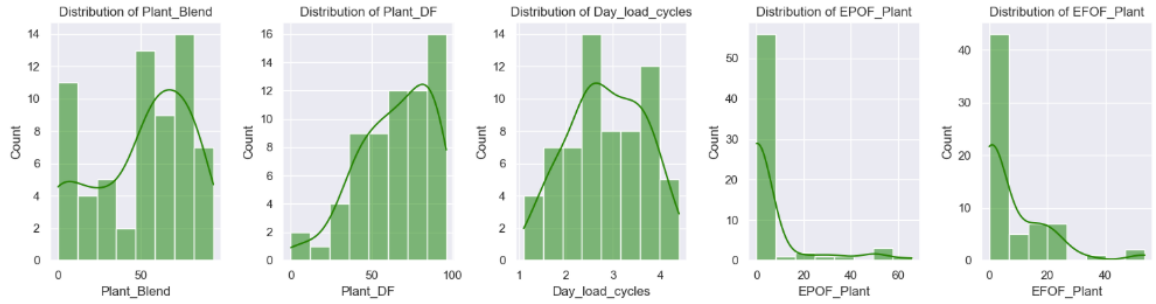


Figure 4-10. Distribution of Predictors.

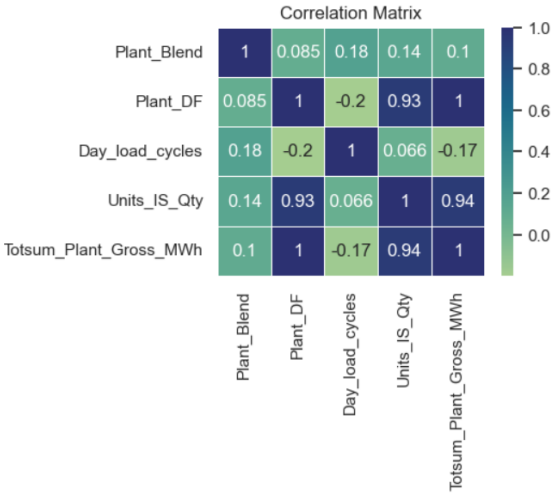


Figure 4-11.a. Initial Features Correlation

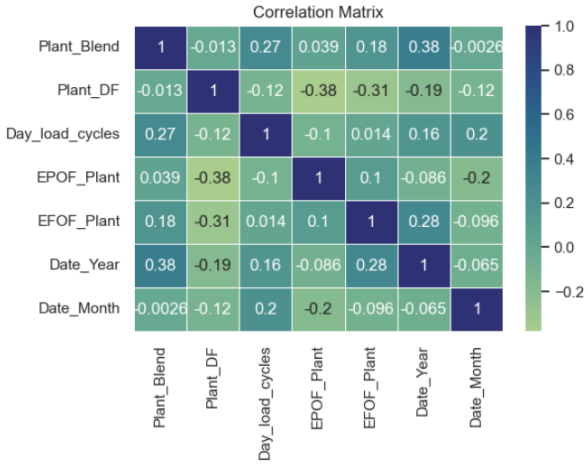


Figure 4-11.b. Revised Features Correlation

The historical trends of maintenance costs suggest the existence of seasonality; therefore, the variables of month and year are included. Furthermore, according to the literature discussed in **Section 2.3**. the values of the Equivalent Planned Outage Factor (EPOF) and the Equivalent Forced Outage Factor (EFOF) were added as possible features, as shown in **Figure 4-11.b**.

The features were further assessed for possible shifts in distribution between the training data and the holdout data. **Figure 4.12** shows a feature distribution analysis. Furthermore, a Feature Drift Analysis reveals significant drifts between training and holdout sets on some features, as shown in **Table 4-2**.

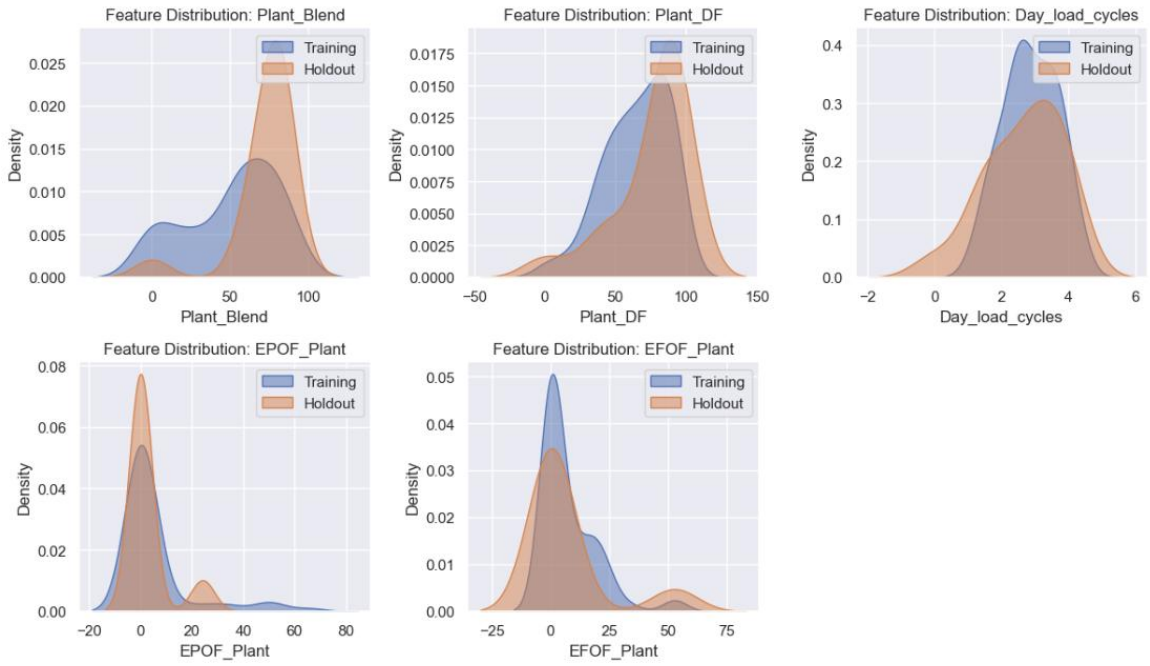


Figure 4-12. Predictors Testing and Holdout Distribution Analysis

Table 4-2. Predictors Drift Analysis

Feature	KS Statistic	P-Value	Interpretation
Plant_Blend	0.525792	0.000581	Severe drift detected
Plant_DF	0.441629	0.006701	Significant drift detected
Day_load_cycles	0.201810	0.564798	No significant drift
EPOF_Plant	0.097738	0.997120	No drift (Distributions are very similar)
EFOF_Plant	0.382805	0.027772	Moderate drift detected
Date_Month	0.203620	0.552201	No drift at all (perfect match)

The features with very low p-values (<0.005) have significantly different distributions in the training and holdout sets, which could impact model performance as the model might not generalize well to new (holdout) data. This suggests that the model might struggle to predict preventive maintenance costs accurately in the holdout phase. It also implies that robust models like tree-based algorithms (Random Forest, XGBoost) that handle feature shifts better could be a better option over simple linear regression models.

4.3.3 Understanding Time Series

Since ARIMA and SARIMA have been included in the analysis, considering that output data is time series and they are sensitive to seasonality and stationarity, an Augmented Dickey-Fuller (ADF), Auto Correlation Function (ACF) tests, and partial Auto Correlation Function (PACF) tests are included results are shown in **Figures 4-13 to 4-16**, representing the different output target variable transformations.

Significant lags were identified by computing the ACF values for the first 24 lags using the 95% confidence interval (alpha=0.05) and identifying the lags where the ACF values fall outside the confidence interval. The list of ACF values is summarized in **Table 4-3**.

Table 4.3. Statistically significant lags.

Preventive Maintenance

Lag	ACF Value	Lower CI	Upper CI	p-value
0	1.000000	1.000000	1.000000	NaN
1	0.286186	0.043082	0.529289	0.018237
2	0.020277	-0.241982	0.282537	0.060761
3	0.062190	-0.200162	0.324542	0.117940
4	0.039918	-0.223305	0.303140	0.200117
5	0.023750	-0.239830	0.287329	0.303505
6	-0.018325	-0.282031	0.245381	0.417304
7	0.009501	-0.254280	0.273282	0.532813
8	0.041713	-0.222088	0.305515	0.625679
9	-0.014565	-0.278756	0.249626	0.718833
10	-0.054696	-0.318935	0.209542	0.776515
11	0.107678	-0.157228	0.372585	0.767446
12	0.594850	0.327369	0.862331	0.000274
13	0.198995	-0.137710	0.535701	0.000150
14	0.024349	-0.319236	0.367935	0.000272
15	0.047117	-0.296570	0.390804	0.000451
16	-0.062938	-0.407007	0.281131	0.000688
17	-0.074549	-0.419298	0.270199	0.000973
18	-0.087396	-0.433096	0.258304	0.001265
19	-0.042747	-0.389751	0.304256	0.001913
20	-0.018408	-0.365722	0.328907	0.002948
21	-0.085064	-0.432437	0.262308	0.003636
22	-0.100527	-0.449128	0.248074	0.004063
23	-0.064327	-0.414637	0.285983	0.005324
24	0.297337	-0.053670	0.648344	0.000521

Figure 4-13. Preventive Maintenance Costs ADF Test and ACF + PACF plots.

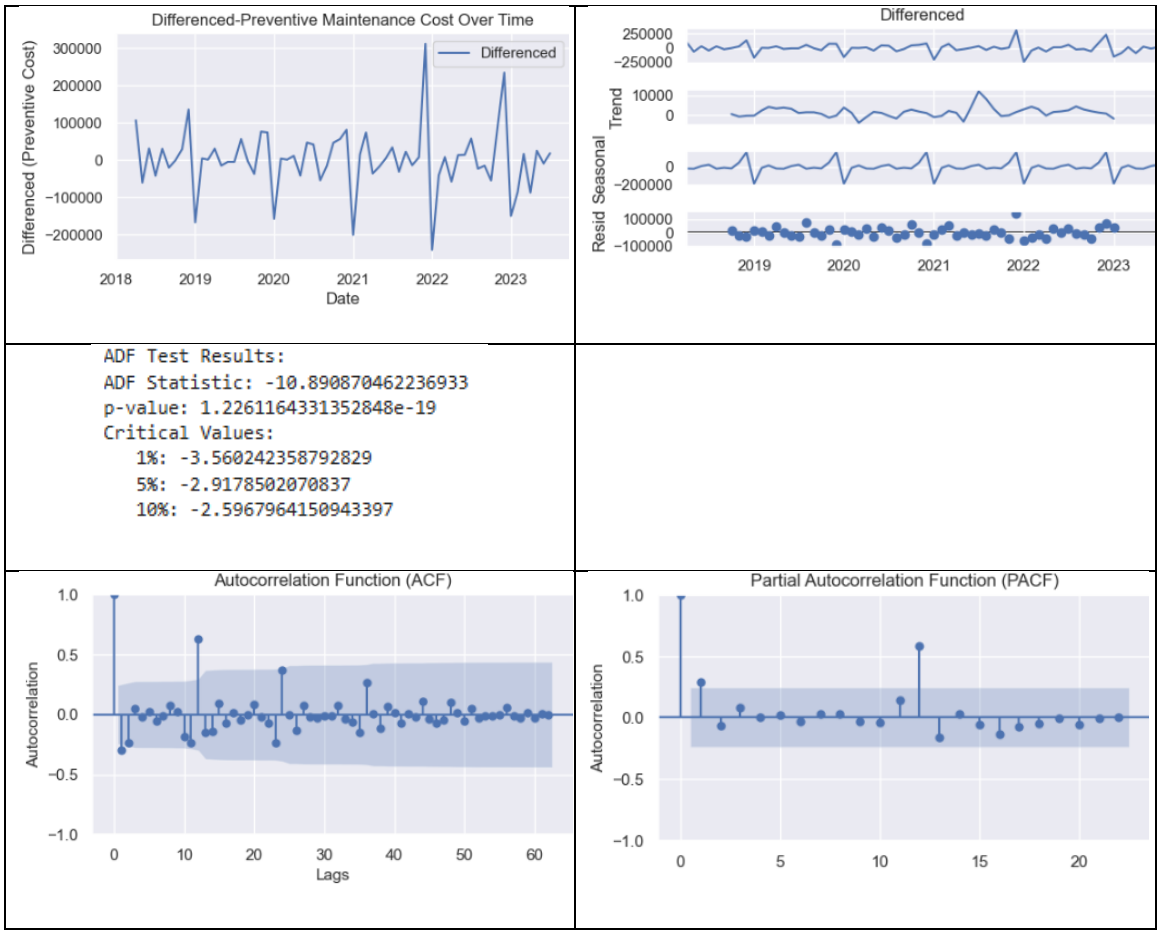


Figure 4-14. Differenced Preventive Maintenance Costs ADF Test and ACF + PACF plots.

Figure 4-15. Log-transformed Preventive Maintenance Costs ADF Test and ACF + PACF plots.

