NYC Taxi Trip Duration Prediction

Data

Training and test data for this project were provided by Kaggle and can be found here: https://www.kaggle.com/c/nyc-taxi-trip-duration/data.

Feature Extraction

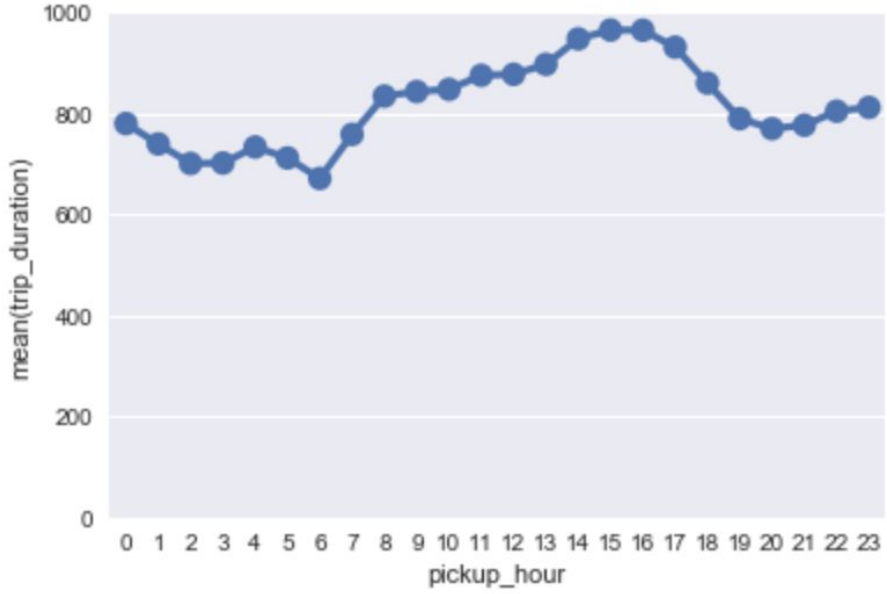
Distance
We calculate trip duration using the haversine distance formula. Haversine formula determines the great-circle distance between two points on a sphere given their longitudes and latitudes. More information about haversine formula can be found here.

Direction
We calculate direction of trip using the bearing distance formula. The bearing of a point is the number of degrees in the angle measured in a clockwise direction from the north line to the line joining the centre of the compass with the point.
http://www.mathsteacher.com.au/year7/ch08_angles/07_bear/bearing.htm

Day of the week
In our exploration, we see that trip duration on Sat and Sun is a lot shorter than trips on Mon-Fri. To capture this variability in the model, we use pickup_day as a feature.

Time of day
We also see variability in trip duration by time of day, and use pickup_hour as a feature to capture this variability.

