NAME: VEERANJANEYULU KONDRAGUNTA

MODULE TITLE: Machine Learning and Pattern Recognition

MODULE CODE: B9DA109

CA_ONE

DBS

1. INTRODUCTION

The aviation industry handles in a dynamic environment where the airfare pricing plays an important role in adapting the consumer choices and airline profitability. In this period of fluctuating volatile market conditions and enormous influencing factors, the ability to accurately predict the airfare prices is important for both industry stakeholders and the consumers. Identifying the importance of this predictive modelling task, this study embarks on a in depth exploration of airfare prediction using R programming and 5 different machine learning models: Linear Regression, Random Forest, Decision Tree, Ridge Regression, and Lasso Regression.

Airfare prediction is a intricate challenge, as it includes the complicated involvement of various attributes such as seasonal trends, fuel prices, competition dynamics, and customer demand. Conventional methods often fall short in identifying the nuances of these relations, needing the utilization of the advanced machine learning methods. R programming, prominent for its credibility in statistical computing and data analysis, provides an ideal platform for implementing and the evaluating these models.

The research uses the three datasets obtained from Kaggle, each including a range of features attributes that contribute to the resolving of airfare prices. Through rigid preprocessing steps, presenting the handling of missing data and standardizing the variables, the datasets are preprocessed for the application of machine learning algorithms. The selected machine learning models are chosen for their different methodologies, starts from the simplicity of linear regression to the complexity of the ensemble methods like Random Forest.

As the aviation industry which continually seeks the innovative solutions to improve the pricing strategies and enhance the revenue management, the results of this research are expected to provide the valuable insights. The comparative analysis of the different machine learning models’ goal to recognize the most effective approach in predicting the airfare prices, contributing to the development of the robust and accurate prediction tools.

This report extends with the detailed exploration of the datasets, followed by the comprehensive discussion of the methodology’s approaches used in the machine learning models. Consequently, the outcomes of the analyses are represented and discussed, straighten out to the strengths and failings of each model. The implications of these outcomes for the aviation industry and strength ways for future research are then discussed in the conclusion. Therefore, this attempt, we goal to offer the meaningful contribution to the ongoing discussion on airfare prediction and its enormous implications for the developing landscape of the airline industry.

2. RELATED WORK

Airfare prices prediction has been a topic of significant interest within the field of aviation and data science, with the researchers and the practitioners utilizing the various methodologies to improve the prediction accuracy. This section reviews relevant literature and research that have contributed to the in depth understanding of airfare prediction, outlining the importance of machine learning models and the utilization of R programming.

- **Belobaba, P.P. (1987). Airfare Determination under Yield Management.**

Belobaba's influential work represents the early applications of yield management in airfare pricing, highlighting the significance of robust pricing strategies. While not interested on predictive modeling, the research represents the foundational understanding of factors influencing airfare pricing.

- **Hansen, M., & Yu, B. (2001). Model Selection and the Principle of Minimum Description Length.**

Hansen and Yu's work delves into the selection of model techniques, providing the insights into the elements of selecting the effective models. This is related to our research as it shows the path for selection of machine learning model algorithms based on their capability to reduce the description length and enhance the predictive performance.

- **Fernandes, K., & Ferreira, A. (2016). Predicting Airfare Prices with Artificial Neural Networks.**

Fernandes and Ferreira investigated about the use of artificial neural networks in airfare prediction. While different from our chosen machine learning models approaches, their research offers a foundation for the extensive scope of predictive modeling in the aviation sector for prediction of fare prices.

- **James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning. Springer.**

James et al.'s thorough investigation offers as a foundational resource for the understanding statistical learning methods. The research is instrumental in showing the path for implementation and evaluation of machine learning models, relates with our study's aims on predictive modeling.

- **Wickham, H., & Grolemund, G. (2016). R for Data Science. O'Reilly Media.**

Wickham and Grolemund's book offers a practical guide to using the R for data analysis and visualization. Given our dependence on R programming, this book is instrumental in navigating the complexities of data manipulation and the model implementation.

By employing these works, our researcher's objective to contribute to the changing landscape of airfare prediction, mainly focusing on the comparative evaluation of various multiple machine learning models using R programming. The combination of these extensive approaches informs our research methodology and improves the robustness of our predictive modeling approach.

3. DATASETS

In order to start our research study, we depend on three datasets accessible on Kaggle, each providing different viewpoints on the variables attributes that impacts the airfare pricing.

