RESEARCH

Fare, Distance and Duration Prediction: Deploying an Interactive Model Based on New York City Taxi Rides

Marcelo J. Rovai[1*], Heriberto Briceño-Fuenmayor[1] and Manuel A. Sacasa[1,2]

*Correspondence: mrovair@udd.cl
[1]Data Science Institute, Faculty of Engineering, Universidad del Desarrollo, Santiago, Chile
Full list of author information is available at the end of the article

Abstract

In big cities, the correct estimation of a taxicab ride is affected by many factors as time of the day, distance to be covered, driver decision, blocked roads, etc. In order to predict a distance, duration and fare of a trip, a Machine Learning model was developed, using as input only data which would be available at the beginning of a ride, as pickup and drop-off coordinates and its start time. Different datasets and several models were tested, been the Random Forest Regressor the choice for a final interactive model to be deployed. To training the model, a more complete dataset which also includes the data captured at end of a trip as drop-off coordinates, trip distance, duration and detailed rate, number of passengers, and a rate code detailing whether the standard rate or the airport rate was applied.

Keywords: Urban mobility; Geospacial location; Prediction; Taxi Trip Fare; Taxi Trip Distance; Taxi Trip Duration; Machine Learning

1 Introduction

Taxi trip is a consistent variable for urban behavior comprehension. Luckily IoT, apps and other technologies channels have been recording those trips during the last decades. For the first time, this huge amount of data gives researchers the possibility to built models, predictions and the capability to understand the behavior of a city. Taxi trip data is a well-known source of urban behavior because of:

- **Geospatial data:** Drop-on and drop-of coordinates, for describing attraction zones, gravity from interesting points (POI) or commuter type of trips
- **Date-time data:** Date and duration, describing labor day trip, holiday months, rush hours bottlenecks or regular morning-afternoon peak traffic hours
- **Fare data:** Price is a perfect correlation variable seeking for patterns between Geospatial and datetime trips features

This data lake gives researchers, product/service designers, entrepreneurs, etc., the possibility of building new TICs solutions for inhabitants searching quality life, less time lost, saving money or just taking the right decision at the right time (For example: knowing the trip fare between two spatial points in rush hour in a particular date).

Fare prediction is one of the most valued features for inhabitants in urban mobility because of its variability range, impact over personal finance and constant changing price [1]. The goal of this research is finally deploy a prototype model for fare,

duration and distance prediction used as a first step service for a mobility taxi trip info. The prototype fare prediction model must be adaptable. This means to be built not as a specific city model, but as an adjustable model for any city with the right amount of specific data to train it.

2 Information and Related Work

For a rate prediction it is important to understand the composition of the target value. Fare, for example, has a codeID for fixed rate trips or standard rate trips. This is the first rule of the final value. Fixed values are used for airport and/or special trips and Standard rate for regular time/distance value.

- **1 = Standard rate**
- **2 = JFK**
- **3 = Newark**
- **4 = Nassau or Westchester**
- **5 = Negotiated fare**
- **6 = Group ride**

Standard rate trips, the fare composition are a raw fare amount based on a time-and-distance value calculated by the meter. The initial charge is USD 2.50. Extras and surcharge including USD 0.50 for 0.2 mile or USD 0.50 per 60 seconds in slow traffic and USD 1 rush hour and overnight charges. USD 0.50 MTA tax that is automatically triggered based on the metered rate in use. USD 0.30 improvement surcharge assessed trips at the flag drop. The improvement surcharge began being levied in 2015.

Even thought fare main composition time and distance have different behavior (linear, no linear) human mobility has a 93 percent of average index of predictability [2] with no relevant variable dependence of age rate or weather. This no linear behavior well known as transit is build up by a main predictable major factor (Social functions, date, time, etc.) and less relevant secondary factors [1] and specifictly for taxi demand [3].

Despite the huge amount of data available (TLC Trip record data)[4], learning curves of Decision Tree Regression model converges using a 100.000 samples as training set. Decision Tree Regressor, Random Forest Regressor (Historical batch data), ARIMA [4] and Support Vector Regressor (Time series prediction) are popular models that have shown the property to understand the complexity of linear/non-linear prediction as in the work of [5] and [6].

Experiments, preprocessing data and selection of features will be based over Chrisophoros et Al.[7] work for fare and duration prediction [1], logical design for a deployed model will follow the concept of Ferreira et Al study exploration work [8].

3 Data

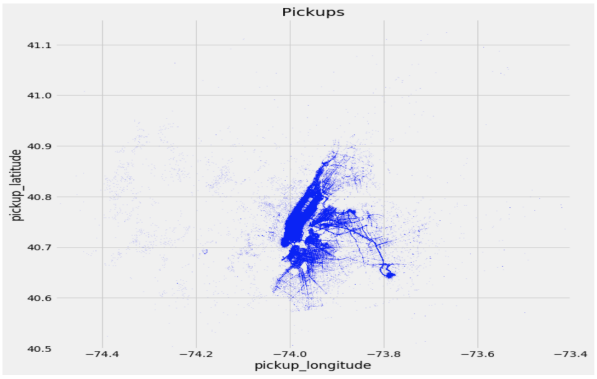
The raw dataset used for building the models was downloaded from NYC Taxi and limousine commission (TLC Trip record data). Only yellow medallion taxicabs were considered for the research. Over 50.000 taxi drivers and 13.000 licensed taxi cabs generate more than 12.000.000 trips during a single month. This research focus in 2015 dataset, where Uber and other App-based companies" influence were not significant.