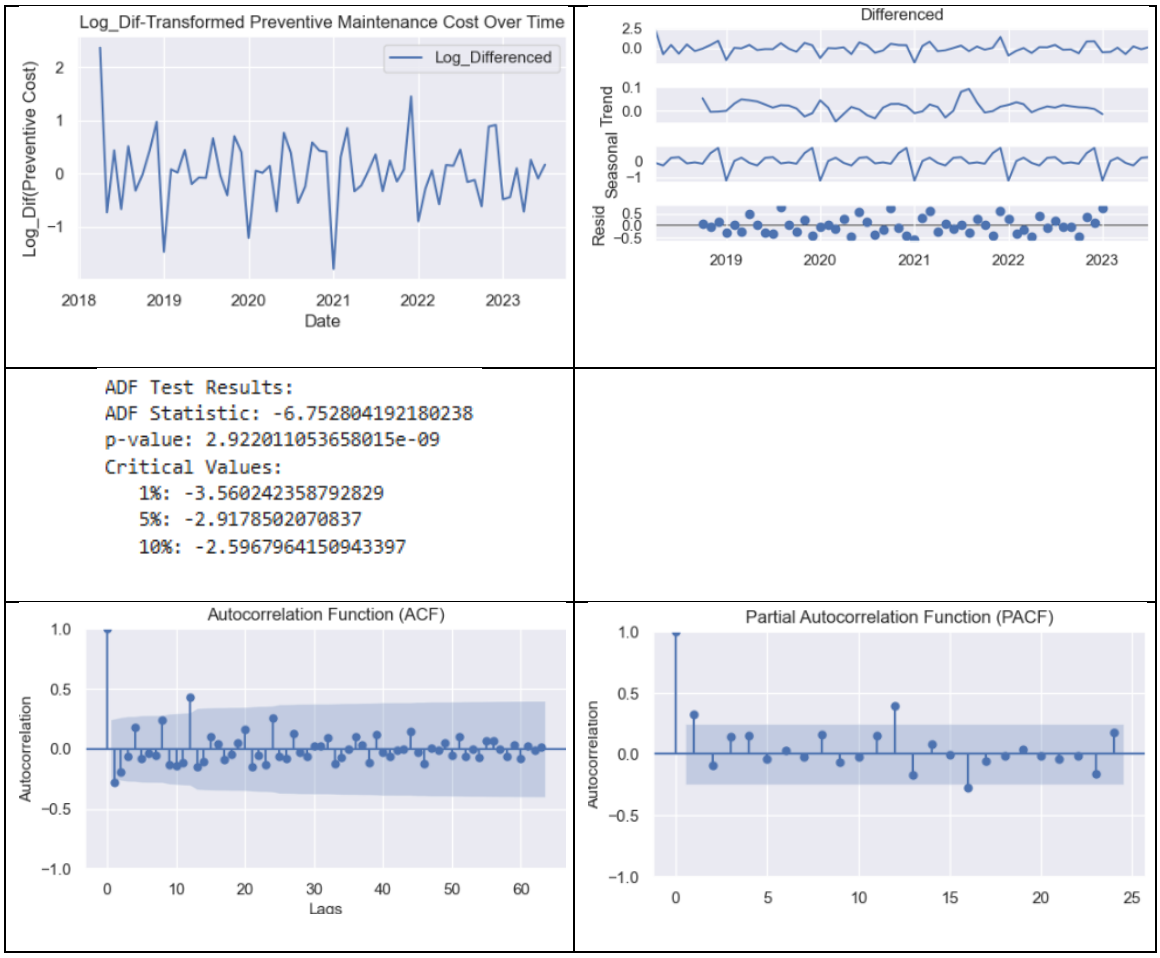


Figure 4-16. Dif+Log-transformed Preventive Maintenance Costs ADF Test and ACF + PACF plots.

4.4 Analysis of Preventive Maintenance Costs Models

Each model was configured and tested, fitting the main algorithm requirements in terms of engineered features, target metric, data split, and tuning to find optimized hyper-parameters. This section summarizes the results obtained for the preventive maintenance forecasting models that support answering each hypothesis.

4.4.1 Hypothesis 1 and Results

H1. *The Main Factors (Independent Variables) identified in this praxis significantly influence the maintenance costs of a thermal power plant*.

Since most models are regression models, the SHAP (SHapley Additive exPlanations) values analysis has been selected to provide insights into the contribution of individual features to the preventive maintenance costs forecasting model's (Lundberg & Lee, 2017). **Figures 4-17.a., 4-17.b** and **4-17.c.** show the top 10 most important features for ElasticNet models, while **Figures 4-18.a, 4-18.b** and **4-18.c.** shows the SHAP value plots for the same models, and **Figures 4-19.a and 4.19.b** display the important features for XGBoost models. ARIMA and SARIMA do not have contributions other than the target variable time series lagged features.

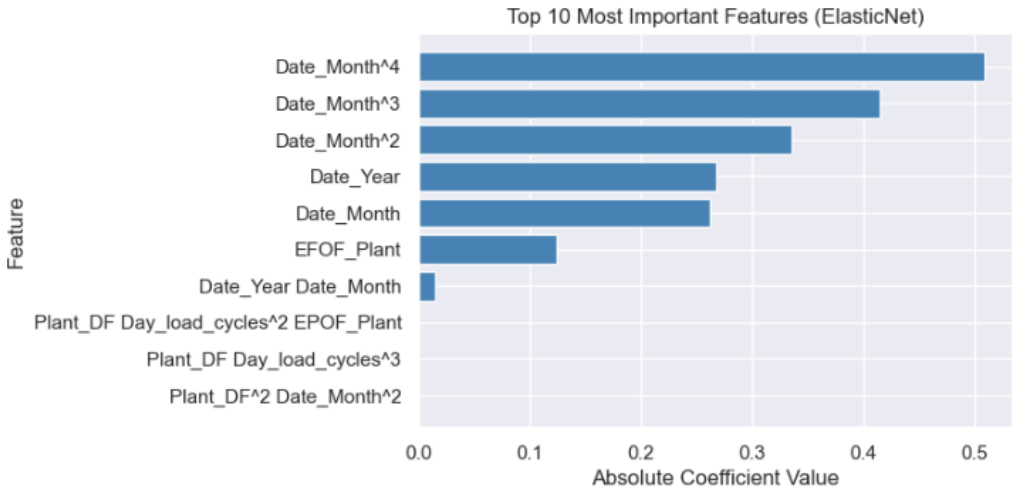


Figure 4-17.a. Top 10 Important Features – ElasticNet Time Series Validation

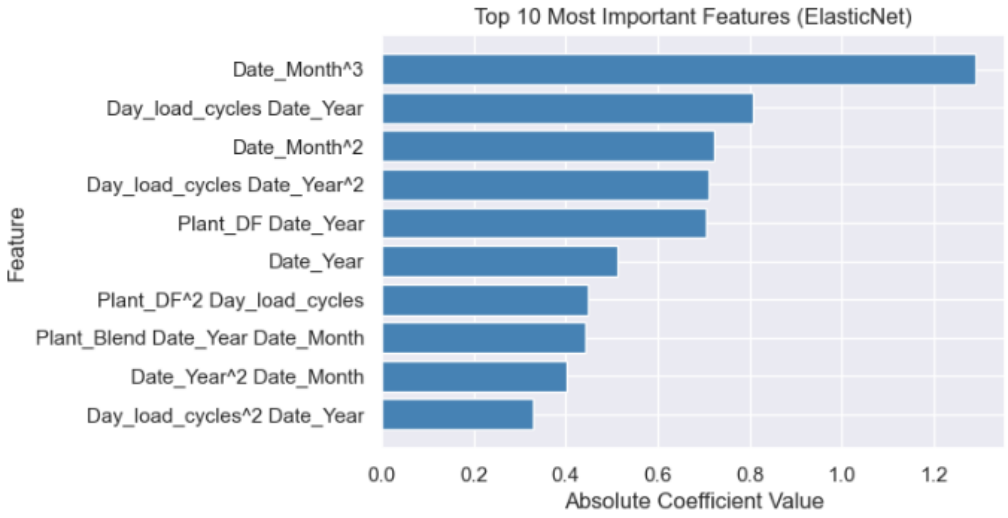


Figure 4-17.b. Top 10 Important Features – ElasticNet Random Series (K-fold validation)

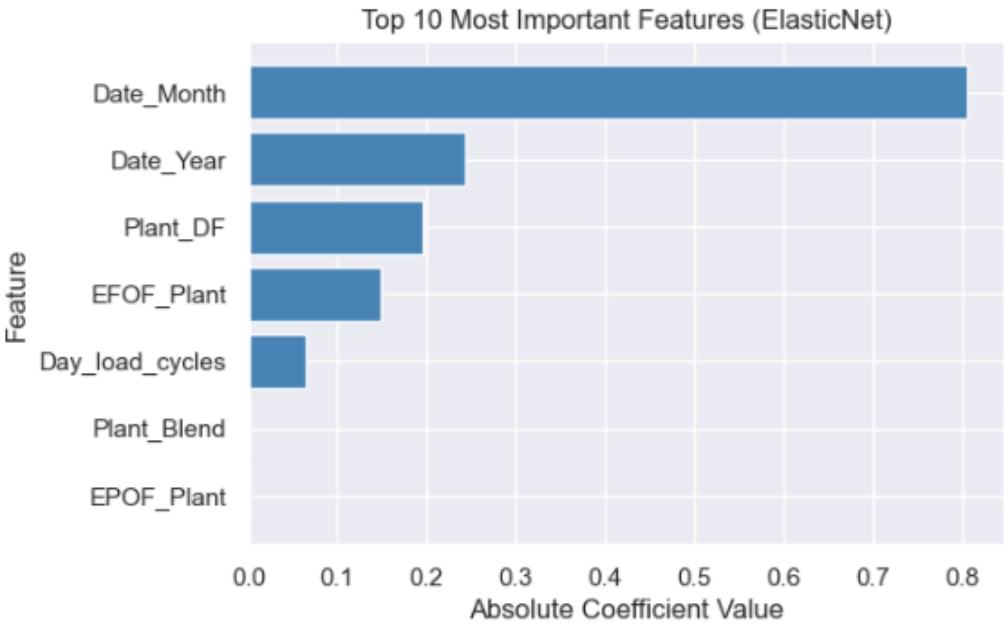


Figure 4-17.c. Top 10 Important Features – ElasticNet Random Series (Walk Forward Validation)

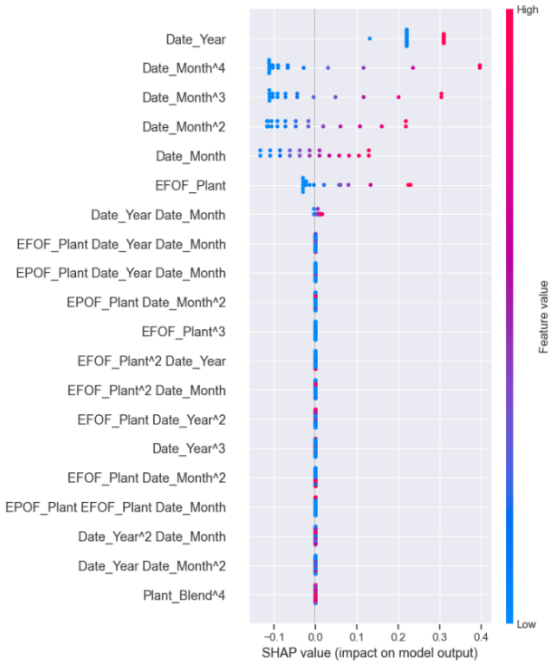


Figure 4-18.a. SHAP Values Plots for Elastic Net with Time Series Validation

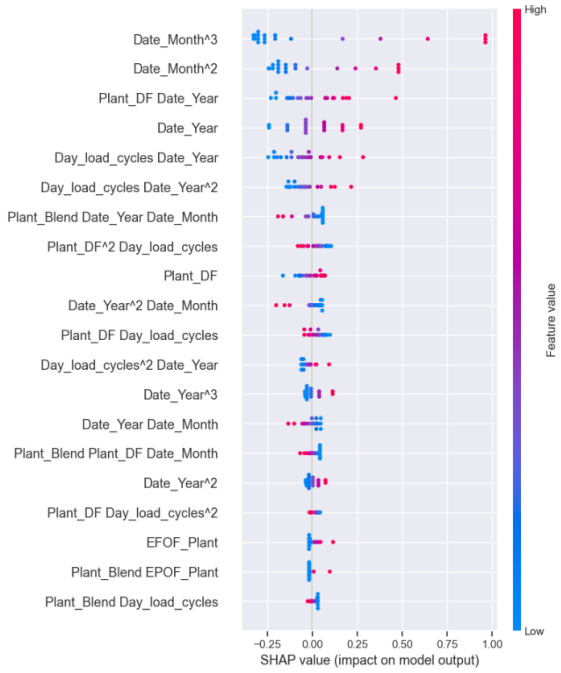


Figure 4-18.b. SHAP Values Plots for Elastic Net with KFold Validation

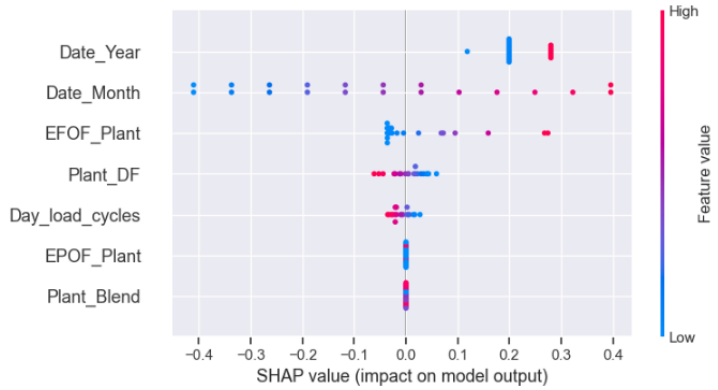


Figure 4-18.c. SHAP Values Plots for Elastic Net with Walk Forward Validation

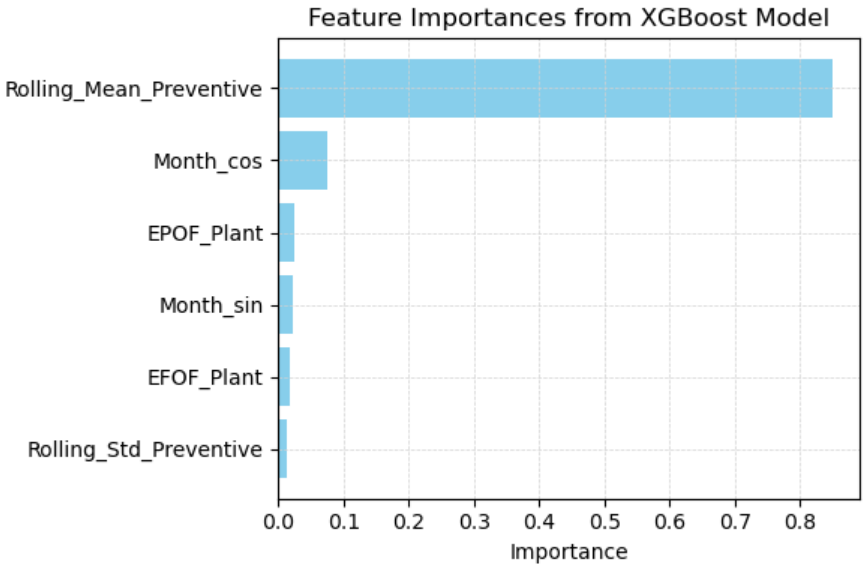


Figure 4-19.a. Important Features – XGBoost Model - GridSearchCV

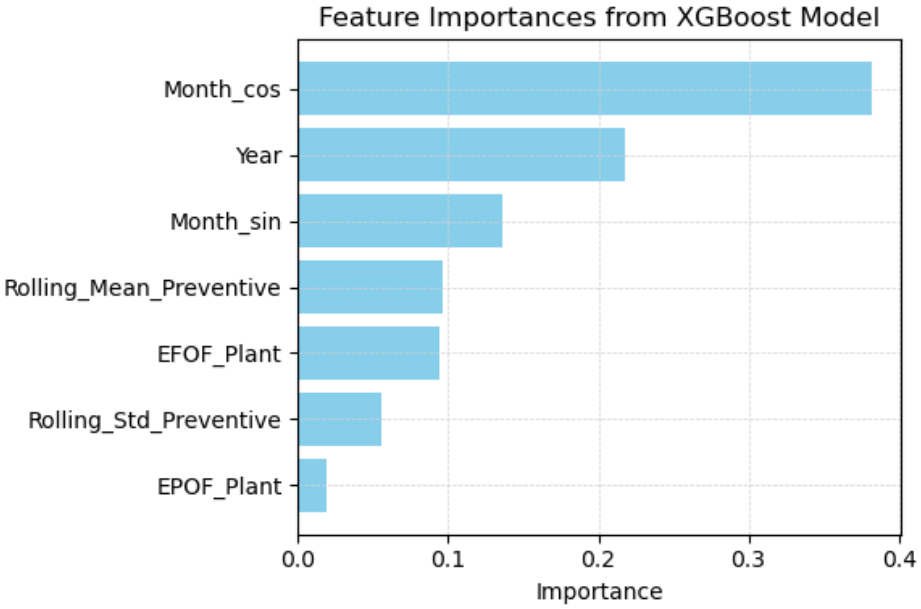


Figure 4-19.b. Important Features – XGBoost Model - Optuna

4.4.2 Hypothesis 2 and Results

H2. *The Time Series models (ARIMA/SARIMA with and without log transformation) will achieve a mean absolute percentage error (MAPE) below 30%, demonstrating their predictive effectiveness in estimating preventive maintenance costs.*

