VIKASH YADAV (B20AI061)

CSL2050 PRML – BONUS PROJECT REPORT

AIRLINE PRICE PREDICTION

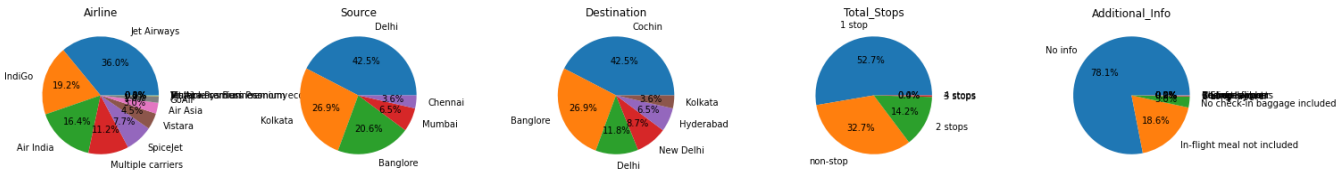
Table of Contents

Aim

The aim of this project is to utilise the Airline Dataset and come up with a ML pipeline for predicting the price of flights.

Data-Loading and Pre-Processing

- **Loaded** the data and **visualised** the distribution of categorical columns.



- **Split** the columns with string entries in English.
- **Encoded** the columns (some label and some one-hot as per suited)
- **Scaled** the Data
- **Split** the data into train (70%), validation (20%) and test set (10%).

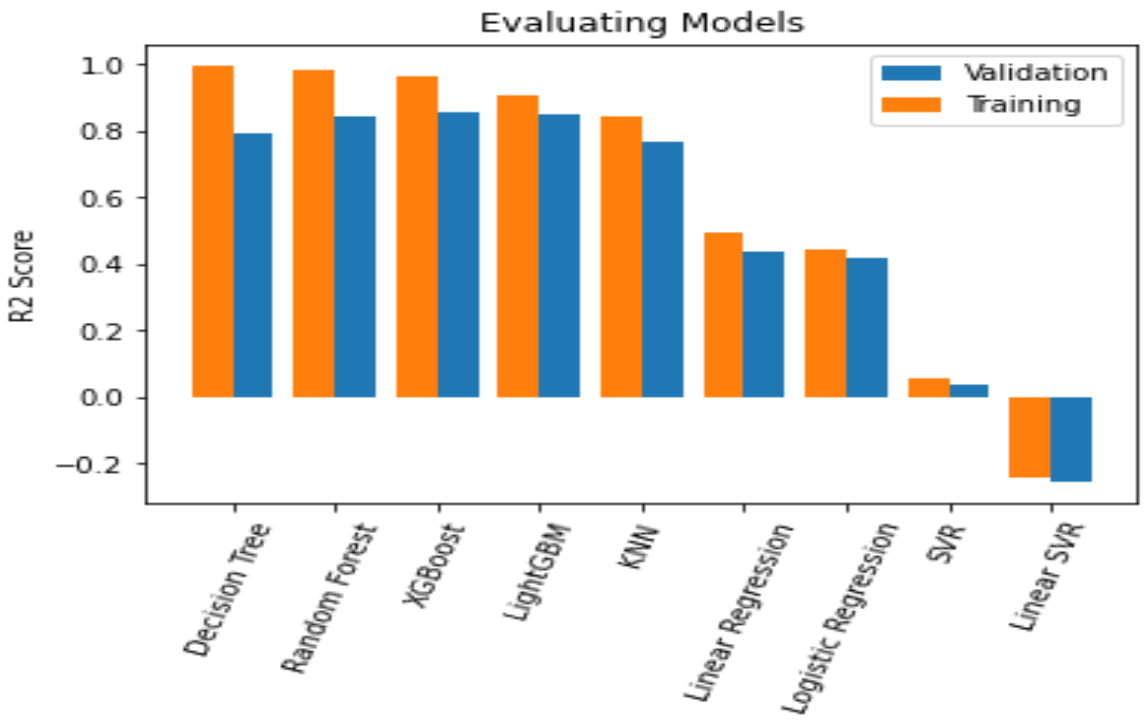
Trying and Testing Models

Firstly, I ran a plain run on the following models and evaluated on validation set to see which model work for this data:

```
models = {
    'Decision Tree': DecisionTreeRegressor(),
    'Random Forest': RandomForestRegressor(),
    'XGBoost': XGBRegressor(),
    'LightGBM': LGBMRegressor(),
    'KNN': KNeighborsRegressor(),
    'Linear Regression': LinearRegression(),
    'Logistic Regression': LogisticRegression(),
    'SVR': SVR(),
    'Linear SVR': LinearSVR()
}
```

The results are:

Model	R2 Score		MSE	
	Training	Validation	Training	Validation
Decision Tree	0.9963	0.79	75288.5600	4629971.52
Random Forest	0.9788	0.84	434133.3003	3602066.06
XGBoost	0.9619	0.86	780860.6267	3224506.35
LightGBM	0.9064	0.85	1916199.0771	3371929.49
KNN	0.8432	0.76	3210392.4841	5290864.31
Linear Regression	0.4909	0.44	10421827.9526	12602985.56
Logistic Regression	0.4440	0.42	11381439.8538	13053869.61
SVR	0.0527	0.04	19392609.5502	21591013.88
Linear SVR	-0.2414	-0.26	25412577.1300	28270140.86



From these results I chose 3 best models: XGBoost, LightBGM & RandomForest and tuned their parameters.

Parameter Tuning

I used the technique of Halving Grid Search to tune the parameters of these 3 models. In this technique, we start evaluating the candidates with a small number of resources and we keep selecting the better candidates and give them more resources. This makes the process of grid search faster.