1.Plane Ticket Price Dataset: This dataset provides the extensive perspective on various attributes that impact the ticket prices. The attributes include details about the airline, departure and arrival locations, flight duration, and flight type (e.g., one- way or round trip). Therefore, the dataset contains the information such as the total number of stops, the specific times of departure and arrival, and the price fare of the ticket. The dataset's loads make it an important asset for developing the robust predictive model.

Dataset: https://www.kaggle.com/datasets/ibrahimelsayed182/plane-ticket-price

2. Airfare ML: Predicting Flight Fares: This data specifically contains the attributes such as the airline, source and destination airports, flight route, and the overall duration of the travel. Therefore, the dataset has the data regarding to departure and arrival dates, the class category (economy, business), and the ticket fare price, along with its core characteristics. A Comprehensive analysis of this dataset improves our understanding of the various factors that impact airfare prices.

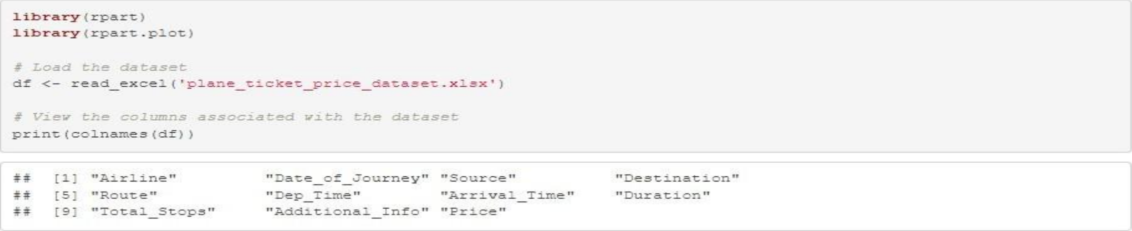
Dataset: https://www.kaggle.com/datasets/yashdhar me36/airfare-ml-predicting-flight-fares

3.Flight Price Prediction Dataset: This dataset contains the attributes on various characteristics, including the airline specifics, flight origin and destination, number of layovers, departure and arrival times, and route duration. Importantly, the dataset includes the features such as the journey date and the travel class. These further attributes provide a nuanced viewpoint on the time-related factors which impacts the airfare prices, thereby enhancing the overall analysis.

Dataset: https://www.kaggle.com/datasets/shubhamb athwal/flight-price-prediction

4. METHODOLOGY

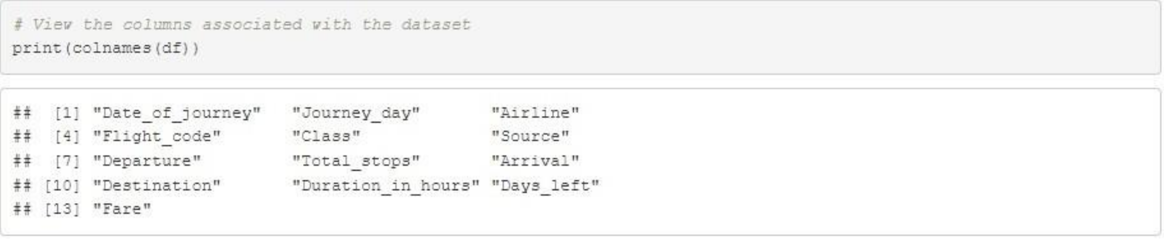
1. Data Loading and Exploration: The study begins with the loading of important libraries, which includes the "tidyverse" for data manipulation and visualization, and specific machine learning libraries such as "randomForest", "e1071", "glmnet", “rpart”, and “rpart.plot”. The datasets, named are 'airfare_ml_dataset.csv', ‘flight_price_data’ is loaded using the read_csv function and “plane_ticket_price” is loaded through the read_excel fucntion because this dataset format was .xlsx, and its structure and content are examined.



```
library(rpart)
library(rpart.plot)

# Load the dataset
df <- read_excel('plane_ticket_price_dataset.xlsx')

# View the columns associated with the dataset
print(colnames(df))
```

```
##  [1] "Airline"         "Date_of_Journey" "Source"          "Destination"
##  [5] "Route"           "Dep_Time"        "Arrival_Time"    "Duration"
##  [9] "Total_Stops"     "Additional_Info" "Price"
```

