***Project: Predicting Airfare on New Routes***

**Predicting Airfares using Analytic Solver**

**Objective:**  
The aim of this project is to investigate the effect of the presence or absence of **Southwest Airlines (SW)** on the airline fare (**FARE**).

### *****Step 1: Data Sampling*****

* **Sampling Method:**  
  We sampled the data using all available input variables or predictors.
* **Sample Size:**  
  A sample of **600 records** was selected, with the seed set to **12345** for reproducibility.

### *****Step 2: Data Preprocessing/Cleaning*****

#### **Typecasting:**

* We examined each input variable and categorized them as either Numerical or Categorical.

#### **Handling Missing Data:**

* Various imputation methods were used based on the type of predictor:
  + **Mode Imputation** was applied to categorical variables
    - S\_CODE, S\_CITY, E\_CODE, E\_CITY, VACATION, SW, SLOT, GATE
  + **Mean Imputation** was used for numerical predictors
    - COUPON, NEW, HI, S\_INCOME, E\_INCOME, DISTANCE, PAX
  + **Median Imputation** was used for variables with outliers:
    - S\_POP, E\_POP
  + **Deletion of Records** was done for the FARE variable with missing data.
* **Result of Missing Data Handling:**
  + All **600 records** were retained, and no records were deleted due to missing data.

#### **Standard Partitioning:**

* The data was partitioned into training, validation, and test sets, with the following distribution:
  + **50% for Training:** 300 records
  + **30% for Validation:** 180 records
  + **20% for Testing:** 120 records
* Partitioning was done using all input variables and with the seed set to **12345** for reproducibility.

***Step 3: Building Predictive Models***

#### **Linear Regression Model:**

* **Choice of Model:**
  + A **linear regression** model was used, as the dependent variable (**FARE**) is numerical.
* **Input Variables:**  
  All input variables were included in the model, assuming they all contribute to predicting the airline fare.
* **Rescaling:**  
  No normalization or standardization was applied to the data. Linear regression models can handle different feature scales, so rescaling was not necessary.
* **Feature Selection:**
  + The **best subset option** was used for feature selection, with the following specifications:
    - **Maximum number of subsets:** 15
    - **Best number of subsets:** 1
  + This method helped identify the most relevant predictors for the model, improving its accuracy, reducing the risk of overfitting, and enhancing the interpretability of the results.\

***Step 4: Model Evaluation***

**Check for Green Tabs**

* **R²: 0.803**
  + **Interpretation:** This means that **80.3%** of the variance in the target variable (FARE) is explained by the model's predictors and this indicates a strong fit.
* **Standard Error Estimate: 36.58**
  + **Interpretation:** On average, the predicted fares deviate from the actual fares by about **36.58** units, showing a reasonable degree of precision in the model's predictions.
* **Predictor Summary on Southwest Airlines (SW):**
  + **Interpretation**: The variable **Treated\_SW\_Yes** (indicating the presence of Southwest Airlines on a route) shows a large positive coefficient of **7.893**, suggesting that Southwest’s presence significantly impacts the fare.
* **Coefficient Summary:** 
  + **Interpretation** The **SW\_Yes** coefficient in the regression model is **-44.0704**, further reinforcing the idea that routes operated by Southwest Airlines lead to a decrease in fare by approximately **44.07 units**.
* **Output of detailed Summary Report:**

A screenshot of a graph

Description automatically generated A screenshot of a graph

Description automatically generated

|  |  |  |  |
| --- | --- | --- | --- |
| **Dataset** | RMSE | MAD | R2 |
| Training | 34.83515 | 27.05033 | 0.803312 |
| Validation | 34.04423 | 26.22374 | 0.797051 |
| Testing | 38.45344 | 29.97937 | 0.700911 |

A screenshot of a graph

Description automatically generated

* **Interpretation:**

1. **RMSE (Root Mean Squared Error):**
   * The model does well on both Training and Validation datasets (around 34), but the Testing dataset shows slightly larger errors (around 38). This means the model is a little less accurate when predicting new data.
2. **MAD (Mean Absolute Deviation):**
   * Errors are smallest for the Validation dataset (26.22), slightly higher for Training (27.05), and largest for Testing (29.98). This pattern matches RMSE and shows the model struggles a bit more with new data.
3. **R2:**
   * R2 is high for Training (0.803) and Validation (0.797), showing the model works well. It drops for Testing (0.701), meaning the model is less reliable when making predictions on new data.