The **ARIMA** models are displayed in **Figure 4-20** and **4-21**. Performance parameters are listed in **Table 4-4**, and ARIMA results will be discussed in **Section 5.2.2.**

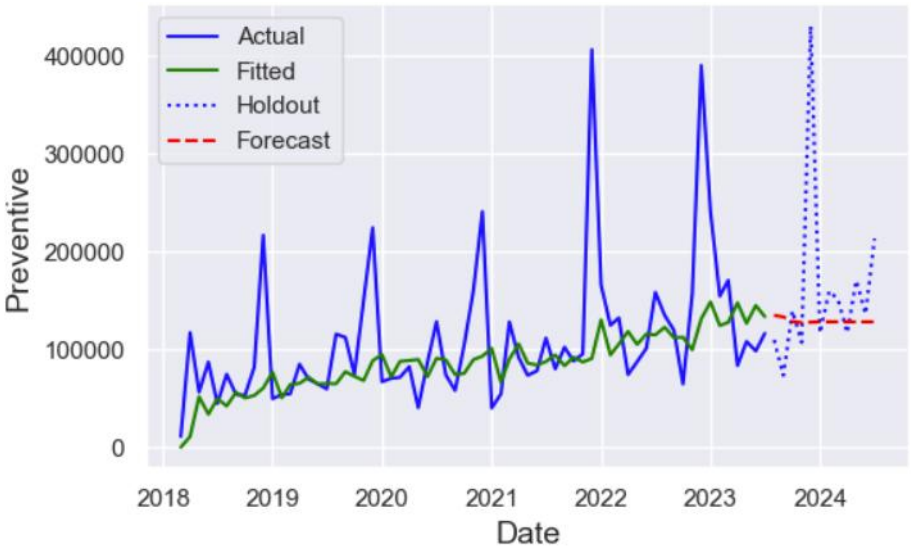


Figure 4-20. ARIMA model without transformation results.

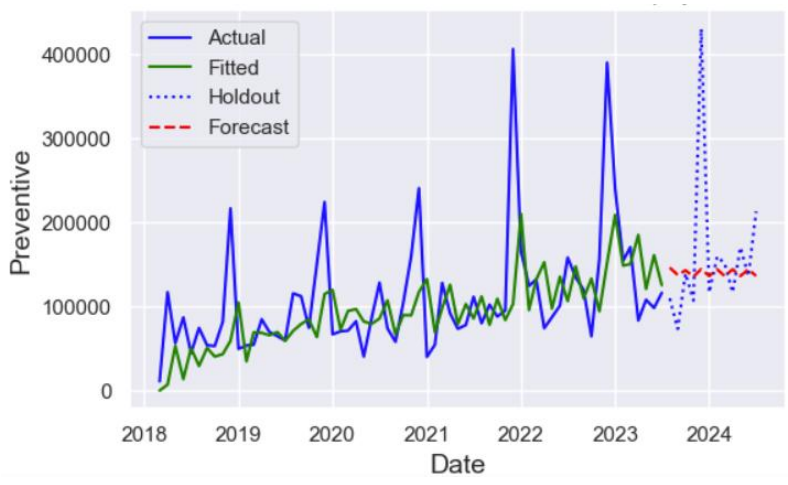


Figure 4-21. ARIMA model with log transformation results.

Table 4-4. ARIMA Models Performance Results

Model	Configuration	Set	MAE	RMSE	MAPE
ARIMA w/o transformation	P= 1, d= 1, q= 1	Test	41542.03	68878.95	38.14
		Holdout	49861.81	90904.02	26.70
ARIMA w/log transformation	P= 3, d= 0, q= 3	Test	40700.53	68992.93	33.4
		Holdout	52641.23	94638.74	27.16

The **ARIMA** models with and without log transformation error plots are displayed in **Figures 4-22.a** and **4-22b**

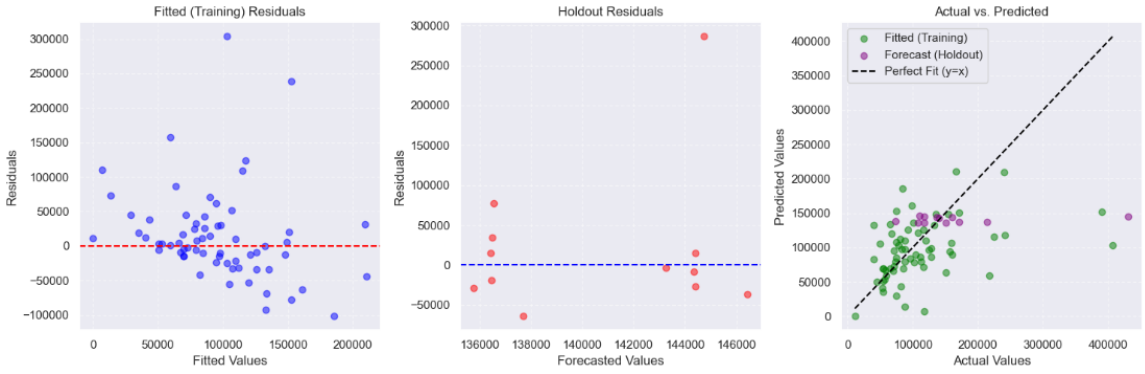


Figure 4-22.a. ARIMA model without log transformation error plots

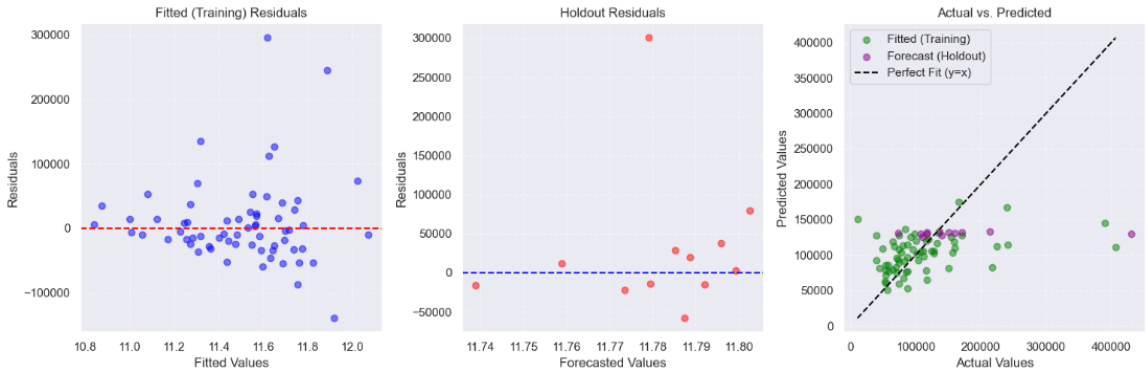


Figure 4-22.b. ARIMA model with log transformation error plots.

Results from the **SARIMA** model without any transformation are displayed in Figure 4-23 and with log transformation in Figure 4-24; the performance parameters are listed in **Table 4-5.** SARIMA results will be discussed in **Section 5.2.2**.

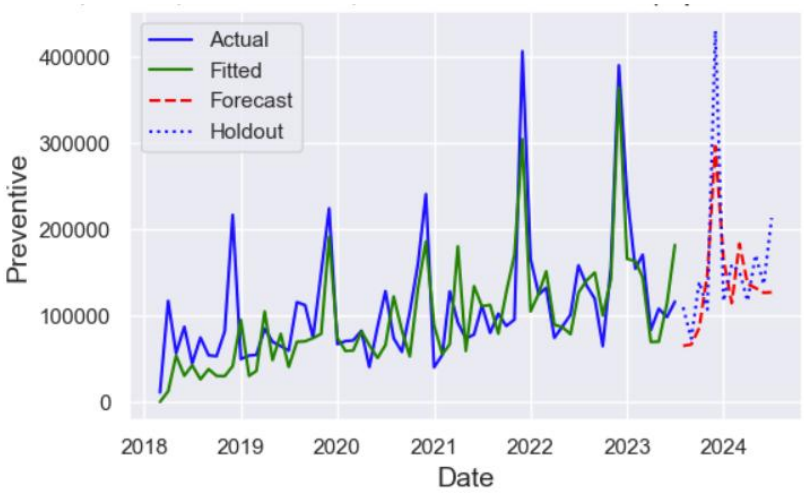


Figure 4-23. SARIMA Model without transformations Results.

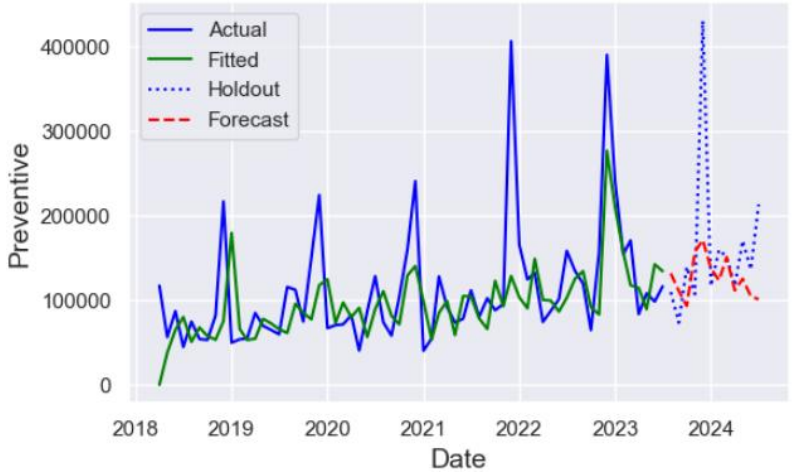


Figure 4-24. SARIMA Model with log transformation Results.

Table 4-5. SARIMA Model Test Set Performance Results.

Model	Configuration	Set	MAE	RMSE	MAPE
SARIMA w/o transformation	p=0, d= 1, q=2, P= 2, D=1, Q=1, s=12	Test	34759.49	46236.59	35.70
		Holdout	46705.64	57362.98	27.98
SARIMA w/log transformation	p=1, d= 1, q=2, P= 0, D=0, Q=1, s=12	Test	38551.17	58779.54	36.03
		Holdout	55633.93	87284.21	29.90

The **SARIMA** models with and without log transformation error plots are displayed in **Figures 4-25.a** and **4-25.b.**

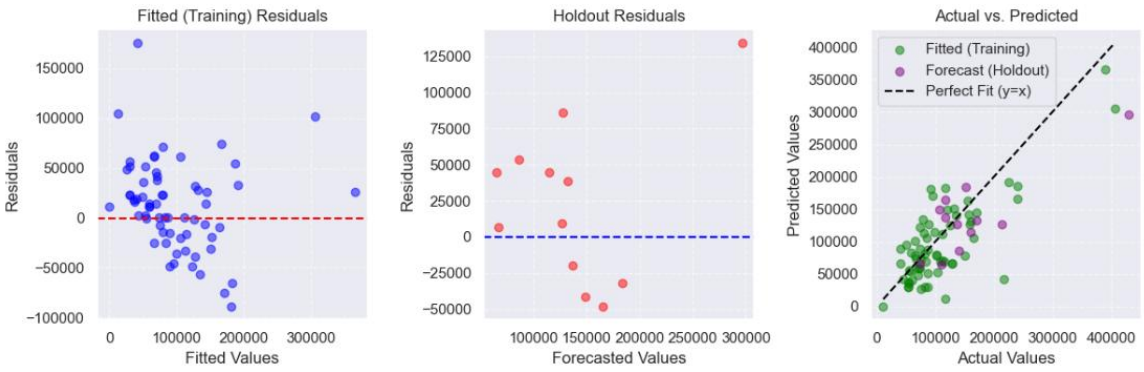


Figure 4-25.a. SARIMA model without log transformation error plots

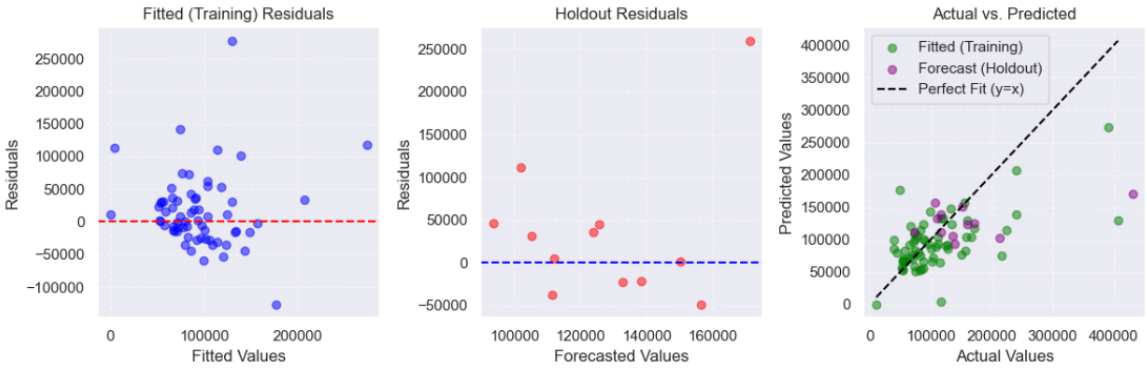


Figure 4-25.b. SARIMA model with log transformation error plots.

