**Final Project report for forecasting and Prediction**

**Course: BUSI650**

**Team Name:**

**Section:**

**Team Members:**

|  |  |
| --- | --- |
| **Student Name** | **ID** |
|  |  |
|  |  |
|  |  |
|  |  |

**Introduction**

This final report depicts the results of two different forecasting and prediction analyses performed on different sets of data. In the first part, we focused on forecasting power usage for furnaces using Microsoft Excel. On the second part we delved into predicting miles per gallon (mpg) for cars using a set of libraries in python.

**Part A Power Usage Forecasting for Furnaces (Excel)**

### **a) Overview**

In this part, we were tasked with forecasting power usage for two furnaces located in Calgary and Mississauga and are both operated by ArcelorMittal Dofasco which is a steel company based in Canada. The aim was to optimize power consumption whilst minimizing extra charges that are incurred when the contracted power limit is exceeded.

**b) Methodology and Tools**

Microsoft Excel was used to perform the forecasting on historical furnace data. To achieve these three forecasting methods were applied:

1. 3-Month moving average
2. Weighted Moving Average
3. Exponential smoothing

The above techniques were evaluated based on its accuracy in predicting power usage for the different furnaces.

To achieve this performance rating, we calculated Root Mean Squared Error (RMSE) and Means Squared Error (MSE) for each forecasting method to provide insights on their accuracy (Lawrence, Klimberg, & Lawrence, 2009). The recommendations for furnace usage for the month of September 2023 were also based on this forecasting results with an aim to reduce extra power costs.

**Part B and C: Car MPG Prediction (Python)**

**a) Overview**

This analysis focused on prediction of miles per gallon (mpg) for cars on the mtcars.csv dataset. The main aim is exploration and contrasting the performance of different models in predicting the fuel efficiency of a car.

**b) Methodology and Tools**

Python was used for data analysis, visualization and modelling. The regression models used include:

1. Multiple Linear Regression
2. Random forest Regression
3. Simple Linear Regression

The performance of each model was then measured through metrics such as MAE, MSE, RMSE and R-squared for each regression model. The choice of the model was based on one which provided the most accurate predictions of car mpg.

**Hypothesis**

**a) Hypothesis for Forecasting Furnace power usage**

**i) Null Hypothesis (HO)**

The 3-month Moving average, Exponential smoothing and Weighted Moving Average techniques for projecting power utilization for the furnaces will not significantly differ in their accuracy.

**ii) Alternate Hypothesis (H1)**

At least one of the forecasting methods will out perform the others in terms of performance and thus provide a more accurate and precise forecast of power usage.

b) Hypothesis for Prediction of car MPG

**i) Null Hypothesis (HO)**

The performance metrics of Simple linear Regression, Random Forest Regression models and Multiple Linear Regression with selected predictors in predicting MPG will not have a significant difference and thus their predictions will be similar.

**ii) Alternate Hypothesis (H1)**

That no less than one of the regression models will prove more predictive performance, yielding more accurate predictions of car mpg compared to others.

The hypotheses drawn will lay foundation for evaluation of both the forecast and prediction methods employed and thus provide guidelines for analysis by providing expected relative performance (Lebanon & El-Geish, 2018).

**Data Specifications**

**Part A: Data on Furnace Power Usage (Excel)**

**Source:** ArcelorMittal Dofasco, a steel firm situated in Hamilton, Ontario, Canada, provided the data.

**Location:** The furnaces under consideration are in Calgary and Mississauga, respectively.

**Frequency:** The data covers the monthly power usage of the furnaces from January 2020 to September 2022.

**Variables:** Date, ID, and power demand (in MWh) for both Furnace 1 (Calgary) and Furnace 2 (Mississauga).

The goal of studying this data is to precisely estimate power demand in order to optimize power use and reduce extra charges paid while exceeding contracted power limits.

**Part B: Dataset mtcars**

**Source**: The mtcars dataset is a classic in data science and statistics, and it is frequently used for instructional and research reasons.

**Content:** The dataset includes data on numerous car models, including specifications and performance indicators.

**Variables**: These are attributes that describe the features and performance of a car, such as miles per gallon (mpg), horsepower (hp), weight (wt), and others.

The goal of this section of the analysis is to create prediction models that estimate miles per gallon (mpg) depending on car specs. This enables us to comprehend and evaluate various regression algorithms for prediction.

**Background**

**Algorithms used for forecasting (Furnace Power Usage)**

3-Month Moving Average Weighted Moving Average, Smoothing on an Exponential Scale.

**Metrics for Evaluation**

**MSE,** Calculates the average squared difference between predicted and actual power consumption. **RMSE** is the square root of MSE, which provides a readable metric of forecast error (Lawrence, Klimberg, & Lawrence, 2009).

**Algorithms used for prediction (car MPG prediction)**

First will be the **Simple Linear Regression** which uses a one-predictor linear regression model. Secondly is the **Multiple Linear Regression** with several predictors (hp, wt, qsec) is known as multiple linear regression and Lastly **Random Forest Regression** a machine learning ensemble approach for regression challenges (Montiel et al., 2021).

