Trip-duration Prediction Report

NYC TAXI

Objectives

1 Understanding CRISP-DM

2 Example CRISP-DM Implementation

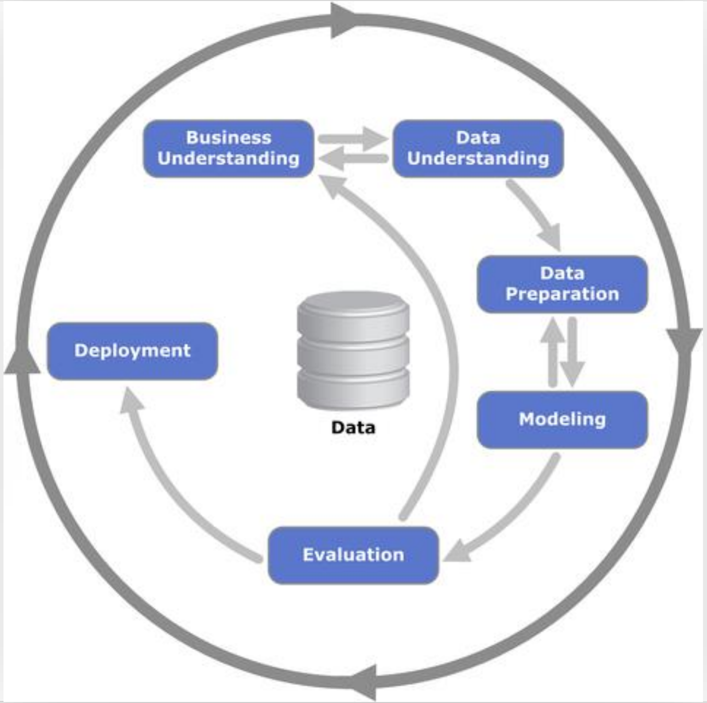
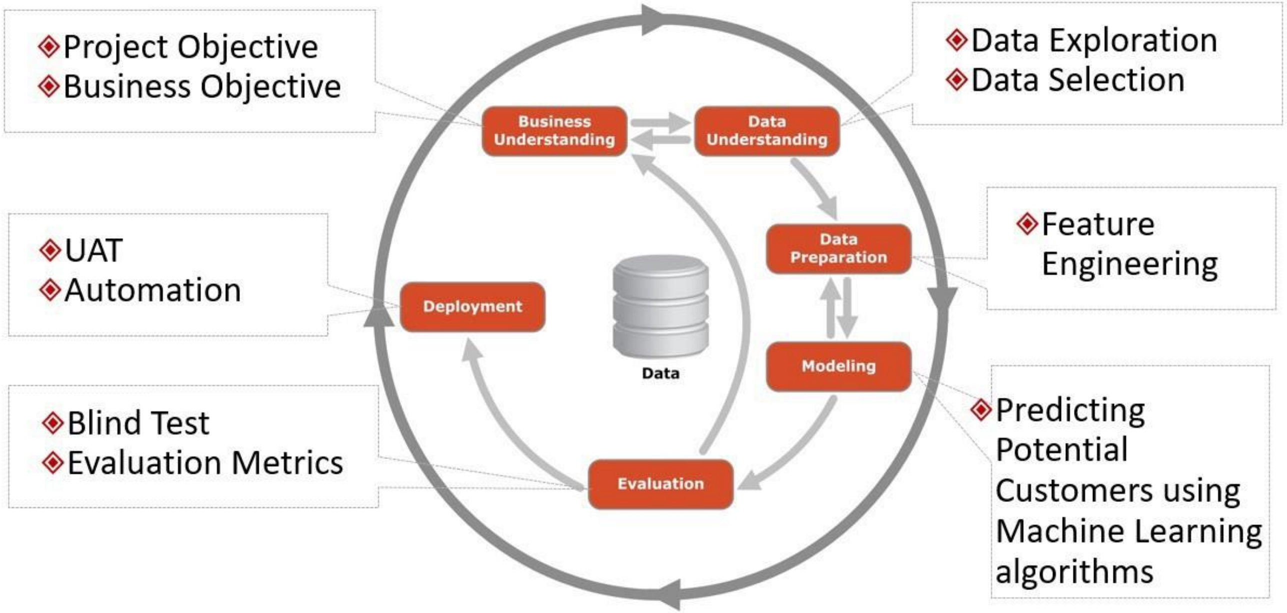


Table of Contents

CRISP-DM

What is CRISP-DM?

CRISP-DM Framework



- ◈Project Objective
- ◈Business Objective

- ◈Data Exploration
- ◈Data Selection

- ◈Feature Engineering

- ◈Predicting Potential Customers using Machine Learning algorithms

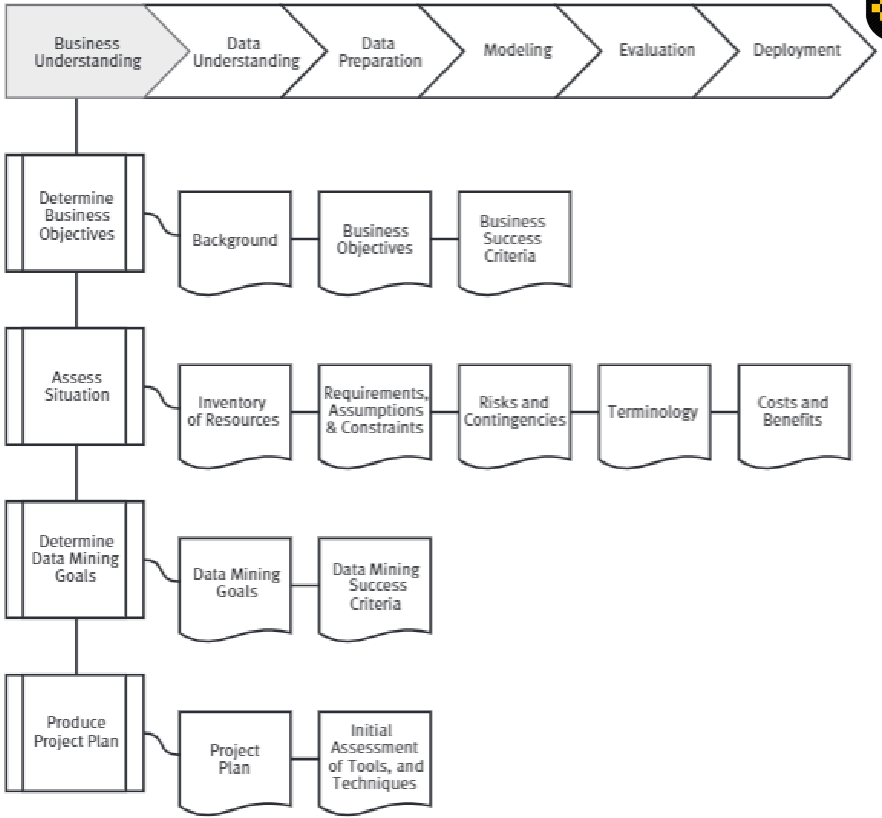
- ◈UAT
- ◈Automation

- ◈Blind Test
- ◈Evaluation Metrics

Cross-Industry Standard Process for Data mining

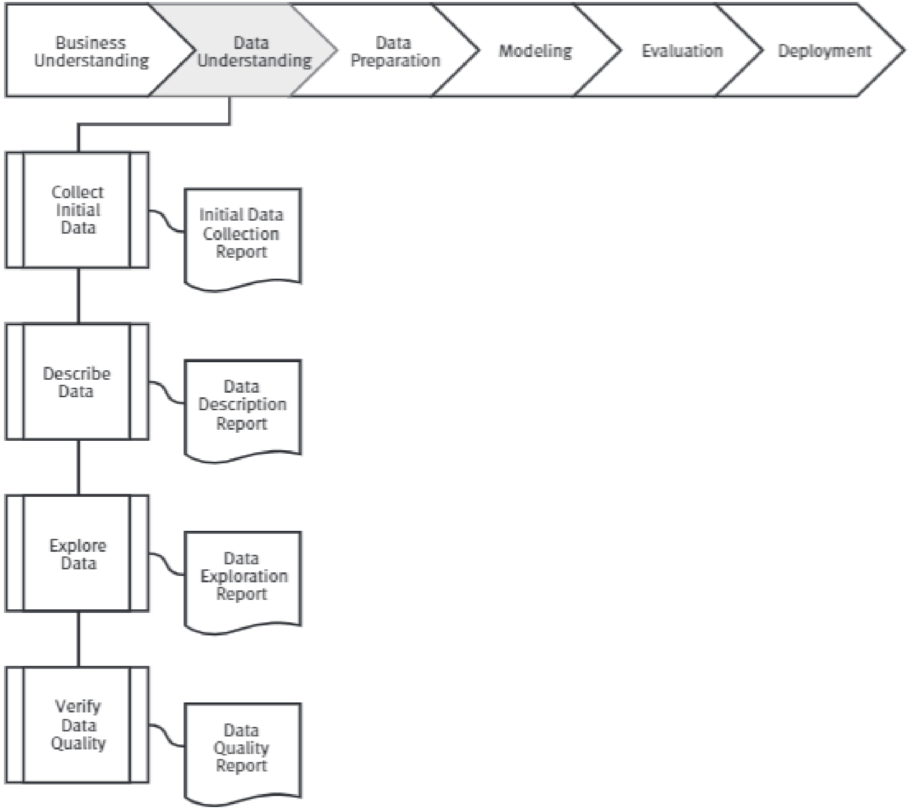
Business Understanding

1. Looking for problem base on business perspective.
2. Define decision base on problem that you find
3. Looking for information that needed to inform those decision
4. Determine type analysis can provide the information needed to inform those decisions



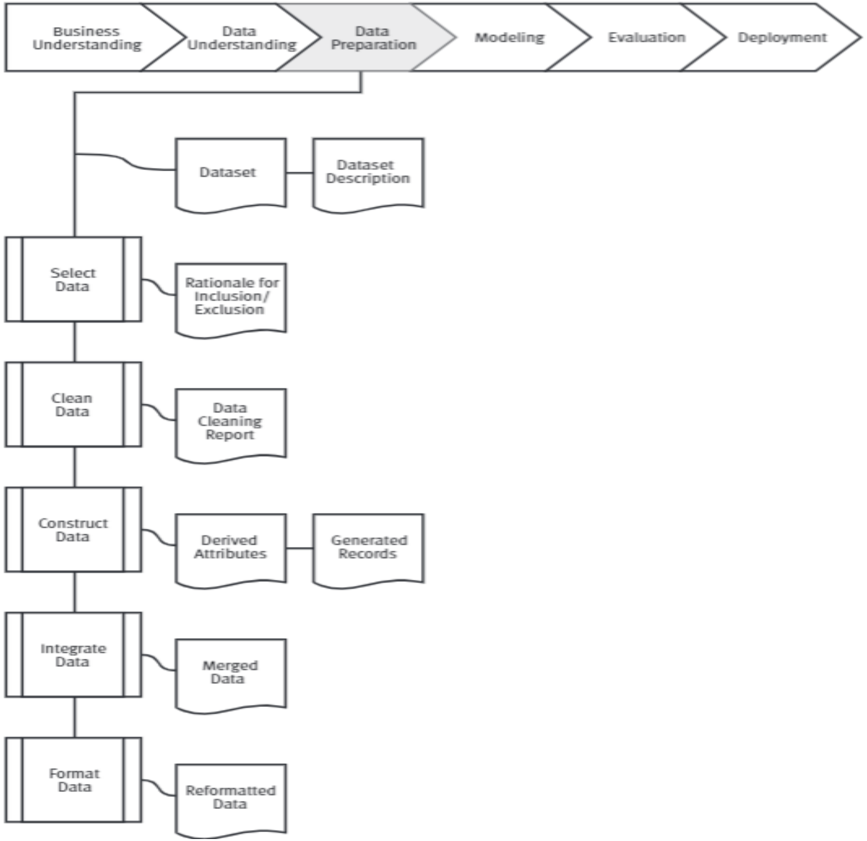
Data Understanding

1. Initial data collection and proceeds with activities in order to get familiar with data,
2. To identify data quality problems and that’s characteristics
3. Discovered first insights into the data or to detect interesting subset to form hypotheses for hidden information



Data Preparation

1. In this step we build dataset to make modeling from raw data.
2. We can iterate step until data is clean and great to make modeling.



