Flight Price Prediction

Abstract – This paper reports our experience with building a Glass vs. No-Glass classifier. We have a dataset generated using Generative Adversarial Neural Network (GAN) which consists of feature vectors of images of people wearing or not wearing glasses. The GAN network creates these images using a 512 number latent vector. We use various classification algorithms and compare their results in this report.. Apart from the GAN dataset , another image dataset was used to implement CNN and VGG for classification.

INTRODUCTION

Flight Price is very hard to guess, because the price shown today for a specific flight may not be the same tomorrow. The goal of this project is to predict the price of a flight ticket. Many features affect the price like date of journey, destination, the starting point. This dataset contains many features that can be used to predict the price of the flight.

Dataset:

The dataset contains 10684 rows and 10 features.

METHODOLOGY

Preprocessing of the data:

(TO MAKE THE DATASET IDEAL FOR APPLYING VARIOUS ML MODELS)

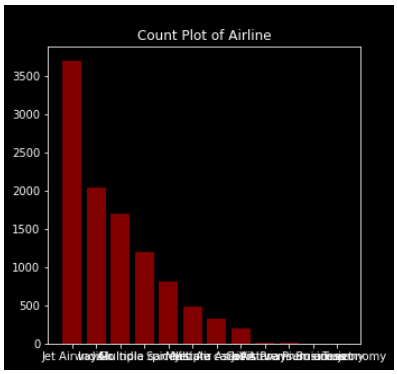
1. All the important and required libraries were imported
2. The excel sheet in .csv format containing the dataset was opened, read and loaded to form a panda dataframe.
3. The data is analyzed i.e., all the attributes are checked for null values

4. Attributes with string values are converted into integers. Eg: Date of Journey is in dd/mm/yy format which is split into a separate date column with dd, month column with mm and year column with yy.
5. Due to the changes made, we found some NULL values hence replaced them/ deleted some unwanted rows.
6. Label encoding is done to convert all of them into categorical values.
7. The target and the features are separated.

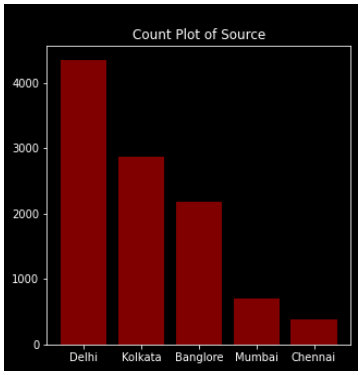
Visualization:

We plotted count plots of all the categorial features.