As kick off data two datasets was used:

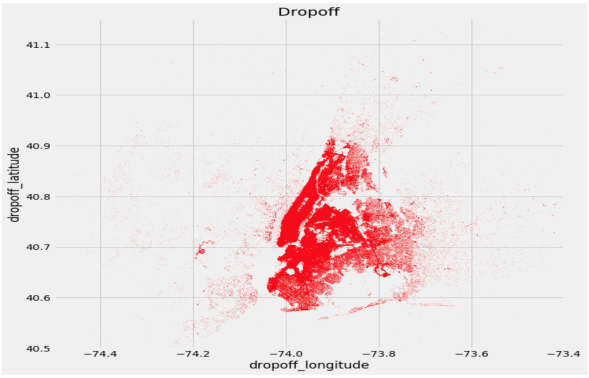
- **Dataset A (DS-A):** One raw (Without preprocessing) from 2015 (June) with 12.300K rows and 19 features (VendorID, tpep-pickup-datetime, tpep-dropoff-datetime, passenger-count, trip-distance, pickup-longitude, pickup-latitude, RateCodeID, Store-and-fwd-flag, Dropoff-longitude, Dropoff-latitude, Payment-type, Fare-amount, Extra, Mta-tax, Tip-amount, Tolls-amount, Improvement-surcharge, total-amount) described in data annexed. The full dataset will be used for EDA and a random sample of 100,000 observations for the ML Models training.
- **Dataset (DS-B):** A preprocessed dataset from all year 2015. It was built from merging 12 samples of 100K rows randomly picked from each month. Each sample was concatenated to build the final 12M rows data set of trips.

4 Exploratory Data Analysis - EDA

Exploration data, cleaning, and transformation is a dual process. Both data sets were explored and cleaning with the same process seeking for possible different results at the end of the pipeline.



(a) (Origin)



(b) (Destiny)

Figure 1: Scatter plots of Pickup and Drop-off inside geofencing.

EDA began with the definition of a research geofencing to filter origin-destiny trips (Used technique because of computational cost/time and quality initial data [6]). The Geofencing was defined as a 2 degree square boundary where the centroid

position is equal to NYC geofencing defined with coordinates 40.7127, -74.0059. With an exploration boundary defined, both datasets were verified, looking for rows with missing data (This was just a quality assurance technique because datasets were in fact, structured and clean).

The independent features cleaning and filtering began with changing DateTime format to date and time features. This will be crucial for the analysis. Then trips were filtered for passengers equal to "0", for both datasets. This will returned both datasets cleaned from trips with no passengers.

Next, fare data were filtered and cleaned. Both dataset were filtered from trips with less than USD 2.5 fare. This baseline is selected because of the initial constant cost of a taxi trip. Type fare number is another filter of both datasets. Type 1 to 4 are fixed fare cost functions, 5 or higher numbers are negotiable fares so very difficult to predict.

Finally, origin and destiny coordinates were filtered by the geofencing defined (fig. 1) in the first step. There were 4 filters implemented: pickup latitude inside the geofencing, pickup longitude outside the geofencing, drop off latitude inside the geofencing and drop off longitude outside the geofencing.

After general and specific features cleaning and filter, each specific target feature (variable to be predicted) was analyzed.

4.1 Fare Analysis

Regarding Fare, there is no difference between labor day or weekends, as well among labor days.

For airport trips, there are different cases. JFK can be a fixed fare independent of pick up origin trip, with around USD 50 fare. Different fare cases are "Newark (EWR)" and "Laguardia (LGA)". Distributions of airports and special destination trips are compared in the figure 2.

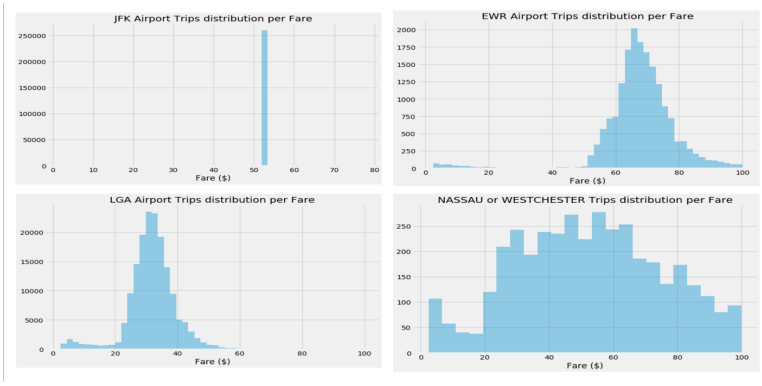


Figure 2: Airport Trip distributions per fare for different destinations

4.2 Distance Analysis

Distance inside the geofencing are 5 miles average rides. It seems that the longest trips are airport drop offs. This assumption may be verified with JFK distance trip distribution. The general distance distribution plot a peak around 18 miles while in

the JFK distance distribution, trip distance average is about 18 miles as shown in figures 3 and 4.