***Step 5: Refining the Model***

Check for Red Tabs

* **Feature Selection Process:**
  + Distance and PAX were included in most subsets, confirming their strong relationship with airfare, and Southwest Airlines (SW\_Yes) also appeared frequently, showing the significant effect of Southwest’s low-cost model on airfares.
  + **Subset Analysis**
    - *Subset 1 (1 coefficient):*
      * **R²:** Close to 0, indicating poor model fit.
      * **RSS:** Very high, suggesting poor performance.
      * **p-value:** Extremely small, indicating statistical significance but poor model fit due to the lack of predictors.
    - *Subset 7 (2 to 4 coefficients):*
      * These subsets show performance improvement, with Subset 7 particularly strong, achieving an R² of 0.76 and a relatively low RSS.
    - *Subset 12 and Beyond (12 to 14 coefficients):*
      * As the number of predictors increases, Adjusted R² improves, reaching around 0.79 to 0.80, which suggests a better fit while maintaining predictive power.

### ****Best Subset Selection:****

### ****Subset 13**** is the best subset to refine the model, as it offers a strong fit with a high ****Adjusted R²**** (0.7916), low ****RSS****, and a well-balanced number of predictors. This subset effectively captures the key factors influencing airfare while avoiding overfitting. It’s very low ****p-value**** ensures the model is statistically significant and robust for predicting airline fares.

* Since the **RMSE** values are reasonably low, the goal is to achieve an even lower RMSE for better accuracy in predicting airfare. To enhance the model's performance, **Subset 13** has been selected from the feature selection process.

***Step 6: Model Comparisons***

* **Output Summary Comparison**

|  |  |  |
| --- | --- | --- |
| **Partition set** | **RMSE: All variables** | **RMSE: Subset 13** |
| Training set | 34.874 | 34.835 |
| Validation set | 34.178 | 34.044 |
| Test set | 38.635 | 38.453 |

* Refining the model by selecting **Subset 13** from the feature selection process has not significantly improved performance. In fact, **refining the model** has slightly increased the RMSE in the test set, which is counterproductive. This suggests that **Subset 13** is not necessarily leading to a better generalization of new data.
* Therefore, we will proceed with the previous model, which includes all variables, as it provides better overall accuracy for predicting airfare.

**NOTE**:

* The RMSE values for the Decision Tree model are higher in both the training and testing phases compared to the Linear Regression model. A higher RMSE indicates that the Decision Tree model's predictions deviate more from the actual values than those of the Linear Regression model. This suggests that the Linear Regression model performs better and provides more accurate fare predictions.

| **Partition** | **RMSE** | **MAD** | **R-squared** |
| --- | --- | --- | --- |
| Training | 40.053 | 31.851 | 0.739977 |
| Testing | 47.224 | 33.940 | 0.548926 |

* Based on the performance evaluation, the neural network model had a worse RMSE of about 52%, making it less effective for fare prediction compared to other models. With such high RMSE, the neural network's predictions deviate significantly from the actual values, suggesting it cannot be reliably used for fare prediction in this case.
* **Thus, linear regression model remains the best choice among the models we tried, providing more accurate and consistent predictions.**

**Predicting Airfares using Rstudio**

**Steps involved:**

**1. Data Collection and Loading:**

* The dataset (Airfare) is loaded into R, and the initial rows are displayed using the head() function.

**2. Data Preprocessing:**

* **Typecasting:**
  + Relevant columns are converted to appropriate data types (e.g., categorical variables like S\_CODE, VACATION, etc.).
* **Sampling:**
  + A subset of 600 rows is randomly selected for analysis to ensure computational efficiency.
* **Handling Missing Values:**
  + Mean imputation is applied to numeric columns without outliers (e.g., COUPON, NEWHI, DISTANCE, PAX).
  + Median imputation is applied to numeric columns with outliers (S\_POP, E\_POP).
  + Mode imputation is used for categorical variables (e.g., SW, SLOT, VACATION, GATE).
* **Data Partitioning:** The dataset is split into training (50%), validation (30%), and testing (20%) sets.