Modelling

1. Determine what methodology to use to solve the problem
2. Determine the important factors or variables that will help solve the problem
3. Build a model to solve the problem
4. Run the model and move to the evaluation phase

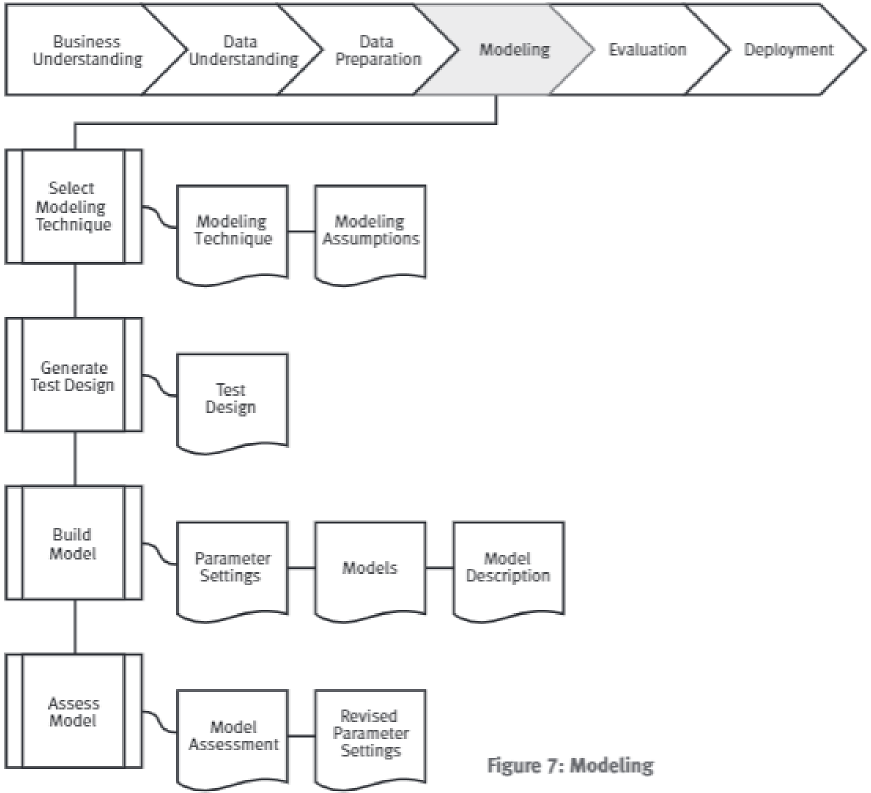
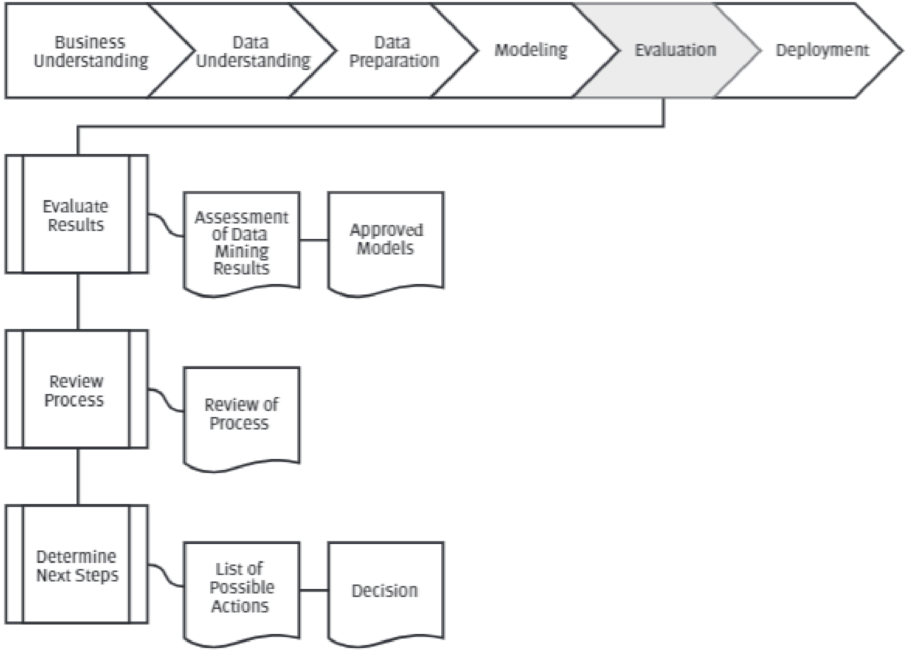


Figure 7: Modeling

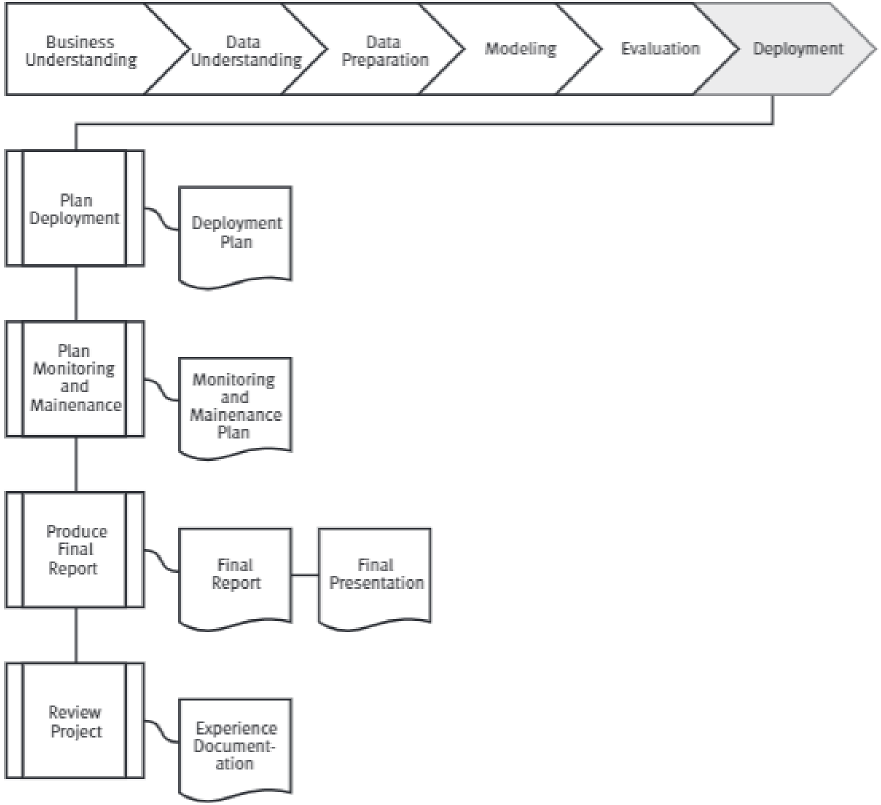
Evaluation

1. To checking the quality model objectively and how effective model to solve the problem
2. Observe the key results on the model
3. Ensure the results make sense within the content of the business problem
4. Determine whether to proceed to the next step or return to a previous phase
5. Repeat as many times necessary



Deployment/Presenting

Knowledge gained will need to be **organized** and **presented** in a way that the **customer can use** it. However, depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise.



IMPLEMENTATION CRISP-DM

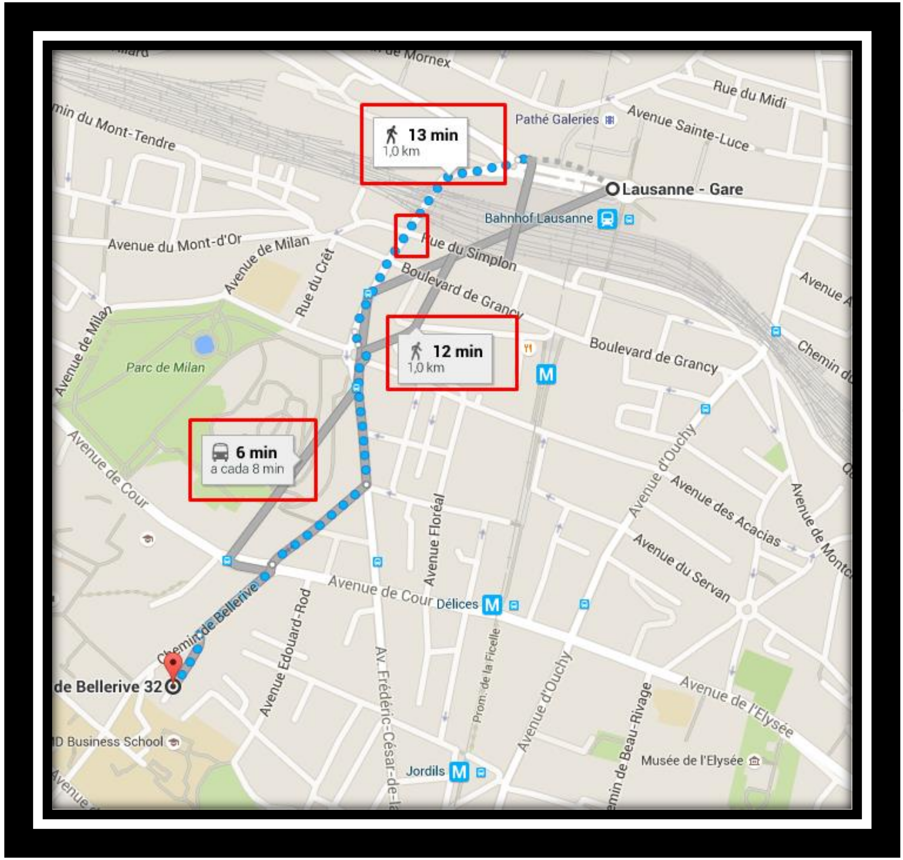
Business Understanding

Business Problems

The New York City taxicab had humble beginnings. When the “traditional” metered, gasoline-powered taxicabs began operating in October 1907, they were in tough competition with other forms of transportation throughout the city.

Nowadays, in the era of online transportation and application, **one of the important thing that passenger need is the accurate trip duration information in the application**. Providing accurate information about ETA is very challenging given road conditions and uncertain circumstances.