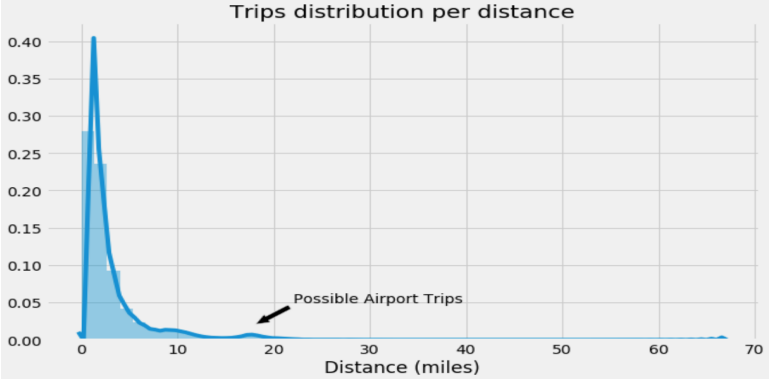


Figure 3: Trips distribution per distance. Month dataset DS-A.

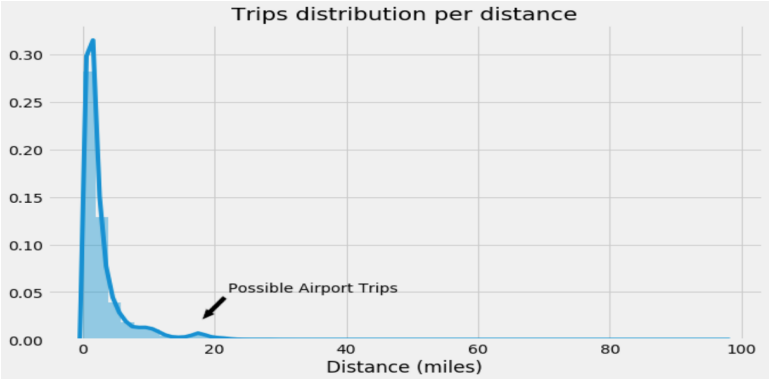


Figure 4: Trips distribution per distance. Year dataset DS-B.

Plotting fare and distance shows how fare is a result from a linear function of distance and a no linear function [7] [2] of time (Here is the problem because prediction may resolve the first duration).

Both figures 5 and 6 (Month dataset and Year dataset), from the fare/distance scatter plot, show a linear correlation between distance and fare, except airport trips fixed fare. Despite the correlation, the fare is variable because of the time trip.

This is the traffic effect over the fare. Comparing both images, monthly scatter has more variability fare because its huge amount of data for only one month describing better all cases. Annual dataset plot has less variability fare because of fewer rows per month [6].

Trips distance distribution for trips less than 7 miles are practically the same [2] for both data sets reinforcing the idea of the difference between linear fare/distance correlation and not linear fare/duration relation as shown on figures 7 and 8

4.3 Trip duration analysis.

Trip duration is the main variable for variability in fare results. The following plots show how despite the obvious correlation between time and distance city has not a

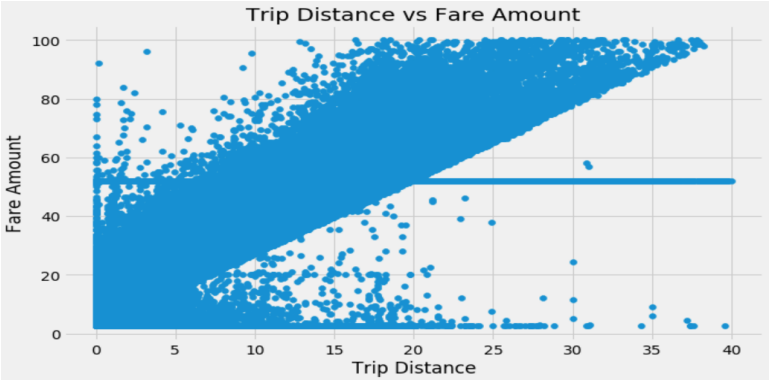


Figure 5: Scatter plot distance vs fare. Month dataset DS-A.

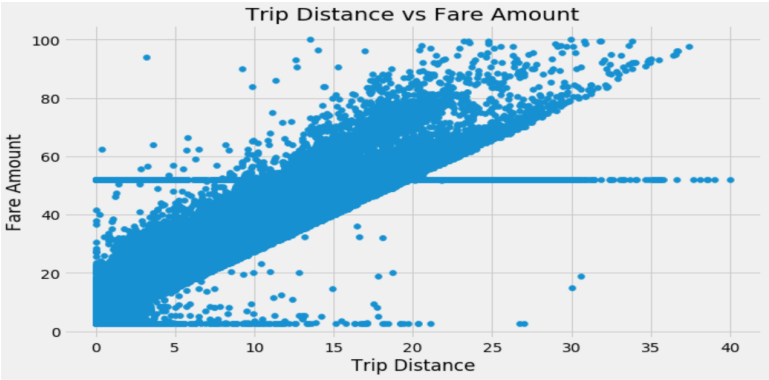


Figure 6: Scatter plot distance vs fare. Year dataset DS-B.

constant speed per hour. The traffic effect results in variability final trip duration depending in date, day type (labor or weekend), hour, pick up and route.;

Figures 9 and 10, show the variability of trip duration over the basic linear correlation between time and distance in a trip. The diagonal imaginary line in the bottom part of the images shows the linear relation, over the line is the space of duration variability for each distance trip. The month dataset has huge variability compared to the annual dataset related to the number of cases randomly selected for DS-B [6].

