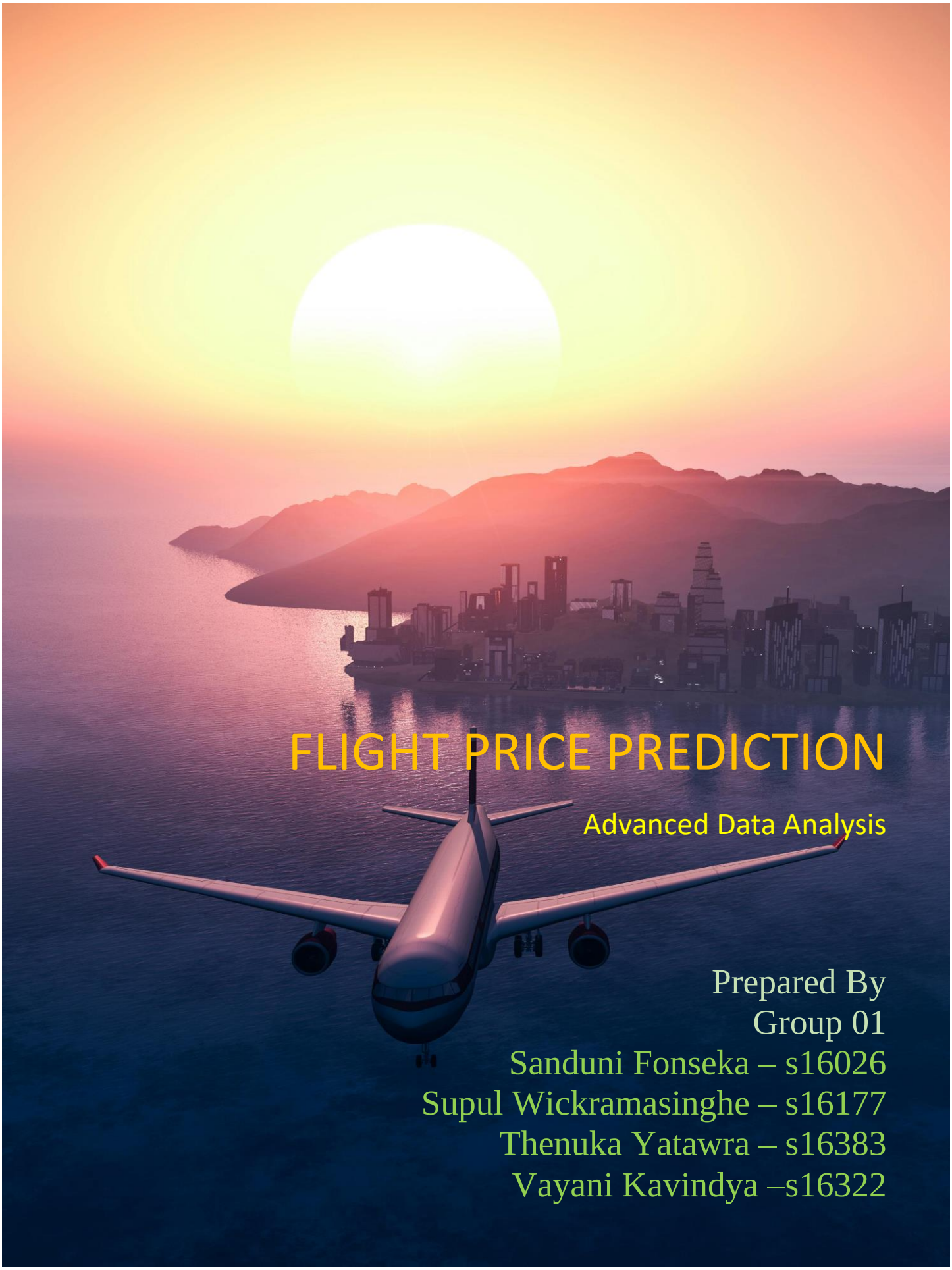

FLIGHT PRICE PREDICTION

Advanced Data Analysis

Prepared By
Group 01
Sanduni Fonseka – s16026
Supul Wickramasinghe – s16177
Thenuka Yatawra – s16383
Vayani Kavindya –s16322

ABSTRACT

This report presents an advanced analysis conducted by utilizing machine learning algorithms. It was built upon insights gleaned from a prior descriptive analysis with the aim of constructing a model to predict the actual price of the flight ticket. For this task, we used a flight price dataset sourced from Kaggle. This innovative approach provides valuable insights for airlines, travelers, travel agencies, airport authorities & tourism boards and researchers in the airline industry. Moreover, it contributes to a deeper understanding of the dynamic factors influencing efficiency and profitability in the airline industry, enabling data-driven decision-making for optimal pricing strategies and resource management.

