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.

Dataset Description

Attributes of the dataset:

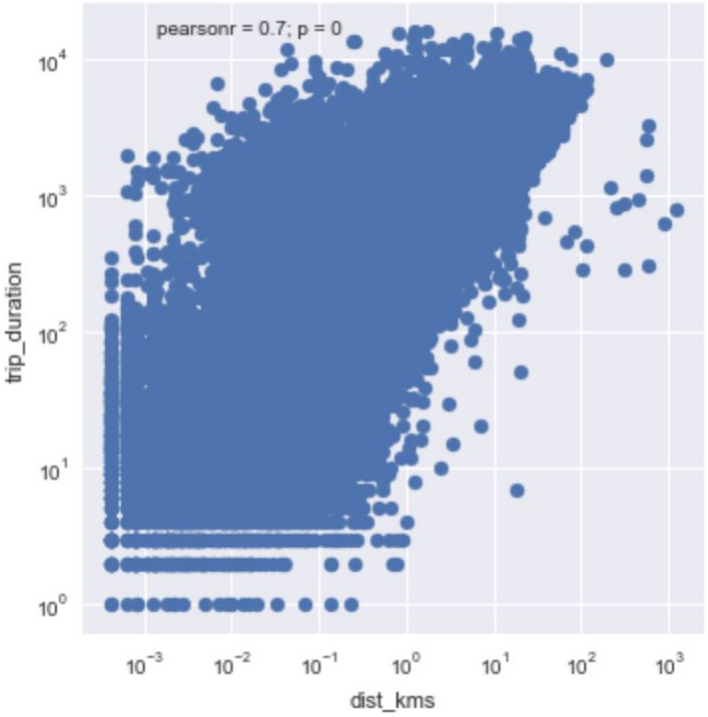
- id - a unique identifier for each trip
- vendor_id - a code indicating the provider associated with the trip record
- pickup_datetime - date and time when the meter was engaged
- dropoff_datetime - date and time when the meter was disengaged
- passenger_count - the number of passengers in the vehicle (driver entered value)
- pickup_longitude - the longitude where the meter was engaged
- pickup_latitude - the latitude where the meter was engaged
- dropoff_longitude - the longitude where the meter was disengaged
- dropoff_latitude - the latitude where the meter was disengaged
- store_and_fwd_flag - This flag indicates whether the trip record was held in vehicle memory before sending to the vendor because the vehicle did not have a connection to the server - Y=store and forward; N=not a store and forward trip
- trip_duration - duration of the trip in seconds

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. We see that the correlation b/w distance and trip duration is really high (0.7), hence distance should be an important feature for prediction.

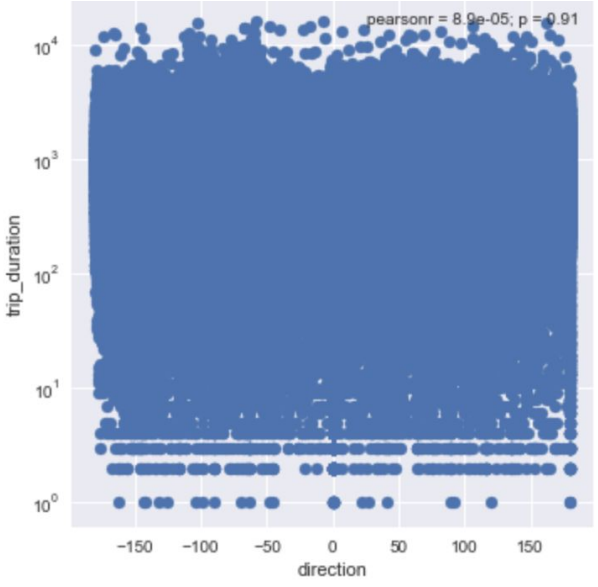
Corr Between Trip Duration and distance (kms)



Direction

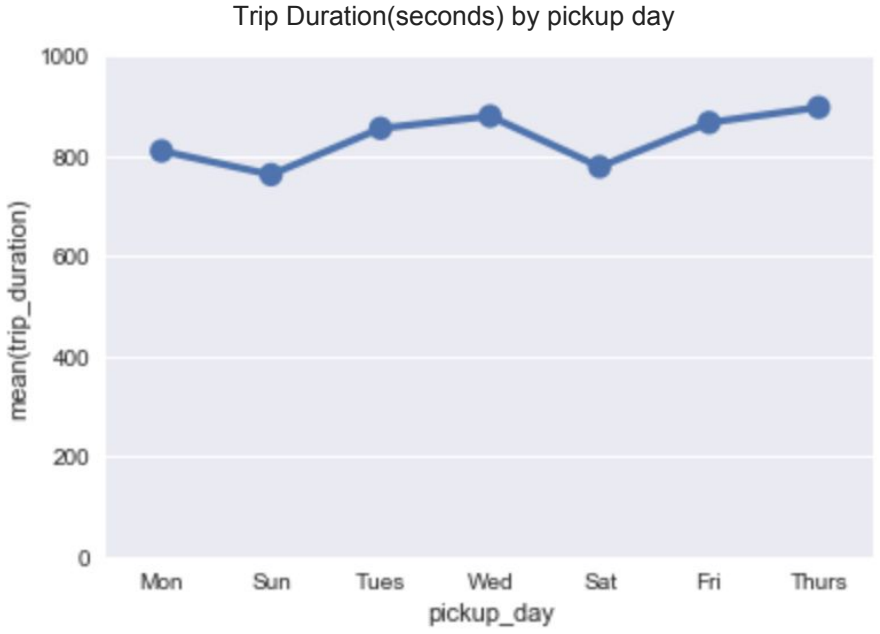
We calculate direction of trip using the bearing distance formula. The bearing of a point is the number of degrees in the angle measured in a clockwise direction from the north line to the line joining the centre of the compass with the point. http://www.mathsteacher.com.au/year7/ch08_angles/07_bear/bearing.htm. We see that there is no correlation b/w duration and direction, but we'll still keep it in the model and the reason for that shall be seen later.