Hour duration distribution shows how time difference can be increased at least 30 percent. Distributions haven't relevant [2] difference (figures 11 and 12).

Finally, the target analysis concludes, sampling Manhattan trips to understand specifically the island behavior (figure 13).

For the final EDA analysis, two additional binary features was added. If the pick up coordinates were inside the island the feature input was set to "1", if were outside, to "0". The same process imputes the binary feature to drop off coordinates too.

For origin-destiny Manhattan trips, all activity concentrates from uptown (110th street) to downtown (Battery Park) (DS-A and DS-B have no relevant difference). Comparing distribution from trips exclusively inside or outside the island shows how

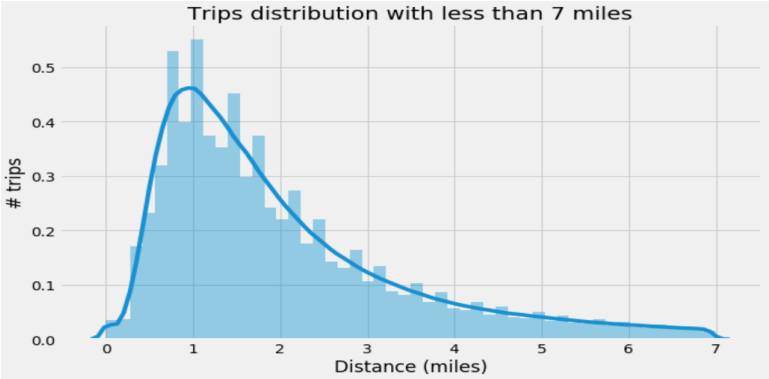


Figure 7: Under 7 miles trip distribution. Month dataset DS-A.

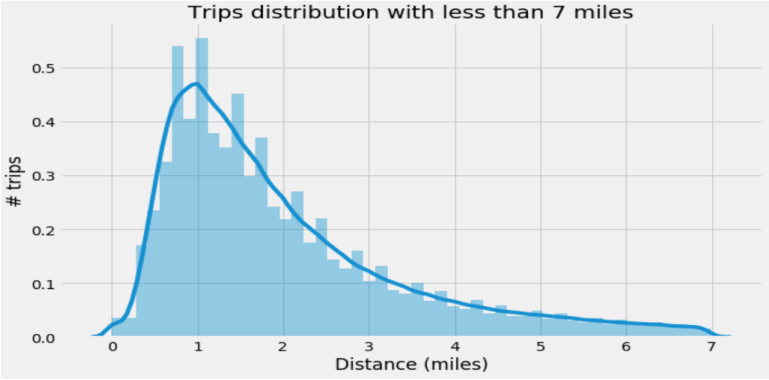


Figure 8: Under 7 miles trip distribution. Month dataset DS-B.

inside trips last labor day activities in a hour resolution plot. Outside trips show a well-defined commute activity with a morning and afternoon peak (fig. 14-15).

For the distribution of others in the second image is not clear any activity, we can make an assumption of not commute trips after work hours (this is just an assumption).

4.4 EDA Conclusion

- Taxi Fare is a linear base related with distance
- Taxi Fare is a linear base related with distance
- Taxi Fare has a completely no linear relation with travel duration that gives variability to a Fare amount at the end of a not fixed trip (Airport)
- Hours can play an important role, for example on rush hour, where trips can be longer and fare goes high, night trips, morning and mid-afternoon (airport trips)

4 ML Models and Methodology

As discussed, fare, distance, and trip duration estimation can be affected for several factors, being the more obvious a) start location, b) end location and c) start time of the trip. This is the start variables of every real trip so they must be enough

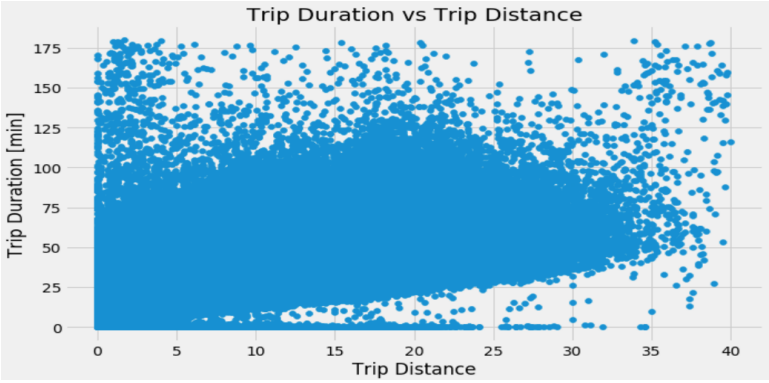


Figure 9: Scatter plot of trip duration vs trip distance. Month dataset DS-A.

