JP Morgan Quant Challenge

Team - Mathminers
Ashay Walke 14MA20050
Nayan Raju Vysyaraju 14MA20049

Question 1

The Research and development wing of Express Transport Company has designed a new transportation system which reduces the time travel by half, compared to the air travel. In the first phase they plan to connect 7 cities in India.

To understand the pricing of their tickets, they recently conducted a survey of passengers travelling by air asking them the details of their existing flights and the price they were willing to pay to reduce the travel time. The data collected from the survey has been shared with you.

You need to design a dynamic pricing model which given the passenger and fare details, computes the price the passenger is willing to pay.

Data set	Length
Training Data set	7500
Testing Data set	2500

Feature Engineering

The features given in the dataset can be classified into 2 categories that are, Customer attributes and Flight Attributes.

Attributes
From and To city
Date of Flight
Date of Birth
Date and Time of Flight
Date of Booking
Class

Target Variable - Fare
The target variable is Fare.

As we can see from the above features most of the features are either categorical or Date/Time. We use 2 approaches to tackle categorical features namely, label encoding and one-hot encoding. We also change the datatype of dates/times into datetime. Following is the list of newly engineered features.

Feature	Details
Age_at_Booking	The difference between booking date and Date of Birth
Booking_to_flight	The difference between booking date and date of flight
Flight_Is_weekend	Is the flight on a weekday
Flight_Is_Month_End	Is the flight on a month end
Booking Year	The year of Booking
Flight time	The time is divided in 5 bins throught the day
Distance	The distance between 2 cities
Flight Year	The year of Flight

One-hot vs Label Encoding
We have used one-hot encoding as well as label encoding for various categroical variables in the data-frame depending upon the model used.

Evaluation Metric

The Evaluation metric used in this case is R^2**.** We train the models to minimize their loss functions and leading.

Models
We trained models rangning from Neural Networks, Random Forest , lightGBM and XGBoost. We have seen the variation in the R^2 after using different models as well as different encoding techniques and adding and removing few varable, we have pepared a table comparing all the models, techniques and variables used and their respective R^2

Model	Variables/Encoding	R^2
Multiple Linear Regression	Categorical Encoding,	0.49
Multilayer Neural Network	Cateogorical Encoding	0.59
Multilayer Neural Network	One-Hot Encoding for class, time bins, To and From	0.62
Random Forest Model	Encoding as above and added distance variable	0.79
Random Forest Model	Same as above with removing top 2.5% and bottom 2.5% entries for every city-city flight entires	0.83
XGBoost Model	Same as above without removing outlier	0.86
Tuned XGBoost Model	Tuned XGBoost model using GridSearchCV	0.893

Following are the variables used in the final model which has given the best results.

```
Age_at_Booking              int64
Booking_to_flight           int64
Booking Year                int64
Booking Month               int64
Booking Dayofweek           int64
Booking Is_month_end         bool
Booking Is_month_start       bool
Flight Month                int64
Flight Dayofweek            int64
Flight Is_month_end          bool
Flight Is_month_start        bool
distance                  float64
Class_0                     int64
Class_1                     int64
From_1                      int64
From_2                      int64
From_3                      int64
From_4                      int64
From_5                      int64
From_6                      int64
From_7                      int64
To_1                        int64
To_2                        int64
To_3                        int64
To_4                        int64
To_5                        int64
To_6                        int64
To_7                        int64
Time Label_0                int64
Time Label_1                int64
Time Label_2                int64
Time Label_3                int64
Time Label_4                int64
```

Hence the airfares are predicted using a Tuned XGBoost model having R^2 0.893