Figure 1: Loading of the Plane Ticket Dataset



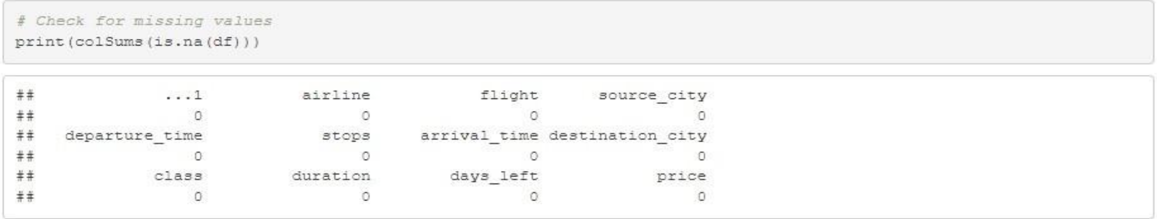
```
# View the columns associated with the dataset
print(colnames(df))
```

```
##  [1] "Date_of_journey"   "Journey_day"       "Airline"
##  [4] "Flight_code"       "Class"             "Source"
##  [7] "Departure"         "Total_stops"       "Arrival"
## [10] "Destination"       "Duration_in_hours" "Days_left"
## [13] "Fare"
```

Figure 2: Column Attributes in Airfare_ml Dataset

2. Data Preprocessing:

Handling Missing Values: Missing values in the datasets which are identified and then consequently removed them using the ‘na.omit’ function, assuring the data integrity for further analyses.

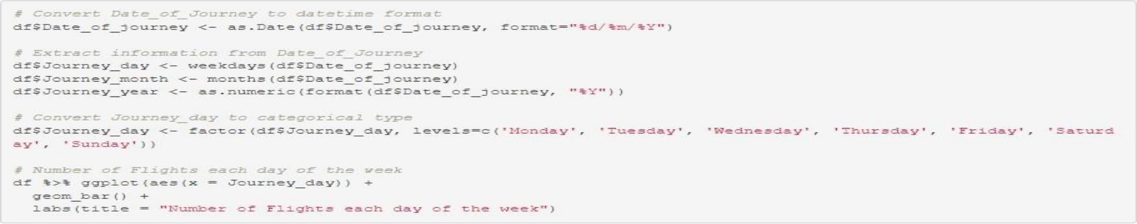


```
# Check for missing values
print(colSums(is.na(df)))
```

```
##              ...1          airline           flight      source_city
##                 0                0                0                0
##    departure_time            stops     arrival_time destination_city
##                 0                0                0                0
##             class         duration        days_left            price
##                 0                0                0                0
```

Figure 3: Missing Values

Feature Engineering: The date-related attribute variable “Date_of_journey” is transformed to a datetime format, and additional features (Journey_day, Journey_month, and Journey_year) are extracted to recognize the temporal patterns. Categorical characteristics are encoded numerically using a LabelEncoder method approach and dropped the unnecessary columns from the dataset.



```
# Convert Date_of_Journey to datetime format
df$Date_of_journey <- as.Date(df$Date_of_journey, format="%d/%m/%Y")

# Extract information from Date_of_Journey
df$Journey_day <- weekdays(df$Date_of_journey)
df$Journey_month <- months(df$Date_of_journey)
df$Journey_year <- as.numeric(format(df$Date_of_journey, "%Y"))

# Convert Journey_day to categorical type
df$Journey_day <- factor(df$Journey_day, levels=c('Monday', 'Tuesday', 'Wednesday', 'Thursday', 'Friday', 'Saturd
ay', 'Sunday'))

# Number of Flights each day of the week
df %>% ggplot(aes(x = Journey_day)) +
  geom_bar() +
  labs(title = "Number of Flights each day of the week")
```