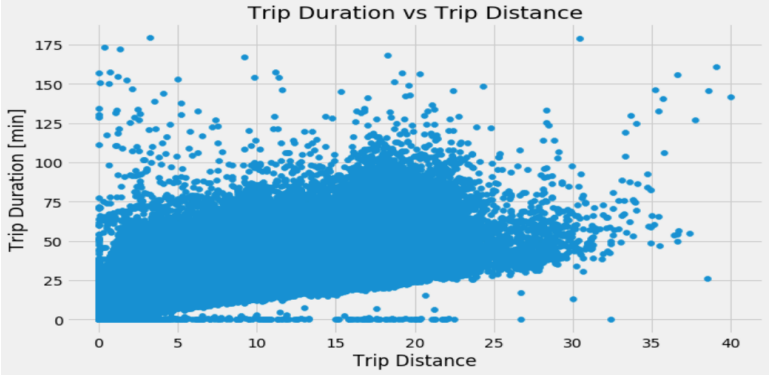


Figure 10: Scatter plot of trip duration vs trip distance. Year dataset DS-B.

to predict. The target features or predicted variables will be fare, distance, and duration. Spite the fact that our dataset has real distance, for example, we can not use it on fare or duration prediction, once the real distance will be a result of the elected route by the driver and known only at end of the trip. So, we should calculate a reference distance feature and for that, a Haversine distance formula will be used. The new feature should be imputed to both data sets.

For each target, was run a 5 models methodology:

- **Baseline.** Defining the worst prediction for each of the target features (Example: average fare).
- **Decision tree regressor.** With a test size of 0.3 from the DS-A/DS-B and depth = 11 (0.7 Training set - 0.3 Test set [4]).
- **Random forest regressor** with default hyper parameters.
- **Random Forest Regressor** with less features. Same model with default hyperparameters but creating a new test dataset with less features.
- **Random Forest Regressor** with less features. Same model with default hyperparameters but creating a new test dataset with less features.; dt = test[['passenger-count', 'pickup-longitude', 'pickup-latitude', 'RateCodeID', 'dropoff-longitude', 'dropoff-latitude', 'fare-amount', 'weekday', 'hour', 'in-man', 'out-man', 'haver-dist']].

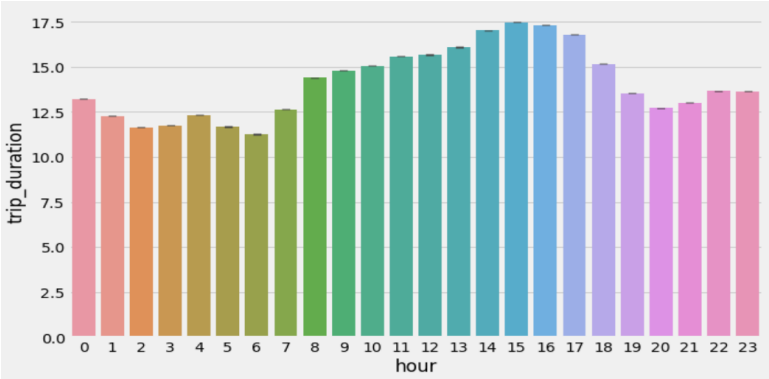


Figure 11: Trip duration per hour on an aggregate average day DS-A.

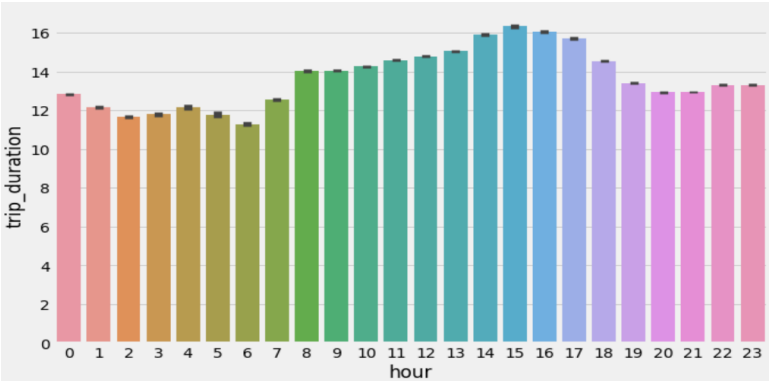


Figure 12: Trip duration per hour on an aggregate average day DS-B.

- **Random Forest Regressor with hyperparameter tuning** Creating a random grid searching for the best hyperparameters comparing the results of possible models.

The results of the three target features for the described 5 models for both datasets (DS-A and DS-B) was compared with RMSE (Root Mean Squared Error) to seek the best model and evaluate from a benchmark.

5 Prediction results and discussion

Comparative results

Different results of prediction are shown in the table included on figure 16. For fare and distance, annual dataset returns better results than 1-month dataset. Instead, trip duration target feature prediction has worst results for the annual dataset. This reflects the traffic effect during a specific period of time. To train this no lineal behavior is better to use an specific month datasets [7][4].

