FLIGHT FARE PREDCTION - MACHINE LEARNING

DETAILED PROJECT REPORT

ZAFAR MAHMOOD WARIS

Contents

Objective

The goal of this project is to analyse the flight prices of different airlines. Based on our analysis we are going to build a Machine Learning model to predict flight fares on future dates.

Problem Statement

- Travelling through flights has become an integral part of today’s lifestyle as more and more people are opting for faster travelling options. The flight ticket prices increase or decrease every now and then depending on various factors like timing of the flights, estination, and duration of flights. Therefore, having some basic idea of the flight fares before planning the trip will surely help many people save money and time.
- The main goal is to predict the fares of the flights based on different factors available in the provided dataset.

PROJECT DETAIL

01

Project Title

FLight Fare Prediction

02

Technolgy

Data Science

03

Domain

Aviation

04

Tools Used

Colab, scikit-learn, Pandas, Numpy

Data Summary

The dataset contains the following features:

- **Airline** - The name of the airline company.
- **Date_of_Journey** - The date on which the passenger is planning to take a flight.
- **Source** - City from where flight takes off.
- **Destination** - City upto where the flight is going.
- **Route** - The route a particular flight is going to follow.
- **Dep_Time** - Departure time of flight from source.
- **Arrival_Time** - Arrival time of flight on the destination.
- **Duration** - The total time of the flight.
- **Total_Stops** - Total nunber of stops in between source and destination.
- **Additional_Info** - Any additional information about the flight.
- **Price** - Price of the flight.

Approach Overview

Data Cleaning

- Find Information on documented column values
- Clean data to get it ready for Analysis