**3. Model Building:**

* A linear regression model is built to predict airfare (FARE) using the training data, incorporating all variables.

**4. Model Refinement:**

* The model excludes irrelevant columns (S\_CODE, S\_CITY, etc.) for simplicity.
* Stepwise selection refines the model further by selecting the most significant predictors.

The stepwise selection process refines the model by removing less significant variables and retaining those that contribute to lowering the AIC:

* **Initial AIC**: 2101.6
* **Removing 'NEW'**: Reduced AIC to 2099.8
* **Removing 'COUPON'**: Reduced AIC to 2098.5

**5. Model Evaluation:**

* Metrics such as RMSE (Root Mean Square Error), MAE (Mean Absolute Error), and R-squared are calculated for the training, validation, and testing sets to evaluate model performance
* MAPE (Mean Absolute Percentage Error) and benchmark errors are calculated to assess prediction accuracy and compare against a baseline.

**6. Results and Metrics Table:**

* A table summarizing RMSE, MAE, and R-squared across all partitions (training, validation, testing) is generated.

**Output interpretation:**

## **Metric Comparison**

| **Metric** | **Without Feature Selection** | **With Feature Selection** | **Better Model (Why)** |
| --- | --- | --- | --- |
| Residual Standard Error | 28.73 | 34.36 | Without (Lower error, indicating more accurate predictions) |
| R-Squared | 0.92 | 0.8151 | Without (Explains 7% more variance in fares, indicating better fit) |
| Adjusted R-Squared | 0.8657 | 0.8079 | Without (Better explanatory power, considering number of predictors) |
| F-Statistic | 16.92 | 113.4 | With (Stronger relationship between predictors and fare, indicating better feature selection) |
| P-Value | < 2.2e-16 | < 2.2e-16 | Both (Highly statistically significant, indicating reliable models) |

**Coefficients summary:**

## **Factors Affecting Airfare Comparison**

| **Variable** | **P-Value (Without)** | **P-Value (With)** | **Impact on Fare (Without)** | **Impact on Fare (With)** |
| --- | --- | --- | --- | --- |
| VACATIONYes | 0.001316 | 3.14e-11 | Lowers fare by about $59.73 | Lowers fare by about $38.14 |
| SWYes | 2.18e-05 | 3.19e-14 | Lowers fare by about $33.58 | Lowers fare by about $42.66 |
| HI | 9.61e-05 | 6.21e-06 | Slightly increases fare | Slightly increases fare |
| SLOTFree | 0.994105 | 0.000974 | No significant impact | Lowers fare by about $17.42 |
| GATEFree | 0.320392 | 3.69e-06 | No significant impact | Lowers fare by about $26.90 |
| DISTANCE | < 2e-16 | < 2e-16 | Increases fare with longer distances | Increases fare with longer distances |
| PAX | 0.000141 | 4.89e-07 | Slightly lowers fare | Slightly lowers fare |
| COUPON | 0.730848 | Omitted | No significant impact | NIL |
| NEW | 0.346835 | Omitted | No significant impact | NIL |

* **Impact of SW on fare:**The presence of SW (Southwest Airlines) improved significantly after refining the model. The p-value decreased from 2.18e-05 to 3.19e-14, indicating stronger statistical significance. The impact on fare also became more specific, with a $42.66 decrease.

**Conclusion:**

1. **Residual Standard Error (RSE):**

* The model without feature selection has a lower RSE (28.73 vs. 34.36), meaning it predicts airfare more accurately.

2. **R-Squared and Adjusted R-Squared:**

* The model without feature selection explains more about airfare variations (R² = 0.92) and does a better job even when considering the number of predictors (Adjusted R² = 0.8657). This makes it a stronger model for understanding how factors affect airfare.