Figure 4: Convert the Journey Day to Categorical

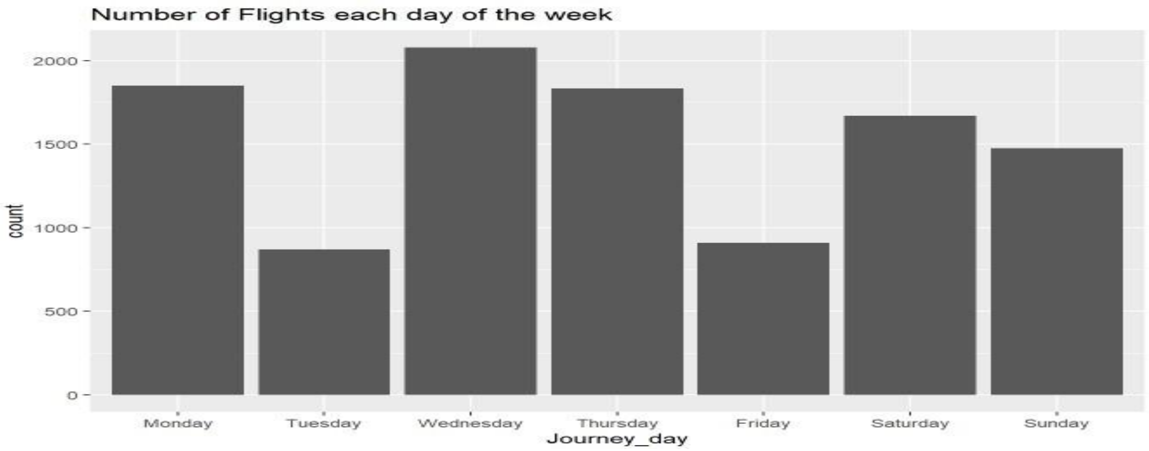


Figure 5: Frequency of Flights (Day Wise)

3. Exploratory Data Analysis & Visualization: Exploratory Data Analysis is directed to obtain the important insights into the distribution of airlines, airfare prices, and the relationship between airfare and airlines which all not visible without visualizations. Where the visualizations, including the polar bar chart for airline distribution, a histogram for airfare distribution, and a boxplot for airfare vs. airline, are generated.

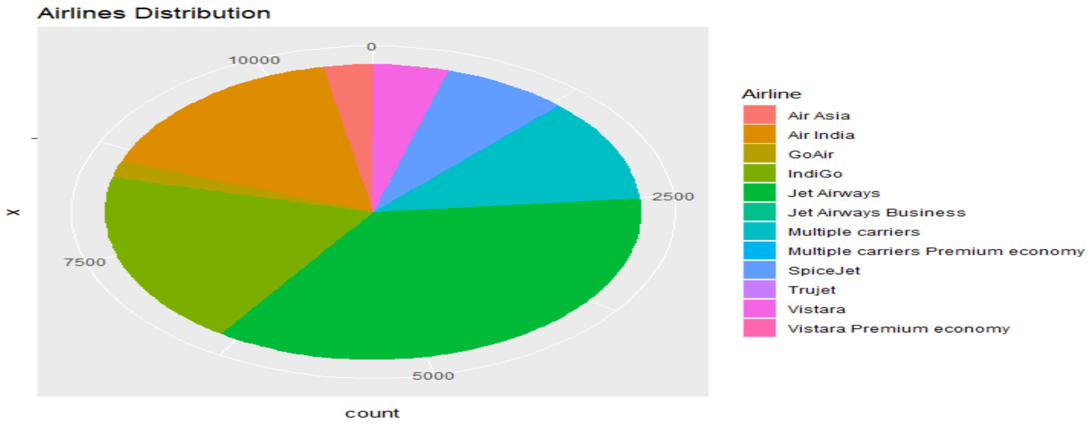
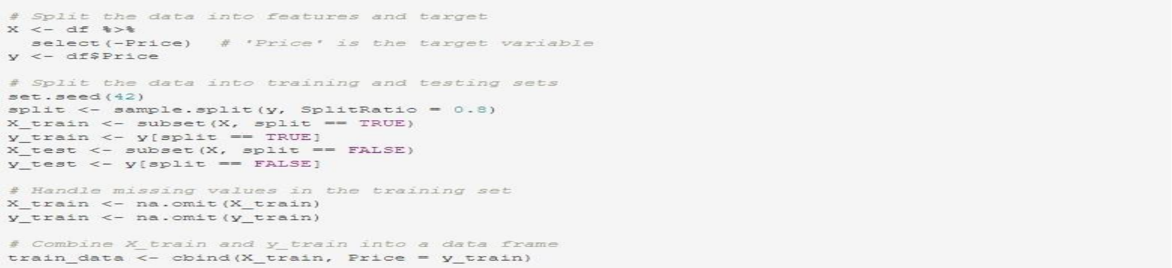


Figure 6: Representations of Airlines Distribution

4. Model Training and Evaluation:

Train-Test Split: The dataset is get split into training (80%) and testing (20%) sets using the sample.split function. Missing values in the training set which are left during the preprocessing step is get handled using the na.omit function.



```
# Split the data into features and target
X <- df %>%
  select(-Price)   # 'Price' is the target variable
y <- df$Price

# Split the data into training and testing sets
set.seed(42)
split <- sample.split(y, SplitRatio = 0.8)
X_train <- subset(X, split == TRUE)
y_train <- y[split == TRUE]
X_test <- subset(X, split == FALSE)
y_test <- y[split == FALSE]

# Handle missing values in the training set
X_train <- na.omit(X_train)
y_train <- na.omit(y_train)

# Combine X_train and y_train into a data frame
train_data <- cbind(X_train, Price = y_train)
```

Figure 7: Splitting the Dataset into Training and Testing

Model Implementation: Five different machine learning models are implemented and then trained on the training dataset for all three airfare datasets:

a) Linear Regression: Implemented using the lm function.
b) Decision Tree: Constructed using the rpart function, with a subsequent visualization using rpart.plot.