4.4.3 Hypothesis 3 and Results

H3. *The linear regression models (ElasticNet with time series validation, k-fold validation, and walk-forward validation) will achieve a mean absolute percentage error (MAPE) below 30%, demonstrating their predictive effectiveness in estimating preventive maintenance costs.*

The three linear models were tested using Elastic Net as the base ML algorithm. Additional features were created using the polynomial features function, with categories up to the 4th degree. The main difference is that one model used a time series-based data split, while the other used a data split based on a random sample.

Elastic Net with Time Split-based models are displayed in **Figure 4-26.a to Figure 4-26.c,** and performance parameters are listed in **Tables 4-6**. Results will be discussed **Section 5.2.3**.

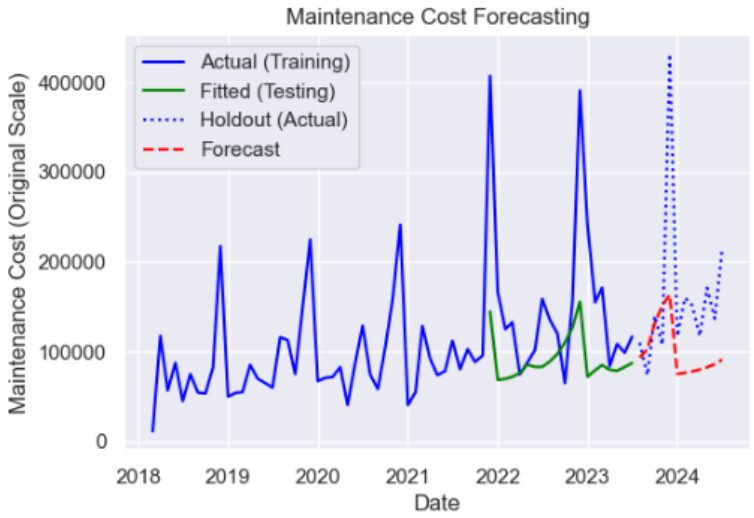


Figure 4-26.a. Elastic Net with Time Split-Validation.

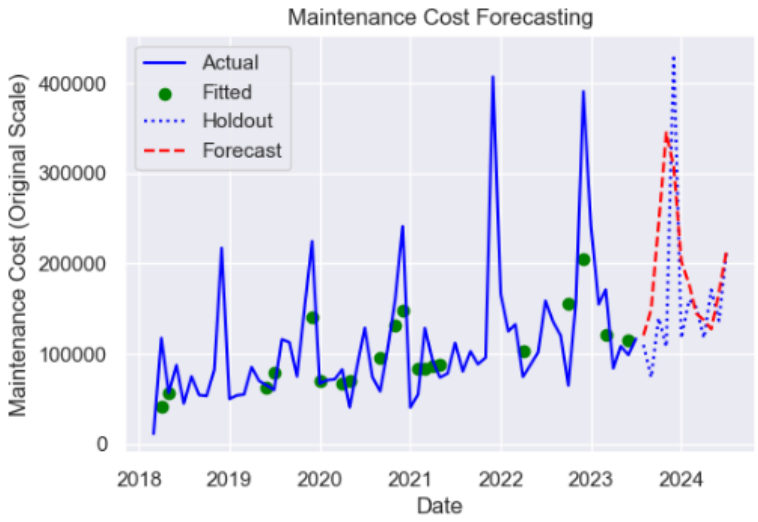


Figure 4-26.b. Elastic Net with K-Fold Cross Validation.

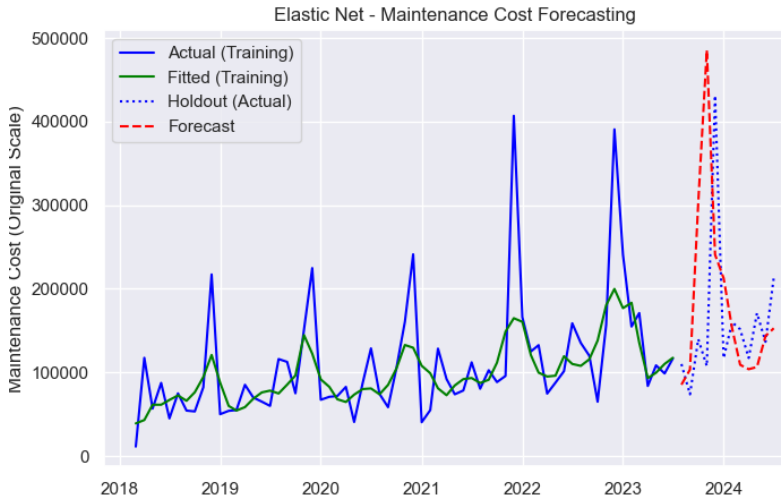


Figure 4-26.c. Elastic Net with Walk-Forward Validation.

Table 4-6. Elastic Net Models Performance Results – Test Set.

Model	Configuration	Set	MAE	RMSE	MAPE
Time series validation	degree= 4, alpha= 0.0668, l1_ratio= 0.55128, max_iter= 3927	Test	67922.73	98667.01	36.35
		Holdout	71793.82	98096.11	39.75
Kfold validation	degree= 3, alpha= 0.0117, l1_ratio= 0.6124, max_iter= 2862	Test	42016.19	60090.79	36.85
		Holdout	63371.35	91505.54	49.12
Walk forward validation	degree= 2, alpha= 0.3640, l1_ratio= 0.0015, max_iter= 3732"	Test	69614.193	106491.14	35.95
		Holdout	71596.49	103385.72	38.65

The **ElasticNet** model's error plots are displayed in **Figures 4-27.a** to **4-27.c**

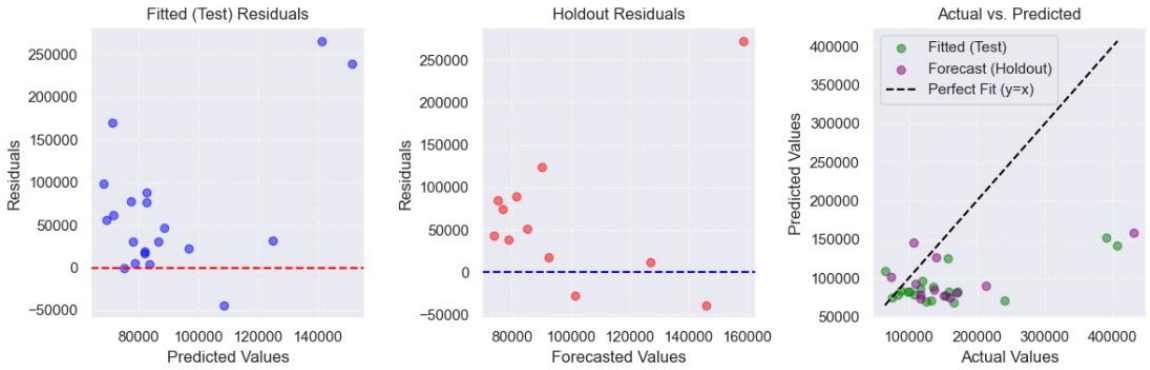


Figure 4-27.a. ElasticNet Time Series Validation model error plots

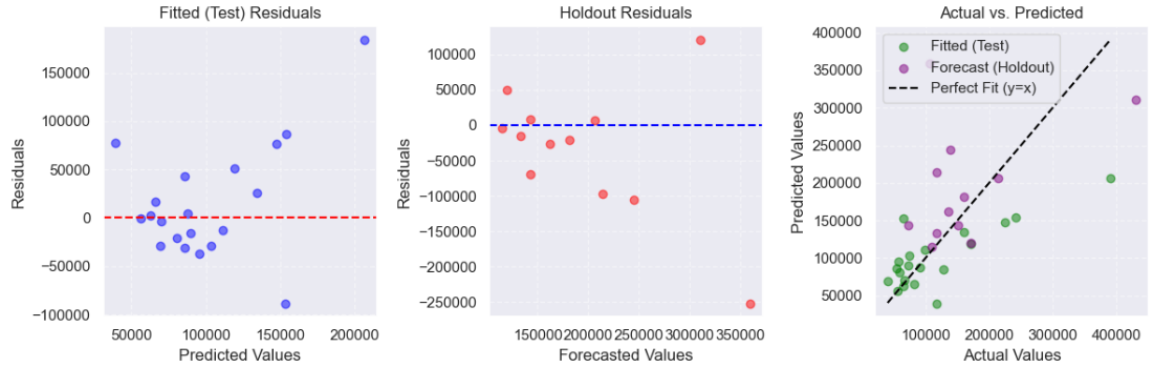


Figure 4-27.b. ElasticNet k-Fold Validation error plots.

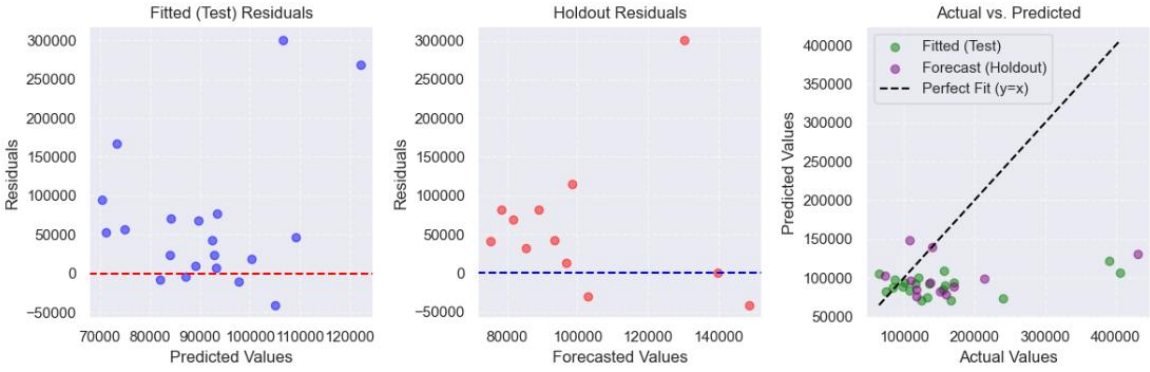


Figure 4-27.c. ElasticNet Walk-Forward Validation error plots

4.4.4 Hypothesis 4 and Results

H4. *The XGBoost models (with GridSearchCV or with Optuna) will achieve a mean absolute percentage error (MAPE) below 30%, demonstrating their predictive effectiveness in estimating preventive maintenance costs*.

XGBoost model has been tested with two configurations including optimization via GridSearchCV (Pedregosa et al., 2011) and OPTUNA. **XGBoost** optimized models are displayed in **Figure 4-28** and **4-29**, and performance parameters are listed in **Tables 4-8**. The results will be discussed in **Section 5.2.4**.

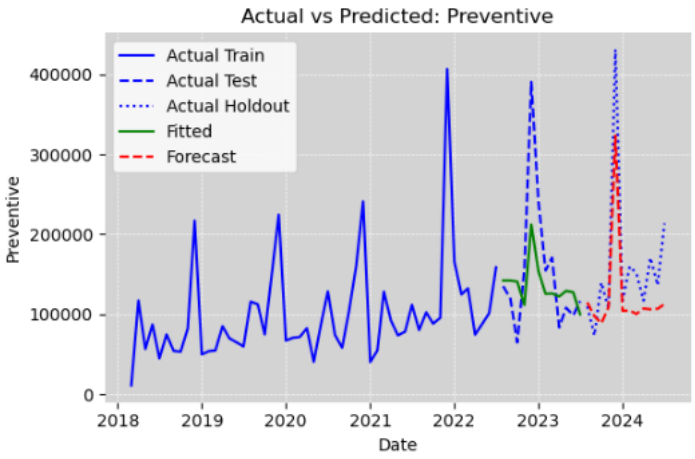


Figure 4-28. XGBoost Grid Search CV Model

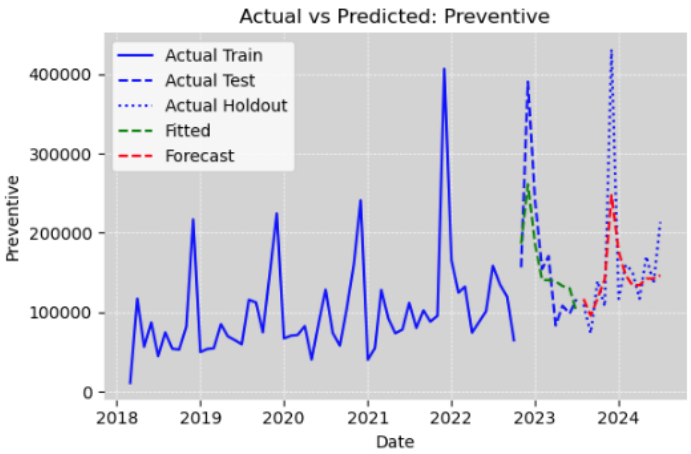


Figure 4-29. XGBoost Optuna Model.

Table 4-7. XGBoost Model Performance Results – Test Set.

Model	Configuration	Set	MAE	RMSE	MAPE
Time series split, Data augmentation, log-transformed, Reduced features, Grid search CV	n_estimators=500, learning_rate=0.01, max_depth=5, reg_alpha= 0.1, reg_lambda=1.0	Test	42720.50	49748.47	29.18
		Holdout	40342.45	63743.37	19.95
Time series split, Data augmentation, log-transformed, Reduced features, Optuna	n_estimators=300, learning_rate=0.03, max_depth=2, reg_alpha= 0.05, reg_lambda=8	Test	42826.52	54364.99	26.27
		Holdout	39791.26	61590.67	22.05

The **XGBoost** models error plots are displayed in **Figures 4-30.a** and **4-30.b.**

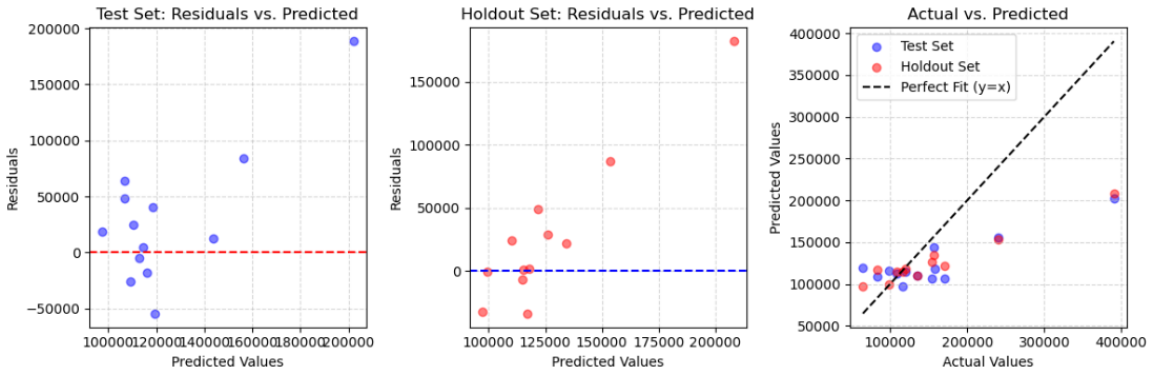


Figure 4-30.a. XGBoost with Grid Search error plots.