Contents

List of Figures

List of Tables

Introduction

The airline industry, a cornerstone of global transportation, plays a pivotal role in economies worldwide, contributing significantly to commerce, tourism, and connectivity. As this industry thrives on efficiency and demand-driven pricing, the importance of accurate flight price prediction cannot be overstated. This report embarks on a journey to unravel the intricate dynamics of airfare pricing, with a focus on constructing a predictive model. Designed to offer comprehensive insights into the factors influencing flight prices, this model serves as a valuable resource for travelers, airlines, travel agencies, researchers, and industry stakeholders alike.

Description of the question

The airline industry is a crucial component of global transportation, driving commerce, tourism, and connectivity across nations. Particularly in regions with growing travel demands, such as South Asia, Southeast Asia, and Africa, the ability to predict flight prices accurately can significantly impact economic decisions. Despite advancements in technology and data analytics, airfare pricing remains highly dynamic due to multiple influencing factors, including seasonal demand fluctuations, airline pricing strategies, fuel costs, and market competition.

Unlike industries with fixed pricing structures, airline ticket prices are subject to real-time changes driven by demand-supply mechanisms, route popularity, and booking time. This complexity makes it challenging for travelers, travel agencies, and airlines to navigate pricing trends efficiently. Sudden price surges and fluctuations can lead to financial inefficiencies and missed opportunities for both consumers and service providers.

To address this challenge, flight price prediction models play a crucial role in enabling better decision-making. By leveraging historical data and key influencing factors, an accurate predictive model can help travelers book flights at the best possible fares, assist airlines in optimizing revenue strategies, and empower travel agencies with valuable insights for improved customer service.

Therefore, our primary objective is:

- **To construct a reliable model to predict flight prices by identifying the key factors that significantly influence airfare variations.**

Description of the dataset

The dataset consists of 300,153 observations and 11 variables related to flight bookings from the kaggle website. The data includes flight information for travel between India's top 6 metro cities. There were no missing values or repeated values detected in the data set therefore we didn't use any pre-processing for it. The variables in the dataset are as follows:

Table 1: Description of variables

Variable	Description	Data type
Airline	The airline which operates the flight.	Categorical
Flight	A concatenated code representing the flight code	Categorical
Source_city	The city from which the flight departs.	Categorical
Departure_time	This is a derived categorical feature obtained by grouping time periods into 6 bins.	Categorical
Stops	A categorical feature with 3 distinct values that stores the number of stops between the source and destination cities.	Categorical
Arrival_time	This is a derived categorical feature created by grouping time intervals into bins. It has six distinct time labels and keeps information about the arrival time.	Categorical
Destination_city	The city to which the flight arrives.	Categorical
Class	Indicates whether the flight is economy or business class.	Categorical
Duration	The total time the flight takes from departure to arrival in hours.	Numerical
Days_left	This is a derived characteristic that is calculated by subtracting the trip date by the booking date.	Numerical
Price	The cost of the ticket for the flight.	Numerical

Table 1: Description of variables

Important Results of the Descriptive Analysis

The distribution of the response variable

The response variable, 'price,' exhibited significant positive skewness as the histogram indicates a non-normal distribution. This characteristic is important to acknowledge as it can potentially influence the performance and assumptions of certain modeling techniques. Specifically, skewed data can lead to models that are overly sensitive to extreme values or violate the assumption of normally distributed residuals in linear regression. While we chose not to apply a log transformation in this analysis, it's crucial to be aware of this skewness and consider its potential impact on the models.

