****

**AIR FARE PREDICTION PROJECT WORK 2020**

**SUMMARY:**

Built a model to predict the price fluctuations over time by exploring previous Air Fare Data so that the consumer could benefit from it . We looked at different models from the space of forecasting and regression.

**PROJECT DETAILS:**

**Project Name :** Air Fare Prediction

**Project Code :** P23

**Group :** G4

**Submitted Date :** 5 July 2020

**Place :** Bangalore, India

**Project Duration :** 26th May 2020 to 5th July 2020

**PROJECT MENTORS:**

Project Design, Guidance and conceptualization of the activities carried out by the following members:

* **Munmun Bhagat**
* **Varun**

**PROJECT MENTORS:**

Development of the project and the deployment of the Model carried out by the following members:

* **Raja Ram Chaudhary**
* **Umesha B.K**
* **Gowthami R.**

**DATASET OVERVIEW:**

* **Dataset Name :** air\_fair.csv
* **Contains :** 278466 rows and 4 columns
* **Data Range :** -497324 to 497678
* **Missing Values :**

Invoice Date : 2

Product Type : 2

Itinerary Type : 32777

NetFare : 60890

**EDA:**

1. **How we handled missing data?**

**Invoice Date:** Deleted the 2 values as two observations among large dataset are not so big

deal.

**Product Type:** Deleted the 2 values as two observations among large dataset are not so big

deal.

**Itinerary Type:** 32777 missing data and they are totally empty. So, dropped them.

**NetFare:** 60890 missing data and they are totally empty. So, dropped them.

1. **How we handled Outliers?**

Net count post deletion of missing values: 184800 and Skewness for NetFare is 11.1324 due to outliers. Identified Outliers with Interquartile Range (IQR) and Used IQR method to drop outliers in the dataset. Total rows using for building the model is **167132**

**Feature Selection & Model Building:**

* Dataset contains categorical variables i.e. Product Type and Itinerary Type. Created dummy variables for each category type.
* Used Correlation Heatmap for the dataset to identify the correlation between the variables.
* Tried with forecasting but actual NetFare and forecasted values are almost same. Might be overfitting.
* Tried with Regression concept and built Predictive Linear Model using Pycaret, Statsmodel and Scikit Learn packages.
* **Pycaret** Package provided **23.72%** and **Stats Model** package provided **24% R-Squared value** with 7 features
* Finally, built model using **Scikit Learn** and got **31% R-Squared value**. Predictive Model shows good prediction.

**Scientific Tools Used For Model Building:**

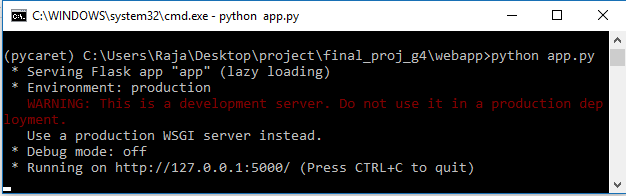
* Jupyter Notebook-6.0.3
* Python-3.6.10
* iPython-7.13.0
* Numpy-1.18.4
* Pandas-1.0.3
* Matplotlib-3.2.1
* Seaborn-0.10.1
* Scipy-1.4.1
* Scikit-learn-0.22
* PyCaret-1.0.0

**FINAL MODEL AND DEPLOYMENT:**

* Our Final model built using **Predictive Linear Model** using **Scikit Learn** Package.
* Deployed the models using below 3 methods.

1. PICKLE : 4july\_lr\_model.pickle
2. FLASK : http://127.0.0.1:5000/
3. HEROKU : https://aerofare.herokuapp.com/

**FLASK WEB SERVER:**



**HEROKU WEB VIEW:**