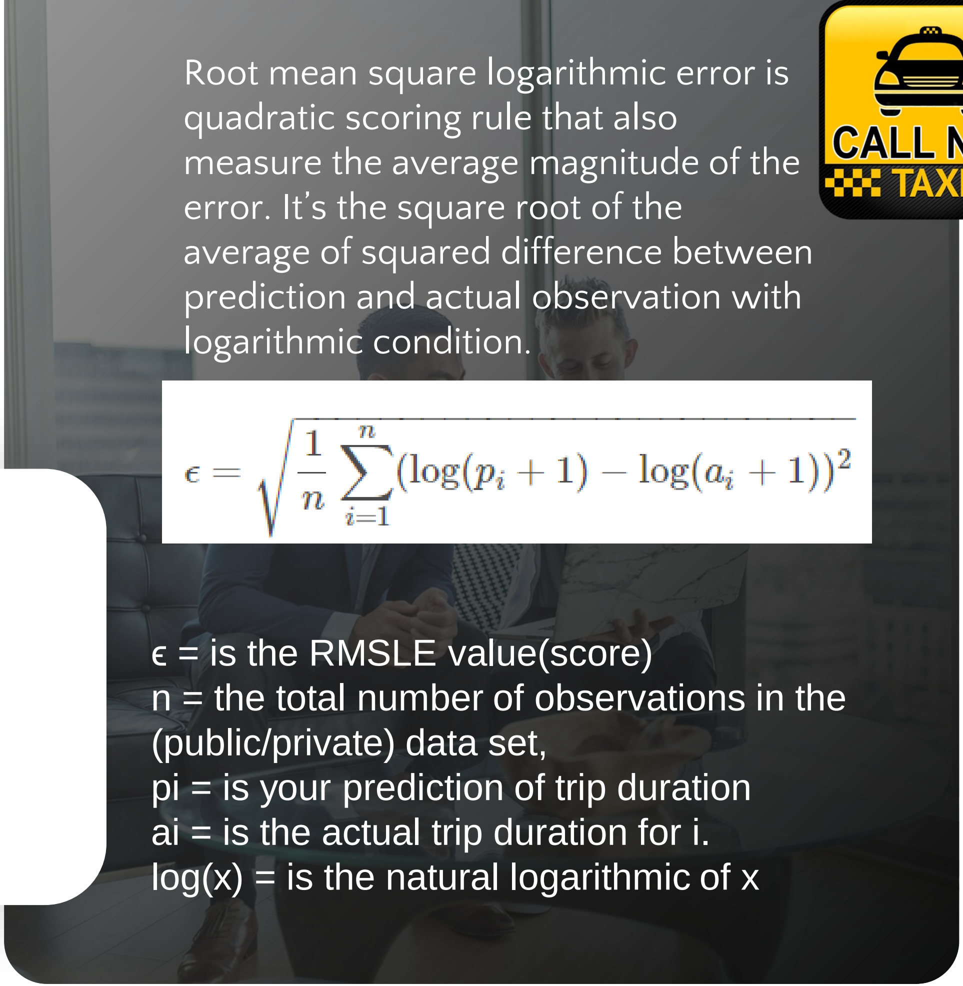
Objectives



Predict trip duration in real time

Success Criteria

Root Mean Square Logarithmic Error (RMSE)



Root mean square logarithmic error is quadratic scoring rule that also measure the average magnitude of the error. It's the square root of the average of squared difference between prediction and actual observation with logarithmic condition.

$$\epsilon = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(p_i + 1) - \log(a_i + 1))^2}$$

ϵ = is the RMSLE value(score)
n = the total number of observations in the (public/private) data set,
pi = is your prediction of trip duration
ai = is the actual trip duration for i.
log(x) = is the natural logarithmic of x

https://medium.com/human-in-a-machine-world/mae-and-rmse-which-metric-is-better-e60ac3bde13d

Data Understanding

Original Data

Data train

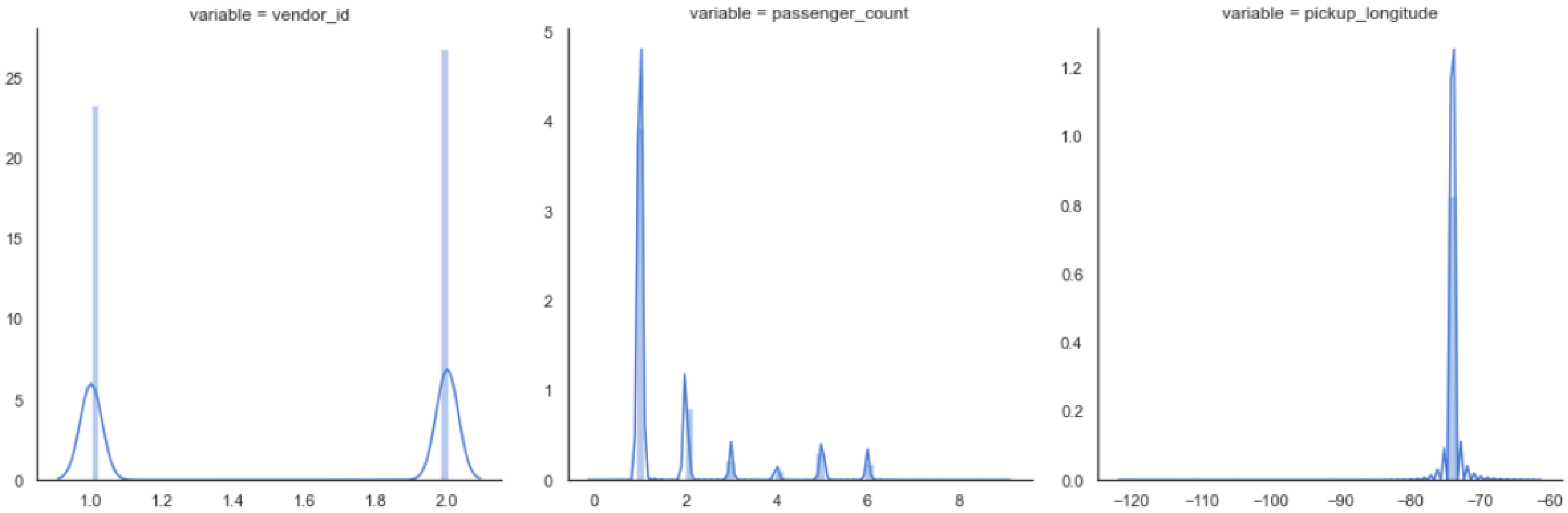
- id
- Vendor_id
- Pickup_time
- Dropoff_time
- Passenger_count
- Pickup_latitude
- Pickup_longitude
- Dropoff_longitude
- Dropoff_latitude
- Store_and_fwd_flag
- Trip_duration

Data test

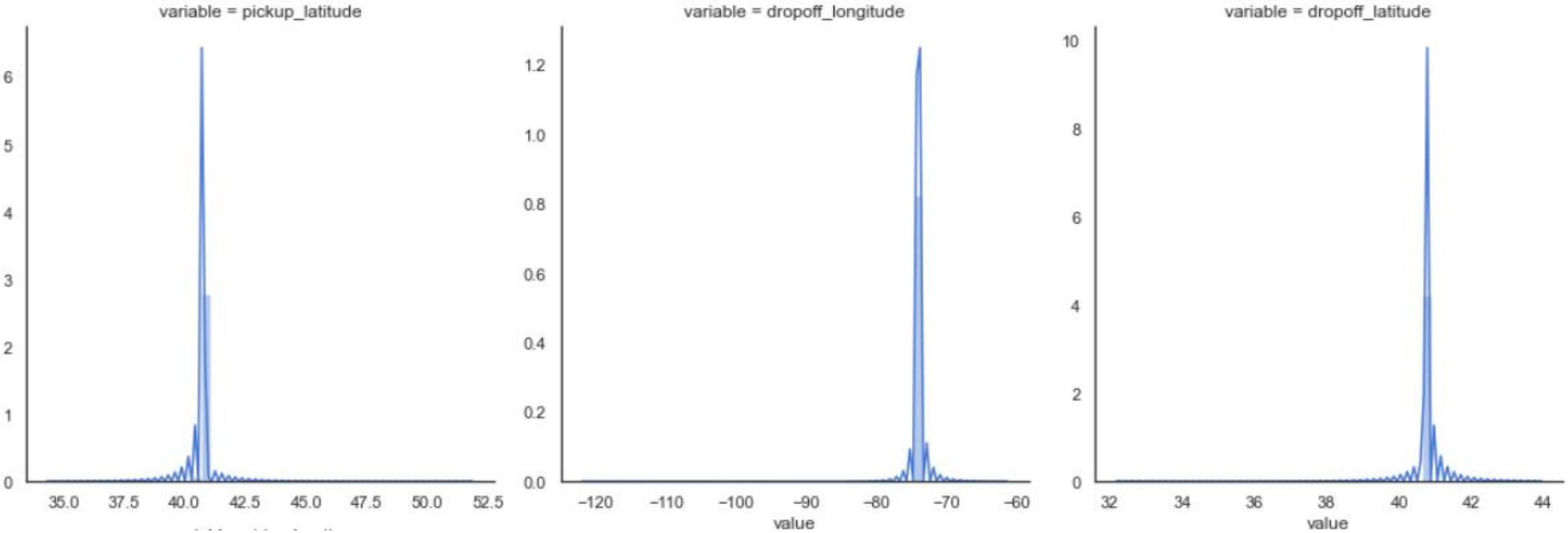
- id
- Vendor_id
- Pickup_time
- Passenger_count
- Pickup_latitude
- Pickup_longitude
- Dropoff_longitude
- Dropoff_latitude
- Store_and_fwd_flag
- Trip_duration

Distribution of Variables (train)

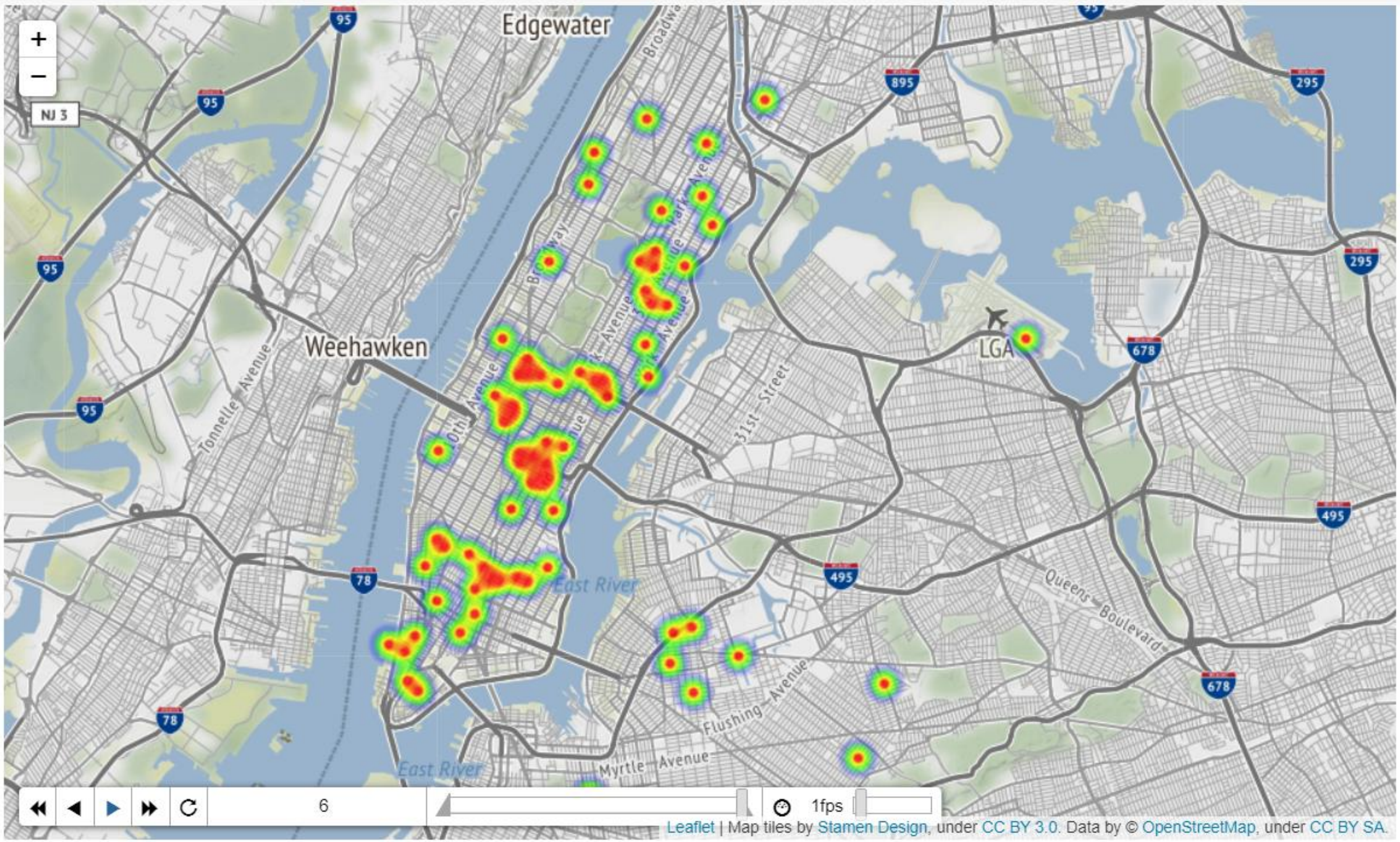
After use np.log



Distribution of Target Variables (train)