**Metrics for Evaluation**

1. **Mean Absolute Error (MAE).** Calculates the average absolute difference between expected and actual car performance.
2. **MSE.** The average squared difference between projected and actual automobile mpg readings.
3. **RMSE.** The square root of MSE, which provides a readable measure of prediction error.
4. **R2.** A metric used to assess of how effectively the model explains the variation in mpg values in cars. A greater R2 suggests a more accurate fit.

**Details Random Forest Regression**

This is an ensemble learning technique that makes predictions by combining many decisions tree regressors. It includes the following important details:

Random Forest generates an ensemble of decision trees, trained on a distinct subset of the data (bagging). This reduces overfitting while improving forecast accuracy. It also has feature randomization that adds unpredictability to the model by selecting a random subset of characteristics for each tree, making it more resilient and less susceptible to individual features. Predictions from individual trees are pooled to produce the final forecast, which frequently results in improved generalization. The model is capable of capturing complex, non-linear interactions between predictors and the target variable (Cutler, Cutler, & Stevens, 2012).

In the analysis of the mtcars dataset, Random Forest Regression is used to forecast automobile miles per gallon (mpg) based on various car parameters. Because of its capacity to handle non-linear interactions and avoid overfitting, it is an excellent tool for regression problems. To test the model's predictive performance, MAE, MSE, RMSE, and R-squared are used (Montiel et al., 2021).

**Methodology**

**Furnace Power Usage Forecasting (Excel)**

1. **Preparation of Data**
2. Import historical power consumption data for Furnace 1 (Calgary) and Furnace 2 (Mississauga).
3. Make a new column for each furnace's 3-month moving average.
4. Define weights for the weighted moving average.
5. Make a new column for each furnace's weighted moving average.
6. Create a new column for the anticipated values and set the initial values for exponential smoothing.
7. **Forecasting Techniques**
8. For each furnace, use the 3-Month Moving Average, Exponential Smoothing and Weighted Moving Average techniques on separate sheets whilst copying the original historical furnace data.
9. Forecasted values for each method should be calculated.
10. **Evaluation**

To evaluate the accuracy of each forecasting approach, compute MSE and RMSE.

1. **Analyze the results** and make a recommendation on the furnace to be used for September 2022 to minimize extra power charges.

**Predicting Car MPG (Python)**

**a) Exploration of Data**

1. Load the data and explore the structure and properties of the mtcars dataset.
2. Use scatter plots and correlation matrices to visualize variable relationships (Lebanon & El-Geish, 2018).

**b) Data Preparation**

1. Choose the appropriate predictors (horsepower, wt, qsec) and the target variable (mpg).
2. To evaluate the model, divide the dataset into training and testing sets.

**c) Models of Regression**

Simple Linear Regression, Random Forest Regression and Multiple Linear Regression are the three regression models to use. Then on the training dataset, train each model.

**d) Model Assessment**

Analyze each model's performance using MAE, MSE, RMSE, and R2. Finally, we compare the models to see which one delivers the most accurate car mpg prediction.

**Results**

**Part A: Furnace Power Usage Forecasting (Excel)**

**a) 3-Month Moving Average**

Calgary: MSE = **19042.23868,** RMSE = **137.9936182**

Mississauga: MSE = **14516.2346,** RMSE = **120.4833374**

Average RMSE = **129.2384778**

**b) Weighted Moving Average**

Calgary: MSE = **33358.28036,** RMSE = **182.6424933**

Mississauga: MSE = **27993.91141,** RMSE = **167.3138112**

Average RMSE = **174.9781522**

**c) Exponential Smoothing**

Calgary: MSE = **22168.11,** RMSE = **148.8895942**

Mississauga MSE = **15727.65146,** RMSE = **125.4099337**

Average RMSE = **137.149764**

**Part B: Predicting Car MPG (Python)**

The following are the findings for predicting automobile mpg using several regression models (already provided)

**a) Simple Linear Regression**

**MAE:** 3.23, **MSE:** 14.43, **RMSE:** 3.80

R-squared (R²): -0.21

**b) Multiple Linear Regression (hp, wt, qsec)**

**MAE:** 2.85, **MSE:** 11.98, **RMSE:** 3.46

**R-squared (R²):** -0.01

**c) Random Forest Regression**

**MAE:** 1.91, **MSE:** 4.77, **RMSE:** 2.18

**R-squared (R²):** 0.60

**Discussion**

This project included two independent parts: electricity usage predictions for steel industry furnaces using Excel and car miles per gallon (MPG) prediction using Python. Here are some of my primary thoughts and observations about the project:

## **i) Furnace Power Usage Forecasting**

Methodology Evaluation. We forecasted power usage for two furnaces (Calgary and Mississauga) using three different forecasting approaches. In terms of MSE and RMSE, 3-month moving average surpassed the other approaches. This strategy produced the most accurate projections of power demand having the lowest overall RMSE of 129.24.

Based on the forecasting results, it is proposed that 3-month moving average be used in September 2023 for both Furnace 1 (Calgary) and Furnace 2 (Mississauga). This technique has the lowest MSE and RMSE for both furnaces averaging and RMSE of 129.24 against 174.98 and 137.15 for Weighted moving average and Exponential smoothing respectively. This shows improved accuracy in power usage prediction.

