New York Taxi Fare Prediction

Chieh-Yin (Jenny), Liao

Abstract

In this poster, I would like to share how I generate more features and how I leverage different Machine Learning techniques by working on New York Taxi Fare Prediction. I try to predict the fare amount for a taxi ride in New York City given the pickup and dropoff locations and times. I apply five models (Linear Regression, Regression tree, Random Forest, Boosted Tree, and Neural Network) to forecast.

Data Preview

The dataset contains the following fields:

- key - a unique identifier for each taxi ride
- fare_amount - the cost of each taxi ride in USD
- pickup_datetime - timestamp value when the taxi ride started
- passenger_count - the number of passengers in the vehicle
- pickup_longitude/ pickup_latitude/ dropoff_longitude/ dropoff_latitude - longitude or latitude coordinate of where the taxi ride started and ended

```
taxi_train.head()
```

	fare_amount	pickup_datetime	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count
0	5.0	2014-07-16 10:57:00+00:00	-73.996147	40.741890	-73.992203	40.739425	6.0
1	3.7	2010-01-31 10:53:00+00:00	-74.001633	40.730766	-73.997108	40.737533	1.0
2	7.7	2010-12-04 14:26:13+00:00	-73.996597	40.736568	-73.982155	40.744322	1.0
3	5.7	2010-08-19 16:33:00+00:00	-73.973831	40.763718	-73.989418	40.771522	1.0
4	12.5	2011-08-31 08:21:47+00:00	-73.917397	40.746487	-73.973755	40.763836	1.0

Feature Engineering

a. Calculate the Haversine distance

Using Haversine distance to calculate the distance between pickup and dropoff points.

$$d = 2r \arcsin\left(\sqrt{\sin^2\left(\frac{\phi_2-\phi_1}{2}\right)+\cos(\phi_1)\cos(\phi_2)\sin^2\left(\frac{\lambda_2-\lambda_1}{2}\right)}\right)$$

- φ_1, φ_2 are the latitude of point 1 and latitude of point 2,
- λ_1, λ_2 are the longitude of point 1 and longitude of point 2.

b. Extract parts of dates

Extracting several columns [pickup_year, pickup_month, pickup_day, pickup_hour, pickup_weekday] from pickup_datetime.

c. Create Base Fare

Wikipedia illustrates that as of June 2006, fares begin at $ 2.50, 3.00 after 8:00 p.m., and $3.50 during the peak weekday hours of 4:00 - 8:00 pm.

d. Add distance from popular landmarks

Calculating the difference between the popular landmarks and dropoff points by using Haversine distance.

- Times Square: (40.7580° N, 73.9855° W)
- JFK Airport: (40.6413° N, 73.7781° W)
- Statue of liberty: (40.6892° N, 74.0445° W)

```
taxi_train.head()
```

passenger_count	distance	pickup_year	pickup_month	pickup_day	pickup_hour	pickup_weekday	basic_fare	time_square_distance	jfk_distance	statue_distance
6.0	2.342486	2013	11	6	11	Sunday	2.5	0.709867	21.226512	9.564284
1.0	0.946365	2011	3	4	18	Friday	3.5	2.551654	20.391082	7.079799
6.0	0.430317	2014	7	2	10	Wednesday	2.5	2.139922	21.079769	7.110075
1.0	0.842717	2010	1	6	10	Sunday	2.5	2.475470	21.328510	6.691976
1.0	1.490212	2010	12	5	14	Saturday	2.5	1.545802	20.655997	8.068041