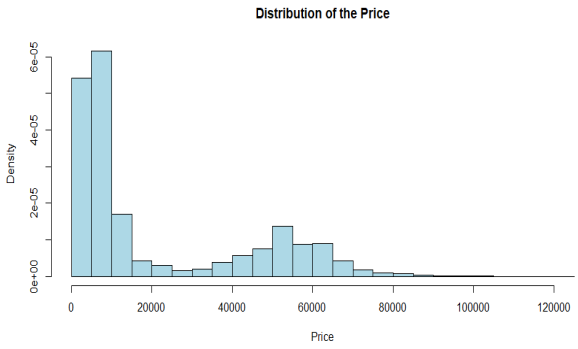

Figure 1: Histogram of Price distribution

Correlation between the predictor variables

Multicollinearity, the high correlation between predictor variables, can pose challenges for regression models. It can inflate the variance of coefficient estimates, making them unstable and difficult to interpret. To assess multicollinearity, we examined the correlation matrix of the predictor variables.

Table 2: Correlation value between variables

Variable 1	Variable 2	Correlation
stops - Zero	duration	-0.5140764
duration	stops - zero	-0.5140764
arrival_time - Night	arrival_time - Evening	-0.3941955
arrival_time - Evening	arrival_time - Night	-0.3941955

These correlations suggest potential redundancy in the predictor variables, which could affect the stability of our models. To address this issue, we employed regularization techniques, specifically Ridge, Lasso, and Elastic Net regression, which are known to be effective in handling multicollinearity.

Score plot of the FAMD

To gain insights into the underlying structure of our data, we performed Factor Analysis of Mixed Data (FAMD) and examined the score plot of the first two dimensions.

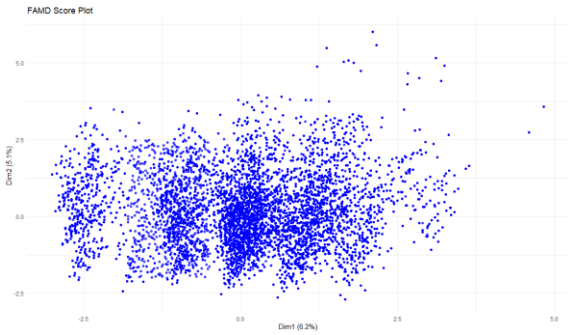

Figure 2: Score plot of first two dimension using FAMD

The FAMD score plot reveals a relatively uniform distribution of data points, with no clear, distinct clusters or groupings evident. This observation suggests that the observations are not naturally separated into distinct categories based on the first two principal components. The lack of clear clustering supports our decision to build a single model for the entire dataset, as there is no apparent need to create separate models for different subgroups.

The exploratory analysis provided valuable information about the characteristics of the data, including skewness, multicollinearity, and clustering patterns. This information directly influenced the choice of modeling techniques and informed the interpretation of the model results.

Important Results of the Advanced Analysis

Multiple Linear Regression (MLR)

MLR as our first option emerged because it is the simplest and fundamental model which align with the characteristics of our data set. We utilized best subset selection method to select the best variables for our model. Best subset selection procedure gave us the model with all the variables. Below is the result we obtained.

Table 3: Evaluation matrix for best subset selection

	RMSE	**R^2**
Training	6752.5028	0.9115
Test	6760.2889	0.9113

Furthermore, to check the validity of the given model we conducted a residual analysis,

1. Multicollinearity:

   The variables arrival_time.Night , arrival_time.Evening , and arrival_time.Morning exhibit Variance Inflation Factor (VIF) values exceeding2, indicating the presence of very lower multicollinearity

2. Linear relationship:

In our plot, we see two distinct curved patterns, suggesting **non-linearity** in the data. And also the spread of residuals increases with larger fitted values (right side of the plot). This suggests that variance is not constant, violating another assumption of linear regression (homoscedasticity)

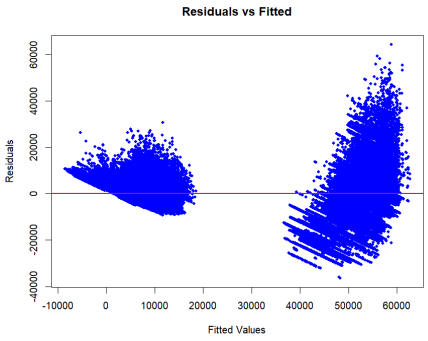

Figure 3: Residual Vs Fitted value plot

3. Independence & Homoscedasticity:

   Durbin-Watson test

   - Test statistic = 0.39171,
   - p-value < 0.05

This confirms that there is significant positive autocorrelation in the residuals. Positive autocorrelation means that errors (residuals) are correlated over time or across observations. This violates the assumption of independent errors**,** which is required for an MLR model to be valid.