1. **P-Values:**

* Both models are statistically reliable (P-Value < 2.2e-16), but the model without feature selection does a better job of including all the important factors that impact airfare.

## **Model Performance Metrics**

| **Metric** | **Training** | **Validation** | **Test** |
| --- | --- | --- | --- |
| RMSE | 33.65 | 35.42 | 36.66 |
| MAE | 26.33 | 27.23 | 29.12 |
| R-Squared | 0.8151 | 0.8042 | 0.7934 |

* The model's performance decreases slightly from training to validation to test, indicating some overfitting. However, the model still performs reasonably well on the test set. The R-Squared values indicate that the model explains around 79-82% of the variation in fares.

## **Model Performance vs. Benchmark**

* Test Mean Absolute Percentage Error (MAPE): 20.84% (model) vs. 53.24% (benchmark)
* Test Root Mean Squared Error (RMSE): 36.66 (model) vs. 80.75 (benchmark)

The model significantly outperforms the benchmark, with a 65.6% reduction in MAPE and a 54.5% reduction in RMSE. This indicates that the model is a substantial improvement over the simple benchmark of using the mean fare.

**Predicting Airfares using SAC**

**Dataset Summary:**

* We built a linear regression model using SAP Analytics Cloud with the following dataset characteristics:
  + **Initial Number of Variables:** 18
  + **Number of Kept Variables:** 13
  + **Number of Records:** 638

1. **Model Performance**

**A screen shot of a computer error

Description automatically generated**

* **Root Mean Square Error (RMSE: 25.49):** On average, the model’s predictions are off by about 25.49 units. A lower RMSE is better.
* **Prediction Confidence (95.59%):** The model is very reliable, with a 95.59% confidence in its predictions. This shows the model is accurate for the given dataset.

1. **Target Variable Statistics**

A screenshot of a computer

Description automatically generated

|  |  |  |
| --- | --- | --- |
| **Metric** | **Training Data** | **Validation Data** |
| Minimum Value | 42.47 | 45.11 |
| Maximum Value | 402.02 | 374.4 |
| Mean | 160.61 | 161.52 |
| Standard Deviation | 75.7 | 76.59 |

The close alignment between training and validation statistics indicates a well-generalized model with no significant overfitting.

1. **Influencer Contributions**

A screenshot of a graph

Description automatically generated

The top five variables contributing to the model’s predictions are:

|  |  |
| --- | --- |
| **Influencer** | **Contribution** |
| DISTANCE | 26.37% |
| SW | 14.46% |
| VACATION | 9.04% |
| COUPON | 6.95% |
| E\_INCOME | 6.35% |

* + **Role of SW in the Model**
    - The variable SW significantly influences fare predictions, accounting for 14.46% of the model’s outcomes, making it the second most impactful factor after distance. SW likely indicates the presence of Southwest Airlines as a competitor or service provider.
    - When SW is present, fares tend to decrease, reflecting the competitive pressure Southwest applies in the market due to its reputation for affordable air travel. This aligns with its strategy of driving down prices in the industry.

1. **Influence of Grouped Category: SW**

A graph with blue squares

Description automatically generated

* The variable SW has an average importance of 14.46% in the model. However, for a specific case where SW is set to "YES," its importance decreases by 54%. This reduction implies that SW's adjusted importance in this scenario becomes (1−0.54)×14.46=6.65%(1 - 0.54) \times 14.46 = 6.65\%(1−0.54)×14.46=6.65%.

1. **Grouped Category Statistics: SW**

* In the validation partition, the presence of SW (SW=YES) has notable effects on fare prices. The average fare price in these cases is **103.67**, and SW airlines are present in **29.57%** of the instances within the validation partition.

**Conclusion:**

The model achieves reliable fare predictions with a **Root Mean Square Error (RMSE)** of 25.49, indicating average deviations of approximately 25.49 units. It demonstrates a high prediction confidence of **95.59%**, reflecting strong accuracy. The variable SW plays a significant role, contributing **14.46%** to the predictions, but its influence reduces to **6.65%** when SW=YES. Overall, the model effectively predicts fares with well-generalized performance across training and validation datasets.