Data Exploration

- Examining the data with visualization

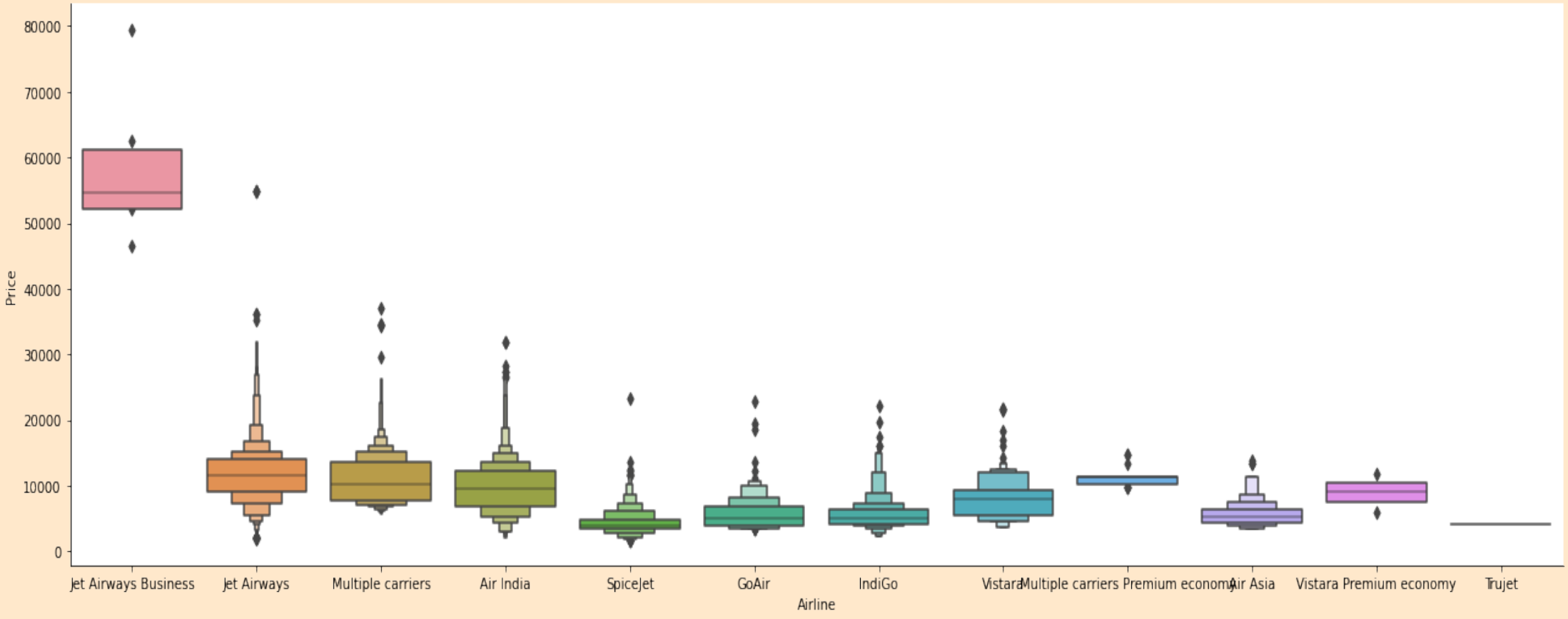
Modelling

- ExtraTressRegressor
- Random Forest Regressor
- Hyperparameter tuning

Exploring Data

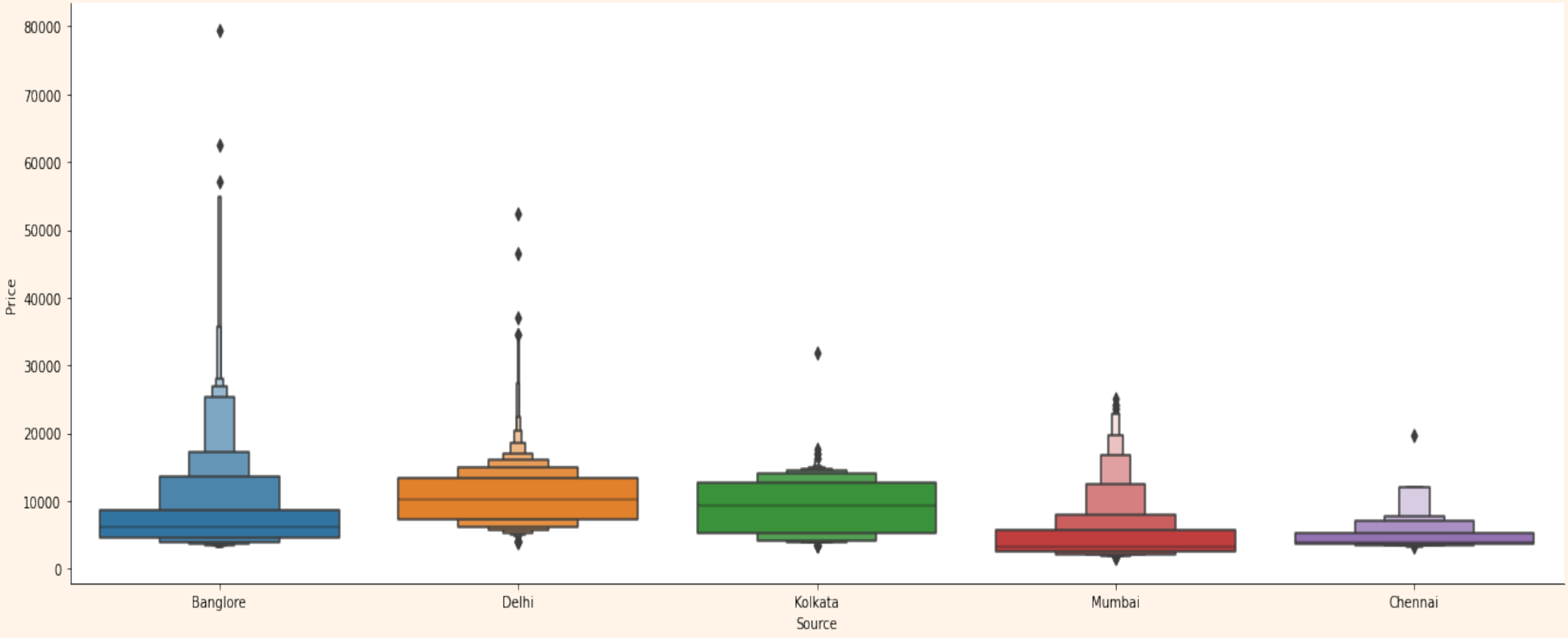
- Dataset contains 10682 instances for different airlines.
- Dataset contains airlines from India for for major cities.
- Very few null values.
- 4 categrical values present.
- 4 months of data available

Airline vs. Price plot



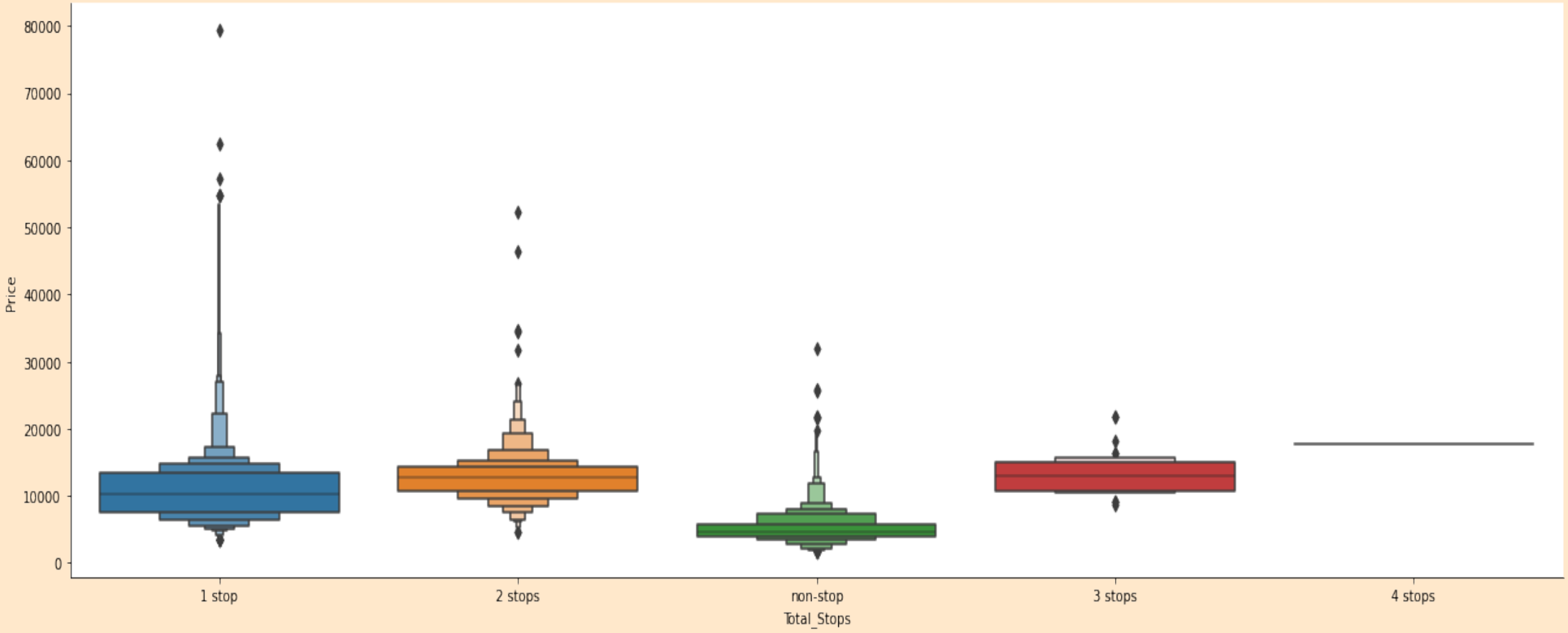
From this plot we see that almost every airline has same median except Jet Airways Business.