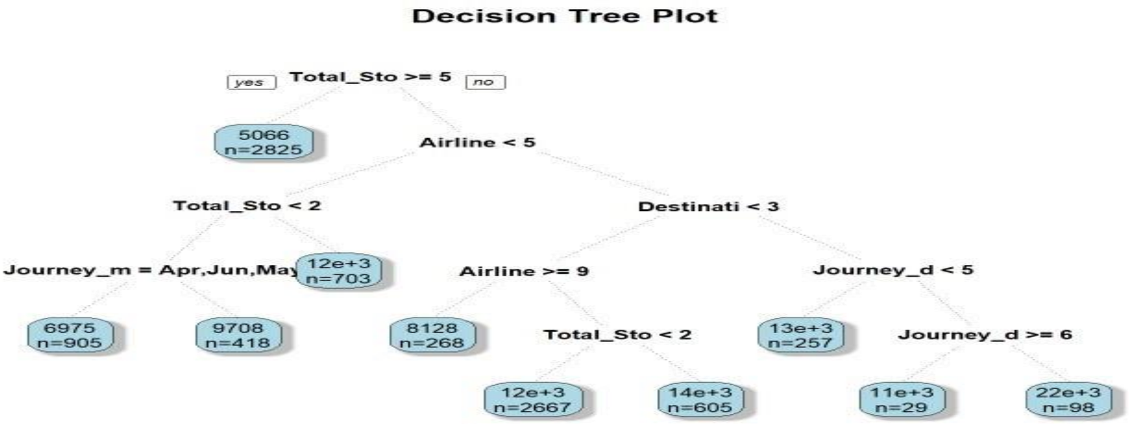
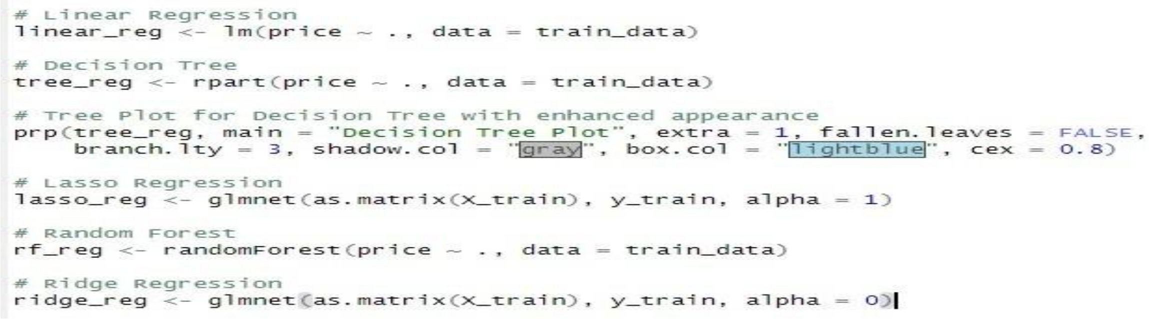