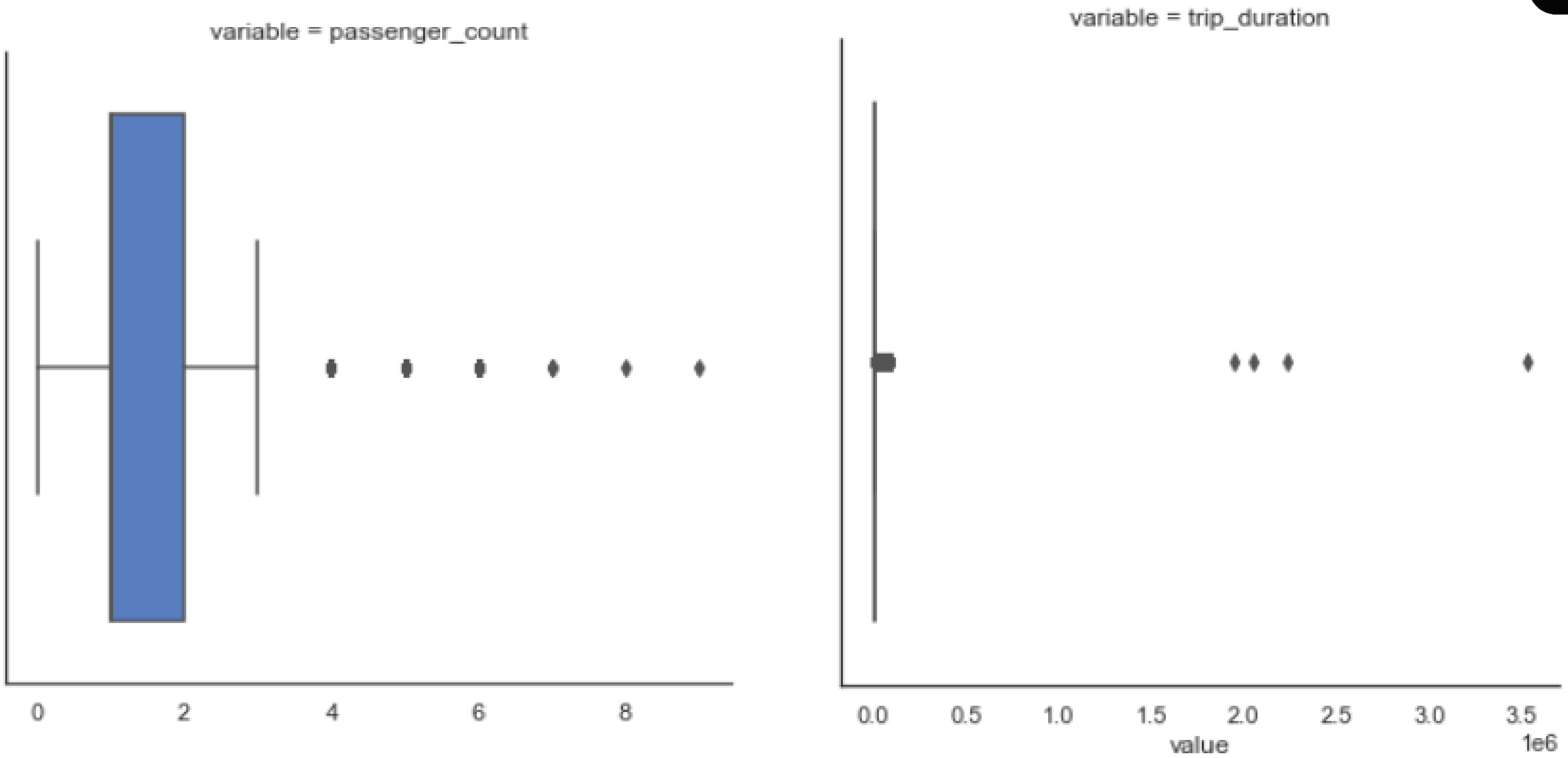


Distribution of Target Variables (train)



Leaflet | Map tiles by Stamen Design, under CC BY 3.0. Data by © OpenStreetMap, under CC BY SA.

Outliers of Target Variables (train)



Adding New Data

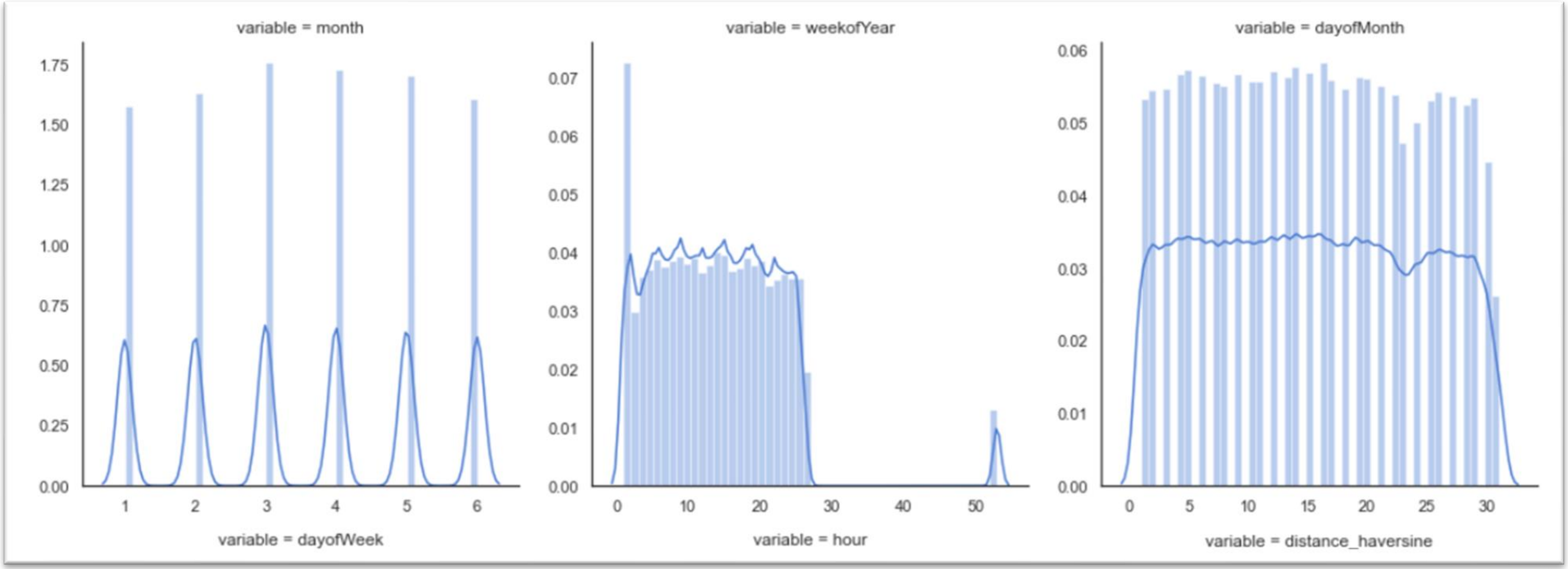
Data train

- Hour
- Month
- weekofYear
- dayofMonth
- dayofWeek
- distance_haversine
- distance_dummy_manhattan
- direction

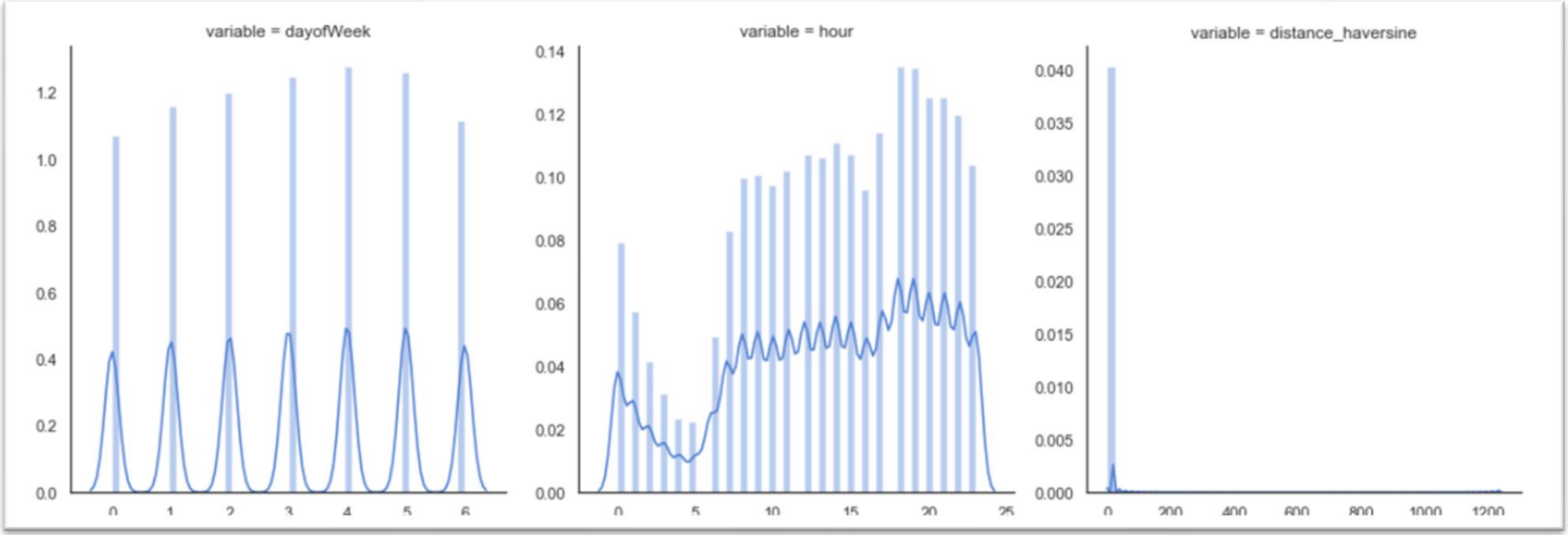
Data test

- Hour
- Month
- weekofYear
- dayofMonth
- dayofWeek
- distance_haversine
- distance_dummy_manhattan
- direction

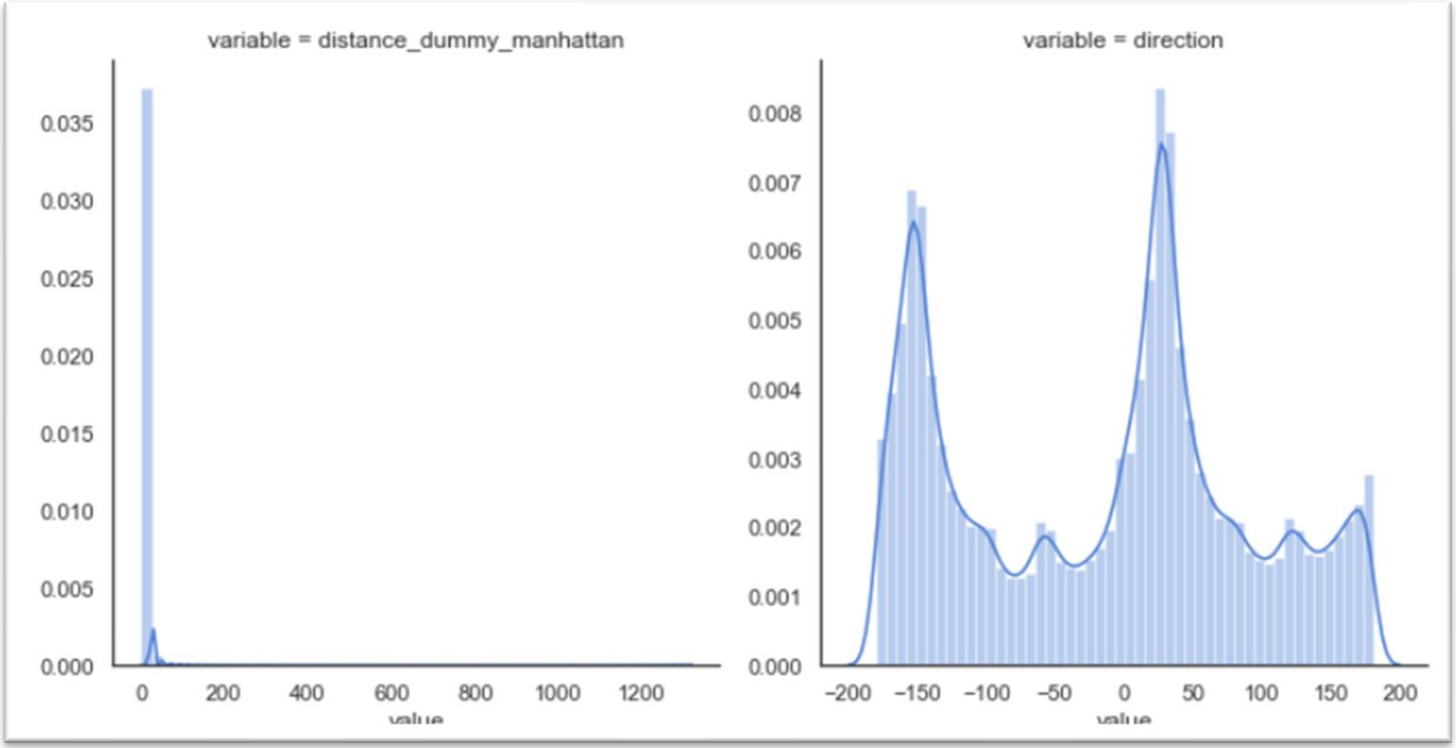
Distribution of Target Variables (train)