Corr Between Trip Duration and Direction



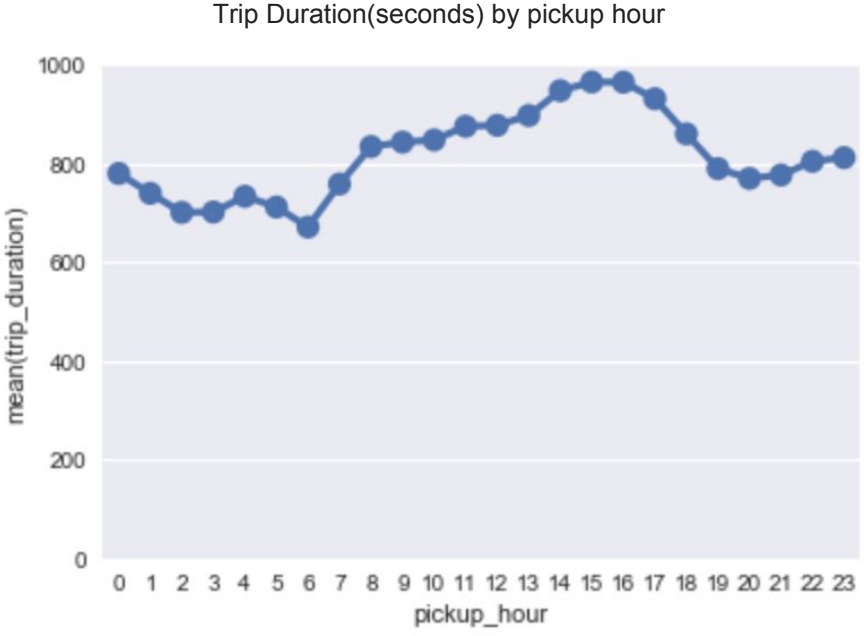
Day of the week

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.



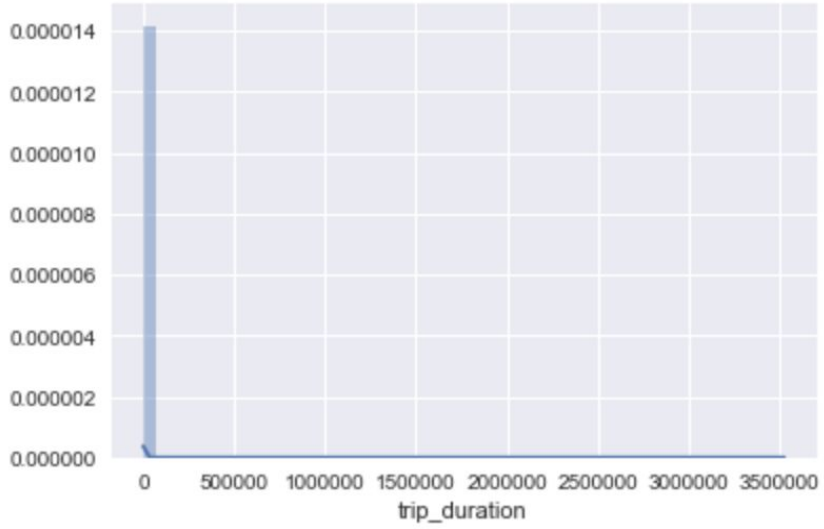
Features Selected For Model

Vendor_id, pickup_hour, pickup_day, distance, direction, passenger_count, store_and_fwd_flag_is_N, pickup_longitude, pickup_latitude, dropoff_longitude, dropoff_latitude.

