Ahsanullah University of Science and Technology

Department of Computer Science and Engineering

CSE 4108

Artificial Intelligence Lab

Project Name: Flight Ticket Price Prediction

Submitted To

Mr. Faisal Muhammad Shah
Associate Professor, CSE, AUST

Md. Siam Ansary

Lecturer, CSE, AUST

Submitted By

Sadia Khanam Arni	180104104
Fariha Alam	180104115
Kishowloy Datta	180104117

Description of the Problem

Artificial intelligence is rapidly evolving across all industries, and by utilizing AI models, we can solve or predict many aspects for a given dataset. In this project, we have tried to predict the price of flight ticket for a given dataset. It is mainly a regression problem .Here we have tried to predict this with some features.

Dataset of the Problem

Mainly we have used Ticket price of flight Dataset as regression dataset for the prediction model. The dataset has a total of 11 columns and 582 rows including the target column. The target columns the last column and rest of the columns are the features**.**

The target column is Price column and the feature columns are :

- **Airline** (Name of the airline used for traveling)
- **Date of Journey** (Date at which a person traveled YYYY-MM-DD)
- **Source** (Starting location of flight)
- **Destination** (Ending location of flight)
- **Route** (This contains information on starting and ending location of the journey in the standard format used by airlines
- **Dep_time** (Starting time of flight from starting location)
- **Arrival_time** (Arrival time of flight at destination)
- **Duration** (Duration of flight in hours/minutes)
- **Total_Stops** (Number of total stops flight took before landing at the destination.)
- **Additional_Info** (Shown any additional information about a flight)
- **Price** (Price of the flight)

Model Description

For predictive modeling of the features, we have used 5 regression algorithms. They are:

1. Linear Regression
2. KNN Regression
3. Decision Tree Classification
4. Random Forest
5. Support Vector Machine

Linear Regression: Linear Regression is a linear model, e.g. a model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x).

KNN Regression: KNN Regression is a non-parametric method that, in an intuitive manner, approximates the association between independent variables and the continuous outcome by averaging the observations in the same neighborhood.

Decision Tree Classification: Decision Tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.

Random Forest: Random Forest is a supervised learning algorithm. It can be used for both classification and regression. It is also the most flexible and easy to use algorithm. Random forest classifier selects decision trees on randomly selected data samples, gets prediction from each tree and selects the best solution by means of voting. It also provides a pretty good indicator of feature importance.

Support Vector Machine: Support Vector Machines (SVMs) are a set of supervised learning methods used for classification, regression and outliers detection. The advantages of support vector machines are: Effective in high dimensional spaces. Still effective in cases where number of dimensions is greater than the number of samples.

Performance Comparison

Linear Regression:

Performance Metrics	R2 Score	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error
Performance Score	0.6017229420068	1954.029801882117	7676682.538512529	2770.682684558542

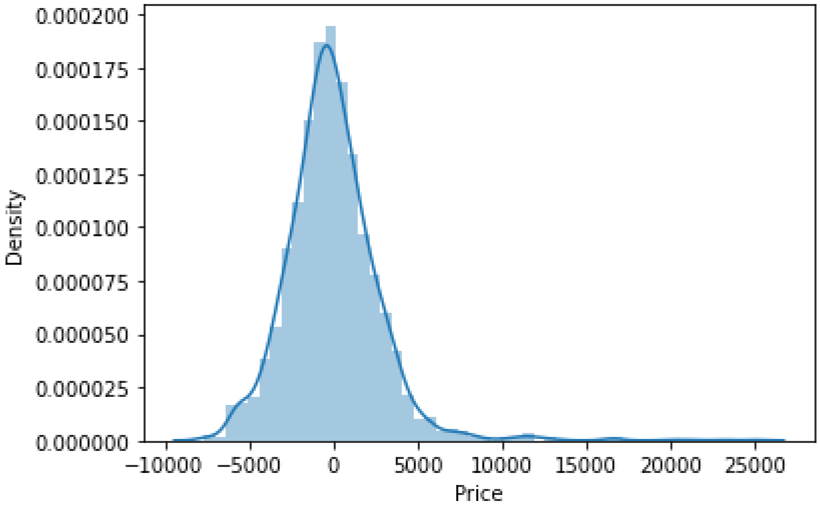

Figure : Linear regression Model

KNN Regression:

Performance Metrics	R2 Score	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error
Performance Score	0.6250530968708452	1771.8035563874591	7227000.115507721	2688.3080395497313

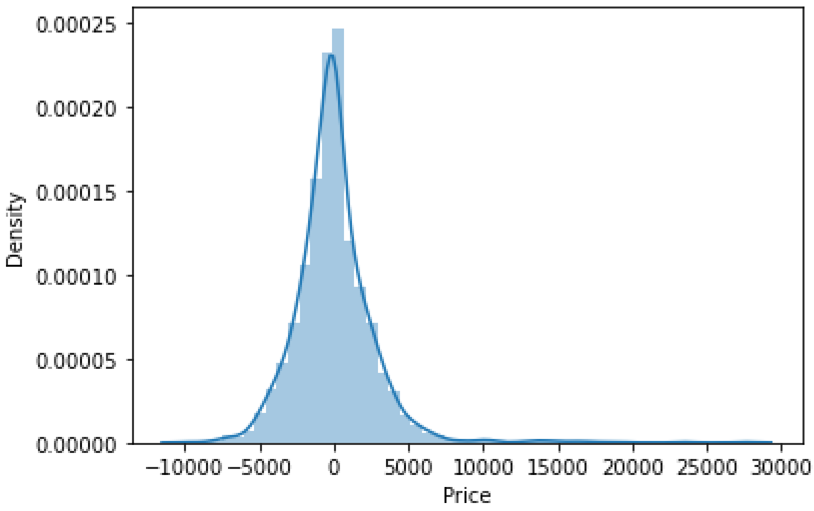

Figure : KNN regression Model

Decision Tree Classification:

Performance Metrics	R2 Score	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error
Performance Score	0.6840943400547228	1374.514272344408	6088996.12681329	2467.589132496188

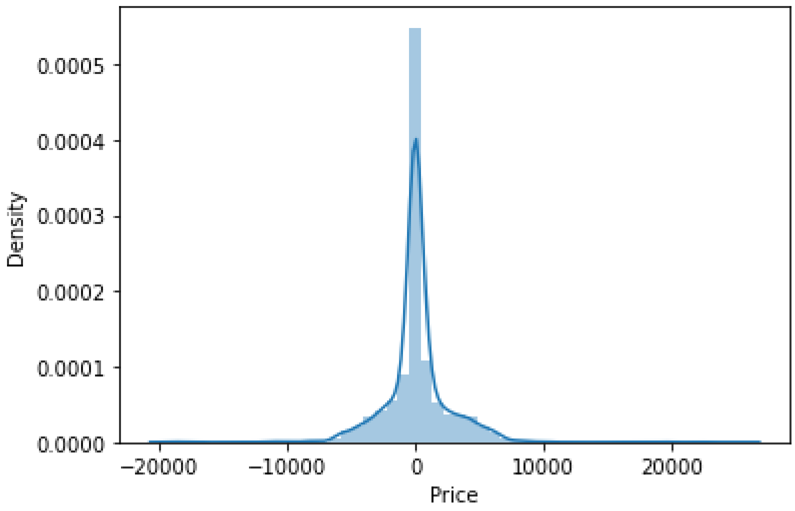

Figure : Decision Tree Model

Random Forest:

Performance Metrics	R2 Score	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error
Performance Score	0.8131447447078719	1156.6015760096893	3601584.491856098	1897.7841004329491