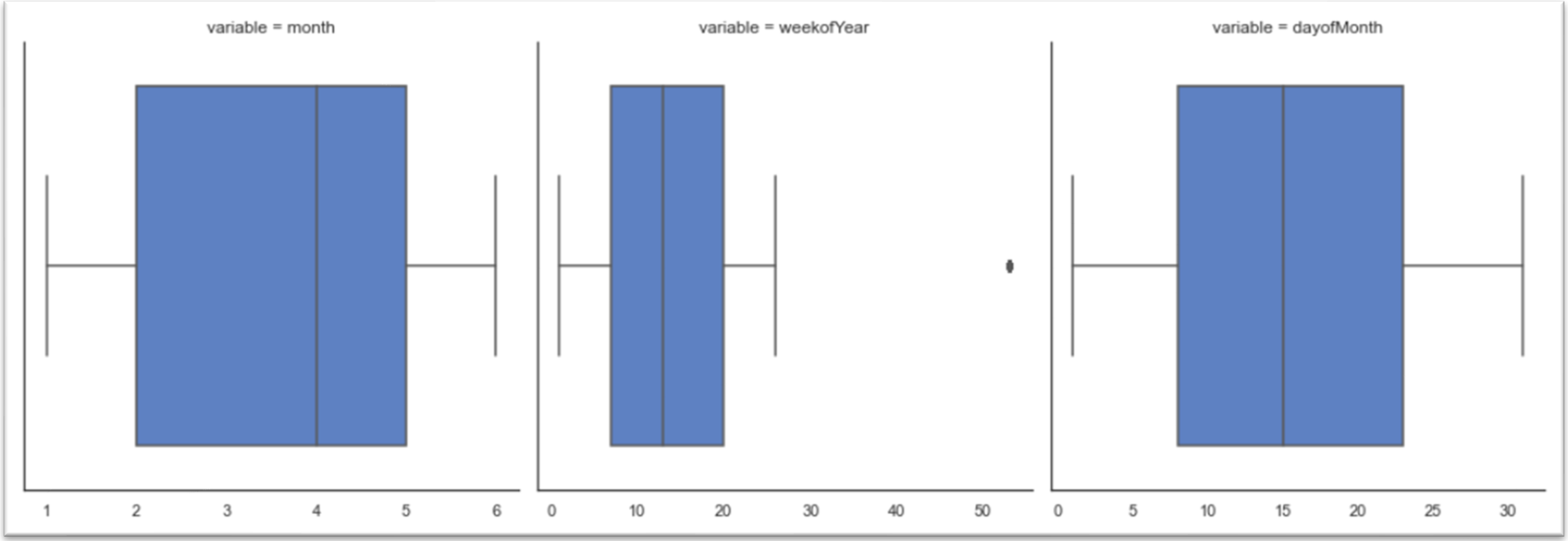
Distribution of Target Variables (train)



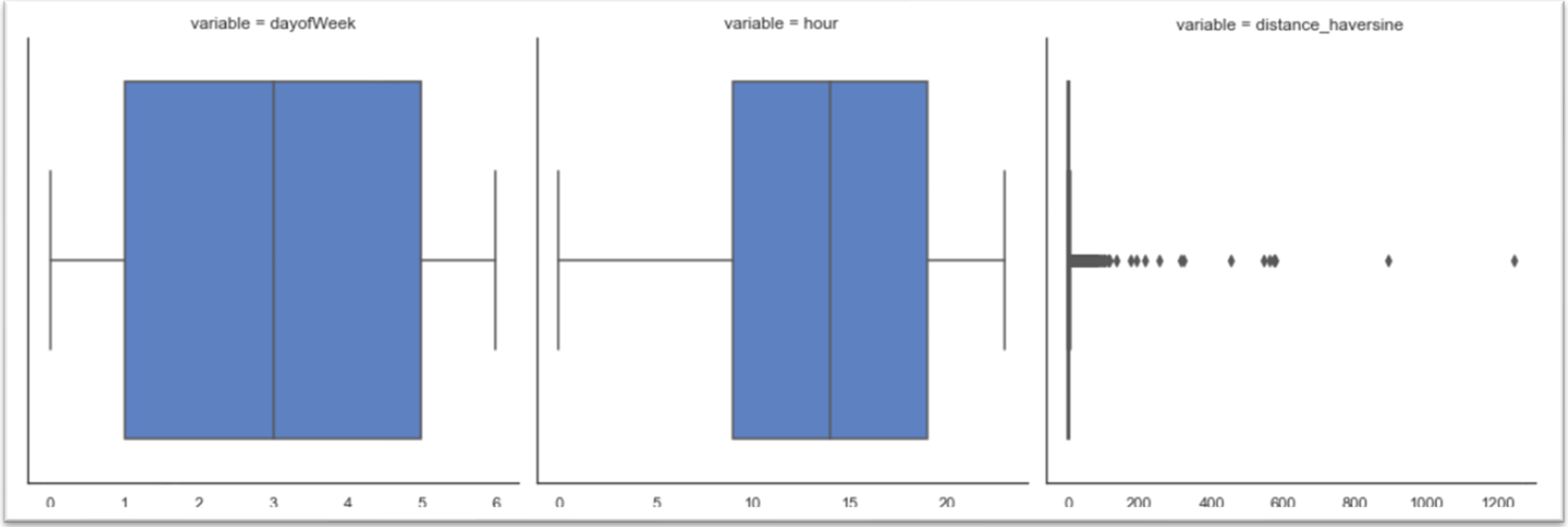
Distribution of Target Variables (train)



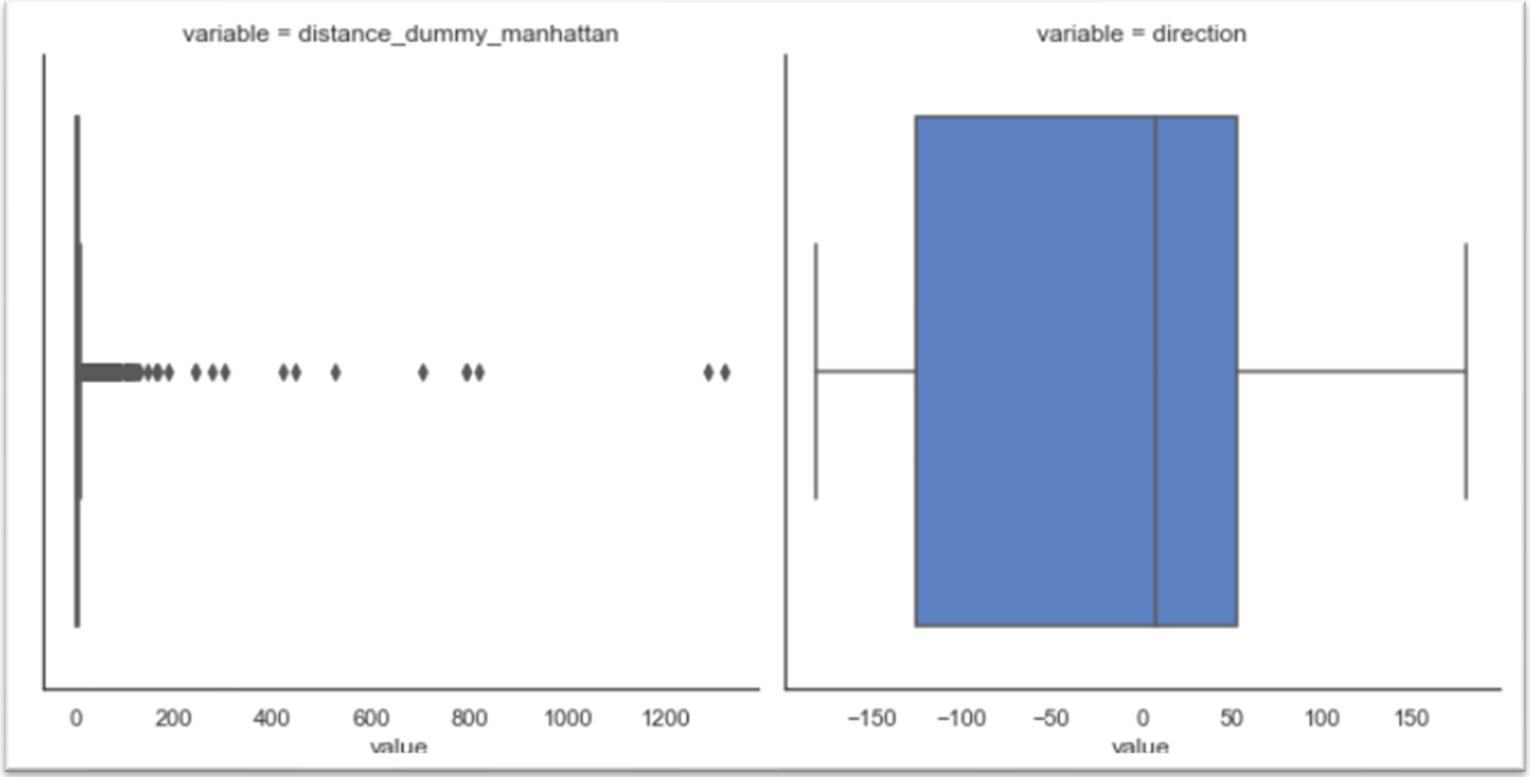
Outliers of Target Variables (train)



Outliers of Target Variables (train)