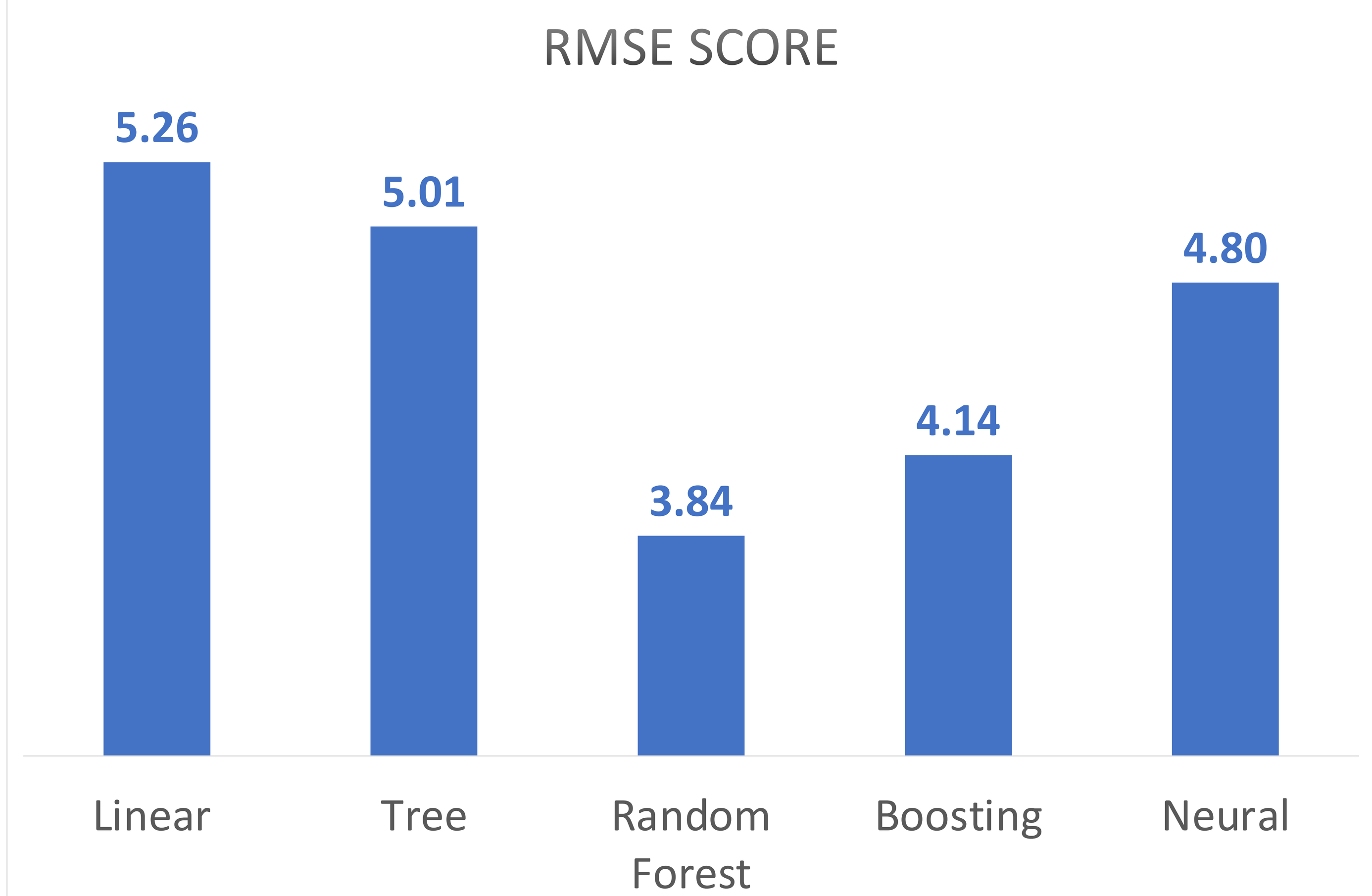
Modeling

Model	Programming	Train sample RMSE*	Test sample RMSE*
Linear Regression	LinearRegression(fit_intercept=True)	5.27	5.26
Regression Tree	DecisionTreeRegressor(max_leaf_nodes=7)	5.04	5.01
Random Forest	RandomForestRegressor(n_estimators=50, max_depth=10, max_features=0.5)	3.78	3.84
Boosting	GradientBoostingRegressor(n_estimators=100, learning_rate=0.05,random_state=1)	4.17	4.14
Neural Network	Layer (type) / Output Shape / Param # dense_1 (Dense) (None, 10) 110 dense_2 (Dense) (None, 5) 55 dense_output (Dense) (None, 1) 6	4.80	4.80