Source vs. Price plot



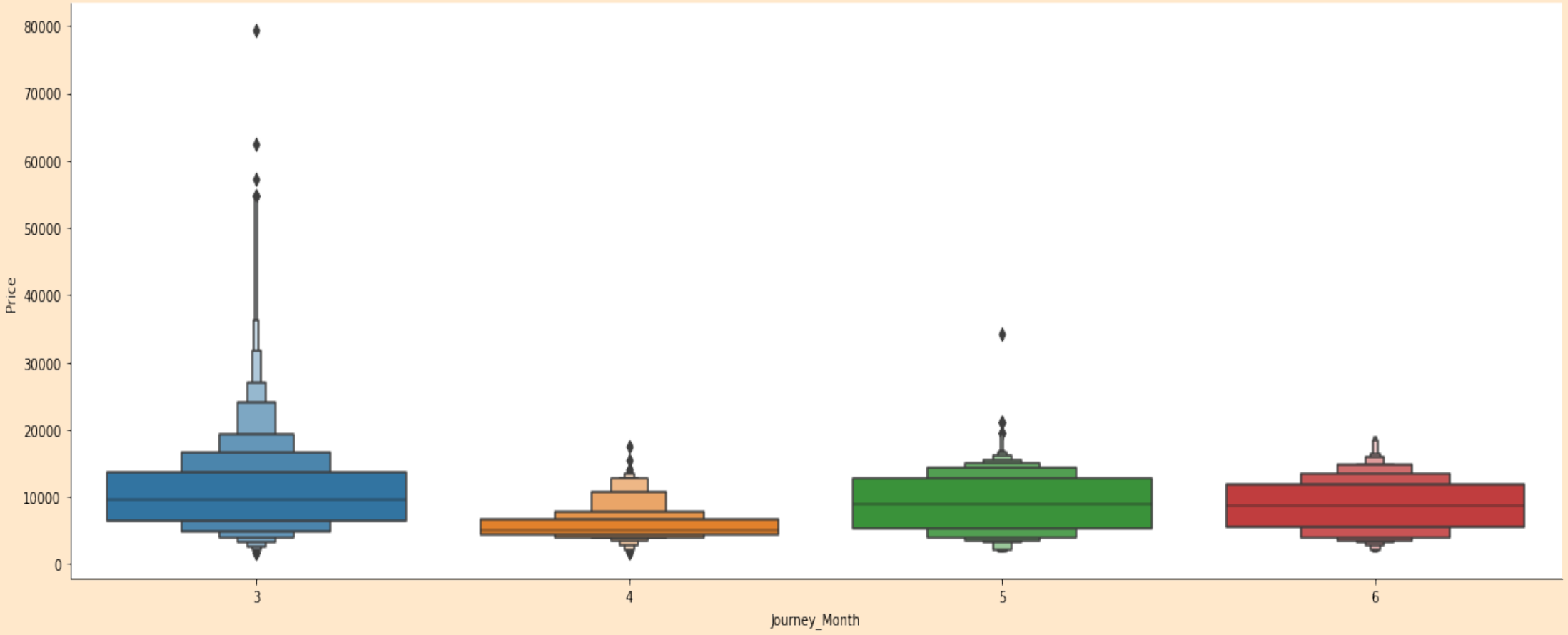
The plot shows that source coulmn does not deviate the dataset much.

Stops vs. Price plot



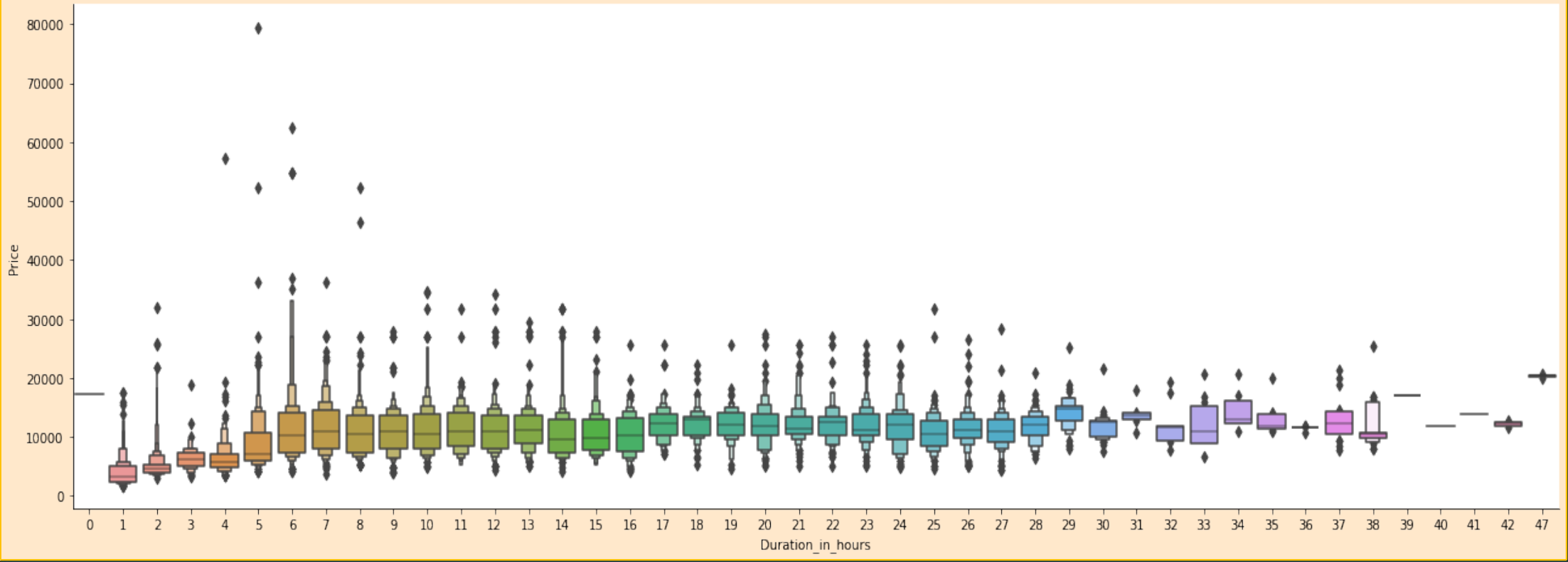
Total Stops affect in the variation of price.