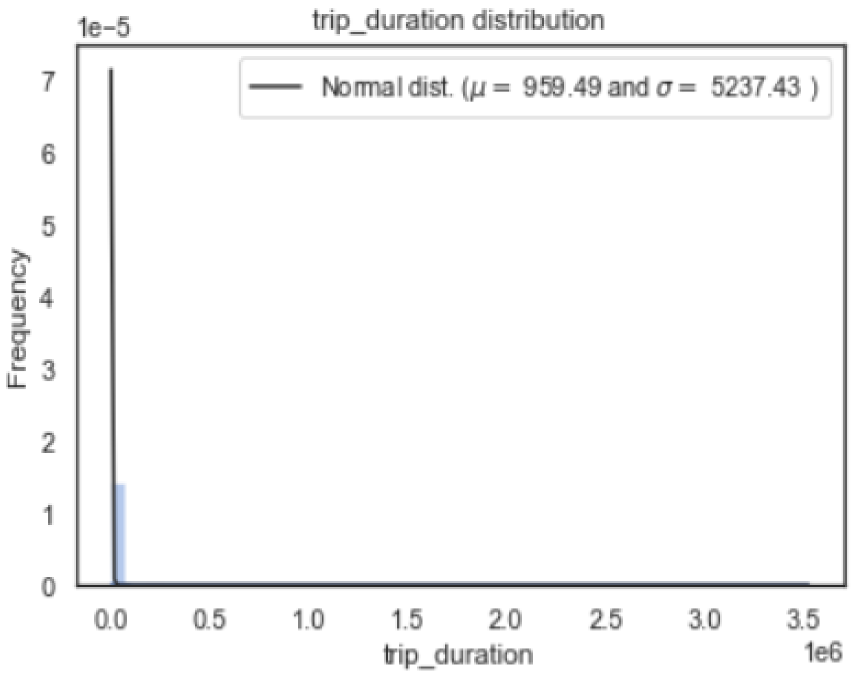


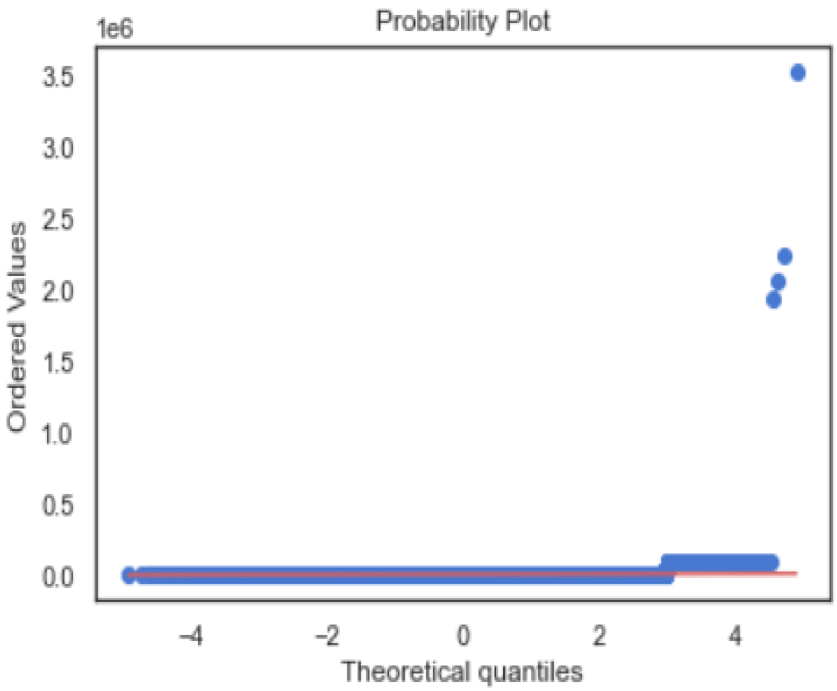
Outliers of Target Variables (train)



Data Preparation

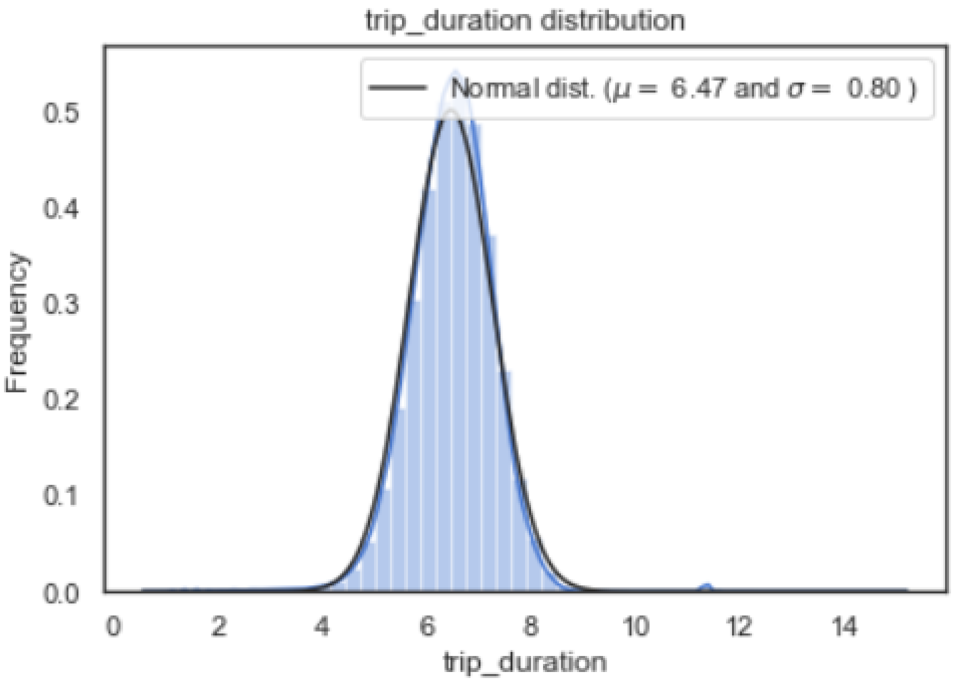
Distribution of Target Variables (train)

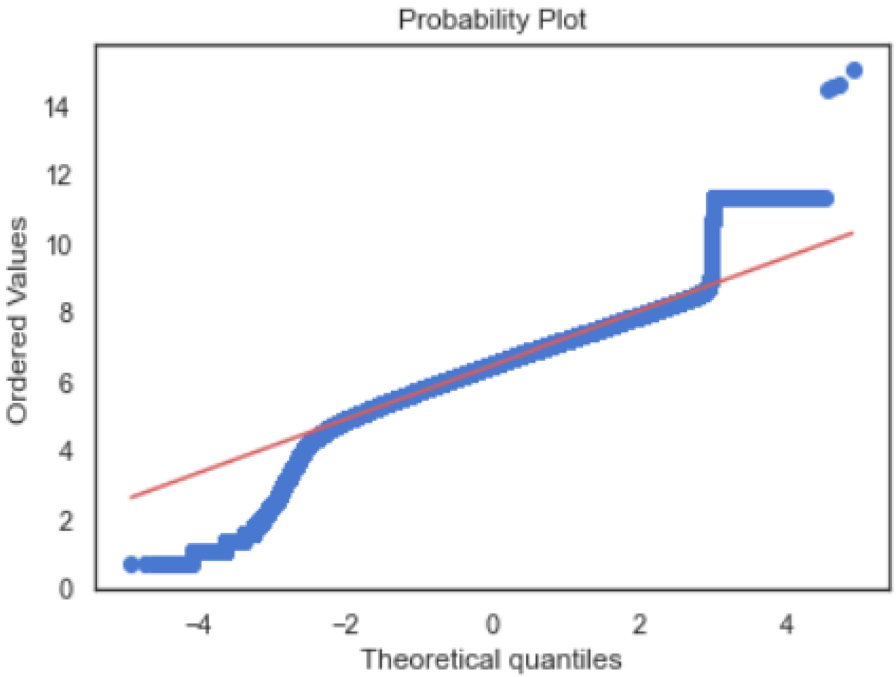




Distribution of Target Variables (train)

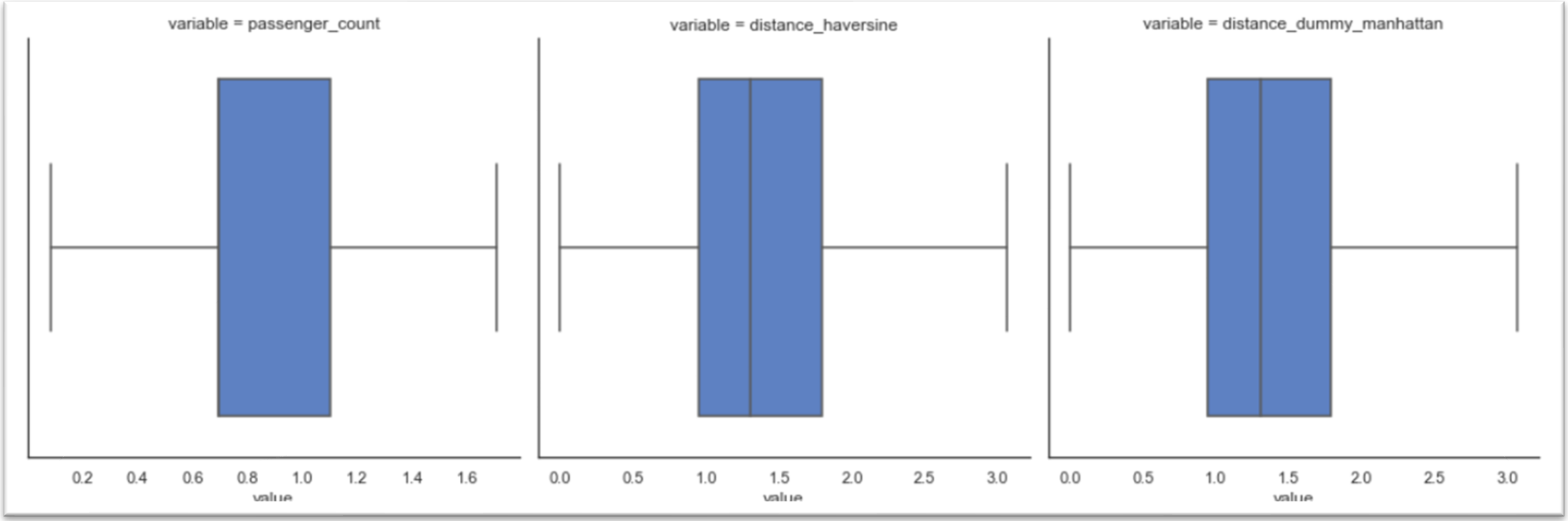
After use np.log to normalize



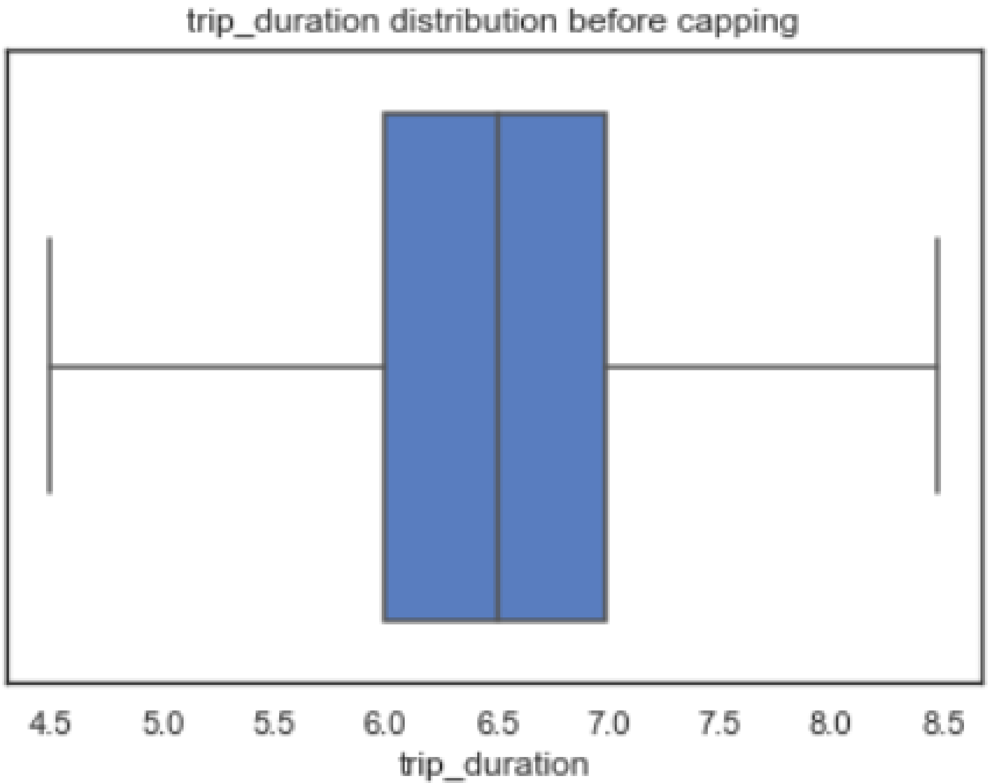


Outliers of Target Variables (train)