Figure 8: Trained Decision Tree Model

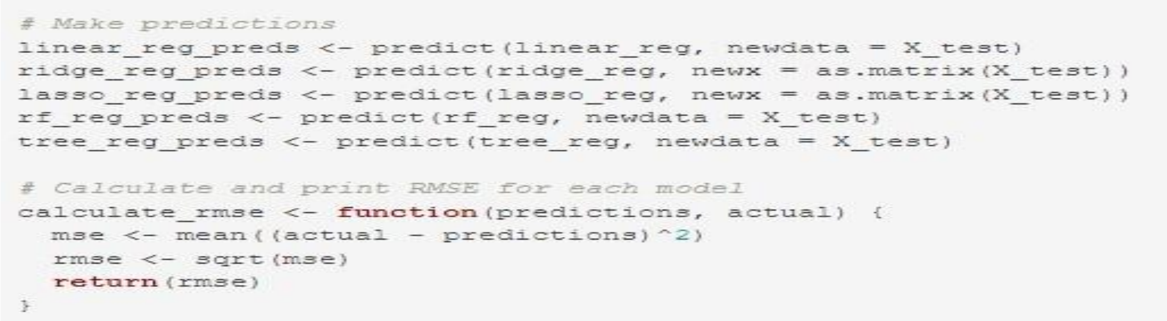
c) Lasso Regression: Utilized the glmnet function with alpha set to 1.
d) Random Forest: Employed the randomForest function.
e) Ridge Regression: Utilized the glmnet function with alpha set to 0.



```
# Linear Regression
linear_reg <- lm(price ~ ., data = train_data)

# Decision Tree
tree_reg <- rpart(price ~ ., data = train_data)

# Tree Plot for Decision Tree with enhanced appearance
prp(tree_reg, main = "Decision Tree Plot", extra = 1, fallen.leaves = FALSE,
    branch.lty = 3, shadow.col = "gray", box.col = "lightblue", cex = 0.8)

# Lasso Regression
lasso_reg <- glmnet(as.matrix(X_train), y_train, alpha = 1)

# Random Forest
rf_reg <- randomForest(price ~ ., data = train_data)

# Ridge Regression
ridge_reg <- glmnet(as.matrix(X_train), y_train, alpha = 0)
```

Figure 9: Model Implementations

Model Evaluation: Predictions are generated using each model on the testing datasets, and Root Mean Squared Error (RMSE) is calculated to evaluate the model performance. The calculate_rmse function is defined to promote this process.



```
# Make predictions
linear_reg_preds <- predict(linear_reg, newdata = X_test)
ridge_reg_preds <- predict(ridge_reg, newx = as.matrix(X_test))
lasso_reg_preds <- predict(lasso_reg, newx = as.matrix(X_test))
rf_reg_preds <- predict(rf_reg, newdata = X_test)
tree_reg_preds <- predict(tree_reg, newdata = X_test)

# Calculate and print RMSE for each model
calculate_rmse <- function(predictions, actual) {
  mse <- mean((actual - predictions)^2)
  rmse <- sqrt(mse)
  return(rmse)
}
```

Figure 10: Prediction of Trained Models

5. RESULT & EVALUATION

The calculated results of RMSE values for each models are represented and then compared using a bar chart. Furthermore, a variable importance plot is created for the Random Forest model, providing the insights into the importance of different features attributes in predicting airfare prices.

Root Mean Square Error Metrics			
Models	**Flight Price Data**	**Plane TicketPrice Data**	**Airfare ML Data**
Linear Regression	21809.35	3295.692	19511.17
Ridge	22202.5	3987.486	19917.31
Lasso	21900.03	3471.315	19667.87
RandomForest	20048.47	2391.032	17949.9
DecisionTree	19689.06	2600.988	18295.56

Table 1: Comparative Results

Flight Price Dataset: In this dataset, the Random Forest model displays the lowest RMSE, identifying the dominant predictive performance compared to other models. The Decision Tree model also performs well, outstanding the Linear and Ridge Regression.

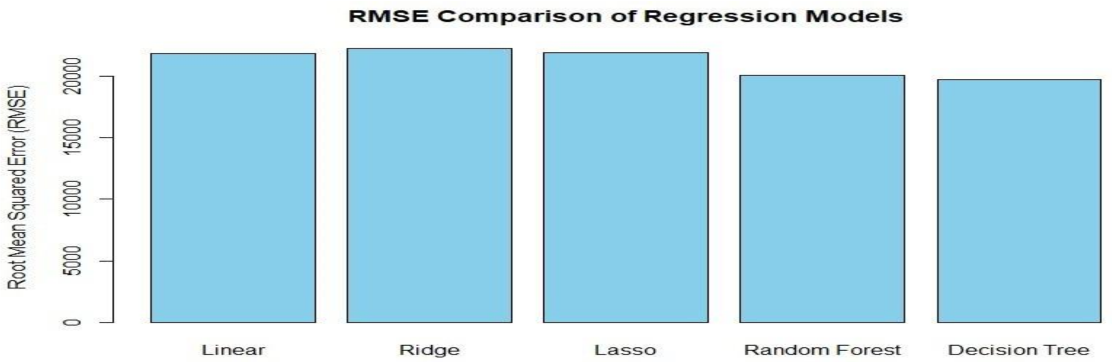


Figure 11: RMSE Results (Flight Price Data)

Plane Ticket Price Dataset: The Plane Ticket Price dataset displays the notable performance differences among the models. Random Forest again represents the lowest RMSE, highlighting the strong predictive capabilities. Linear Regression and Lasso Regression perform some well, while Ridge Regression weakens behind.