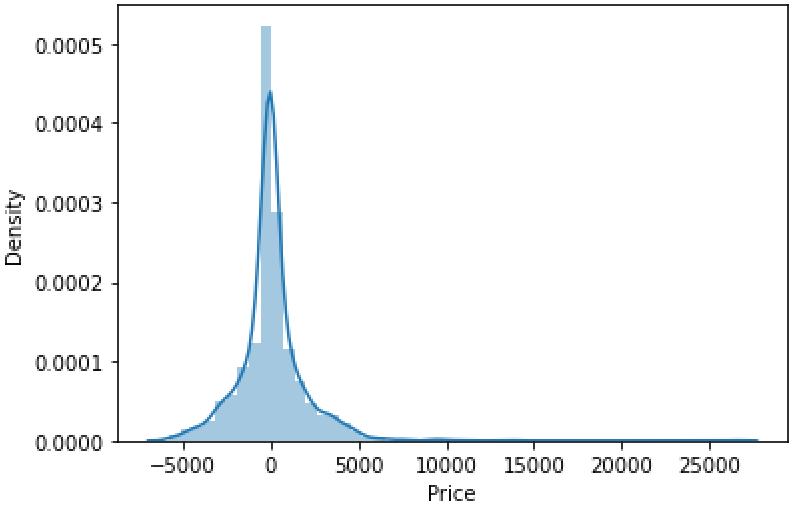

Figure : Random Forest Model

Support Vector Machine:

Performance Metrics	R2 Score	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error
Performance Score	0.05738434846621354	3346.536598485241	18168661.65758806	4262.471308711421

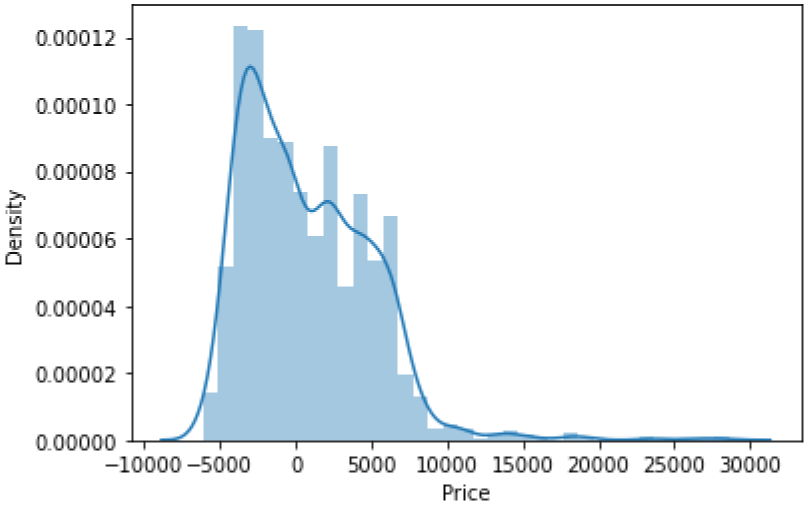

Figure : Support Vector Machine Model

Discussion

After analyzing the five models, we have come to a conclusion that the Random Forest Model has performed better than the rest of the models.

Random Forest works pretty smooth and compute better results for datasets. The accuracy of the Random Forest is 81% that means the model is predicting 81% of the data correct, which is an ideal value.

On the other hand, the accuracy of the Decision tree is 68% that means the model is predicting 68% of the data correct, which is an ideal value. The accuracy of the rest of the models are below than 60%.

So, in compare to these 5 model for predictive dataset, we can say that Random Forest model is best suited than other for this dataset.

Contribution

ID-180104104 [33.33%]:

- Data pre-processing
- Linear Regression Model
- Dataset Documentation
- Report

ID-180104115[33.33%]:

- Data pre-processing
- Random Forest Model
- Dataset Documentation
- Report

ID-180104117[33.33%]:

- Decision Tree Classification Model
- KNN Regression Model
- Support Vector Machine Model
- Report.