Kemal Enes Akyüz - 22003521
Efe Tarhan - 22002840

EEE485 - Statistical Learning and Data Analytics Term Project Proposal

Flight Price Prediction Using Linear Regression, Random Forest and Deep Learning.

Introduction

The aim of the proposed project described within this report is to investigate and analyze the dataset titled "Flight Price Prediction" [1] that contains an extensive list of ticket prices for flights (300,261 data points) and information about the airline company, origin and destination, number of days a ticket has been bought before the flight and some more features (11 in total). To analyze this data, Linear Regression , Random Forests and Deep Learning methods will be used with, if necessary, adequate additions to the methods to better represent the data and to avoid under and overfitting.

Project Details

Through the application of this project, three differing methods will be used to analyze the same data set, which will enable the possibility of comparing the performances of these methods and decide which one is the best option among them to represent the Flight Price Prediction data set. The first of the methods that is proposed for this purpose is Linear Regression that assumes a linear relation between the response and the predictors and then tries to find the coefficients that best suit the data set. This method, if necessary, can also be supplemented by using additional penalty terms (ridge, lasso regression) or feature expansion to cover more relations missed by a basic linear relation. The second method that is proposed is random forests which makes use of and improves the decision tree method. Here, a number of decision trees are constructed on the bootstrapped versions of the data set (repeated samples from the data set) and for each node only some percentage of all predictors are considered so that the correlation between trees is lowered. Lastly, a deep learning method will be applied to the same data set which will use the predictors as inputs to the model and the flight price as the output that will be given by the model prediction.

Dataset Description

The "Flight Price Prediction" Dataset has been constructed by an individual with name Shubam Bathwal by gathering information from a website called "EaseMyTrip" with the help of a web snippet tool and has been uploaded to Kaggle. Data was collected for 50 days, from February 11th to March 31st, 2022.

The data contains information about the prices of flights between 6 major cities of India with additional information as features of the dataset. Feature columns of the dataset include information about the airline, flight code, departure location, departure time of the flight with 6 categorical values, number of stops with 3 categorical values, arrival time with 6 categorical values, arrival location, the "class" of the flight (economy or business), duration of the flight and how many days before the flight does the ticket is taken and lastly the price of the given flight. These features will be explained in further detail.

1. **Airline**: 6 categorical values that include the names of the airline companies.
2. **Flight**: The planes' flight code which is a categorical value
3. **Source City**: 6 categorical values that denote the names of the cities that the flight is being departed
4. **Departure Tim**e: 6 categorical time labels which is the quantized version of the 24 hours into 6 intervals
5. **Stops**: 3 categorical values that address the number of stops of the flight
6. **Arrival Time**: 6 categorical time labels which is the quantized version of the 24 hours into 6 intervals.
7. **Destination City**: 6 categorical values that denote the landing location of the flight.
8. **Class**: 2 categorical values which indicate if the flight is "economy" or "business" class.
9. **Duration**: A continuous feature that represents how long the flight is expected to last.
10. **Days Left**: This feature indicates how many days earlier the ticket has been bought before the flight.
11. **Price**: This final feature is the price of the flight which is a continuous variable and target of the prediction task of this dataset.

Expected Challenges

Main challenge of this term project is to do analysis on the selected dataset by applying the given machine learning methods (Linear Regression , Random Forests and Neural Networks) by designing the functions and algorithms from scratch, by using only the basic libraries of Python like NumPy or Pandas. Adjusting the methods and the hyperparameters such that the data set is best described by the chosen methods may also prove to be difficult to achieve. There are also additional challenges of the dataset like understanding the intuitive relation between the features of the data and the price. Following are some example challenges of the project:

1. What is the relationship between the number of days buying the ticket before the flight and the price of the ticket?
2. How does the source and destination of the flight change the price by considering the bird's eye distance between locations?
3. How does the price of a ticket change with the class of the flight ?
4. How does the time of the day in which the flight occurs affect the price of the ticket?
5. Which airlines provide cheaper tickets with respect to the others?

References

[1] Bathwal, Shubham. "Flight Price Prediction." Kaggle. Accessed October 14, 2023. https://www.kaggle.com/datasets/shubhambathwal/flight-price-prediction.