Journey month vs. Price plot



Journey month also does not create much deviations.

Duration vs. Price plot



From this plot we can verify that Duration of a flight is an important feature related to price

Modelling Overview

- Supervised learning/Regression

  Models Used:
  - Extra Trees
  - Random Forest
  - RandomizedCV

Modelling Steps

01

Data Preprocessing

- Feature Selection
- Feature engineering
- Train test data split(80%-20%)

02

Data Fitting and Tuning

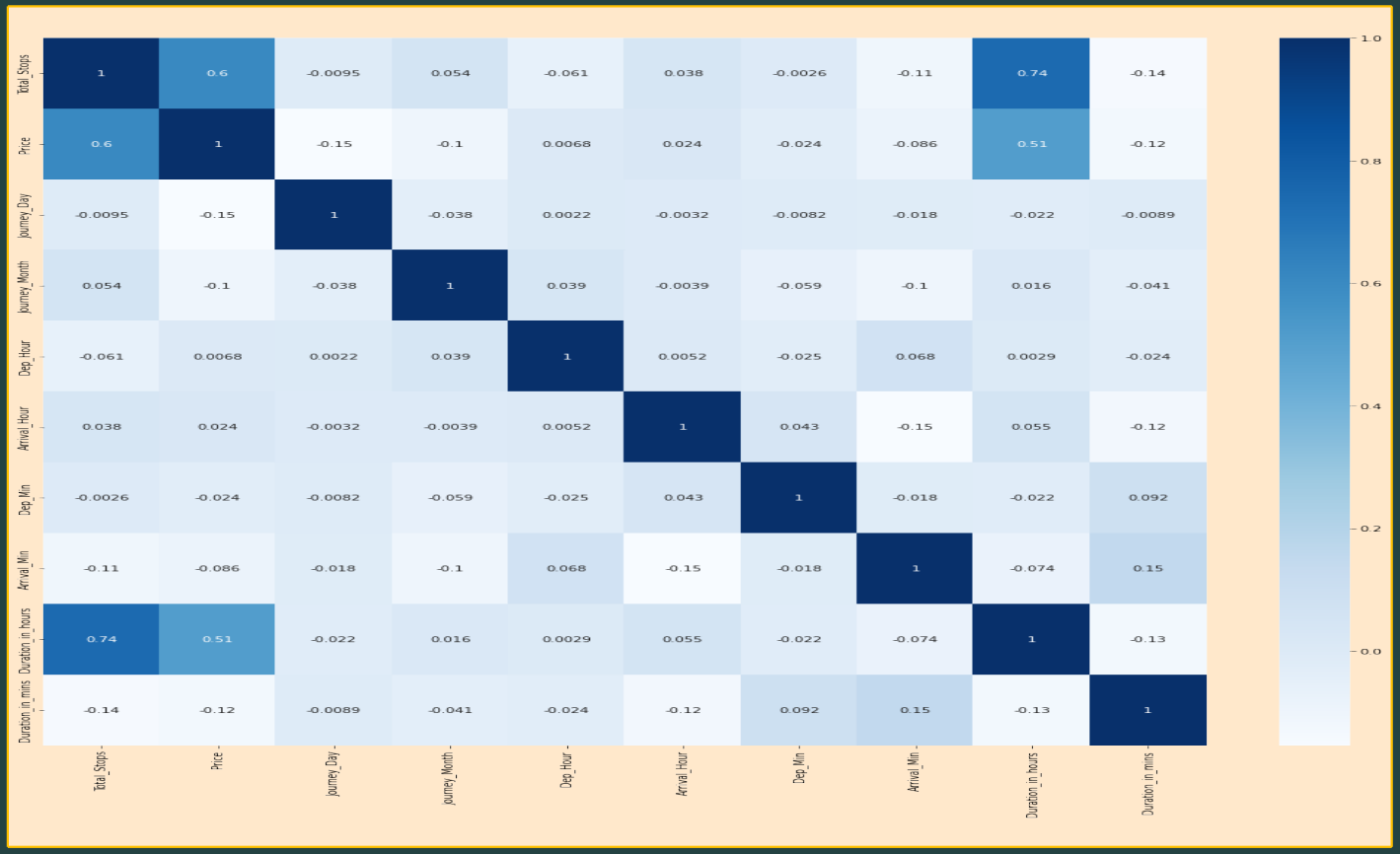
- Start with default model parameters
- Hyperparameter tuning
- Density plot