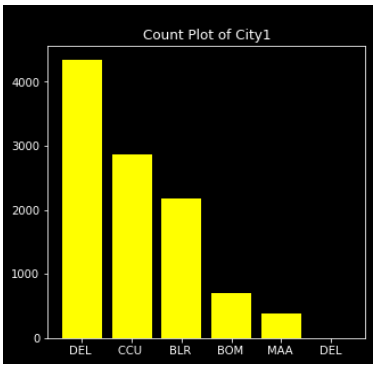
Feature:Airline



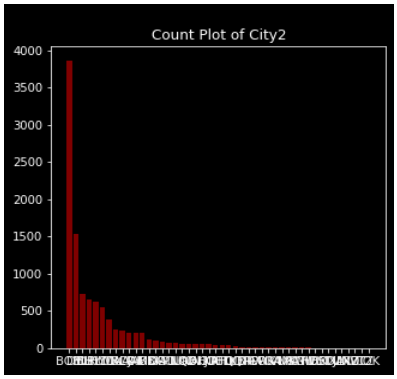
Feature:Source



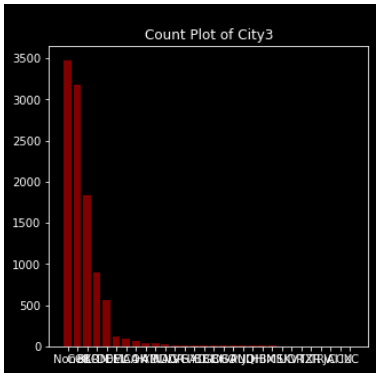
Feature:City1



Feature:City2



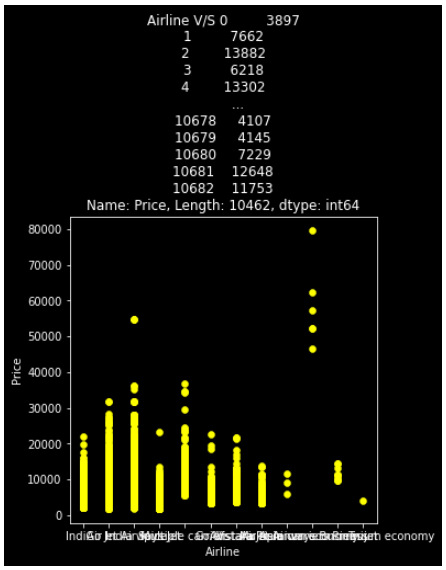
Feature:City3

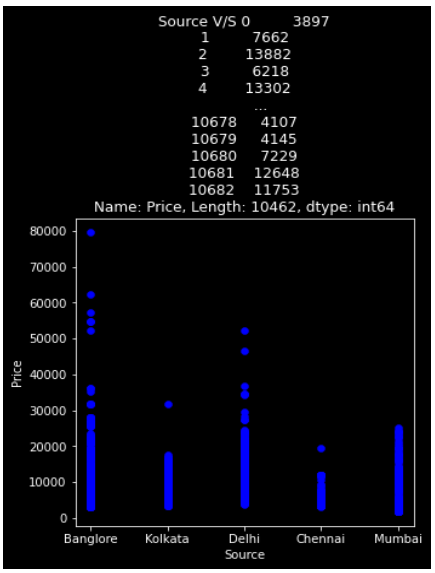


From the count plots we can observe that

- Jet Airways is the most preferred airway.
- Majority of the flights take off from the source Delhi.

Comparison Plots





Next from the comparison plots, we find that

- Jet Airways business class has the highest prices between 50k -80k.
- All the flights with high cost depart from Bangalore.
- All the high cost flights have a destination as Delhi.
- Business class tickets have a high price.

After visualizing the data, it is split into train and test data with test size of 0.2.

Training the data using Regression models:

There are many regression models, but out of them I chose to compare

- Decision Tree Regressor
- Gradient Boosting Regressor
- KNN
- Random Forest Regressor

  First all the training data is trained using all the four models without defining any parameters, and obtaining all their r2_scores and root mean square errors.

  By keeping all those values in a table

Models	Score
Decision Tree Regressor	84.71
Gradient Boosting Regressor	84.87

KNN	80.81
Random Forest Regressor	91.36

Tuning the Hyper Parameters of the models

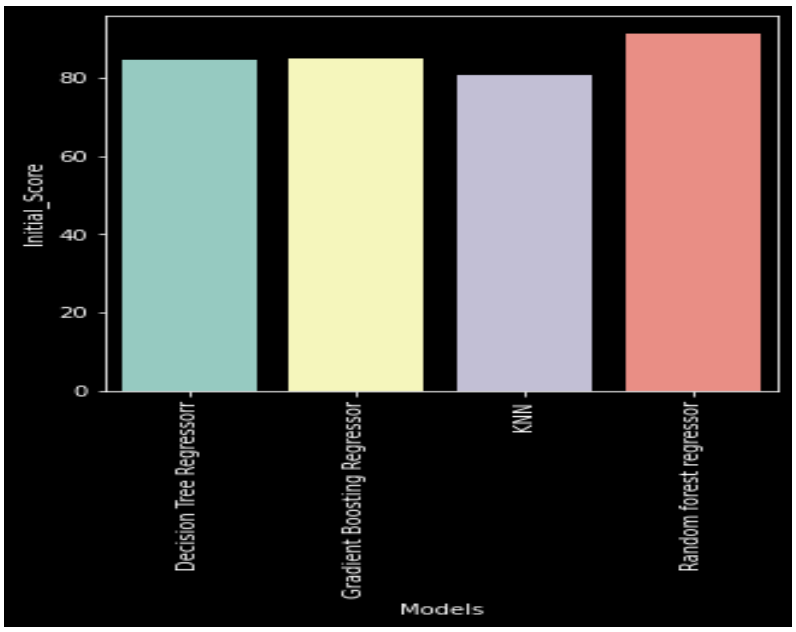
For optimizing our models and for obtaining the best results we tune the hyperparameters of the models using Grid search and obtain the best parameters and then again trained our data with them.