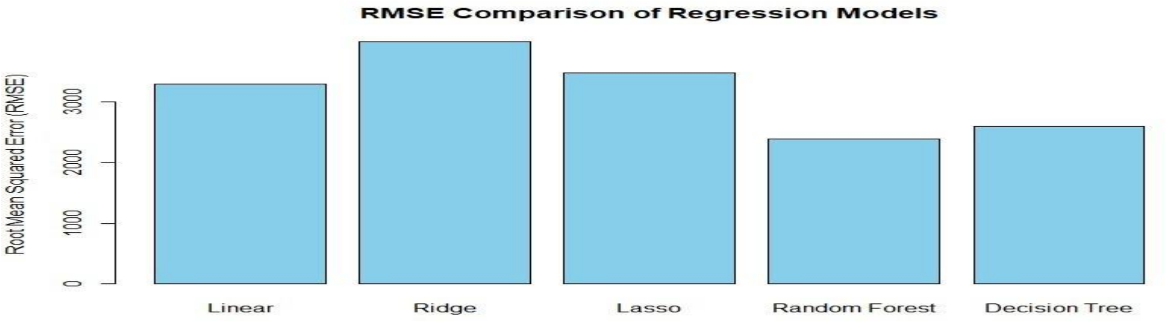


Figure 12: RMSE Results (Plane Ticket Prices Data)

Airfare ML Dataset: In the Airfare ML dataset, Random Forest once again performed well with the lowest RMSE, succeeded closely by the Decision Tree model. Linear, Ridge, and Lasso Regression models highlights the higher RMSE values but still contribute valuable predictions.

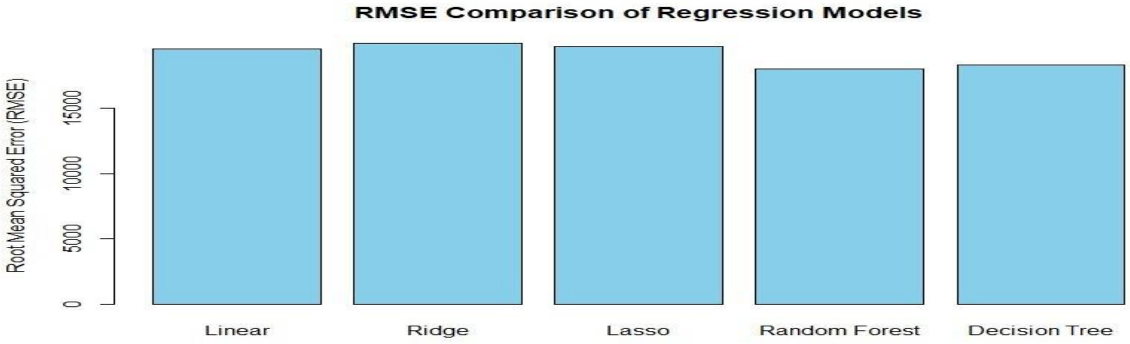


Figure 13: RMSE Results (Airfare ML Data)

This below illustrations demonstrates the features which are playing the major role for performing the prediction of airfare prices in Random Forest.

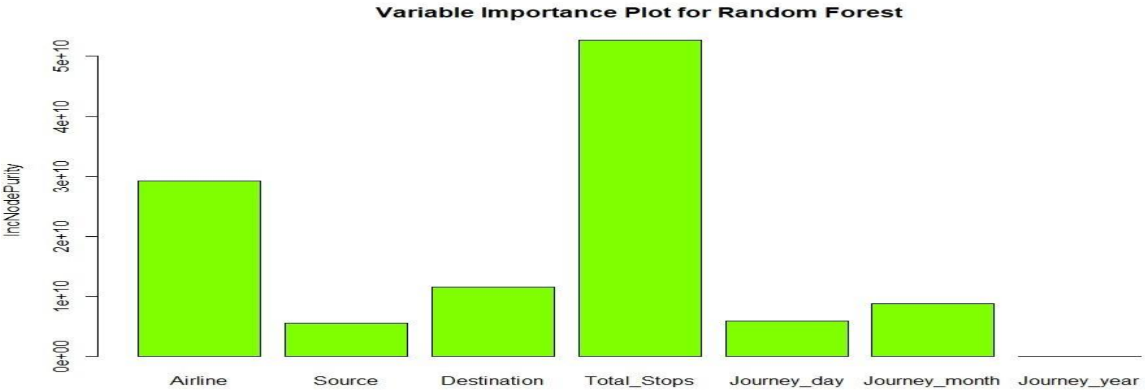


Figure 14: Important Features for Fare Prediction

Summary and Implications:

- **Consistent Model Ranking:** Across all the three datasets, Random Forest constantly performs well, highlighting its robustness in identifying the complex relationships within the data.