03

Model Evaluation

- Model testing
- Comapare with the other models
- Measure RMSE Score and R2 score

Correlation matrix



Correlation matrix can be used for feature selection

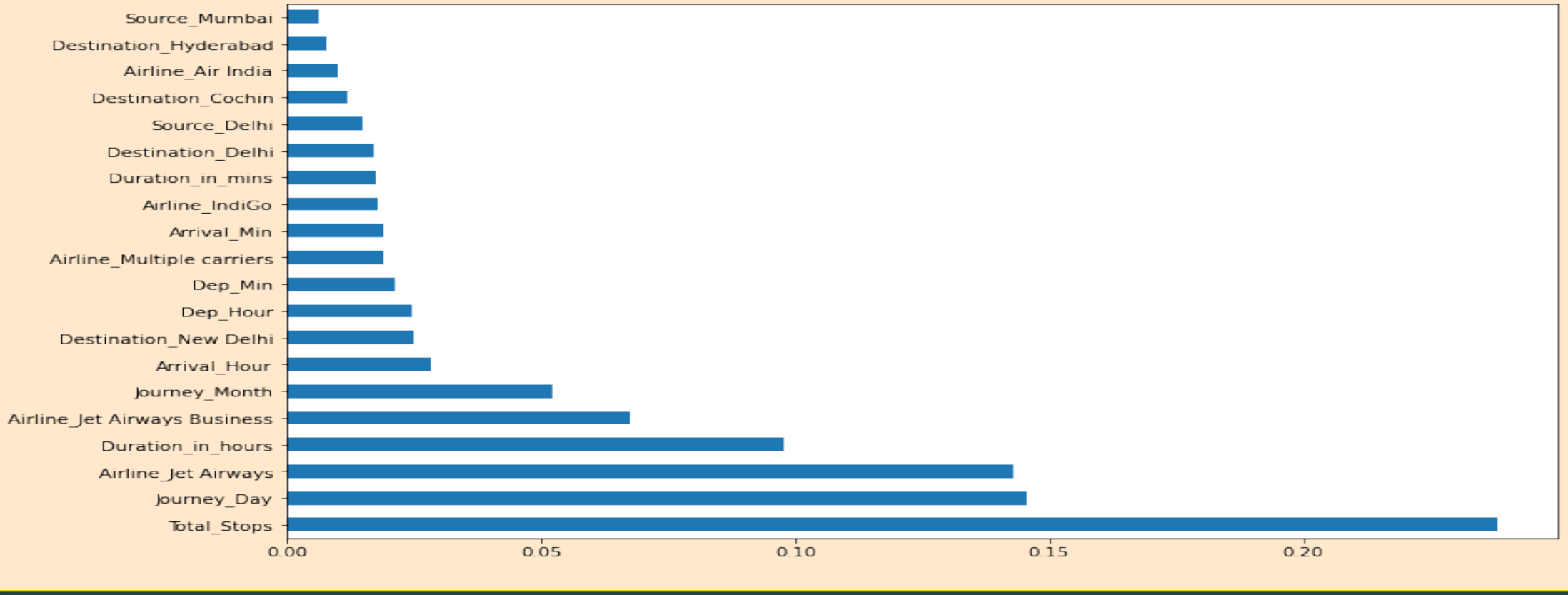
Extra Trees Regression Model

METRICS:

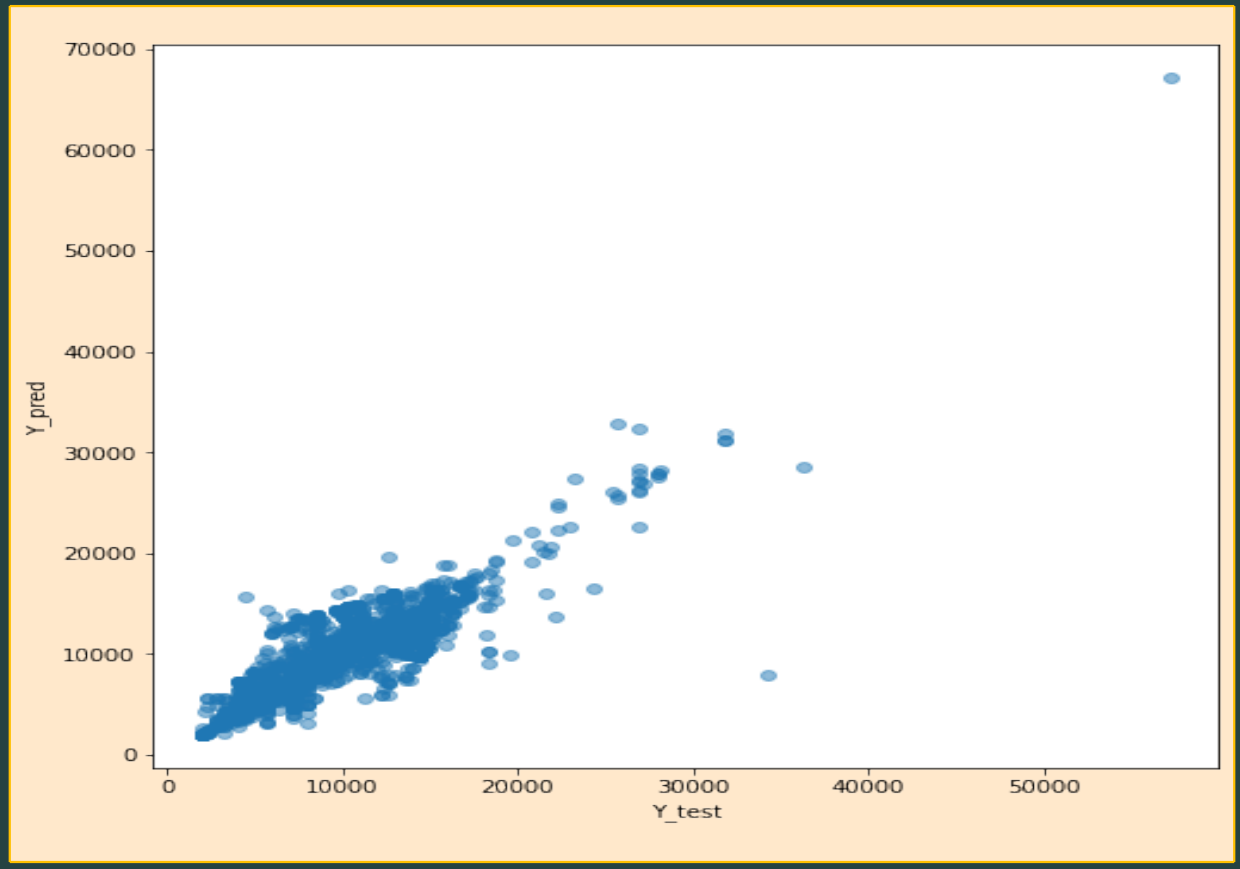
MAE for ExtraTressRegressor: 1221.9764114802683
MSE for ExtraTressRegressor: 4170677.2210174548
RMSE for ExtraTressRegressor: 2042.2235972139424