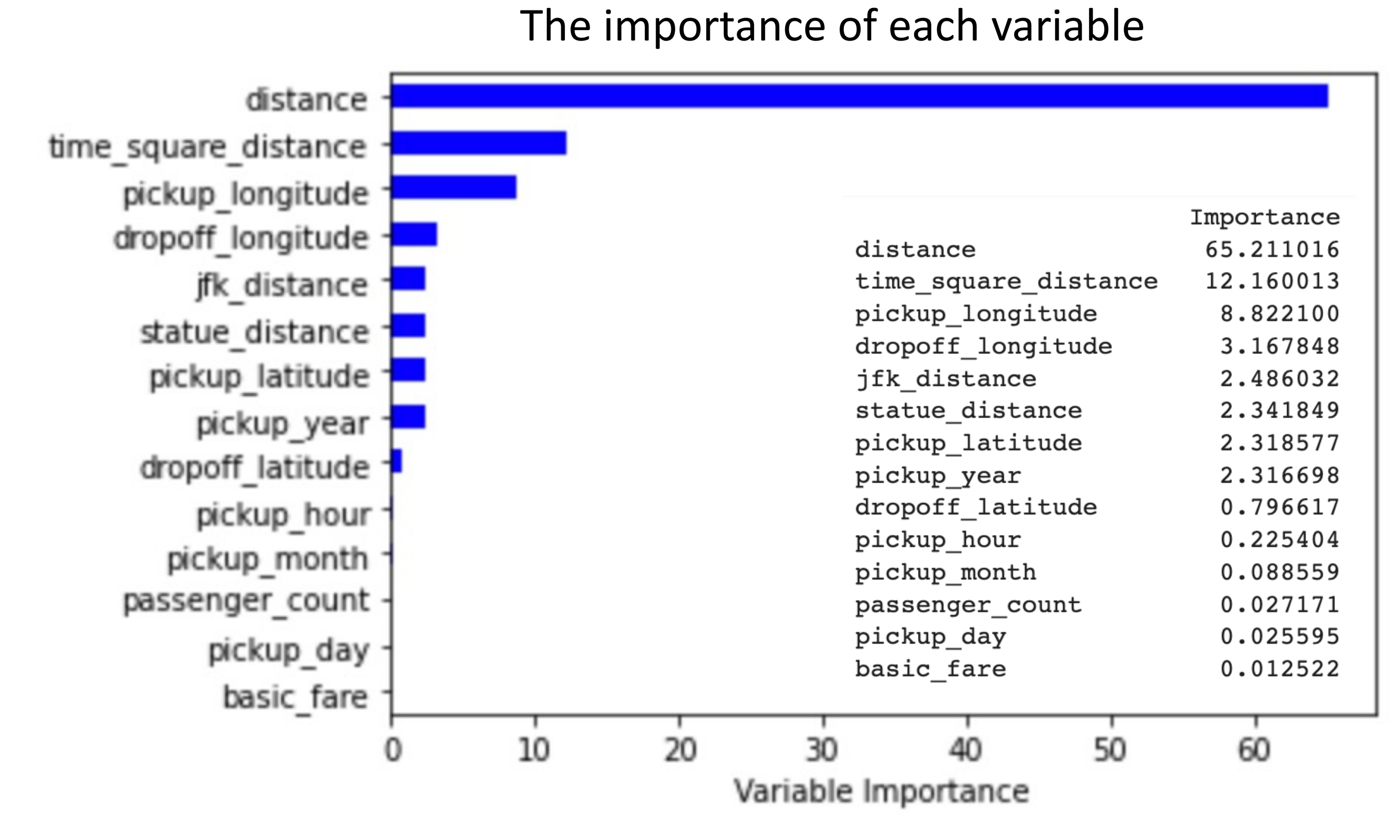
* I split 25% of random samples into test datasets and the remaining into train datasets

Results

In order to evaluate a model's performance, I apply the **Root Mean Squared Error (RMSE)** score in test sample to compare these models. RSME is the standard deviation of the residuals, lower values of RMSE indicate a better fit. As the graph shows as below, RMSE Score in Random Forest (3.84) is smaller than others, which means that the Random Forest model has a better prediction in this taxi fare case



The importance of each variable



Contact

Chieh-Yin (Jenny), Liao
Master of Business Analytics
University of Wisconsin-Madison
Wisconsin School pf Business

cliao46@wisc.edu
(608)320-8167

LinkedIn (Chieh-Yin, Liao)