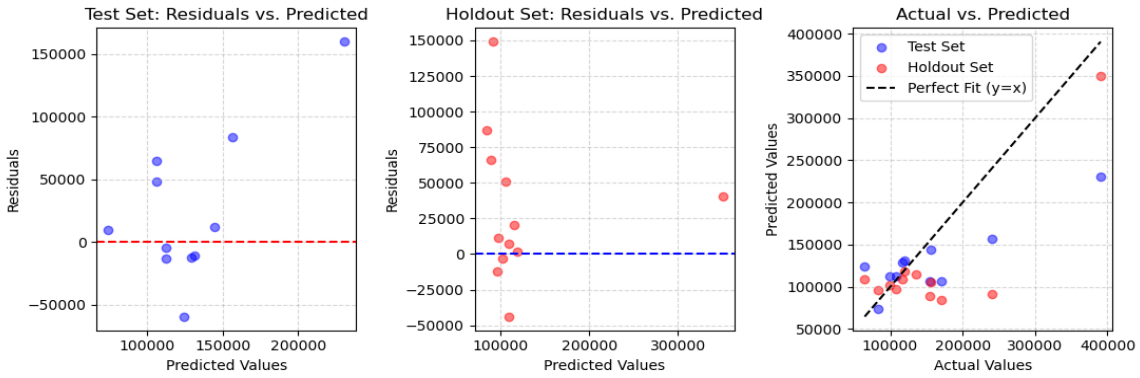


Figure 4-30.b. XGBoost Optuna error plots

4.4.5 Hypothesis 5 and Results

H5. *The LSTM models (one with training/test split and Recursive One-Step Forecasting and other with training/validation/test split and Recursive Multi-Step Forecasting) will achieve a mean absolute percentage error (MAPE) below 30%, demonstrating their predictive effectiveness in estimating preventive maintenance costs.*

The **LSTM** model-A and model-B results plots are displayed in **Figure 4-31** and **4-32,** and performance parameters are listed in **Tables 4-8.** Results will be discussed in **Section 5.2.5.**

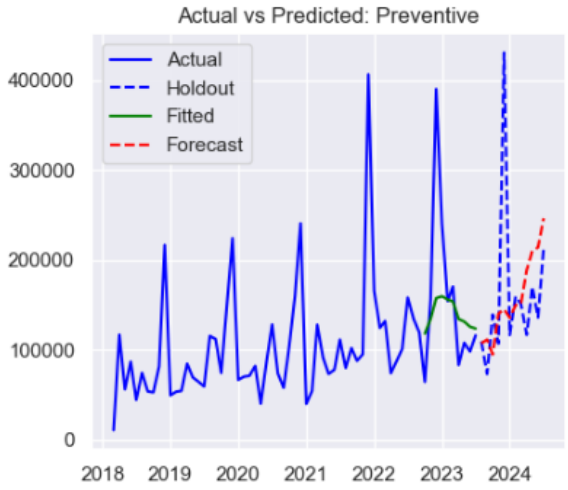


Figure 4-31. LSTM Model Test One-Step Forecasting Results.

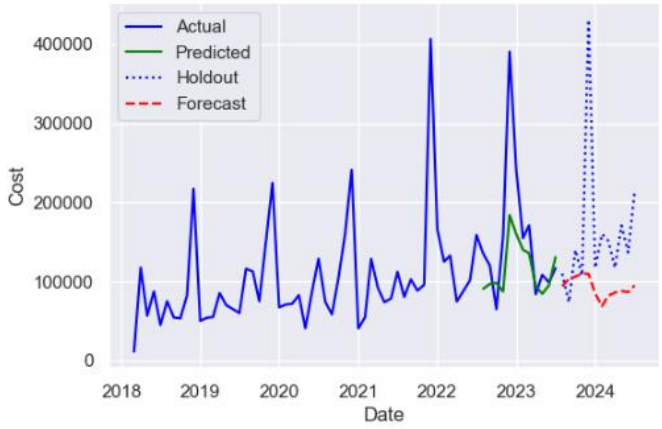


Figure 4-32. LSTM Model with Multi-Step Forecasting Results.

Table 4-8. LSTM Optimized Model Performance Results – Test Set.

Model		Set	MAE	RMSE	MAPE
W/log transformation, reduced feature eng, spike detection, wo exogenous. No data augmentation Time series split Training Recursive One-Step	n_units=38, relu, input_shape=1, Dropout=0.3326, Learning_rate=0.0019, epoch= 100, batch_size=45	Test	54611.49	91479.09	30.25
		Holdout	35605.39	43861.88	28.99
W/ log transformation, extended feature eng, spike detection, wo exogenous. W Data Augmentation Sliding Window Training Direct Multi-Step LSTM	n_units=128, relu, input_shape=12, Dropout=0.3207, Learning_rate=0.0090, epoch= 50, batch_size=8	Test	46689.49	70798.67	26.27
		Holdout	73301.93	109703.65	37.84

The **LSTM** models' error plots are displayed in **Figures 4-33.a** and **4-33.b.**

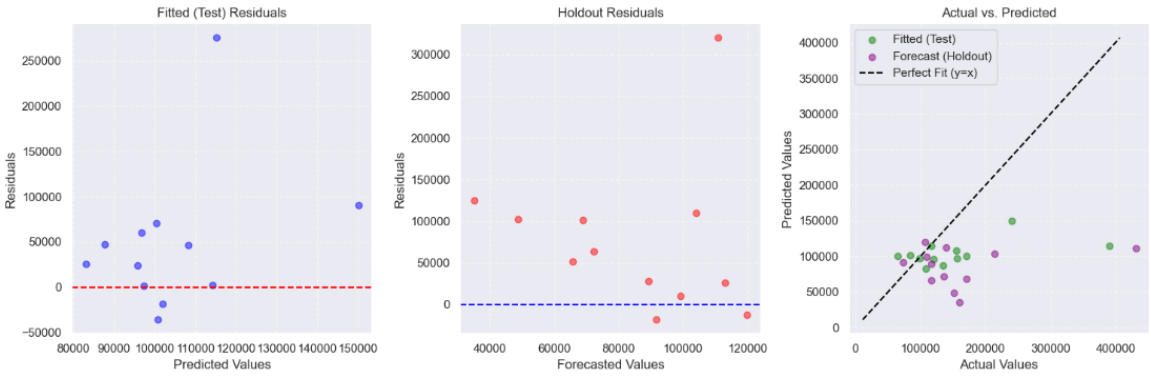


Figure 4-33.a. LSTM Train/Test split and Single-Step error plots.

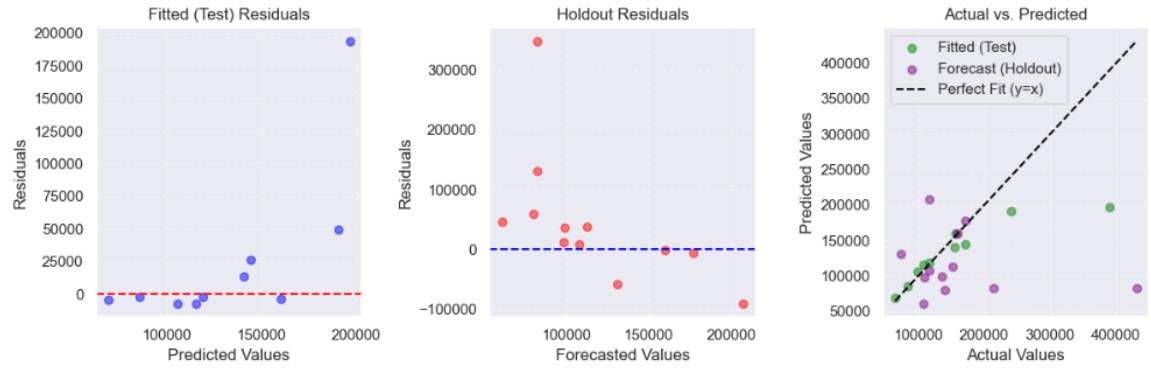


Figure 4-33.b. LSTM Train/Validation/Test split and One-Step forecasting error plots

4.5 Models Results Comparison

Models are ranked by comparing across all performance parameters. **Table 4-9** shows the performance parameters results and ranking for the Test sets for all eleven models, likewise **Table 4-12** shows the performance parameters and ranking for Holdout sets. **Figures 4-34** and **4-35** display the results in a spider graph for Test sets and Holdout set, respectively.

Table 4-9. Preventive Maintenance - Models Performance – (Test Sets).

No	Model	MAE Test	RMSE Test	MAPE Test	Rank
1	ARIMA wo Transf	41542.03	68878.95	38.14	8
2	ARIMA w/Transf	40700.53	68992.93	33.40	5
3	SARIMA wo Transf	34759.49	46236.59	35.70	1
4	SARIMA w/Transf	38551.17	58779.54	36.03	4
5	ElasticNet TS	67922.73	98667.01	36.35	10.5
6	ElasticNet RS	42016.19	60090.79	36.85	7
7	ElasticNet WF	69614.19	106491.14	35.96	10.5
8	XGBoost GridSearch	42720.5	49748.5	29.2	2
9	XGBoost Optuna	42826.5	54365.0	26.3	3
10	LSTM One-step	54611.49	91479.09	30.25	9
11	LSTM Multi-step	46689.49	70798.67	26.27	6

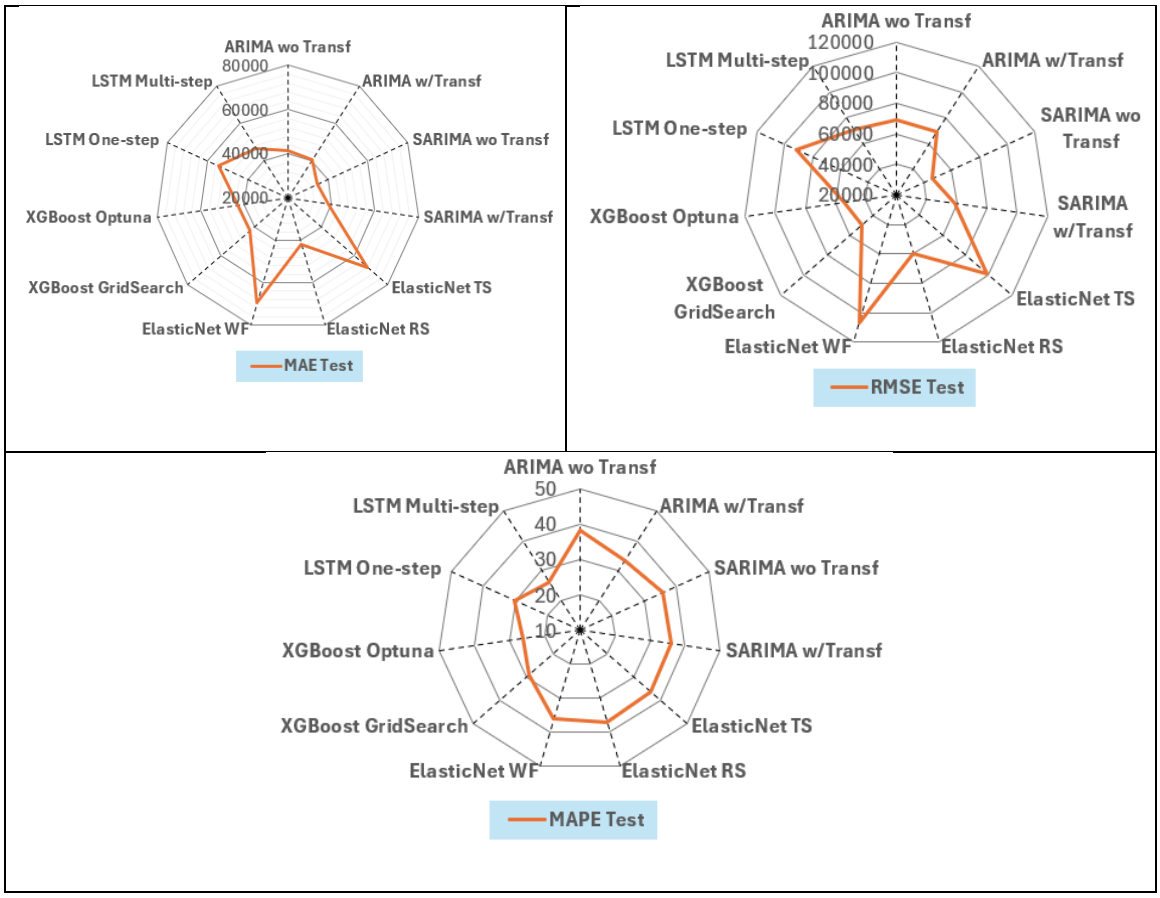


Figure 4-34. Preventive Maintenance - Models Performance – (Test Set).

Table 4-10. Preventive Maintenance - Models Performance – (Holdout Sets).