R2 score for ExtraTressRegressor: 0.8065733083006597

Extra Trees feature importances



Extra Trees Model Scatter Plot

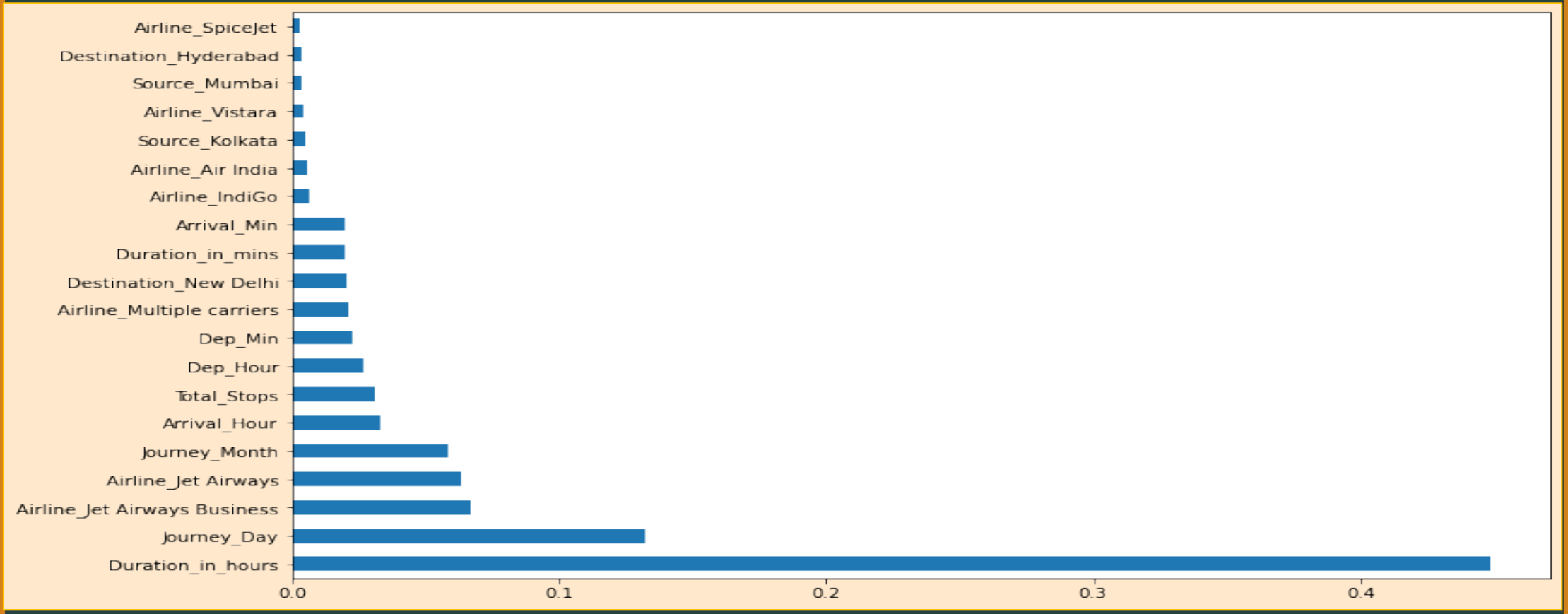


Random Forest Regression Model

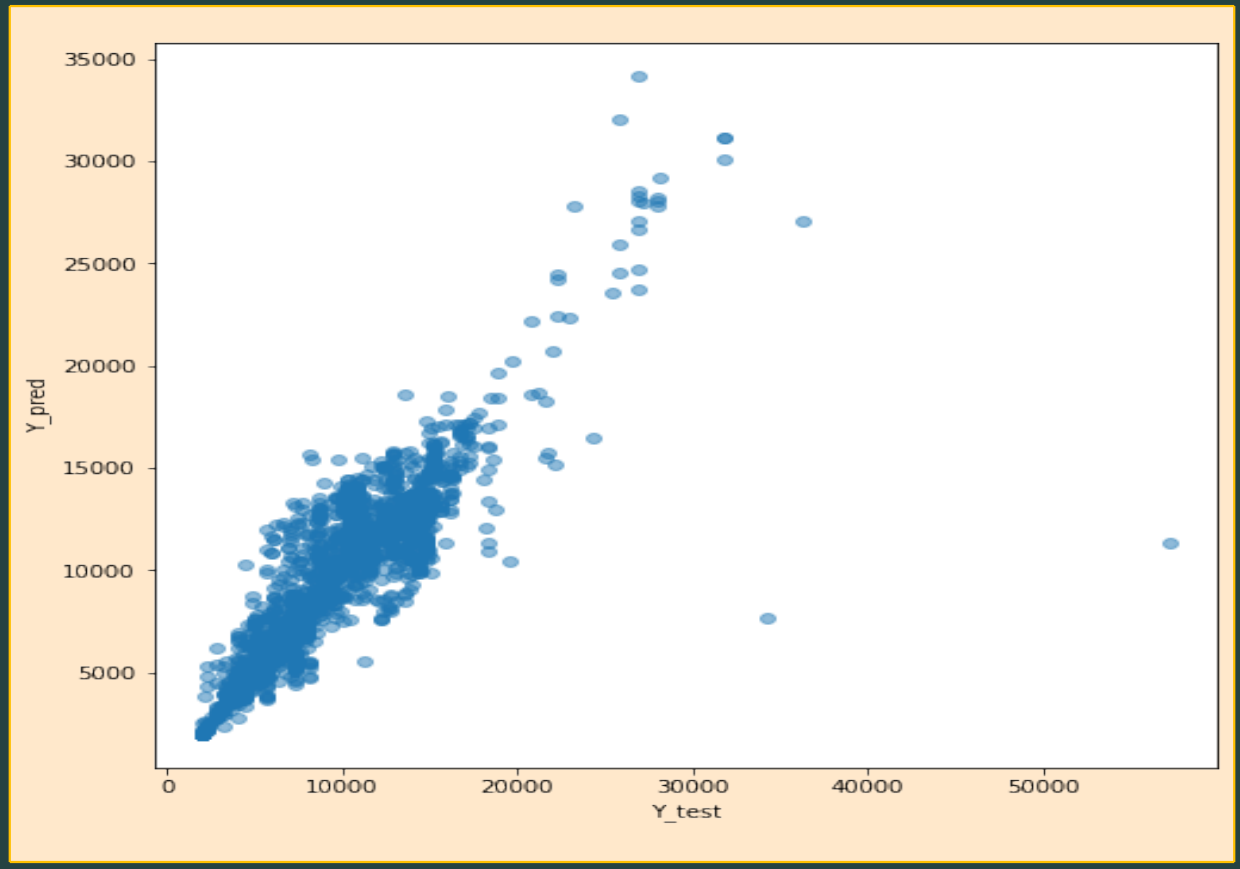
METRICS:

MAE for RondomForestRegressor: 1179.6510737922713