- **Decision Tree Performance:** Decision Tree models also demonstrate the competitive performance, especially in the Flight Price and Airfare ML datasets.

- **Linear Regression Variability:** The performance of Linear Regression varies through the datasets, with higher RMSE values, recommending the limitations in identifying the non-linear patterns present in the data.

- **Regularization Techniques:** Ridge and Lasso Regression, which includes the regularization, shows the competitive but slightly higher RMSE values when compared it to the ensemble methods.

In conclusion, the selection of the dataset which impacts the model performance, and Where Random Forest arises as a reliable choice for airfare prediction across all diverse datasets. The results provide valuable insights for selecting the appropriate models in the field of airline fare prediction.

6. DISCUSSION

The evaluation of airfare prediction models across all three different datasets—Flight Price, Plane Ticket Price, and Airfare ML—discovers the important insights into the performance and adaptability of different machine learning algorithms for this complicated task. The discussion encompasses with the observed trends, model potentials and failures, implications for the aviation field industry, and suggestions for future research.

1. Consistent Model Performance: Decision Tree models display the competitive performance, representing their potential in airfare prediction. Especially, Decision Trees provide the illustrations, making them suitable for understanding the reasoning behind predictions is important.

2. Challenges with Linear Models: Linear Regression models constantly represent the higher RMSE values, highlights the challenges in identifying the non-linear complex relationships involved in airfare data. These models may face consequences to adapt to the various complex dynamics of factors influencing the airfare prices.

3. Implications for the Aviation Industry: The accurate predictive abilities of Random Forest and Decision Tree models have the significant indication for decision support systems within the aviation field of industry. Airlines and travel agencies can utilize these models to enhance the pricing strategies, improves the revenue management, and respond to the market dynamics more adequately.

4. Dataset-Specific Variances: The choice of dataset leverages the model performance, highlighting the importance of understanding the various unique characteristics of the data. While Random Forest constantly excels, the variance in RMSE values across all the datasets highlights the necessary need for dataset-specific model selection.

5. Recommendations for Future Research:

- **Feature Importance Analysis:** Future research could delve deeper into the feature importance analysis to understand the more relevant features, particularly within Random Forest models. Understanding which features significantly impacts on the airfare predictions can provide useful insights for stakeholders.

- **Temporal Dynamics:** Considering the temporal nature of airfare pricing, more studies may explore the integration of time series analysis approaches to catch the evolving patterns and trends in the dataset.

- **Ensemble Model Exploration:** Investigating the potential benefits of combining the various models, such as an ensemble of Random Forest and the linear models, could provide the improved predictive accuracy.

7. CONCLUSION

In conclusion, this study illustrates the effectiveness of the machine learning models in forecasting the airfare prices, with Random Forest and Decision Tree models occurring as top performers. The outcome results provide the important valuable guidance for the stakeholders in the field aviation industry, offering the models that can improve it in decision-making processes and contribute to them for more accurate airfare predictions. As the industry continues to constantly evolve, the ongoing research and development in predictive modeling remain necessary for staying ahead of the extensive dynamic market conditions and meeting according to the customer expectations.

8. REFERENCES

[1] Wang, T., Pouyanfar, S., Tian, H., Tao, Y., Alonso, M., Luis, S., & Chen, S. C. (2019, July). A framework for airfare price prediction: a machine learning approach. In 2019 IEEE 20th international conference on information reuse and integration for data science (IRI) (pp. 200-207). IEEE.

[2] Tziridis, K., Kalampokas, T., Papakostas, G. A., & Diamantaras, K. I. (2017, August). Airfare prices prediction using machine learning techniques. In 2017 25th European Signal Processing Conference (EUSIPCO) (pp. 1036-1039). IEEE.

[3] Vu, V. H., Minh, Q. T., & Phung, P. H. (2018, January). An airfare prediction model for developing markets. In 2018 International Conference on Information Networking (ICOIN) (pp. 765-770). IEEE.

[4] Papadakis, M. (2014). Predicting Airfare Prices. Clerk Maxwell.

[5] Ren, R., Yang, Y., & Yuan, S. (2014). Prediction of airline ticket price. University of Stanford, 1-5.