No	Model	MAE holdout	RMSE Holdout	MAPE Holdout	Rank
1	ARIMA wo Transf	49861.81	90904.02	26.70	5
2	ARIMA w/Transf	52641.23	94638.74	27.16	6
3	SARIMA wo Transf	46705.64	57362.98	27.98	4
4	SARIMA w/Transf	55633.93	87284.21	29.90	7
5	ElasticNet TS	71793.82	98096.11	39.75	10
6	ElasticNet RS	63371.35	91505.54	49.12	8
7	ElasticNet WF	71596.49	103385.72	38.65	9
8	XGBoost GridSearch	40342.45	63743.37	20.0	2.5
9	XGBoost Optuna	39791.26	61590.67	22.0	1
10	LSTM One-step	35605.39	43861.88	28.99	2.5
11	LSTM Multi-step	73301.93	109703.65	37.84	11

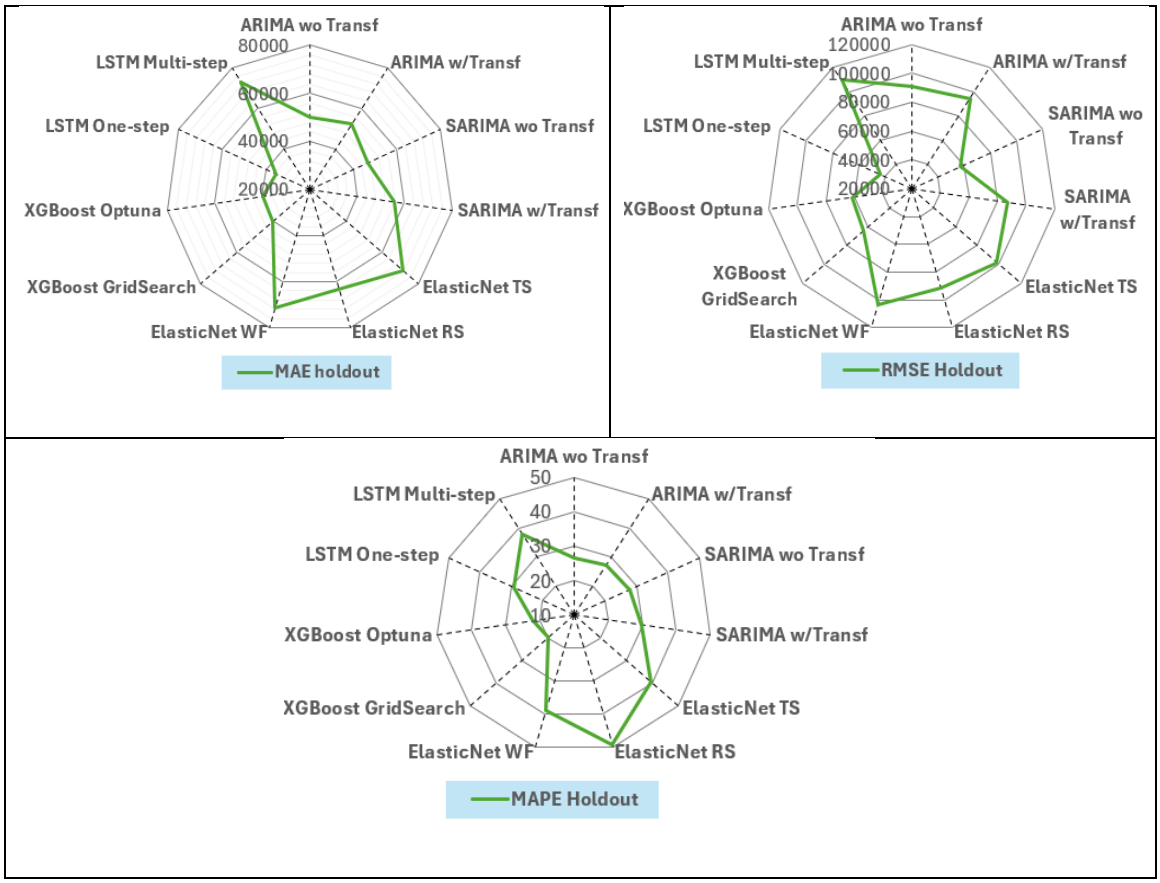


Figure 4-35. Preventive Maintenance - Models Performance – (Holdout Set).

4.6 Selection of the Best Performance Model

The best model selection is done by combining the results of Test sets and Holdout sets as shown in **Table 4-11** and **Figure 4-36**. The results will be discussed in **Section 5.3**.

Table 4-11. Best Models Actual (Unseen) vs Predicted Data – (Holdout Set).

No	Model	Rank Test	Rank Holdout	Final Rank
1	ARIMA wo Transf	8	5	7
2	ARIMA w/Transf	5	6	4.5
3	SARIMA wo Transf	1	4	3
4	SARIMA w/Transf	4	7	4.5
5	ElasticNet TS	10.5	10	11
6	ElasticNet RS	7	8	8
7	ElasticNet WF	10.5	9	10
8	XGBoost GridSearch	2	2.5	2
9	XGBoost Optuna	3	1	1
10	LSTM One-step	9	2.5	6
11	LSTM Multi-step	6	11	9

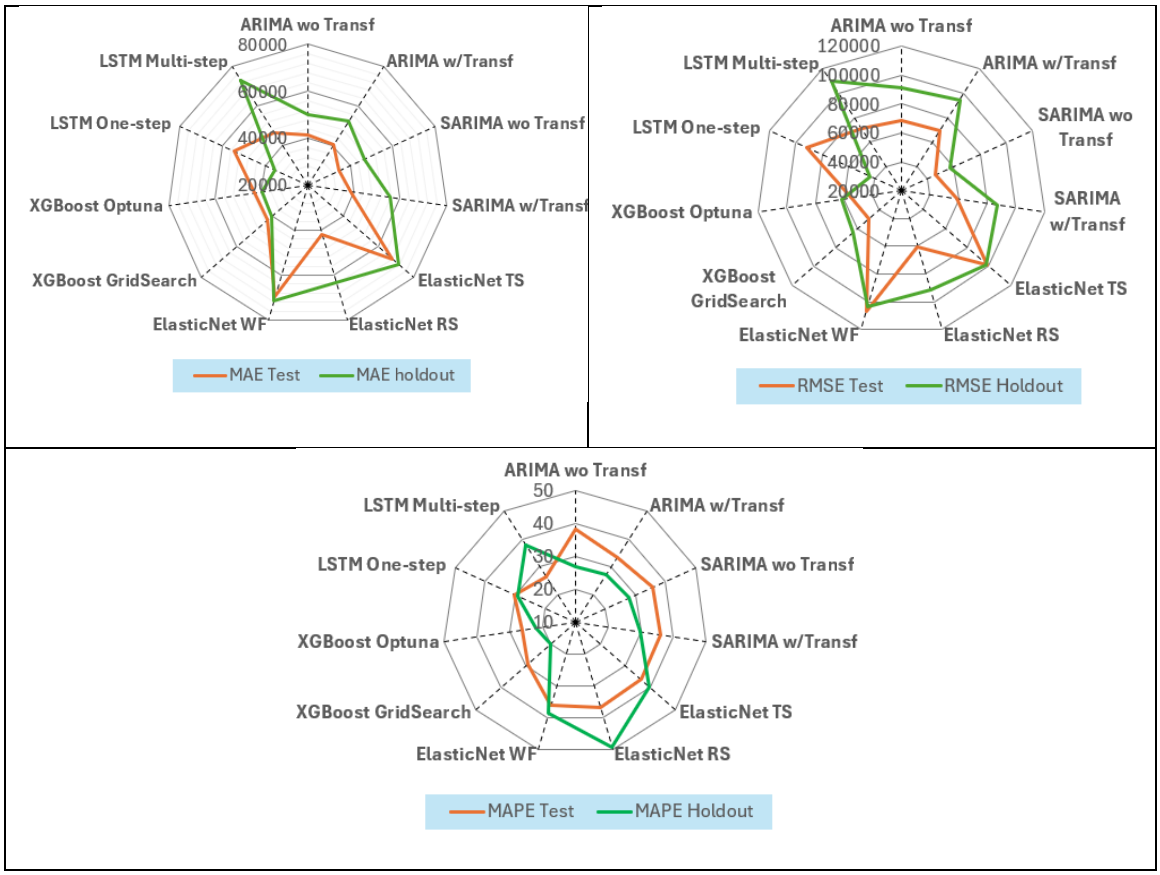


Figure 4-36. Best Models Actual (unseen) vs Predicted data – (Holdout Set).

Chapter 5—Discussion and Conclusions

5.1 Introduction

This research aimed to develop a preventive maintenance cost forecasting model for a thermal power plant, exploring the use of supervised machine learning techniques. This chapter discusses the key findings, examining how the results of Chapter 4 align with the hypothesis as presented in **section 5.2**, while **section 5.3** draw conclusions based on the qualitative and quantitative predictive performance of the machine learning models. Additionally, this chapter outlines lessons learned in **section 5.4**, contribution to the body of knowledge in **section 5.5** and recommendations for future research in **section 5.6**.

5.2 Discussion

5.2.1 Discussion of Hypothesis 1

H1. *The Main Factors (Independent Variables) identified in this praxis significantly influence the maintenance costs of a thermal power plant*.

The hypothesis is validated through a feature importance analysis using the SHapley Additive exPlanations (SHAP) values and Feature Importance plots, which highlight how each factor contributes to the model's preventive maintenance cost predictions. The SHAP analysis revealed that factors such as plant dispatch, load cycling coefficient, and fuel blend have no effect on the preventive maintenance costs, while the equivalent planned outage factor (EPOF) and equivalent forced outage factor (EFOF) have between 4% and 10% influence on cost predictions.

It is notable how the feature importance distribution between GridSearchCV and Optuna varies, reflecting the way different hyperparameter tuning strategies influence feature selection and weighting. This is attributed to the nature of each algorithm execution.

Time-based factors are the ones with the highest contribution to the 12-month forecast. Among time-based factors, “month” was particularly influential, underscoring seasonality in the maintenance expenses. The models benefit from key predictors such as monthly cyclic trends, historical rolling statistics, and temporal variables (Month_cos, Year, and Month_sin). The total contribution of time-base factors add up to 90 to 96% of the weight.

Annual seasonality towards the end of the year confirms the correlation with increased preventive maintenance costs due to planned activities tied to the maintenance calendar. The results affirm that an understanding of the maintenance seasonality is needed for accurate short-term cost forecasting.

Result: Hypothesis confirmed.

5.2.2 Discussion of Hypothesis 2

H2. *The Time Series models (ARIMA/SARIMA with and without log transformation) will achieve a mean absolute percentage error (MAPE) below 30%, demonstrating their predictive effectiveness in estimating preventive maintenance costs*.

The Augmented Dickey-Fuller Test (ADF) and Auto Correlation Function (ACF) and Partial Auto Correlation Function (PACF) values and plots for non-differenced preventive maintenance costs confirmed a non-stationarity and a significant yearly seasonality pattern. Differencing makes the series stationary, but ACF and PACF could

not confirm seasonality, as verified by the statistically significant lags table. This suggests a better fit for SARIMA compared to ARIMA.

Both ARIMA models captured the averages well, applying differencing and using the closest previous lags between 1 to 3, but cannot capture seasonality. The two tested models demonstrated a worse than expected mean absolute percentage error (MAPE) on the testing set, but better than expected in the holdout set, suggesting failure to generalize.

ARIMA model without log transformation MAPE result for the test set is 38.14% and 26.70% for the Holdout set.

Result: Hypothesis not confirmed.

ARIMA model with log transformation MAPE result for the test set is 33.40% and 27.16% for the Holdout set.

Result: Hypothesis not confirmed.

The performance of SARIMA models indicates MAPE results similar to ARIMA models, but approximately 32% better on Root Mean Square Error (RMSE), indicating a better capture of seasonality. Both ARIMA and SARIMA struggle to generalize.

SARIMA model without log transformation MAPE result for the test set is 35.70% and 27.98% for the Holdout set.

Result: Hypothesis not confirmed.

SARIMA model with log transformation MAPE result for the test set is 36.03% and 29.90% for the Holdout set.

Result: Hypothesis not confirmed.

5.2.3 Discussion of Hypothesis 3

H3. *The linear regression models (ElasticNet with time series validation, k-fold validation, and walk-forward validation) will achieve a mean absolute percentage error (MAPE) below 30%, demonstrating their predictive effectiveness in estimating preventive maintenance costs.*

The **ElasticNet - Time Series Split Validation** - linear regression models did not meet the MAPE expectations. The models yielded a MAPE of around 36% on the test and 40% on the holdout sets, while obtaining the worst Mean Absolute Error (MAE) and RMSE among the models. It suggests that, while linear regression methods offer simplicity, their predictive power may be limited in complex, multi-dimensional maintenance cost scenarios.

The **ElasticNet – k-Fold Cross Validation** - model did not meet expectations. It yielded an average MAPE of around 37% on the test set and 49% on the holdout set. This confirms the limitations of linear regression methods, even when polynomial features were introduced to capture potential nonlinear relationships. However, this k-fold random split method obtained better MAE and RMSE results over ElasticNet models, which contradicts the literature that predicted better performance on time splits for time series.

The ElasticNet – Walk-Forward Validation - model did not meet the expectations. It yielded an average MAPE of around 36% on the test set and 39% on the holdout set. While the model captures general trends, its predictive accuracy varies significantly.

Despite enhancements and the use of polynomial features, the ElasticNet models still failed to achieve the expected MAPE, reinforcing the challenges of using linear regression in complex, multi-dimensional, and nonlinear maintenance cost scenarios.

ElasticNet Time Series Validation model MAPE result for the test set is 36.35% and 39.75% for the Holdout set.

Result: Hypothesis not confirmed.

ElasticNet k-Fold Cross Validation model MAPE result for the test set is 36.85% and 49.12% for the Holdout set.

Result: Hypothesis not confirmed.

ElasticNet Walk-Forward Validation model MAPE result for the test set is 35.95% and 38.65% for the Holdout set.

Result: Hypothesis not confirmed.

5.2.4 Discussion of Hypothesis 4

H4. *The XGBoost models (with GridSearchCV or with Optuna) will achieve a mean absolute percentage error (MAPE) below 30%, demonstrating their predictive effectiveness in estimating preventive maintenance costs.*

The **XGBoost - GridSearchCV** model demonstrated strong predictive capability in forecasting preventive maintenance costs, outperforming other models in accuracy metrics. The test set results show a MAPE of 29%, indicating a reasonable fit to the training data. More importantly, the model's performance on the holdout set exhibits a MAPE of about 20%, which suggests that it generalizes well to unseen data.

The **XGBoost - Optuna** model demonstrated strong predictive performance with slight variations compared to the GridSearchCV-optimized version. The test set results show a MAPE of 26%, indicating a reasonable fit to the training data. More importantly, the model's performance on the holdout set have a MAPE of about 22%, suggesting a marginal decrease in accuracy when compared to the GridSearchCV model.

Both XGBoost models obtain also the best results on MAE and RMSE compared to other models. The forecast visualizations display a reasonable alignment with actual preventive maintenance costs, especially in capturing seasonal cost variations. But, despite its strong performance, the model's slight underperformance during high-variance periods suggests that further refinement may be needed.

Notably, XGBoost also showed the most consistent performance across all models evaluated in the study, making XGBoost the most reliable model for capturing recurring patterns and seasonality in preventive maintenance costs.

XGBoost GridSearchCV model MAPE result for the test set is 29.18% and 19.95% for the Holdout set.

Result: Hypothesis confirmed.

XGBoost Optuna model MAPE result for test set 26.27% and 22.05% for the Holdout set.

Result: Hypothesis confirmed.