MSE for RondomForestRegressor: 4370346.765271038
RMSE for RondomForestRegressor: 2090.5374345538607
R2 score for RondomForestRegressor: 0.797313080924765

Random Forest feature importances



Random Forest Model Scatter Plot



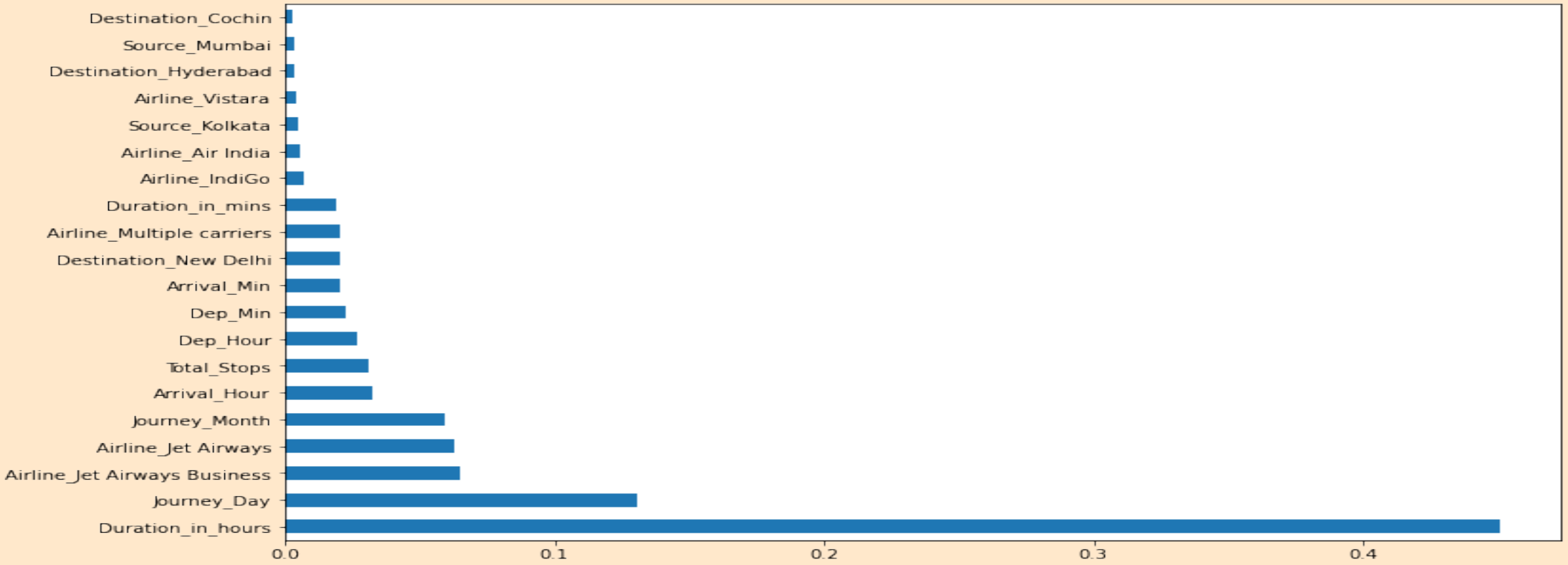
Random Forest Regression Model with Hyperparameter Tuning