After capping data

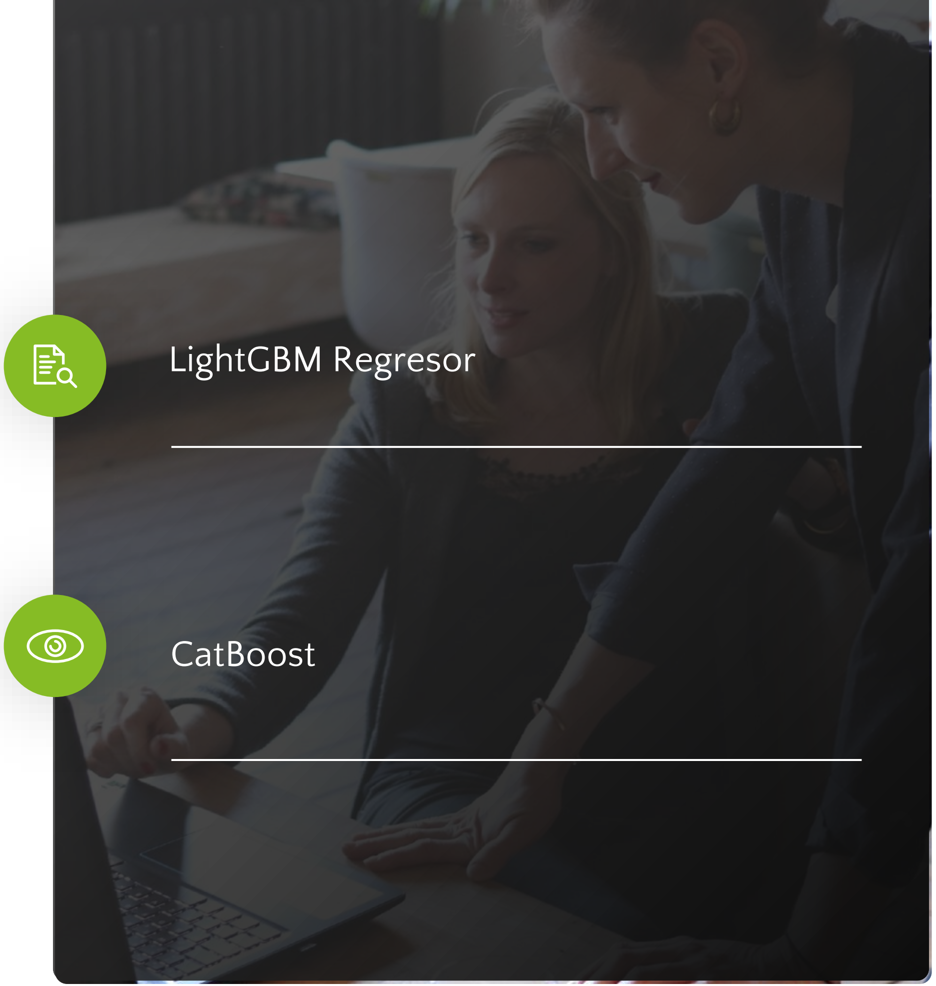


Outliers of Target Variables (train)



Modelling & Evaluation

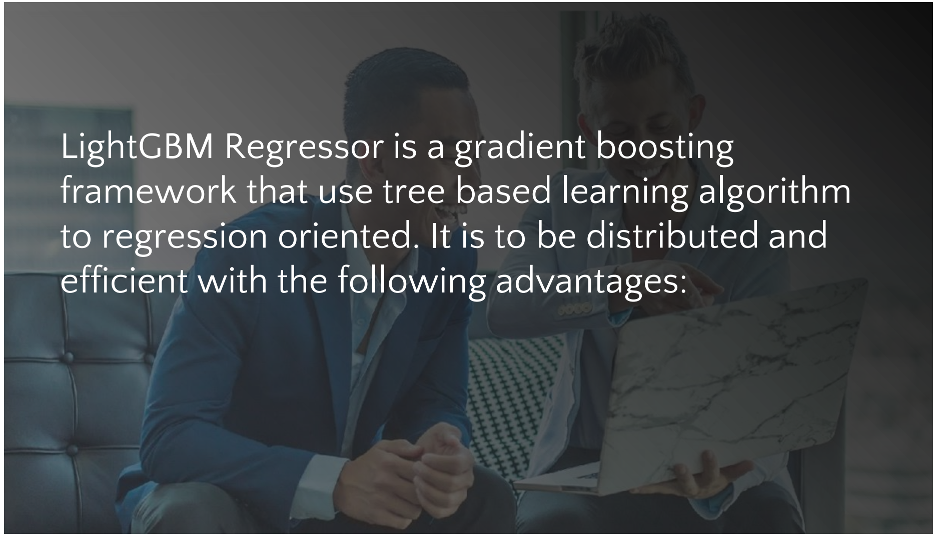
Modelling Technique



LightGBM Regresor

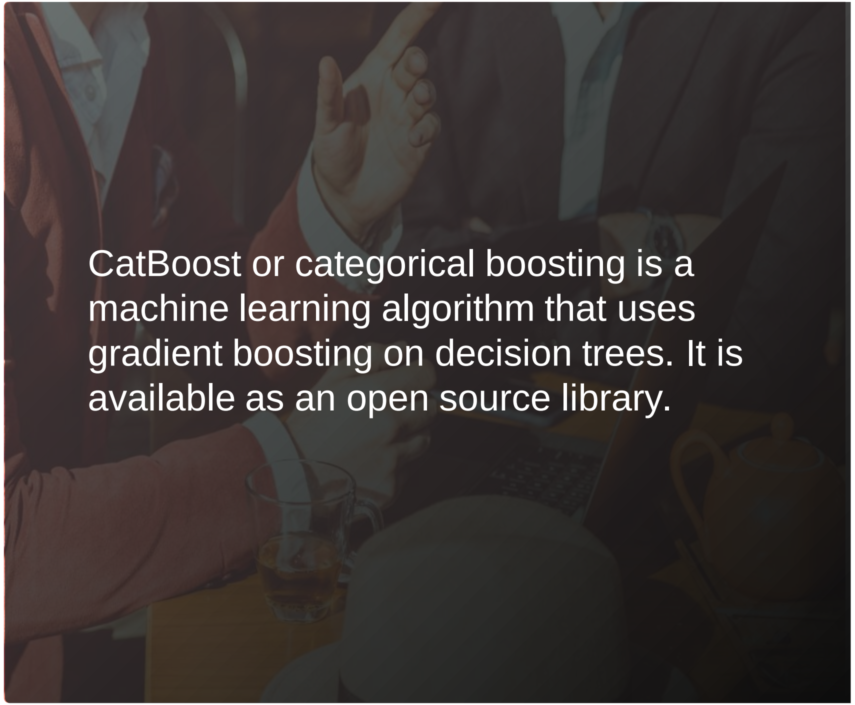
CatBoost

LightGBM Regressor



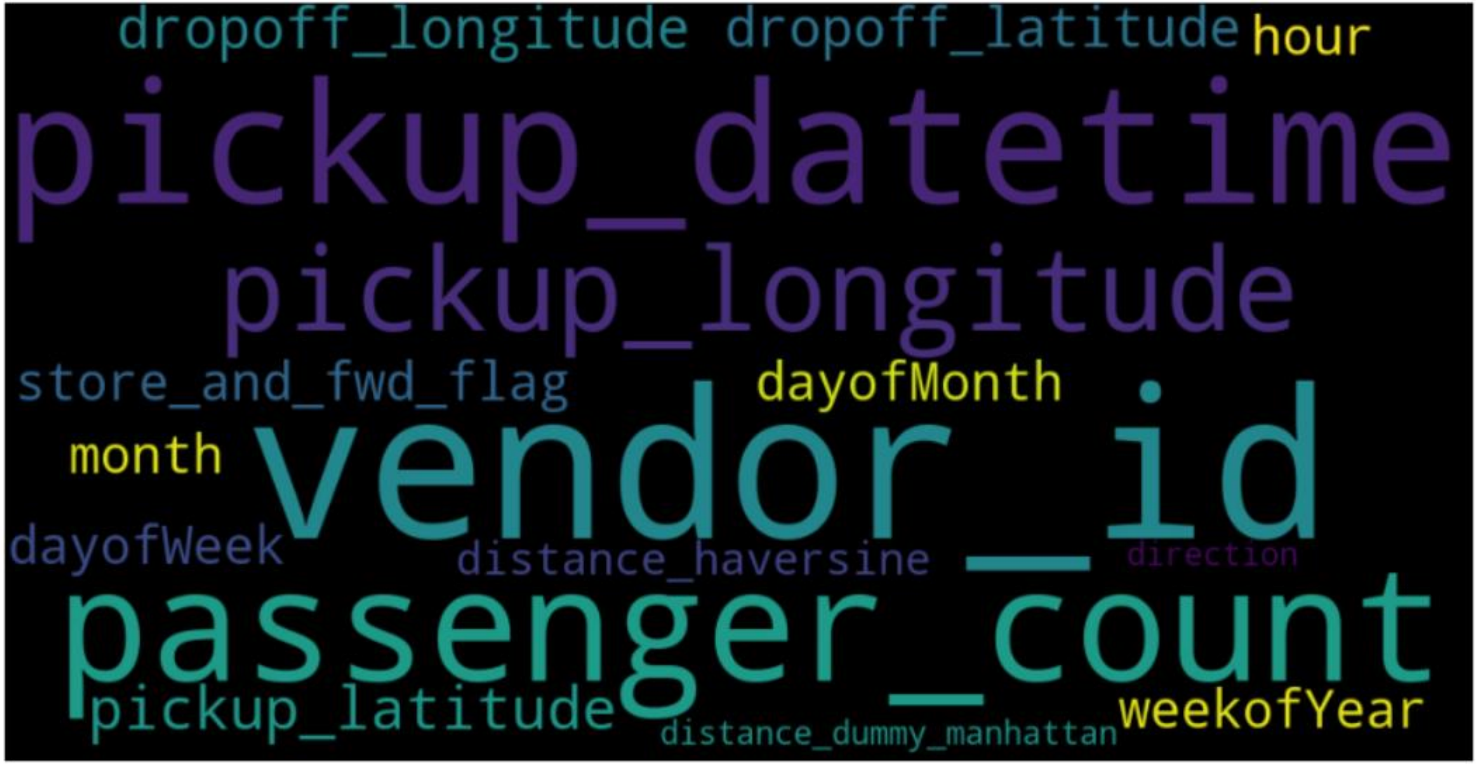
LightGBM Regressor is a gradient boosting framework that use tree based learning algorithm to regression oriented. It is to be distributed and efficient with the following advantages:

CatBoost



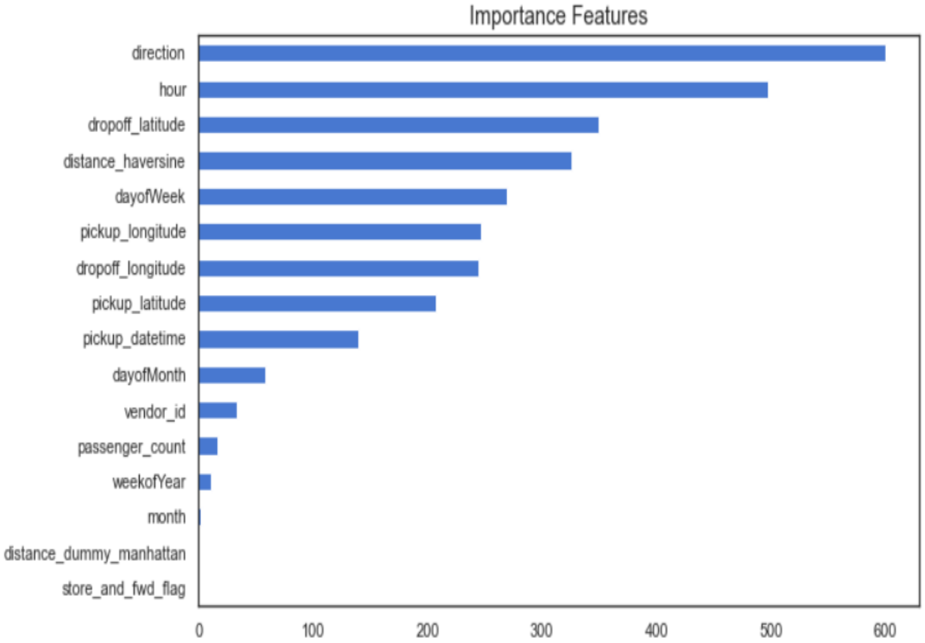
CatBoost or categorical boosting is a machine learning algorithm that uses gradient boosting on decision trees. It is available as an open source library.