4. Multivariate Normality:

From this Q-Q plot clearly see that Normality assumption is violated.

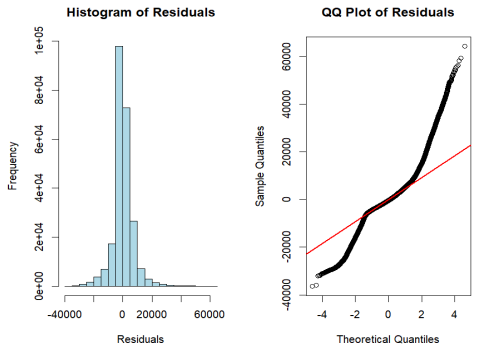

Figure 4: Histogram of Residual plot & Q-Q plot

Regularization Methods:

After conducting the MLR analysis, Ridge regression was used to predict diamond prices accurately by reducing multicolinearity. A variable selection strategy is necessary to choose the most appropriate group of predictor variables to decrease. Complexity and to prevent overfitting. Multiple lasso regression was utilized as a variable selection and prediction method. The Elastic Net model combines Lasso and Ridge regularizations to address their respective shortcomings.

Results obtained for Ridge regression, Lasso regression and Elastic Net Regression:

Table 4: Evaluation matrix for Ridge, Lasso & Elastic net

Model	Best λ	Training		Test	
		RMSE	R^2	RMSE	R^2
Ridge	2128.718	7035.76	0.9039	7047.75	0.9036
Lasso	19.85249	6754.84	0.9114	6762.54	0.9113
Elastic Net	32.96378	6754.15	0.9114	6761.91	0.9113

All three models achieved high R-squared values, indicating a good fit to the data. The training and test errors are very close, suggesting that the models are not overfitting. Since Lasso and Elastic Net has lower RMSE value and Higher R^2 than Ridge those methods are better compare to Ridge. The performance of Lasso and Elastic Net is very similar. Lasso and Elastic Net both performed feature selection by shrinking some coefficients to zero, which can improve model interpretability. The coefficients provide insights into the relationships between the features and the target variable. The results indicate that both models are effective in predicting flight prices and handling the potential challenges of multicollinearity and skewness in the data.

Partial Least Squares Regression (PLSR)

The model has high predictive accuracy with low error and high explanatory power. The minimal difference between training and test performance suggests that the model is not overfitting and generalizes well. However, Flight prices often have outliers. Since Tree models handle outliers better than linear models like PLSR, which are sensitive to extreme values.

Table 5: Evaluation matrix for PLSR

	RMSE	R^2
Training	6732.117	0.9123
Test	6792.198	0.9105

Tree based methods:

Tree-based algorithms are robust to outliers and multicollinearity issues, and they can capture nonlinear relationships between predictors and the response variable. Furthermore, when assessing the correlation between the response variable and predictor variables, most predictors showed weak correlations with the response variable. This suggests a potential non-linear and joint relationship between predictors and the response variable. The specialty of tree-based modelling lies in the fact that there are not many assumptions to satisfy.

Decision tree

After training the decision tree regression model with optimized hyperparameters; the following results were obtained:

Table 6: Hyperparameter values of Decision Tree

Hyperparameter	Value
Max depth	30
Min sample leaf	4
Min sample split	10

Table 7: Evaluation matrix for Decision tree

	RMSE	R^2
Training	2753.83	0.9853
Test	4279.982	0.9644

The training R-squared value is slightly higher than the testing R-squared, indicating that the model fits the training data very well but does not generalize perfectly to unseen data. While the gap is not extreme, it suggests a mild degree of **overfitting**. To improve generalization, further refinements could be applied, such as adjusting the tree depth, increasing the minimum number of samples per leaf, or implementing cross-validation to fine-tune hyperparameters. This analysis emphasizes the importance of balancing model complexity to ensure strong performance on both training and test data while mitigating overfitting.

XG Boost

Traditional machine learning models like decision trees and random forests are easy to understand but may not work well on complex data. XG Boost (short for eXtreme Gradient Boosting) is a powerful machine learning method that is fast, efficient, and gives better results. Since the XG Boost model gave the best results for predicting flight prices, it was used to make predictions on the test data.