5.2.5 Discussion of Hypothesis 5

H5. *The LSTM models (one with training/test split and Recursive One-Step Forecasting and other with training/validation/test split and Recursive Multi-Step Forecasting) will achieve a mean absolute percentage error (MAPE) below 30%, demonstrating their predictive effectiveness in estimating preventive maintenance costs*.

The **LSTM One-Step Forecasting** model demonstrated promising performance in preventive maintenance cost forecasting. The model achieved a MAPE of approximately 30% on the test set and about 29% on the holdout set. While the model performed reasonably well, its forecasting ability appears limited. The forecasted values exhibit a degree of instability and deviation from actual patterns, which may indicate that the model struggles with capturing temporal dependencies beyond the training window. Furthermore, the reliance on recursive one-step forecasting suggests a potential compounding error effect over extended forecasting periods.

The **LSTM Multi-Step Forecasting** model demonstrated moderate predictive performance. On the test set, the model achieved a MAPE of approximately 26% on the test set, indicating a reasonable fit to the training data. However, its performance on the holdout set deteriorated significantly, with a MAPE of about 38%. This decline suggests that the model struggled with generalization when predicting unseen data. The forecasting plot indicates that while the model captured some of the fluctuations in the historical data, its predictions for the future exhibit higher uncertainty.

Both LSTM models struggled with using a small amount of data records. While not consistently meeting the expected accuracy expectations, the LSTM model's performance is promising and has potential for use given sufficient data points.

LSTM Train/Test split and Multi-Step forecasting model MAPE result for the test set is 30.25% and 28.99% for the Holdout set.

Result: Hypothesis not confirmed.

LSTM Train/Validation/Test split and One-Step forecasting model MAPE result for the test set is 26.27% and 37.84% for the Holdout set.

Result: Hypothesis not confirmed.

5.3 Conclusions

The results align with the literature on machine learning's advantages over traditional methods and yet somehow misalign with the literature regarding cost dependency on operation and performance factors by empirically demonstrating that short-term (12-month) maintenance costs are rather highly dependent on time factors.

Consistent with Boulmaiz et al. (2020), this work demonstrated LSTM's (long short-term memory) poor performance related to inconsistent prediction when working with small data sets. The poor performance of ARIMA and linear regression models reaffirms their limitations in non-linear and seasonal contexts, as observed by Ng et al. (2023). This study's results do support Elmousalami (2021) work with XGBoost by effectively handling complex, high-dimensional, and noisy data. Additionally, some potential is identified with the SARIMA model due to its ability to capture seasonality.

This study demonstrates the potential of machine learning models, particularly XGBoost, in providing accurate preventive maintenance cost forecasts for thermal power plants by achieving consistent accuracy levels within the expected MAPE of 30% or better. While the models struggle slightly with high-cost events, this is expected in

maintenance cost forecasting. However, they still generalize well, as seen in the holdout set predictions.

By achieving a MAPE of less than 30%, the XGBoost models offer practical tools for enhancing budgeting precision and resource allocation with minimal effort. The findings contribute to the broader field of maintenance cost estimations, offering a data-driven approach that can adapt to the evolving needs of the energy sector.

The empirical results of this study demonstrate minimal to no relation between preventive maintenance costs and plant performance parameters, suggesting that the relationships found in the literature review are only useful for longer-term comparison of similar plants working at different operating regimes, but do not support short-term budget forecasts. Conversely, the main contributors to the forecasting of preventive maintenance cost are demonstrated to be time-related variables like month and maintenance costs rolling windows, which add up to 90 to 96% of the weight.

The variation in feature importance distribution between two like models with slightly different hyperparameters seen for XGBoost supports the need for careful model validation and interpretation to ensure that the selected features align with domain knowledge and operational relevance in forecasting maintenance costs.

Each model encountered specific limitations during the development process. The ARIMA model struggled with weak seasonality and stationarity. The linear regression models (ElasticNet) showed limitations in handling complex nonlinear relationships. The LSTM model is highly limited when using a small dataset and requires extensive computational resources and longer training times: nevertheless, if a larger dataset is available, it might be a good option for further testing. All models struggled slightly,

especially underestimating high-cost events; this is expected in maintenance cost forecasting.

Data Limitations: The study relied on historical data from a single thermal power plant, potentially limiting the generalizability of the results to other plants or contexts. Model results are also dependent on the quality of the data. Therefore, it is important to have all asset-related information structured and stored in standard formats.

Model Limitations: Results from the application of similar models may vary with different data structures or plant operations, requiring further validation and configuration in diverse settings or environments. Additionally, the results cannot provide granular cost details. As suggested in **Appendix B**, higher dimensionality may lead to lower accuracy; therefore, this method cannot be considered a replacement for deterministic methods yet. Results may also vary when major modifications or improvements are made to assets.

5.4. Lessons Learned

One of the key insights gained during this work was the significance of data preprocessing — one of the study's most complex and time-consuming stages. Although every research project might present unique challenges, several lessons were learned about predictive modeling in industrial contexts.

- **Managing Data from Multiple Sources:** Preprocessing data from various sources (e.g., ERP, CMMS, and process historians) introduced complexity in ensuring consistency across different formats and data structures. This emphasizes the importance of establishing robust data integration practices from the outset.

- **Handling Financial Transactions:** The financial data included negative and positive transactions related to the way external services are accrued, which introduced significant noise to the dataset and therefore needed to be balanced out. Special ML filtering and cleaning techniques were used to ensure that only relevant data was included.
- **Dealing with Outliers**: Maintenance costs are highly skewed by nature, this exposes the modeling to outliers that cannot be filtered out completely due to their weight on the total costs. This significantly impacts model accuracy. Therefore, the use of log transformation helped address these outliers during feature engineering, minimizing distortions in the results.
- **Addressing Data Quality Issues:** Data inaccuracies (e.g., incorrect inputs or incomplete records) were a recurring challenge. Implementing comprehensive validation and correction methods ensured the dataset was accurate and reliable.

These lessons highlight the critical role of data preprocessing in ensuring the reliability and accuracy of machine learning models. Addressing these issues early in the workflow can significantly improve the outcomes of similar industrial projects.

5.5 Contributions to Body of Knowledge

- This work is a novel application in Thermal Power Plants, leveraging commonly available data on industrial setups.
- This work is a structured machine learning-driven forecasting framework that is easily adaptable to thermal power plants and other asset intensive industries.

- This research contributes to the maintenance management field by exploring and demonstrating the efficacy of supervised machine learning models for preventive maintenance cost forecasting in industrial settings.
- This work provides insights into the strengths and weaknesses of different ML algorithms use for maintenance cost forecasting and guides future researchers and practitioners willing to implement predictive analytics in similar industry.

5.6 Recommendations for Future Research

The following areas are recommended for future research .

Enhancing Model Accuracy: Explore ensemble models by combining ML techniques to leverage the models' strengths, e.g. XGBoost, SARIMA, or LSTM with other . This could enhance predictive accuracy, and may improve generalization and response to high-cost predictions typical of seasonal maintenance activities.

Applicability to Other Plants: Investigate the transferability of the models to different types of power plants or similar industrial asset-intensive plants.

Data Augmentation: Incorporate additional data, such as asset-related information, reliability data, process data deviations, or specific maintenance plans to improve feature richness and predictive power; this includes feature augmentation with new variables.

Long-term Model Monitoring: Conduct longitudinal studies to assess the stability and adaptability of the models over extended periods, ensuring sustained accuracy in dynamic operational environments throughout the lifecycle. Additionally, testing higher levels of cost breakdown could provide further granularity in predictions and cover other maintenance categories.

Future research can build on these insights to further refine forecasting methods, supporting more efficient and sustainable operations in the industry.

References

Abbas, F., Afzaal, H., Farooque, A. A., & Tang, S. (2020). Crop yield prediction through proximal sensing and machine learning algorithms. *Agronomy*, 10(7), 1046. https://doi.org/10.3390/agronomy10071046

Abdenour, J. I. (2021). *A cost estimation model for improving the budget estimates of industrial plant construction projects*. (Publication No. 28715536) [Doctoral praxis, The George Washington University]. ProQuest Dissertations & Theses Global. https://www.proquest.com/docview/2568678359?fromopenview=true&pq-origsite=gscholar&sourcetype=Dissertations%20&%20Theses

Aghelpour, P., Mohammadi, B., & Mostafa, S. (2019). Long-term monthly average temperature forecasting in some climate types of Iran, using the models SARIMA, SVR, and SVR-FA. Theoretical and Applied Climatology (2019) 138:1471–1480. https://doi.org/10.1007/s00704-019-02905-w

Alashari, M., El-Rayes, K., & Helaly, H. (2023). Machine-learning model for predicting maintenance costs of EPDM roofing systems. *Journal of Performance of Constructed Facilities*, 37(6), 04023048. https://doi.org/10.1061/JPCFEV.CFENG-4413

Alhamzawi, R., & Ali, H. T. M. (2018). The Bayesian elastic net regression. *Communications in Statistics - Simulation and Computation*, 47(4), 1168-1178. https://doi.org/10.1080/03610918.2017.1307399

Anaconda, Inc. (2023). *Jupyter Notebook (Version 6.5.4) [Computer software]*. Anaconda, Inc. https://www.anaconda.com

Amir, A., & Henry, M. (2023). Reverse engineering of maintenance budget allocation using decision tree analysis for data-driven highway network management. *Sustainability*, *15*(13), 10467. https://doi.org/10.3390/su151310467

ASTM International. (2011). *Standard classification for cost estimate classification system (ASTM E2516-11)*. ASTM International. https://doi.org/10.1520/E2516-11

Badouard, T., Oliveira, D. M., Yearwood, J., & Torres, P. (2018). *Final report cost of energy (LCOE): Energy costs, taxes and the impact of government interventions on investments*. European Commission – DG Energy A.4. ENER/2018-A4/2018-471.

Bahr, C., & Lennerts, K. (2010). Quantitative validation of budgeting methods and suggestion of a new calculation method for the determination of maintenance costs. *Journal of Facilities Management*, *8*(1), 47–63. https://doi.org/10.1108/14725961011019076

Bassi, V., Pereira-Bonvallet, E., Abdullah, M. A., & Palma-Behnke, R. (2018, June). Cycling impact assessment of renewable energy generation in the costs of conventional generators. *MDPI Energies, 11*(7), 1640. https://doi.org/10.3390/en11071640

Ben-daya, M., Kumar, U., & Murthy, D. P. (2016). *Introduction to maintenance engineering: Modeling, optimization, and management*. John Wiley & Sons. https://doi.org/10.1002/9781118926581

Bentejac, C., Csörgő, A., & Martínez-Muñoz, G. (2019). A comparative analysis of XGBoost. Preprint. Retrieved from https://www.researchgate.net/publication/337048557_A_Comparative_Analysis_of_XGBoost

Bezzar, N. E. H., Laimeche, L., & Meraoumia, A. (2022). Time series analysis of household electric consumption with XGBoost model. *Proceedings of the 4th International Conference on Pattern Analysis and Intelligent Systems (PAIS)*, 978-1-6654-6161-0. IEEE. https://doi.org/10.1109/PAIS56586.2022.9946913

Bitencourt, H. V., Orang, O., de Souza, L. A. F., Silva, P. C. L., & Guimarães, F. G. (2023). An embedding-based non-stationary fuzzy time series method for multiple output high-dimensional multivariate time series forecasting in IoT applications. Neural Computing and Applications, 35(9407–9420). https://doi.org/10.1007/s00521-022-08120-5

Boulmaiz, T., Guermoui, M. & Boutaghane, H. (2020, June). Impact of training data size on the LSTM performances for rainfall–runoff modeling. *Modeling Earth Systems and Environment*, 6, 2153–2164. https://doi.org/10.1007/s40808-020-00830-w

Bunnag, T. (2024). Forecasting PM10 Caused by Bangkok's Leading Greenhouse Gas Emission Using the SARIMA and SARIMA-GARCH Model. *International Journal of Energy Economics and Policy*, 14(1), 418–426. https://doi.org/10.32479/ijeep.15275

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. https://doi.org/10.1145/2939672.2939785

Code of Federal Regulations (CFR). (2024, March). *18 CFR Part 101 uniform system of accounts prescribed for public utilities and licensees*. Code of Federal Regulations. https://www.ecfr.gov/current/title-18/chapter-I/subchapter-C/part-101

Coffie, G. H., & Cudjoe, S. K. F. (2023, December). Using extreme gradient boosting (XGBoost) machine learning to predict construction cost overruns. *International Journal of Construction Management*, 1–9. https://doi.org/10.1080/15623599.2023.2289754

Crivellari, A., & Beinat, E. (2020). Forecasting spatially-distributed urban traffic volumes via multi-target LSTM-based neural network regressor. *Mathematics*, 8(12), 2233. https://doi.org/10.3390/math8122233

Drachal, K., & Pawłowski, M. (2024). Forecasting selected commodities' prices with the Bayesian symbolic regression. *International Journal of Financial Studies*, 12(2), 34. https://doi.org/10.3390/ijfs12020034

Draz, M. M., Emam, O., & Azzam, S. M. (2024). A predictive model for software cost estimation using ARIMA algorithm. *International Journal of Advanced Computer Science and Applications*, 15(7), 656-665. https://doi.org/10.14569/IJACSA.2024.0150766

Electric Power Research Institute. (2013, September). *Impact of cycling on the operation and maintenance cost of conventional and combined-cycle power plants* (Report No. 3002000817). Palo Alto, CA: Electric Power Research Institute.

Electric Power Research Institute. (2018, March). *Program on technology innovation: Integrated generation technology options 2017* (Report No. 3002011806). Palo Alto, CA: Electric Power Research Institute.