**Recommendation for the furnace to use for September 2022**

The forecasted power usage for September 2022 using the 3-month moving average is as follows:  
**Calgary furnace** = 1090.82134 MWh

**Mississauga furnace** = 1187.696768 MWh

**Contracted limit** = 900 MWh

**Extra Charge** = $7/MWh

**a) Calgary**

Extra cost calculation

**Excess Usage** = =

**Extra Cost** =

**b) Mississauga**

Extra cost calculation

**Excess Usage** = =

**Extra Cost** =

Based on the calculations for September 2022 and considering the cost efficiency, it is recommended to use Calgary furnace for September 2022. The Calgary furnace with a forecasted power usage of 1090.82MWh had an extra cost of $1335.74. In contrast, the forecasted power usage of Mississauga for the same month is approximately 1187.70MWh which incurs an extra cost of $1971.9. As a result, choosing the Calgary furnace for September 2022 remains the most cost-effective option.

**ii) MPG Prediction for a Car**

We used three regression models in the automobile MPG prediction task: Simple Linear Regression, Multiple Linear Regression (including horsepower, weight, and quarter-mile time) as predictors, and Random Forest Regression. The evaluation measures revealed that the Random Forest Regression model fared better than linear regression models significantly, with lower MAE, RMSE and MSE and a higher R2.

The Random Forest Regression model identified horsepower (hp) as the most important feature in predicting car miles per gallon (mpg). The model's predicted accuracy was also influenced by weight (wt) and quarter-mile time (qsec).

## **Challenges**

1. **Model Selection Obstacles & Difficulties.** The choice of appropriate forecasting and regression models was crucial. We had to select the models that best suited the data and made correct predictions in both tasks. We were able to make more informed decisions thanks to the evaluation metrics.
2. **Interpreting data.** Careful analysis was necessary to interpret the data and determine the most essential features or elements impacting the outcomes. Understanding the significance of these characteristics is critical for making sound business decisions (Lebanon & El-Geish, 2018).

**Conclusion**

Here we summarize the findings and their implications for everyone.

**i) Furnace Power Usage Forecasting**

We were able to accurately anticipate how much electricity two steel furnaces would require in the future.

Our initial hypothesis for forecasting furnace power usage stated that the multiple forecasting methods' accuracy might not differ much (Null Hypothesis, HO), but that at least one of them might beat the others (Alternate Hypothesis, H1).

Our findings align with the alternate Hypothesis. The best method we discovered for making accurate forecasts was known as "3-month moving average" It assists the organization in saving money by using the most suitable furnace to meet their production need while avoiding unnecessary expenditures. We also performed calculations to determine the cost efficiency of operating both furnaces for September 2022 using the forecasted demands and found that Calgary was more cost efficient as it had less extra cost compared to Mississauga.

In the future, this data can be utilized to more efficiently plan electricity usage and decrease expenses even further (Loh, 2011).

**ii) Car MPG Prediction**

We analyzed data to anticipate how many miles per gallon (MPG) an automobile can get depending on factors such as horsepower, weight, and others.

We argued in our hypothesis for forecasting automobile MPG that there might not be a significant difference in the performance metrics of the regression models (Null Hypothesis, HO), but that at least one model might be more predictive (Alternate Hypothesis, H1).

Our findings support the alternative hypothesis (H1), as the Random Forest Regression model outperformed the linear regression models in terms of predictive performance. This link emphasizes the significance of selecting the correct model for the job, as it can lead to more accurate predictions and potentially influence car design decisions to enhance fuel efficiency.

Car makers can utilize this data to create vehicles that are more fuel-efficient and environmentally friendly. This data can be used to design greener, more fuel-efficient vehicles in the future (Loh, 2011).

**Future Prospects**

More advanced forecasting algorithms for electricity usage data can be investigated to make even more precise predictions (Lawrence, Klimberg, & Lawrence, 2009). This can save even more money on electricity while also lowering the environmental impact.

For automobile data, we can gather more information about new car models and utilize it to make more accurate forecasts (Loh, 2011). This can result in cars that use less fuel and emit less pollution.

**References**

Cutler, A., Cutler, D. R., & Stevens, J. R. (2012). Random forests. *Ensemble machine learning: Methods and applications*, 157-175.

Lebanon, G., El-Geish, M., Lebanon, G., & El-Geish, M. (2018). Visualizing Data in R and Python. *Computing with Data: An Introduction to the Data Industry*, 277-324.

Lawrence, K. D., Klimberg, R. K., & Lawrence, S. M. (2009). *Fundamentals of forecasting using excel*. Industrial Press Inc.

Loh, W. Y. (2011). Classification and regression trees. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, *1*(1), 14-23.

Montiel, J., Halford, M., Mastelini, S. M., Bolmier, G., Sourty, R., Vaysse, R., ... & Bifet, A. (2021). River: machine learning for streaming data in python. *The Journal of Machine Learning Research*, *22*(1), 4945-4952.

**Appendix**

**Google Collab Link:**