- **Decision Tree Regressor:** The model improves over 67percent baseline RMSE in fare and distance prediction. Again duration trip behavior didn't follow the same pattern with only 21 percent(DS-A) and 22,9percent (DS-B).
- **Random Forest Regressor (Default Hyperparameters):** The model offers the best results in every target feature prediction. While in fare and

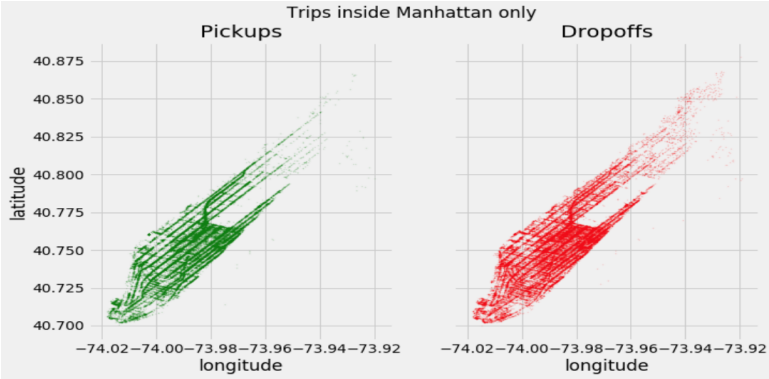


Figure 13: Scatter plot of pickups and drop off inside Manhattan.

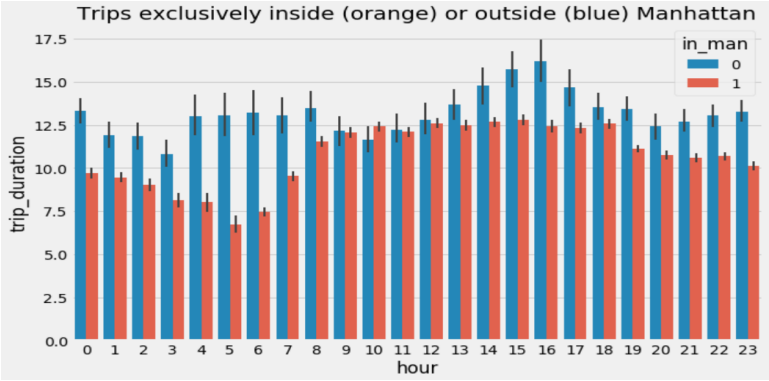


Figure 14: Inside and outside Manhattan trip distribution DS-A.

distance prediction the model improves 2percent average, with duration the model improves 35 percent of average.

- **Random Forest Regressor (Less Features):** The difference between RFR with default hyperparameters and fewer features model is not relevant, except for the time of processing, that is lower with fewer features.
- **Random Forest Regressor (Hyperparameters Tuning):** Searching for the best Hyperparameters returns the best results from the model. In Fare prediction over 70percent of improvement from baseline RMSE, trip duration over 60percent and distance prediction over 75percent of improvement over baseline error.
- **Random Forest Regressor with less features:** Same model with default hyperparameters but creating a new test dataset with less features.
  dt = test[['passenger-count', 'pickup-longitude', 'pickup-latitude', 'RateCodeID', 'dropoff-longitude', 'dropoff-latitude', 'fare-amount', 'weekday', 'hour', 'in-man', 'out-man', 'haver-dist']].
- **Random Forest Regressor with Hyperparameter Tuning:** Creating a random grid searching for the best hyperparameters comparing the results of possible models.

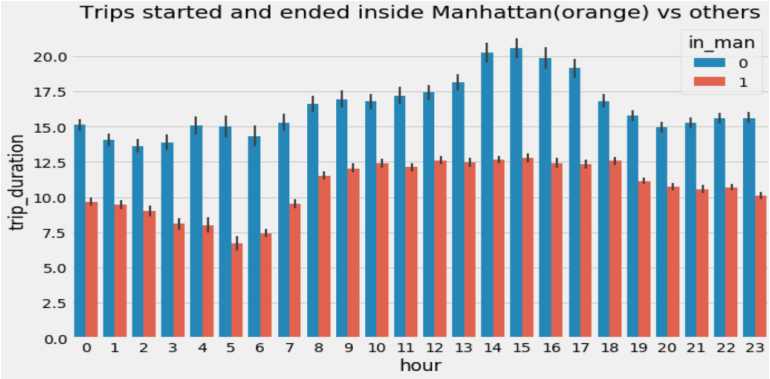


Figure 15: Inside and outside Manhattan trip distribution DS-B.

Finally, as a benchmark, the fare, duration and distance prediction were compare with the taxi fare finder website.

The input data for a new data for testing was:

- **RateCodeID = 1**
- **passenger-count = 1**
- **pickup-longitude = -73.985439**
- **pickup-latitude = 40.748839**
- **dropoff-longitude = -73.998933**
- **dropoff-latitude = 40.707738**
- **in-man = 1**
- **out-man = 0**

DS-A Month Dataset result: Random Forest Regressor with HP tuning trip has 4.48 miles and will last 22.5 minutes, with a basic cost of USD 18.46;

Using https://www.taxifarefinder.com we got the following result:

- **Fare: 21.25 dollars**
- **Dist: 5.1 miles**
- **Time: 14 min**

Both fare and trip distance results are inside the expected error (USD 2.80 dollars and 5.3 miles). Duration is out of error, but checking on Google, trip duration was between 21 and 27 minutes and the distance, from 3.7 to 5.2 miles.