METRICS:

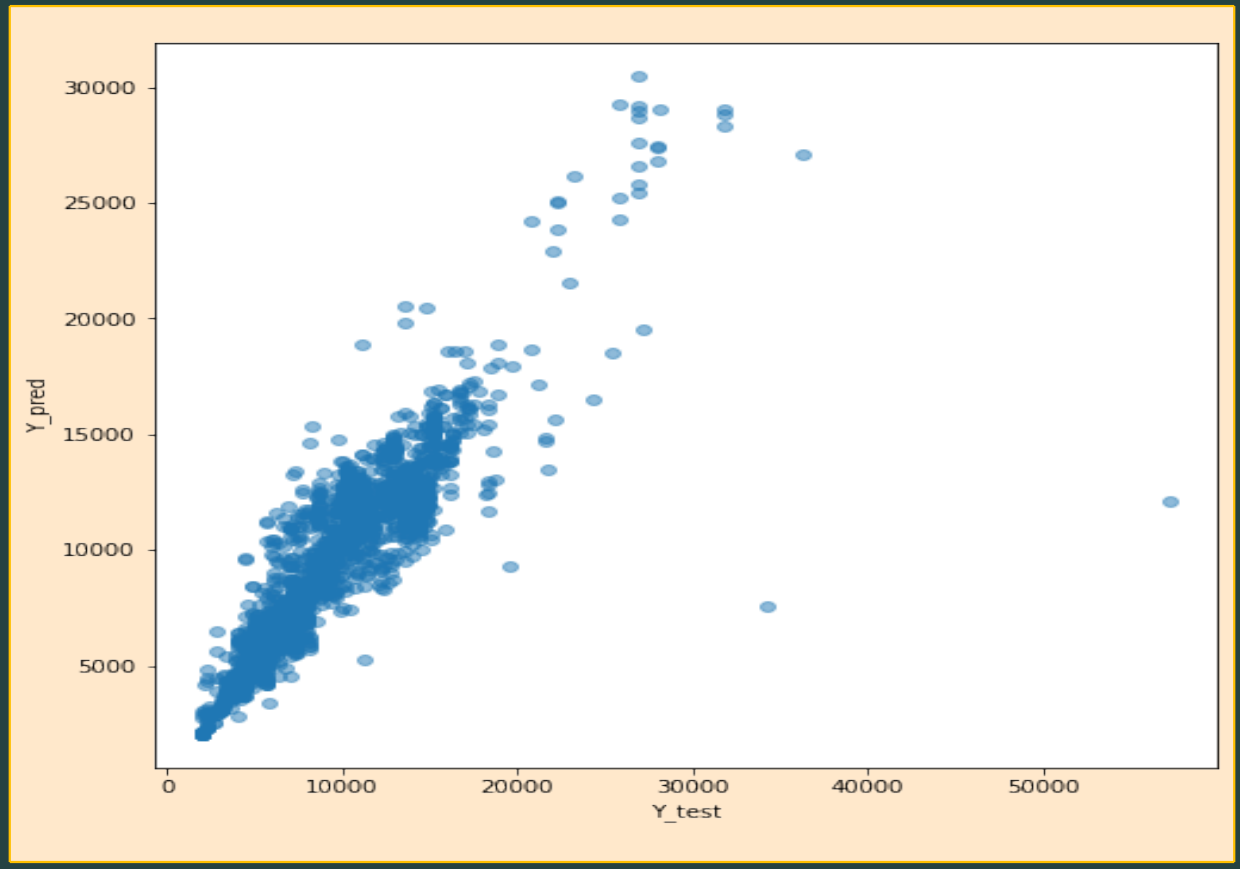
MAE for RondomForestRegressor with Hyperparameter Tuning: 1166.265454002958
MSE for RondomForestRegressor with Hyperparameter Tuning: 4050145.144259414
RMSE for RondomForestRegressor with Hyperparameter Tuning: 2012.4972408079009

R2 score for RondomForestRegressor with Hyperparameter Tuning: 0.8121633167370523

Random Forest with Hyperparameter tuning feature importances



Random Forest Model with Hyperparameter Tuning Scatter Plot



R2 Score Comparison

R2 score for RondomForestRegressor: 0.797313080924765

R2 score for ExtraTressRegressor: 0.8065733083006597

R2 score for RondomForestRegressor with Hyperparameter Tuning: 0.8121633167370523

Challenges

- Understanding the features
- Feature Engineering
- Getting Higher accuracy on the models with less overfitting

Conclusion

- Random forest model with hyperparmeter tuning provides the best result with R2 score of 81.2%
- Extra Trees has also good R2 score but the model seems to overfit.
- Base model of Random forest was least satisfactory with an R2 score of 79%

THANK YOU

FOR YOUR ATTENTION