# Route-Airfares

# Introduction and discovery-

The dataset is based on the problem that occurred when many major cities were facing issues with airport congestion, partly because of the 1978 deregulation of airlines. Both fares and routes were free from regulation, and low-fare carriers such as Southwest (SW) Airline began competing on existing routes and starting non-stop service on routes that previously lacked it. After SW Airlines started flights on some routes, it was contemplating on expanding the routes served. We are going to predict airfare on previous and new routes for SW Airlines based on the several features by creating different linear regression models. We will use our best model to predict the airfare. The dataset consists of 638 rows and 16 columns.

# Data Preparation-

The dataset is about predicting airfare. The dataset consists of different variables. Description of variables in the Airfare Data Set are as follows-

S\_CITY - Starting city name

E\_CITY - Ending city name

COUPON - Average number of coupons (a one-coupon flight is a non-stop flight, a two-coupon flight is a one stop flight, etc.) for that route

NEW - Number of new carriers entering that route between Q3-96 and Q2-97

VACATION - Whether a vacation route (Yes) or not (No); Florida and Las Vegas routes are generally considered vacation routes

SW - Whether Southwest Airlines serves that route (Yes) or not (No)

HI - Herfindel Index – measure of market concentration

S\_INCOME - Starting city’s average personal income

E\_INCOME - Ending city’s average personal income

S\_POP - Starting city’s population

E\_POP - Ending city’s population

SLOT - Whether either endpoint airport is slot controlled or not; this is a measure of airport congestion

GATE - Whether either endpoint airport has gate constraints or not; this is another measure of airport congestion

DISTANCE - Distance between two endpoint airports in miles

PAX - Number of passengers on that route during period of data collection

FARE - Average fare on that route.

A screenshot of a computer screen

Description automatically generated

When the dataset is read it consists of 16 columns with 638 rows. Several of these columns and rows are not required for the analysis for which some data preprocessing is required. Some data preprocessing steps are carried out like Checking and dropping the rows that contains null or random values. The sheet contains S\_CITY and E\_CITY columns that needs to be removed and only 14 columns would be there. Null values are also checked and necessary actions are taken if any null value is present. Checking that all the columns names are in uppercase and replacing spaces with underscore.

A screenshot of a computer

Description automatically generated

The summary statistics are as follows:

A screenshot of a computer

Description automatically generated

# Model Implementation-

Feature Engineering –

We created some dummy features from the categorical data to use them for our code efficiently. Then, we dropped columns VACATION, SW, SLOT and GATE. We used FARE as our target feature and we took the ratio of 75:25 of training and testing of data. We performed 2 feature selections methods.

1> Correlation Based. We selected "COUPON","S\_INCOME","E\_INCOME","DISTANCE" and "SW\_Yes" features because it has higher correlation with the FARE.

A screenshot of a computer screen

Description automatically generated

2>Variance Threshold Selection. We set up the threshold value to 0.15. So, features below this 0.15 won't be selected.

A screenshot of a computer screen

Description automatically generated

For our feature transformation, Log1p transformation and MinMax scaling was used. After creating empty lists for feature selection, transformation and scaling. We created different regression models with several combinations plus the Lasso regression model and calculated their R2 and RMSE values. Out of these models, The best model found out to be Variance Threshold model because it has the highest R2 and Lowest RMSE value.

A screenshot of a computer

Description automatically generated

The best model is:

A screenshot of a computer program

Description automatically generated

A plot between actual and predicted values is being created:

A graph of blue dots

Description automatically generated

# Results Interpretation and Implications-

We used boosting techniques to combine the weak learners in a sequential method, this will improve our observations. We have created empty lists for models, r2 and rmse values, names. We have applied linear regression, SGDRegressor, decision tree, gradientboost, random forest, adaboost, xgboost and catboost techniques for our analysis. We again used variance threshold feature selection method by setting threshold value to 0.15. Then a pipeline is created which is transforming the dataframe using Polynomial Features with degree 2. The resultant will fit into a scalar function with robust scaler for the scaling of data. We then calculated the R2 and RMSE values and then we concluded the best model. The best model that we got is CatBoostRegressor.

A screenshot of a computer

Description automatically generated

A screenshot of a computer code

Description automatically generated

# Concluding Remarks-

Comparing all the regression models we could conclude that the regression model having the lowest RMSE value is the best model.

We can notice that as the distance increases the Fare for that particular route also increases. The highest fare for a particular route is around 400. The count of Southwest Airlines not serving a particular route is more than the count of Southwest Airlines serving that route. The count of Southwest Airlines not serving that route is more than 400 and the count of Southwest Airlines serving that route is approximately 200. The number of non-vacation route are more compared to the number of vacation route. The count of non-vacation route is more than 400 and the count of vacation route is around 180. We can also conclude that Gates having constraints are less compared to gates that are free. Number of passengers on a particular route during period of data collection could be used to determine the number of passengers in the flight when the Airlines serves that route and when the Airlines does not serve that particular route.

# Video Presentation link-

https://drive.google.com/file/d/1yFgL8bLh7-c8y4j3oY-B7SNeu-6G8Uj2/view?usp=sharing