DS-B Year Dataset result: Predicted Trip was "4.24 miles and will last 24.9 minutes, with a basic cost of USD 17.99."

Using https://www.taxifarefinder.com, fare is slightly over error and trip distance results is inside the expected error (2.8 dollars and 5.3 miles). Duration is out of error, but checking on Google, time is between 21 and 27 min and a distance from 3.7 to 5.2 miles.

	Fare DS_A	Fare DS_B		Fare DS_A	Fare DS_B
	RMSE	RMSE	Comparison A-B	Base line comparison	Base line comparison
Base Line	10,25	9,88	3,6%	0,0%	0,0%
Decisión Tree Regresor	3,25	3,21	1,2%	68,3%	67,5%
Random Forest Regressor	3,01	2,92	3,0%	70,6%	70,4%
Random Forest Regressor (Reducing features)	3,00	2,92	2,7%	70,7%	70,4%
Random Forest Regressor (Changing Hyper_param)	2,77	2,72	1,8%	73,0%	72,5%

	Time DS_A	Time DS_B		Time DS_A	Time DS_B
	RMSE	RMSE	Comparison A-B	Base line comparison	Base line comparison
Base Line	14,52	13,94	4,0%	0,0%	0,0%
Decisión Tree Regresor	11,47	10,75	6,3%	21,0%	22,9%
Random Forest Regressor	6,03	6,26	-3,8%	58,5%	55,1%
Random Forest Regressor (Reducing features)	5,60	5,73	-2,3%	61,4%	58,9%
Random Forest Regressor (Changing Hyper_param)	5,26	5,38	-2,3%	63,8%	61,4%

	Dist DS_A	Dist DS_B		Dist DS_A	Dist DS_B
	RMSE	RMSE	Comparison A-B	Base line comparison	Base line comparison
Base Line	3,61	3,46	4,2%	0,0%	0,0%
Decisión Tree Regresor	0,90	0,82	8,9%	75,1%	76,3%
Random Forest Regressor	0,83	0,75	9,6%	77,0%	78,3%
Random Forest Regressor (Changing Hyper_param)	0,78	0,71	9,0%	78,4%	79,5%

Figure 16: Table Comparative results between models

6 Deploying the model

The final prediction deployed tool has a 3 Random Forest Tree Regression based models predicting Fare, Duration and Distance. The tool process an input origin-destiny data name and respond to the prediction for the 3 main target variables [8]. Finally, the deployed tool has the following pipeline:

- **a.-User enter origin and destiny names.**;
- **b.-The main function runs a similar features frame than the experiments with the following variables:.**
- **'passenger-count': passenger-count**
- **'pickup-longitude': pickup-longitude**
- **'pickup-latitude': pickup-latitude**
- **'RateCodeID': RateCodeID,**
- **'dropoff-longitude': dropoff-longitude,**
- **'dropoff-latitude': dropoff-latitude,**
- **'dropoff-longitude': dropoff-longitude,**
- **'weekday': weekday,**
- **'in-man': in-man,**
- **'out-man': out-man,**
- **'haver-dist': haver-dist**
- **'out-man': out-man,** ;

The passenger count is fixed = 1, pickup and dropoff lat/lon is processed by "GeoPy", an API to search Open Street Map (OSM) data and the rate code as an algorithm based in fixed fares conditions of NYC Taxi and limousine commission. The final three features (In Manhattan, out Manhattan and Haversine distance formula) are precalculated in a distance and position function to impute the result in the dataframe of our user trip prediction. The haversine distance is used as a more precise distance than Euclidean distance.

Finally, the inhabitant can use the deployed model entering Addresses or Point of Interest (POI) and returning fare, distance and duration prediction for a precise date-time call. The next examples show the results of the predicted trips and a benchmark from Google Maps and Taxifarefinder site:

Example 1

- **start-loc = "Empire State Building"**
- **end-loc = "World Trade Center"**
- **Trip has 3.87 miles and will last 21.4 minutes, with a basic cost of USD 16.72**
- **[Trip Info]: Trip inside Manhattan; Rate Code: Standard**
- **Google: 20 to 24 minutes and 3.7 to 5.2 miles**
- **TaxiFareFinder: USD 21.25, 5.1 miles and 14 min**

Example 2

- **start-loc = "Empire State Building"**
- **end-loc = "JFK"**
- **Trip has 17.01 miles and will last 67.0 minutes, with a basic cost of USD 52.00**
- **[Trip Info]: Trip partially in Manhattan; Rate Code: JFK**
- **Google: 43 to 52 minutes and 15.2 to 20.4 miles**
- **TaxiFareFinder: USD 52.00 (special Fare), 17.6 miles and 42 min**

7 Future works

Despite the fact that the model was successfully deployed and the difference between baseline RMSE and prediction RMSE are relevant, there are some future test and experiment that might improve the results. The possible future works are:

- **New datasets:** Despite random forest regressor is a solid model adapting to non linear scenarios like traffic, the results of predicting time travel get worse training with the year dataset. Possibly those results confirm the not totally random nature of traffic and the monthly pattern of mobility. New historical random samples data sets of each month may improve the results.
- **City time travel at rush hour:** Time and distance are not constant in a city with bottlenecks congestion and peak demand. The weighted feature depending on demand peaks and rush hour may improve results. Variability of fare over distance is only explained by traffic effect over time (fig. 5-6) so knowing speed of the route is the main feature to predict duration and finally fare [4])[6].
- **Pick up demand:** Despite traffic speed, specific demand in small zones might improve the route and time. The study and prediction of demand in a zone might anticipate rush hour and fare changes.
- **Manhattan distance:** Test prediction a deployed model using manhattan or taxicab distance.
- **ARIMA:** Testing one of the most popular models for predicting taxi demand in time series. It might improve the results of the model used specific in duration or traffic effect over fare. ARIMA's improved model may actually perform better than the Original ARIMA prediction model[1].

- **Nested model:** Testing impute prediction of distance and duration in the fare prediction model. Training distance prediction model with a year dataset and travel time prediction model with an aggregated monthly dataset from several years.

8 Conclusion

Fare, distance and time travel (duration) are main and useful information for passengers and taxi drivers. The accuracy of predicting those target features improves between the time range history of the data, type of data and prediction model. Fare main components, time and distance, have different behavior (Linear and nonlinear) so models and data should have the property to express these distinctions. Random Forest Regressor had a good performance as a model to understand the behavior of fare but there are future experiences to better understand data preprocessing. Apparently, a month dataset is a better picture of the traffic effect pattern over travel time than a year dataset that built an average pattern more aggregate and less specific. In this case, June was a chosen month, because it is a non Holiday month, with normal traffic (not on a snow or rain season). Conversely the year dataset trained performs better to predict fare and distance because of the sample variability. A nested fare model using Random Forest Regressor imputing distance and travel time predicted (trained by a year dataset and a monthly data set) should adapt better to the almost aleatory patterns of traffic.

Despite the future and possible works the deployed model accuracy represent a good first step for inhabitant or driver users to make decisions. Although we use a huge amount of data, 100K samples is enough for training Random Forest Regressor and models in general. This is the upper baseline amount of data to think training the model with another city fare.

Competing interests
The authors declare that they have no competing interests.

Author's contributions
All authors contributed to the design of the study. All authors participated in manuscript preparation. All authors read and approved the final manuscript.

Acknowledgements
The analysis was performed using Jupyter Notebooks, jointly with the scikit-learn, pandas, seaborn, goepy and ipyleaflet. The maps on this study include data from ©OpenStreetMap contributors. We especially thank the valuable advice, considerations and share of experiences provided by professor Loreto Bravo (UDD).

Source Code
Source codes can be found at http://doi.org/10.5281/zenodo.2587932

Cite as
Rovai, Marcelo, Briceño, Heriberto, Sacasa, Manuel. (2019, March 8). Fare, Distance and Duration Prediction: Deploying an Interactive Model Based on New York City Taxi Rides (Version 1). Zenodo. http://doi.org/10.5281/zenodo.2587932

Author details
[1]Data Science Institute, Faculty of Engineering, Universidad del Desarrollo, Santiago, Chile. [2]Telefonica I+D, Santiago, Chile.

References

1. Li, X., Pan, G., Wu, Z., Qi, G., Li, S., Zhang, D., Wang, Z.: Prediction of urban human mobility using large-scale taxi traces and its applications. Frontiers of Computer Science **6(1)**, 111–121 (2012)
2. Song, C., Qu, B.N. Z., Barabási, A.L.: Limits of predictability in human mobility. Science **327(5968)**, 1018–1021 (2010)
3. Zhao, K., Khryashchev, D., Freire, J., Silva, C., Vo, H.: Predicting taxi demand at high spatial resolution: Approaching the limit of predictability. IEEE International Conference on Big Data (Big Data), 833–842 (2016)

4. Grinberg, J., Jain, A., Choksi, V.: Predicting taxi pickups in new york city. (2014). Final paper for CS221
5. Moreira-Matias, L., Gama, F.M. J., Mendes-Moreira, J., Damas, L.: redicting taxi–passenger demand using streaming data. P IEEE Transactions on Intelligent Transportation Systems **14(3)**, 1393–1402 (2013)
6. Wu, C.H., Ho, J.M., Lee, D.T.: Travel-time prediction with support vector regression. IEEE transactions on intelligent transportation systems **5(4)**, 276–281 (2004)
7. Antoniades, C., Fadavi, D., Amon Jr, A.: Fare and duration prediction: A study of new york city taxi rides. (2016). Stanford.
8. Ferreira, N., Poco, J., Vo, F.J. H. T., Silva, C.T.: Visual exploration of big spatio-temporal urban data: A study of new york city taxi trips. IEEE Transactions on Visualization and Computer Graphics **19(12)**, 2149–2158 (2013)