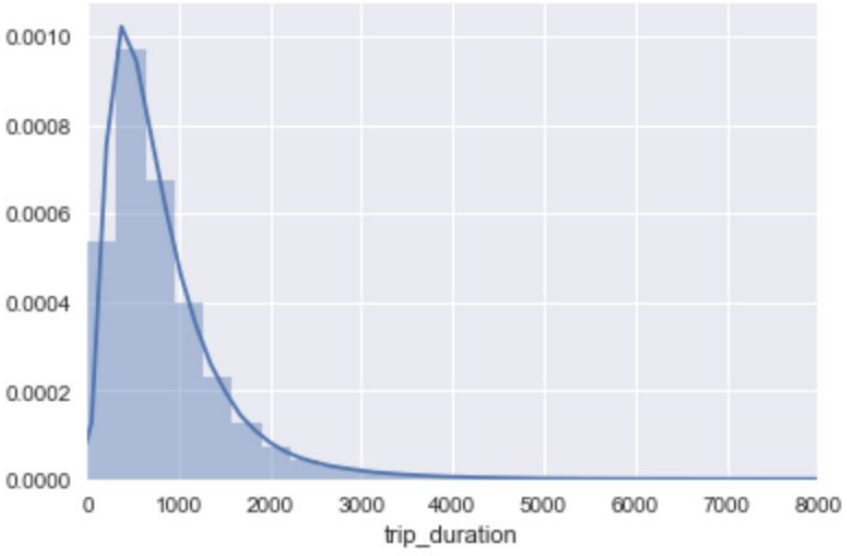
Removing Outliers

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.

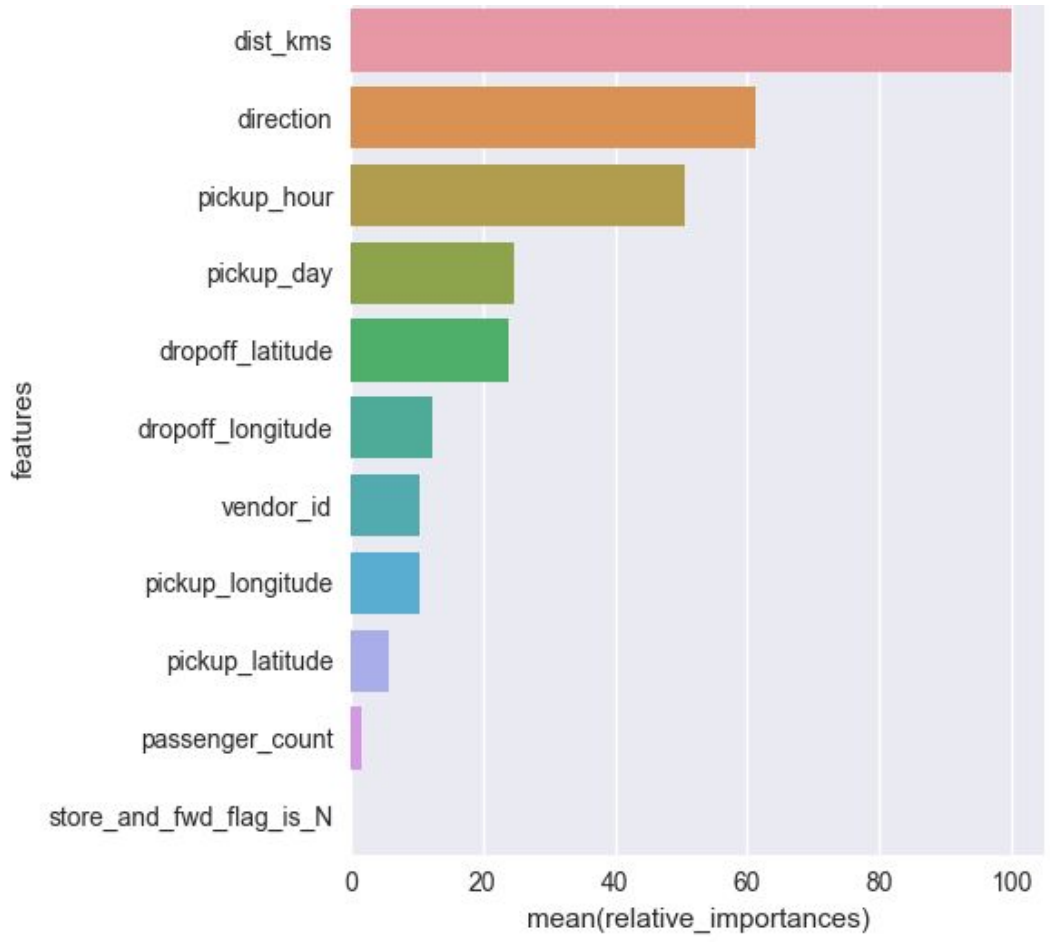
Models

Gradient Boosting Regression

We apply gradient boosting regressor using sklearn to predict trip duration. Gradient boosting regression is an ensemble supervised learning method and builds an additive model of decision trees by selecting a tree that optimizes the objective function.

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. Even though direction wasn't correlated with trip duration, it is the second most important predictor of trip duration after distance. The model is capturing some non linear relationship between trip duration and direction.

Feature importance scores



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

XGBoost

To try to improve our model performance, we use xgboost Regressor since one of the best results in the Kaggle competitions were obtained using xgboost. Xgboost uses a more regularized model formalization than regular gradient boosting to control over-fitting. A great introduction to xgb can be found here.
https://xgboost.readthedocs.io/en/latest/model.html#tree-boosting

We get slight improvements in our model with XG Boost, but since the results are better than GBR, we settle with xgboost as the final model.

XG Boost Results

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