Elmousalami, H. H. (2021, February). Comparison of artificial intelligence techniques for project conceptual cost prediction: A case study and comparative analysis. *IEEE Transactions on Engineering Management*, *68*(1), 183–196. https://doi.org/10.1109/TEM.2020.2972078

Feng, R., Du, H., Du, T., Wu, X., Yu, H., Zhang, K., Huang, C., & Cao, L. (2024). Fault diagnosis for wind turbines based on LSTM and feature optimization strategies. *Concurrency and Computation: Practice and Experience*, 36(1), e7886. https://doi.org/10.1002/cpe.7886

Ferreira, R. S., Barroso, L. A., Feinstein, C. D., & Borges, C. L. T. (2013). A non-linear programming approach to maintenance budgeting for multi-component systems. *2013 IEEE Power & Energy Society General Meeting*, 1–5. https://doi.org/10.1109/PESMG.2013.6672767

Gafurov, A., Mukharamova, S., Saveliev, A., & Yermolaev, O. (2023). Advancing agricultural crop recognition: The application of LSTM networks and spatial generalization in satellite data analysis. *Agriculture*, 13(1672). https://doi.org/10.3390/agriculture13091672

Galar, D., Sandborn, P., & Kumar, U. (2017). *Maintenance costs and life cycle cost analysis*. CRC Press, Taylor & Francis Group. https://doi.org/10.1201/9781315154183

GHD Advisory. (2022, Dec). *Power station and associated costs, Benchmark reserve capacity price 2023*. Economic Regulatory Authority. https://www.erawa.com.au/cproot/23060/2/-BRCP.2023---Power-Station-and-Associated-Costs-Report---GHD-Advisory---Dec-2022.PDF

Gupta, P., Gupta, S., & Gandhi, O. P. (2014). Annual maintenance budget estimation for a plant system. *Journal of Quality in Maintenance Engineering*, *20*(2), 193–210. https://doi.org/10.1108/JQME-04-2013-0015

Gupta, S., & Gupta, P. (2017). Evaluating maintenance budget for an AC plant under its contextual conditions. *Journal of Facilities Management*, *15*(1), 76–89. https://doi.org/10.1108/JFM-04-2016-0010

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... & Oliphant, T. E. (2020). *Array programming with NumPy*. Nature, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

International Energy Agency (IEA). (2023, October). *World Energy Outlook 2023*. https://iea.blob.core.windows.net/assets/86ede39e-4436-42d7-ba2a-edf61467e070/WorldEnergyOutlook2023.pdf

International Organization for Standardization. (2014). *ISO 55000:2014 Asset management—Overview, principles and terminology*. https://www.iso.org/standard/55088.html

International Organization for Standardization. (2014). *ISO 55000:2014 Asset management—Overview, principles, and terminology*. https://www.iso.org/standard/55088.html

International Organization for Standardization. (2019). *ISO/TS 55010:2019 Asset management—Guidance on the alignment of financial and non-financial functions in asset management*. https://www.iso.org/standard/72422.html

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An introduction to statistical learning*. Springer Texts in Statistics. https://doi.org/10.1007/978-1-4614-7138-7

Kumar, N., Besuner, P., Lefton, S., Agan, D., & Hilleman, D. (2012, July). *Power plant cycling costs*. National Renewable Energy Laboratory (NREL). https://www.nrel.gov/docs/fy12osti/55433.pdf

Le, T.-A., Nguyen, N.-H., Huynh, Q.-T., Cao, P.-N., & Nguyen, T.-H. (2020). A method for project completion cost predicting using LSTM in earned value management technique. *2020 4th International Conference on Recent Advances in Signal Processing, Telecommunications & Computing (SigTelCom)*, 87-92. IEEE. https://doi.org/10.1109/SigTelCom49789.2020.9059750

Lemes, L. C., & Hvam, L. (2019, December). Maintenance costs in the process industry: A literature review. *2019 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, 1481–1485. https://doi.org/10.1109/IEEM44572.2019.8978559

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774. https://github.com/slundberg/shap}

Mani, G., & Volety, R. (2021). A comparative analysis of LSTM and ARIMA for enhanced real-time air pollutant levels forecasting using sensor fusion with ground station data. *Cogent Engineering*, 8(1), 1936886. https://doi.org/10.1080/23311916.2021.1936886

Meyers, L. L., & Fuller, L. C. (1979, January). Operation and maintenance costs for large steam-electric power plants. *AACE International Transactions, 1978-01*(22), 240. ISSN: 1528-7106

Microsoft. (2024). *Microsoft Excel (Version 2024) [Computer Software]*. Microsoft Corporation. https://www.microsoft.com/

Mirzaei, J., Amiri Hanzaei, H., Khaleghi, H., & Kashani, H. (2024). Cost estimation models for seismic retrofit of masonry school buildings. *International Journal of Construction Management*, 24(8), 843-853. https://doi.org/10.1080/15623599.2022.2138616

Ng, K. Y., Zainal, Z., & Samsudin, S. (2023). Comparative performance of ARIMA, SARIMA, and GARCH models in modelling and forecasting unemployment among ASEAN-5 countries. *International Journal of Business and Society, 24*(3), 967–994. https://doi.org/10.33736/ijbs.6393.2023

North American Electric Reliability Corporation (NERC). (2023, June). *2023 state of reliability technical assessment*. https://www.nerc.com/pa/RAPA/PA/Performance%20Analysis%20DL/NERC_SOR_2023_Technical_Assessment.pdf

North American Electric Reliability Corporation (NERC). (2023, December). *2023 long-term reliability assessment*.

https://www.nerc.com/pa/RAPA/ra/Reliability%20Assessments%20DL/NERC_LTRA_2023.pdf

Ogunbayo, B. F., Aigbavboa, C. O., Thwala, W. D., & Akinradewo, O. I. (2022). Assessing maintenance budget elements for building maintenance management in Nigerian built environment: A Delphi study. *Built Environment Project and Asset Management, 12*(4), 649–666. https://doi.org/10.1108/BEPAM-06-2021-0080

Optuna Development Team. (2019). *Optuna: A hyperparameter optimization framework (Version 2.0) [Computer software]*. https://optuna.org

Pandas Development Team. (2020). *pandas-dev/pandas: Pandas (Version 1.2.0) [Computer software]*. Zenodo. https://doi.org/10.1038/s41586-020-2649-2

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). *Scikit-learn: Machine learning in Python*. *Journal of Machine Learning Research, 12*, 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html

Peixeiro, M. (2022). *Time series forecasting in Python*. Manning Publications. https://ebookcentral.proquest.com/lib/gwu/detail.action?docID=7127005

Python Software Foundation. (2023). *Python (Version 3.11) [Computer software]*. Python Software Foundation. https://www.python.org

Reliability Web. (2023). *Best practices maintenance benchmarks*. Reliability Web. https://reliabilityweb.com/en/best-practices-maintenance-benchmarks

Tay, J. K., Aghaeepour, N., Hastie, T., & Tibshirani, R. (2023). Feature-weighted elastic net: Using "features of features" for "better prediction". *Statistica Sinica*, 33(1), 259-279. https://doi.org/10.5705/ss.202020.0226

Tong, D., Zhou, S.-L., Li, H., & Zhang, X. (2023). *Prediction method of power grid marketing cost based on improved ARIMA model*. Proceedings of SPIE - Third International Conference on Artificial Intelligence and Computer Engineering (ICAICE 2022), 12610, 126103S. https://doi.org/10.1117/12.2672723

Tran, M.-K., Panchal, S., Chauhan, V., Brahmbhatt, N., Mevawalla, A., Fraser, R., & Fowler, M. (2022). Python-based scikit-learn machine learning models for thermal and electrical performance prediction of high-capacity lithium-ion battery. *International Journal of Energy Research*, 46(2), 786–794. https://doi.org/10.1002/er.7202

U.S. Energy Information Administration (EIA). (2023, March). Annual energy outlook. U.S. Energy Information Administration. https://www.eia.gov/outlooks/aeo/pdf/AEO2023_Narrative.pdf

Waskom, M. L. (2021). *Seaborn [Computer Software]*. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021

Wheeldon, J. M., & Thimsen, D. (2013, December). Economic evaluation of circulating fluidized-bed combustion power generation plants. In *Fluidized bed technologies for near-zero emission combustion and gasification* (pp. 620–638). Woodhead Publishing Limited. https://doi.org/10.1533/9780857098801.2.620

Winalytra, I., Nugroho, A. S. B., & Triwiyono, A. (2018). Cost estimation model for I-girder bridge superstructure using multiple linear regression and artificial neural network. *Applied Mechanics and Materials*, 881, 142-149. https://doi.org/10.4028/www.scientific.net/AMM.881.142

Zhang, Z., Lai, Z., Xu, Y., Shao, L., Wu, J., & Xie, G.-S. (2017). Discriminative elastic-net regularized linear regression. *IEEE Transactions on Image Processing*, 26(3), 1466–1480. https://doi.org/10.1109/TIP.2017.2651396

Zournatzidou, G., Mallidis, I., Farazakis, D., & Floros, C. (2024). Enhancing Bitcoin price volatility estimator predictions: A four-step methodological approach utilizing elastic net regression. Mathematics, 12(1392). https://doi.org/10.3390/math12091392

Appendix A

Given the multiclass nature of maintenance costs, they can be grouped according to the views of different stakeholders. For example, logistics would like to see them by expense type (e.g., consumables, external services, or spare parts), maintenance personnel would prefer to see them as work order types, (e.g., preventive, corrective, improvements, etc.), or specialties (e.g., mechanical, electrical, instrumentation, etc.), and operations would like to see them organized by asset system hierarchy.

This research identifies a maintenance cost structure that supports categorization and analysis. The cost structure is divided into four primary classes, adapted from Galar et al. (2017). **Figure A-1.** shows the plant maintenance costs categorization subject for this study case:

- **Expense Type:** Spare Parts, Services, Tools, Consumables.
- **Maintenance Type:** Preventive, Corrective, Improvement.
- **Specialty:** Mechanical, Electrical, Instrumentation, etc.
- **Asset System:** According to assets tree hierarchy.

This study centers its attention on the Preventive Maintenance Costs category, which is a good example for testing the accuracy of ML models due to known predictable cost and periodic demand. Other categories are left for future research.

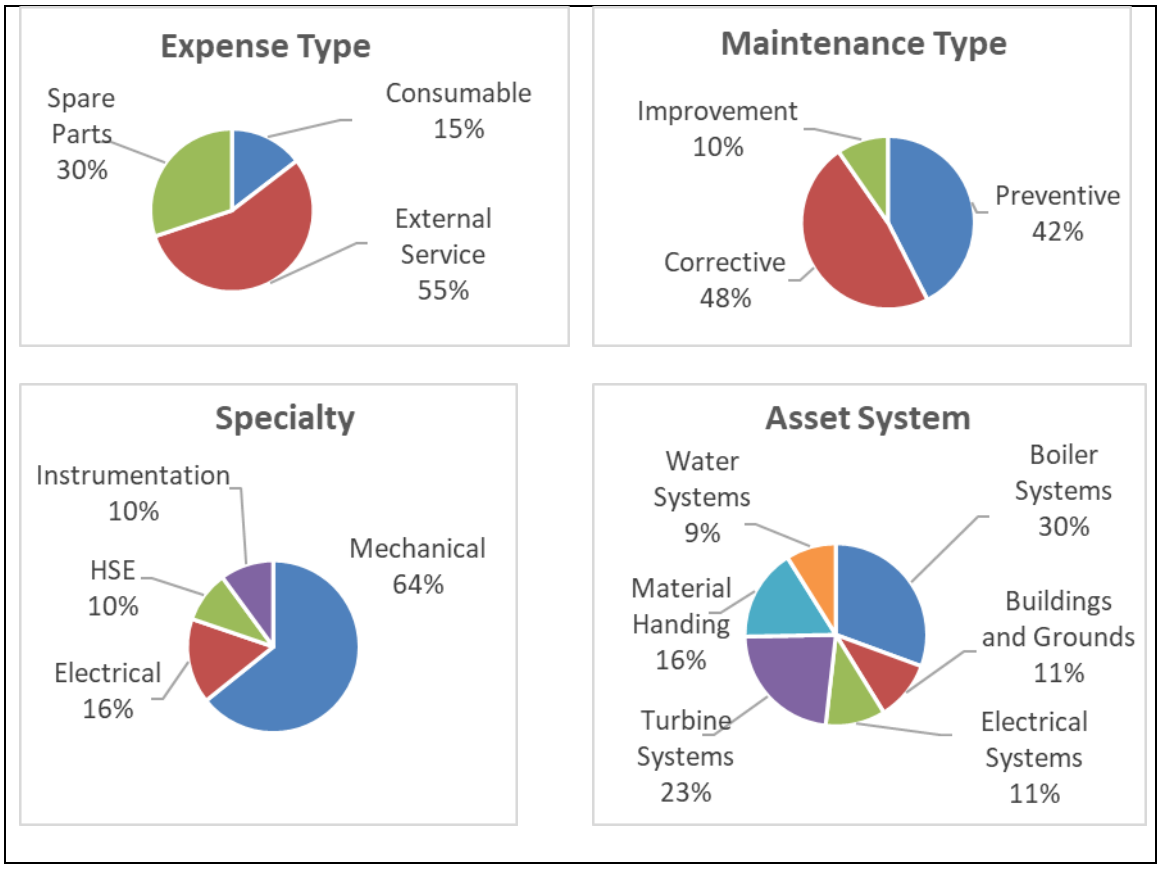


Figure A-1. Power Plant Maintenance Cost Structures.

Appendix B

Table B-1. Cost Transactions Fields

No	Description	Use
1	SAP_RefDate	Reference Date
2	SAP_Account_No	Not used
3	SAP_Transfer_Type	Not used
4	SAP_Transfer_Description	Used as key to extract WO number when missing.
5	SAP_Section_Code	Not used
6	SAP_Equipment_Type_Code	Not used
7	SAP_Expense_Code	Not used
8	SAP_OyM_Cost Center Code	Key to get cost center for filtering out
9	SAP_Expense_Type Code	Not used
10	SAP_WO_No	Key to merge with WO table for indexing
11	SAP_Plant_Area	Not used
12	SAP_Speciality	Not used
13	SAP_Supervisor_Name	Not used
14	SAP_Emergency_ID	Not used
15	SAP_Ammount_USD	Transaction cost in USD
16	SAP_Project_Code	Key for filtering out
17	SAP_Project_Description	Not used

Table B-2. Work Order Report Used Fields

No	Description	Use
1	WO_No	Key to merge with costs transactions
2	WO_Description	Not used
3	WO_Status	Not used
4	WO_Asset_Tag	Key to retrieve asset system structure
5	WO_Asset_Description	Not used
6	WO_Reported_By	Not used
7	WO_Assigned_To	Not used
8	WO_Creation_Date	Not used
9	WO_Completion_Date	Not used
10	WO_Forecast_Date	Not used
11	WO_Type	Key to retrieve the Maintenance Type
12	WO_Priority	Not used
13	WO_Restriction	Not used
14	WO_ID_Outage	Not used
15	WO_Restriction_Comments	Not used
16	WO_Area	Not used
17	WO_Speciality	Key to retrieve the Department Specialty
18	WO_Planned_Start_Date	Not used
19	WO_Planned_Finish_Date	Not used
20	WO_Cost_Center	Not used
21	WO_Responsible_Supv	Not used
22	WO_Project_Code	Not used
23	WO_Unit	Not used
24	WO_PM_Plan_Code	Not used
25	WO_Cost_Est	Not used
26	WO_Creation_Year	Not used
27	WO_Creation_Month	Not used
28	WO_Emergency_code	Not used