Here are the results:

Random Forest:

```
Validation R2 Score after tuning: 0.8505428792607926
Best Parameters for random forest:
 {'max_depth': 32, 'min_samples_leaf': 2, 'min_samples_split': 4, 'n_estimators': 800}
```

XG Boost:

```
Validation R2 Score after tuning: 0.8633919854633768
Best Parameters for XG Boost
 {'learning_rate': 0.01, 'max_depth': 8, 'n_estimators': 1000, 'subsample': 0.6}
```

Light GBM:

```
Validation R2 Score after tuning: 0.8498621698569981
Best Parameters for LightGBM
 {'learning_rate': 0.1, 'max_depth': 16, 'n_estimators': 100, 'subsample': 0.6}
```

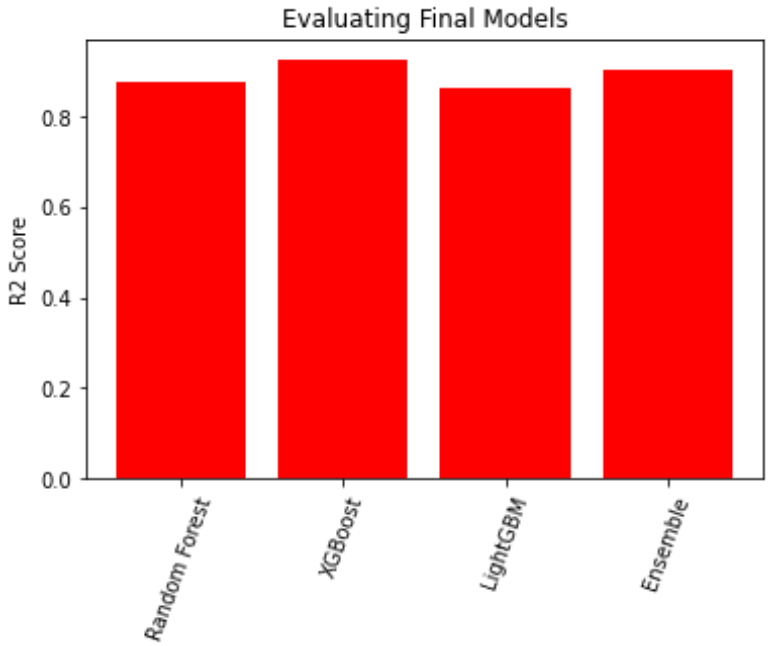
Making Ensemble

I made an ensemble of these 3 models whose prediction is the average of the prediction given by the 3 models. This would make my model more robust.

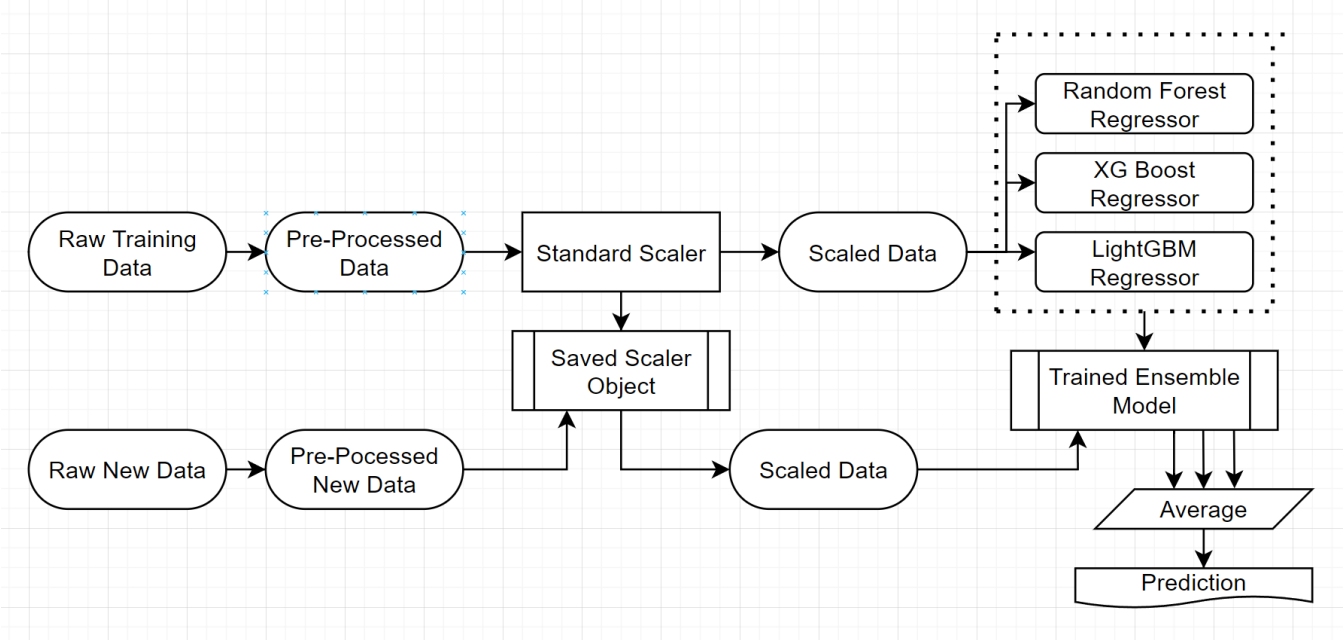
Final Evaluation on Test Set

Here are the results:

```
Final Test Results (on untouched Test Data):

Model                          R2 Score       MSE

Random Forest                  0.8770         2998604.73
XGBoost                        0.9249         1831359.50
LightGBM                       0.8630         3340695.09

Ensemble                       0.9012         2408355.93
```



Final PipeLine



Deployment

I used pickle to store the scaler and the model in files. These files can easily be stored on any server and be easily loaded from python to deploy the ML model.