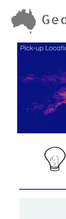
A2B Taxi Fare Predictions

Group 2

Time-based Information

Hourly Pattern

- At on aggregate level, the number of taxi trips peaked at around 8 am.
- The average taxi price was the highest at around 5 am.



Daily Pattern

- Taxi fares experienced a larger fluctuation during a month. Price always moves in the same direction as trip duration.
  - Low taxi fares seemed to appear mainly on non-holidays.
  - Whether it was weekend or not did not drastically change the relationship between trip duration and taxi fares, though the positive relationship of the two was slightly stronger for data on the weekends.

Geographically-based Information

- We can see three clusters of pick-up locations, which based on the geographical context are Sydney, Tornworth and Coffs Harbour.
- Compared to the pick-up point, the drop-off location is more concentrated in the city centre.

Trips that start or end in the Sydney airport have a higher average price than other types of trips

The price per kilometre is higher all times for trips that both started and ended in the city

Prediction Model

Random Forest

- Features: 6
- RMSE: 3.204
- Accuracy: 0.929(20%) 0.828(10%)
- MAE: 1.437

Neural Network

- Features: 15
- RMSE: 4.022
- Accuracy: 0.925(20%) 0.676(10%)
- MAE: 2.345

The gradient boosting model has the highest accuracy measures, low error rate and it can be accurately generalised to the test dataset.

Benchmark KNN

- Features: 15
- RMSE[1] 3.850
- Accuracy[2]: 0.930(20%) 0.758(10%)
- MAE[3] 2.099

Decision Tree regressor

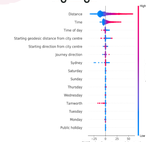
- Features: 15
- RMSE: 3.400
- Accuracy: 0.934(20%) 0.847(10%)
- MAE: 1.472

Gradient Boosting

- Features: 15
- RMSE: 3.265
- Accuracy: 0.936(20%) 0.849(10%)
- MAE: 1.424

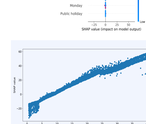
1. RMSE: The Root Mean Square Error is the standard deviation of prediction errors. RMSE is calculated based on validation set the lower RMSE, the better
2. Accuracy: The accuracy within 10% and 20% is the proportion of samples correctly predicted within 10% and 20% of the true value. The higher
3. MAE: The Mean Absolute Error is a measure of errors between paired observations expressing the same phenomenon

Feature Attribution



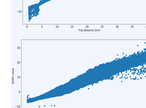
We use **Shapley value** to interpret the importance of various features in our model

- There are a large number of trips for which the fare prediction is not attributed by distance nor time BUT - there are a few trips for which the fare prediction is substantially attributed to distance and time
- Predictably, as trip distance and time increases the attribution to fare prediction increases



- There's a positive correlation between trip distance/time and total Shapley value (respectively)
- The higher feature value (of trip distance and time) of those upper tail observations possibly indicates that the model is appropriately capturing their influence.

Case Studies



To clearly illustrate the performance of our final prediction model we segment our data to 30 clusters based on the pick-up time of day and the pick-up and drop-off locations. The features of each of these clusters as well as the model performance were explored briefly for the purpose of extracting five case studies for further illustration

- Case 1: morning commuters in the North-West area of Sydney
- Case 2: morning commute from the Western Suburbs towards the city
- Case 3: afternoon commute in Sydney's East
- Case 4: early morning trips in the Greater Western Sydney area
- Case 5: afternoon and evening trips in the Penrith area.

- Trips in Case 2 has a higher price due to the longer distance
- Trips in Case 5 is cheaper because they're not in an urban area

Key Findings

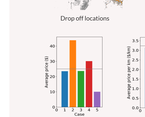
- The prediction error is calculated to evaluate the performance of each case

- Case 1 and 4 have high accuracy and low RMSE, which means a better set of predictions, and vice-versa for Case 3

- Case 2 has higher accuracy, however many outliers draw up the RMSE

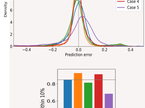
- Case 5 is more spread but with fewer outliers

-In summary, the predictive performance of the final model is varied based on the data's features.



Flagfall

- We notice that the prediction model is hindered by the exclusion of information relating to the type of taxis.

- Figure on the right depicts the distributions of the prediction errors categorized by the flagfall values

- There is a clear indication that the charged price of maxi taxi and handicap trips are being consistently under-estimated.