Table 8:Evaluation Matrix of XG Boost

	RMSE	R^2
Training	3484.37	0.9765
Test	3691.067	0.9736

Following figure shows the predicted flight prices compared to the actual prices. The test data fit well with the XG Boost model, and the predicted prices were very close to the real prices.

Figure 6:Actual Vs Predicted Flight Price

Figure 5: Days left Vs flight price plot

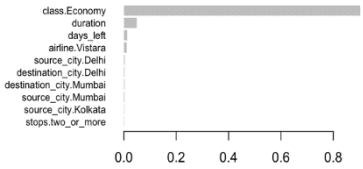

Figure 7:Feature Importance of XG Boost

The actual and predicted flight prices from the test data were plotted based on the number of days left before departure and the flight price (as shown in following Figure). The predicted prices closely matched the actual prices, showing that the XG Boost model performed well.

Key Predictors

To understand the key predictors influencing the response variable, we examined feature importance plot from the top-performing model in our analysis: XG Boost. class.Economy being the dominant feature suggests ticket class is the strongest predictor in your model**.** duration being second makes intuitive sense - flight length often affects other variables**.** The city pairs (Delhi, Mumbai, Kolkata) having lower importance suggests route-specific factors are less critical**.** days_left having some importance indicates booking timing

matters, but less than class and duration**.** stops.two_or_more being low suggests number of stops isn't a major factor in predictions.

Issues Encountered and Proposed Solutions

In the course of our advanced analysis, several challenges occurred. Moving into the analysis, we identified violations of key assumptions in the MLR model. To mitigate this, we implemented diagnostic checks. In regularization techniques, due to unexpected equality in results, we used cross-validation to optimize regularization parameters. Additionally, difficulties in generating a scree plot prevented us from identifying components in PLSR. We explored innovative approaches to address these issues. Finally, in tree-based methods, some variables exhibited zero importance. The solution involved the removal of these variables, followed by model refitting, optimizing the predictive power of the model. This troubleshooting approach,

addressing challenges at each stage of the analysis, demonstrates our commitment to producing reliable and robust insights in the face of analytical hurdles.

Discussion and Conclusions

Table 9:Summary of all R^2 and RMSE values

	Training		Test	
	RMSE	**R^2**	**RMSE**	**R^2**
MLR(Best Subset)	6752.5028	0.9115	6760.2889	0.9113
Ridge	7035.76	0.9039	7047.75	0.9036
Lasso	6754.84	0.9114	6762.54	0.9113
Elastic Net	6754.15	0.9114	6762.54	0.9113
PLSR	6732.117	0.9123	6792.198	0.9105
Decision Tree	2753.83	0.9853	3691.067	0.9736
XG Boost	3484.37	0.9765	3691.067	0.9736

Among the models tested, **Decision Tree**, achieving the lowest RMSE and highest R^2 on both training and test datasets. But there's huge gap between test RMSE and training RMSE value. So this might can happen because of overfitting.

Traditional linear models, including **PLSR and regularized regression (Ridge, Lasso, Elastic Net)**, provide reasonable predictions but fail to match the performance of tree-based models, likely due to the inability to capture complex, non-linear relationships in the data.

Thus, **XG Boost is the most sutaible for this model.**, but further hyperparameter tuning and validation can be explored to enhance its robustness and efficiency. It effectively balances accuracy and generalization, avoiding the severe overfitting observed in the **Decision Tree model** while outperforming traditional regression-based approaches like **PLSR, Ridge, and Lasso**.

References

1. https://www.statology.org/lasso-regression-in-r/
2. https://www.geeksforgeeks.org/elastic-net-regression-in-r-programming/

Appendix

Link for the google Collab Note Book
https://colab.research.google.com/drive/1hxEG_sJjBuCD0Et2q4imE6WFQTIZmWIp?usp=sharing

https://colab.research.google.com/drive/1_Vva0lm2yAPsZEI8wsGs3QKgpdqJ0EAx?usp=sharing
https://colab.research.google.com/drive/1KQkF0XMhHyPVhSE2iif32SBKFuNyOTez?usp=sharing
Link for Google drive
https://drive.google.com/drive/folders/1bzZ1D68gfK1Il66RnX0bIPyLqaMN4Vgk