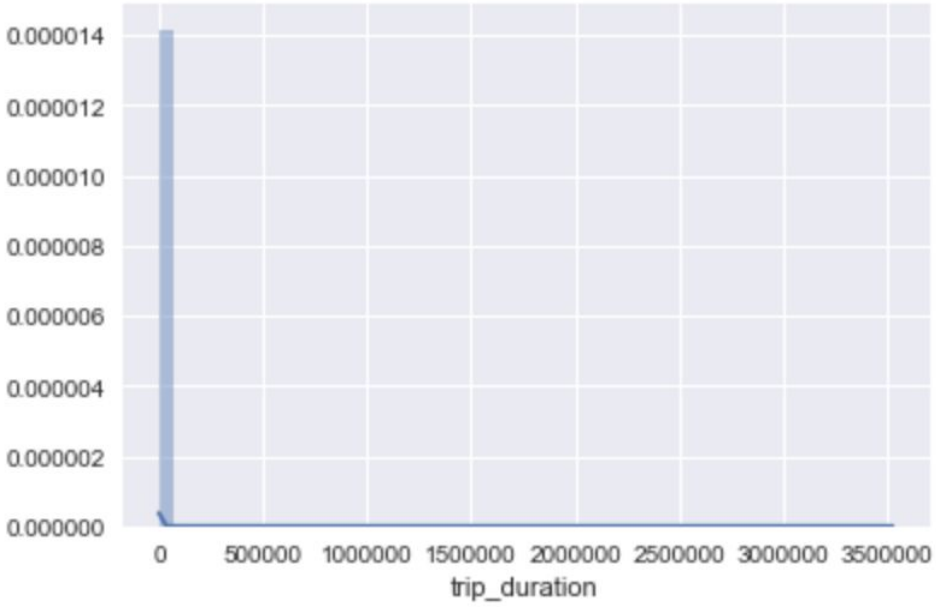
Trip Duration(seconds) by pickup hour



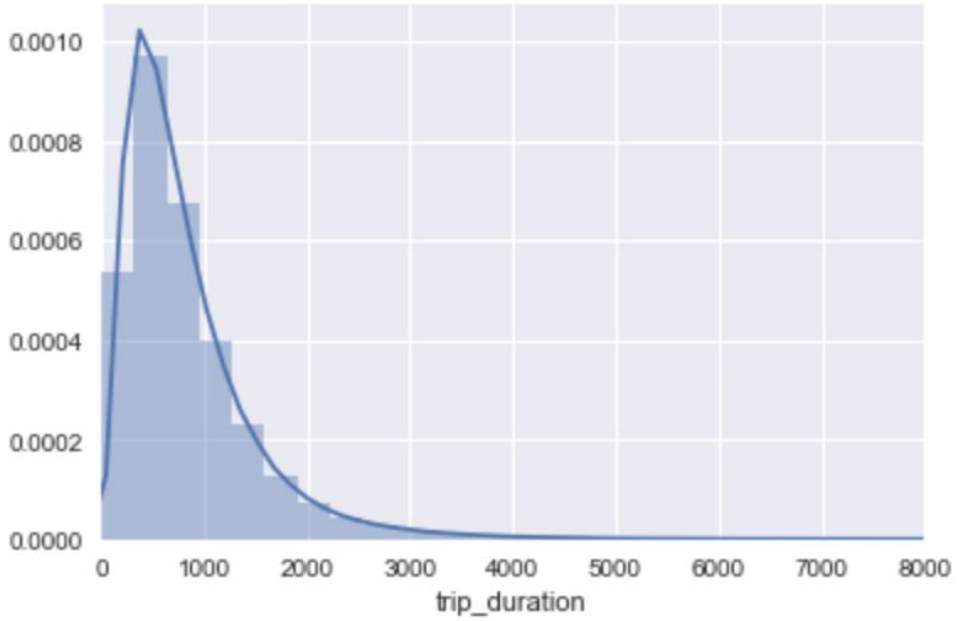
Removing Outliers

In the distribution below, we see that there are some outliers in the dataset (see X Axis is stretched to 3500K seconds). We remove all trip duration values greater than 3 standard deviation from the mean.

Distribution of Trip Duration(seconds) **Before** Removing Outliers



Distribution of Trip Duration(seconds) **After** Removing Outliers



Model Training

Our output variable(Trip duration) is continuous, so we will set up the model as a regression problem. We have three different datasets to train and assess the performance of our model.

- Training data
- Validation data
- Kaggle's test data

Kaggle has its own test data which calculates the RMSLE for our model and idea of having a good score on Kaggle's dataset is to understand if the model generalizes well for unseen data.

Evaluation Metric

Evaluation metric for the kaggle competition is RMSLE (Root Mean Squared Logarithmic Error), so we convert our dependant variable "Trip Duration" into "Log Trip Duration" and use RMSE as the evaluation Metric.