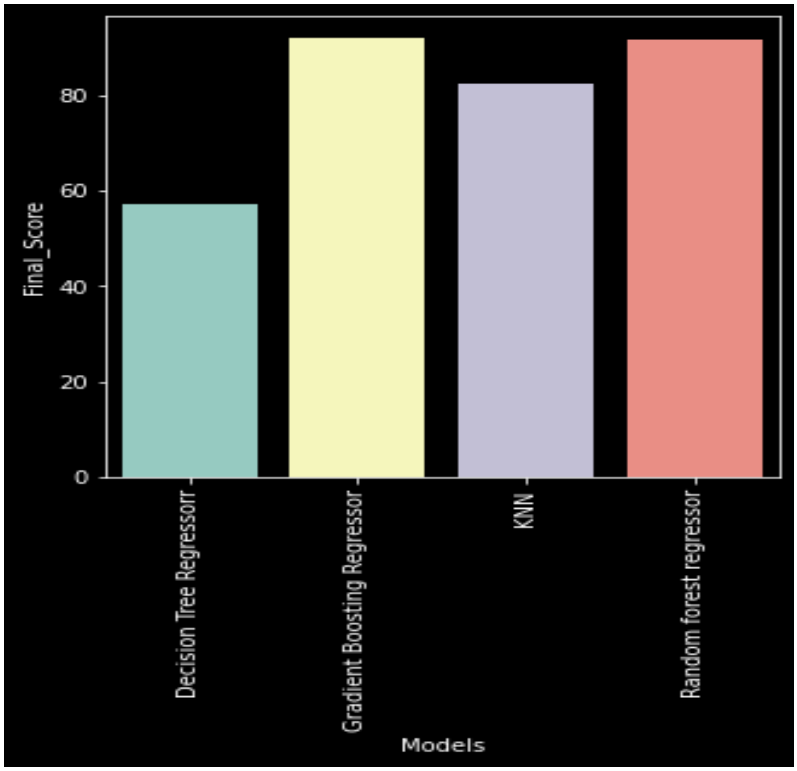
Models after Tuning	Score
Decision Tree Regressor	57.17
Gradient Boosting Regressor	92.16
KNN	82.52
Random Forest Regressor	91.75

Evaluation and Comparison of Models

The models implemented were evaluated using the techniques like r2_score and root mean squared error and are noted in the below table.

Models	Final Score	Initial Score
Decision Tree Regressor	57.17	84.71
Gradient Boosting Regressor	92.16	84.87
KNN	82.52	80.81
Random Forest Regressor	91.75	91.36





Result and Analysis

Finally using all the models-Decision Tree Regressor, Gradient Boosting, KNN, Random Forest Regressor before and after the tuning, we get the maximum score with Gradient Boosting Regressor with its best parameters tuned. So hence we finally chose Gradient Boosting with some mentioned parameters(max_depth=15,min_samples_split=200) and the score we finally obtained is nearly 92%. Random Forest Regressor also gave the second

best accuracy after the Gradient Booster, but we finally only chose Gradient Boosting Regressor.