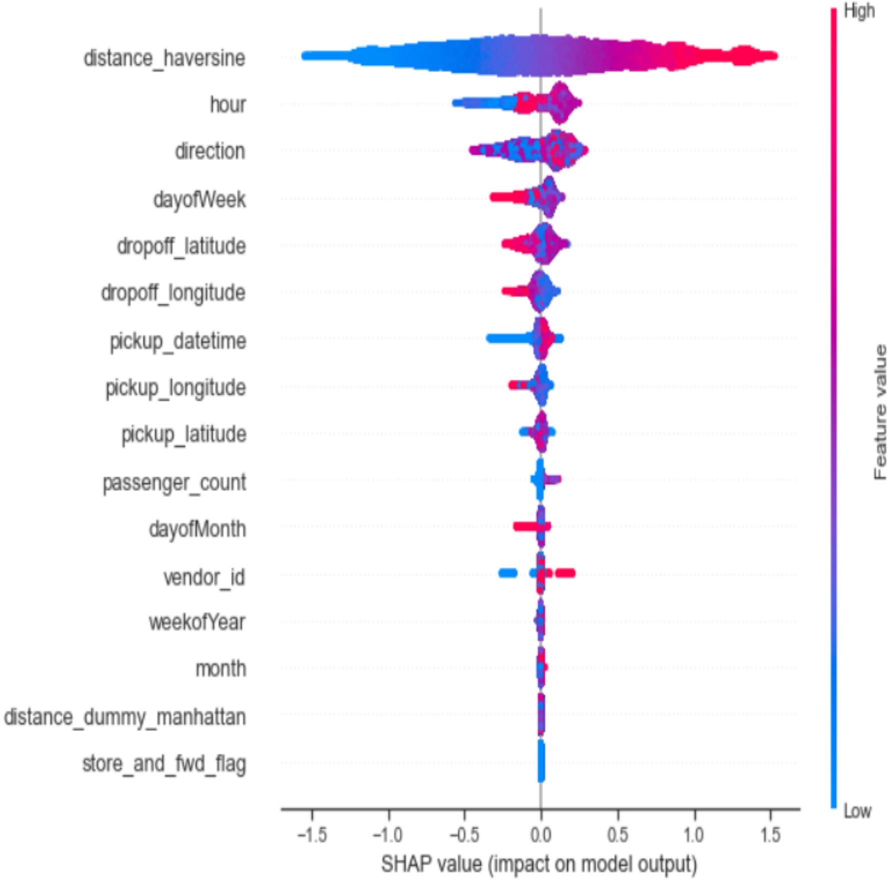
16 Features


dropoff_longitude
dropoff_latitude
hour
pickup_datetime
pickup_longitude
store_and_fwd_flag
dayofMonth
month
vendor_id
dayofWeek
distance_haversine
direction
passenger_count
pickup_latitude
distance_dummy_manhattan
weekofYear

LightGBM Result

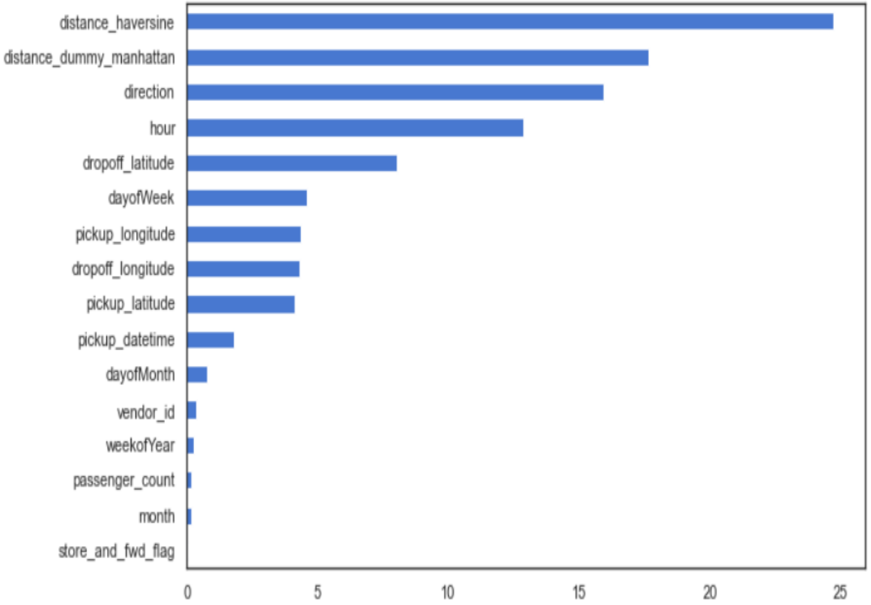
train	Test	MALE	MSLE	RMSLE
0.78	0.78	0.24	0.11	0.34

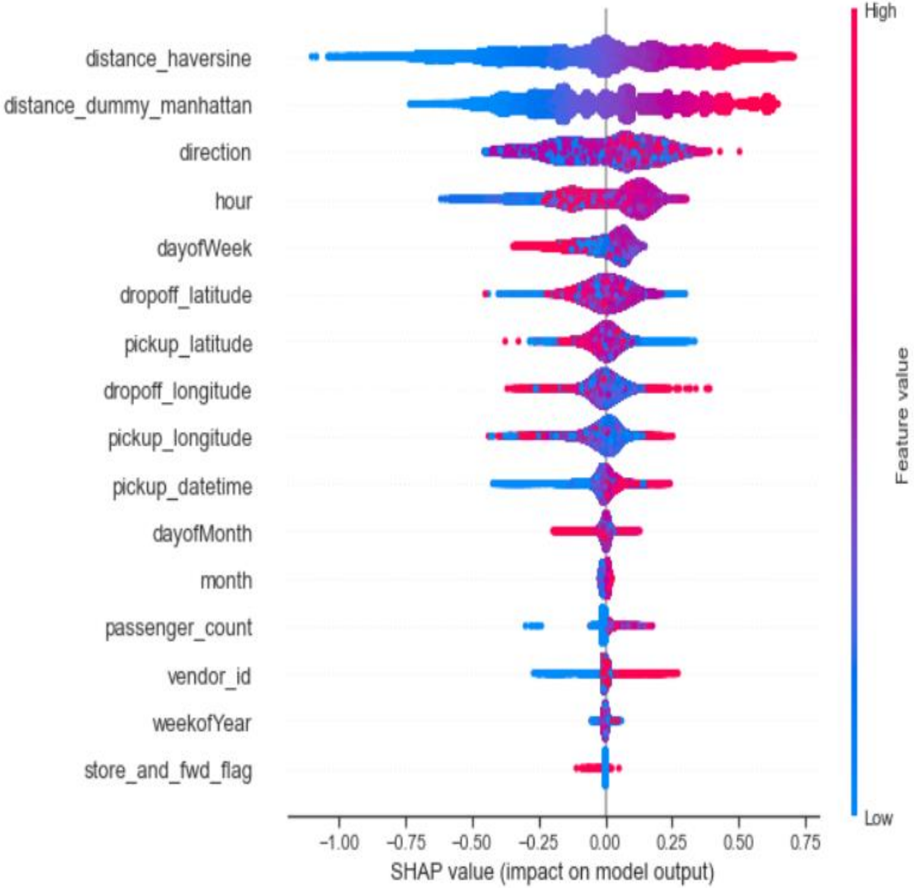




CatBoost Result

train	Test	MALE	MSLE	RMSLE
0.82	0.81	0.22	0.10	0.31





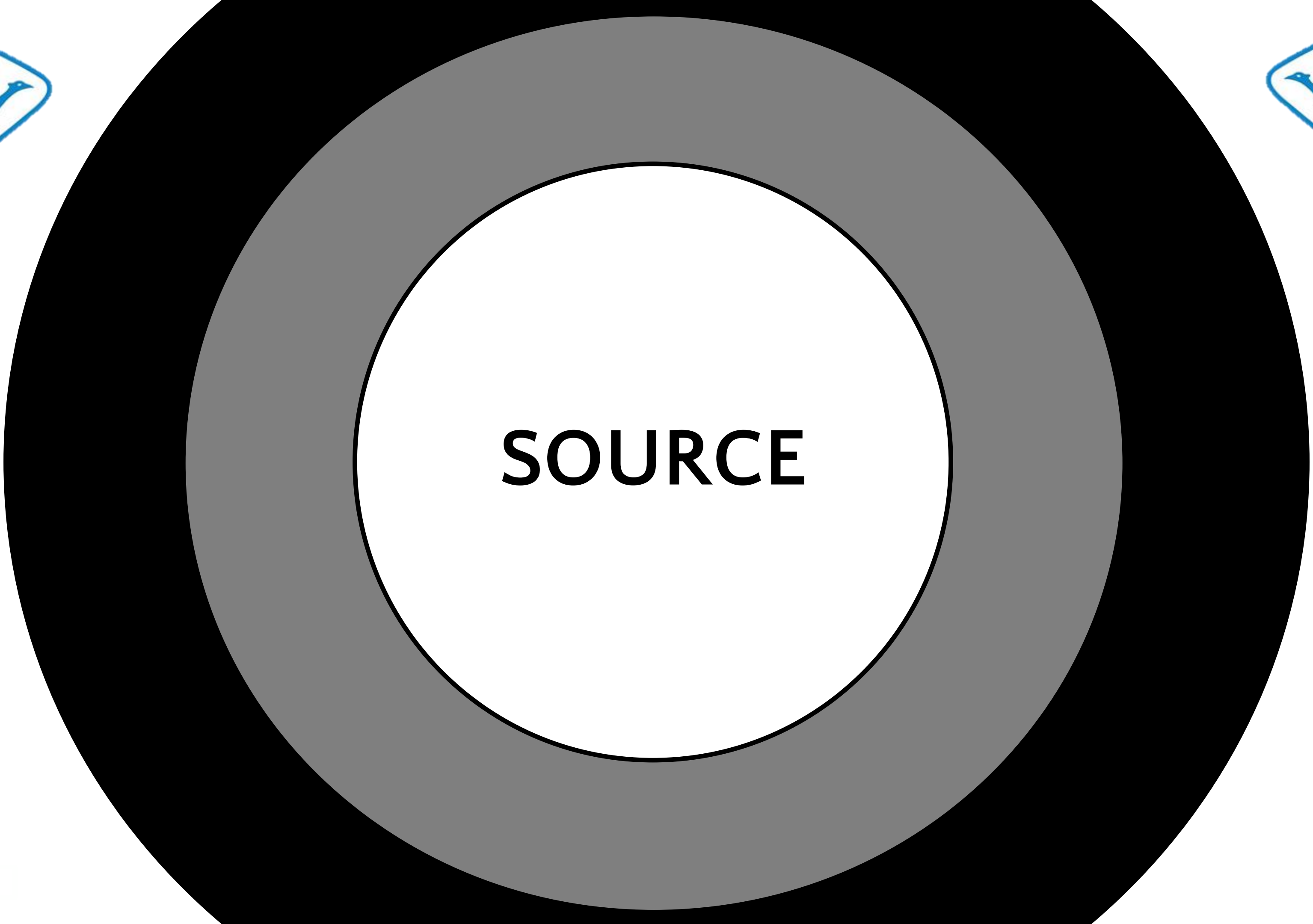
Conclusion

Conclusion

	Train	Test	MALE	MSLE	RMSLE
LightGBM	0.78	0.78	0.25	0.11	0.34
CatBoost	0.82	0.82	0.22	0.10	0.31

We recommend, that CatBoost algorithm to solve in this case.

Thank You



SOURCE

Source:

- https://www.kaggle.com/c/nyc-taxi-trip-duration/notebooks
- https://catboost.ai/docs/concepts/about.html
- https://lightgbm.readthedocs.io/en/latest/
- http://www.nyc.gov/html/media/totweb/taxioftomorrow_history.html