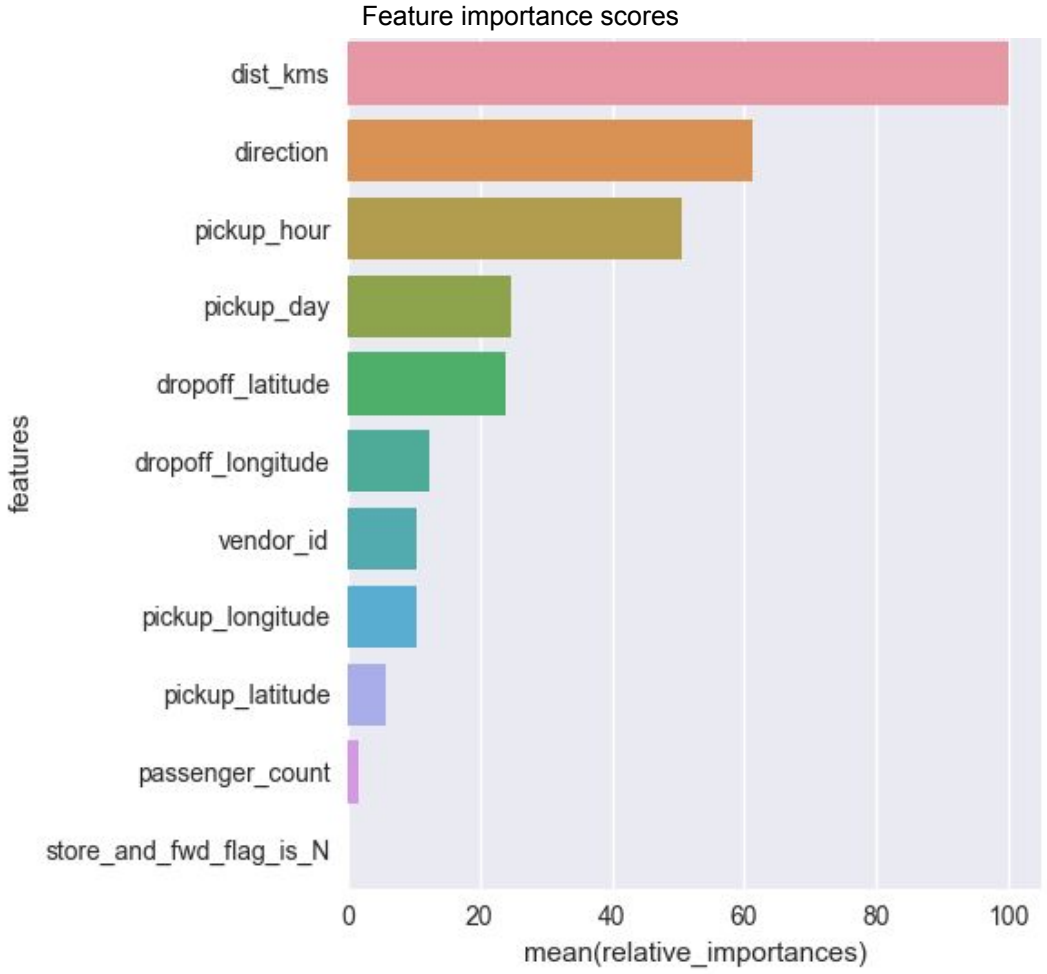
Cross Validation

We perform 4 fold cross validation on our dataset to evaluate model performance. This helps ensure that the model is not overfitting and generalizes well to unseen data.

Models

Gradient Boosting Regression

We apply gradient boosting regressor using sklearn to predict trip duration. More information on Gradient boosting regressors can be found here https://en.wikipedia.org/wiki/Gradient_boosting. A tree based approach gives us feature importance scores which are useful in understanding which features are actually useful to determine variability in the output variable. We see that distance and direction are the two most important predictors of trip duration.



With Gradient boosting regression, we get a RMSLE of 0.39 for the validation set and 0.43 for Kaggle's test dataset.

Gradient Boosting Regression Results

Mean R squared (Cross validated)	**Mean RMSLE (Cross Validated)**	**Kaggle Test Dataset RMSLE**
74.58%	0.3911	0.438

XG Boost

To try to improve our model performance, we use XG Boost Regressor. XG Boost is widely used in Kaggle competitions. We get slight improvements in our model with XG Boost.

XG Boost Results

Training Dataset RMSLE	**Validation Dataset RMSLE**	**Kaggle Test Dataset RMSLE**
0.330	0.375	0.407

We settle with XG Boost model for our final submission.