NYC Taxi Fare Prediction with Machine Learning

A data science case study

By: Adolf Yoshua Marbun

July 2022

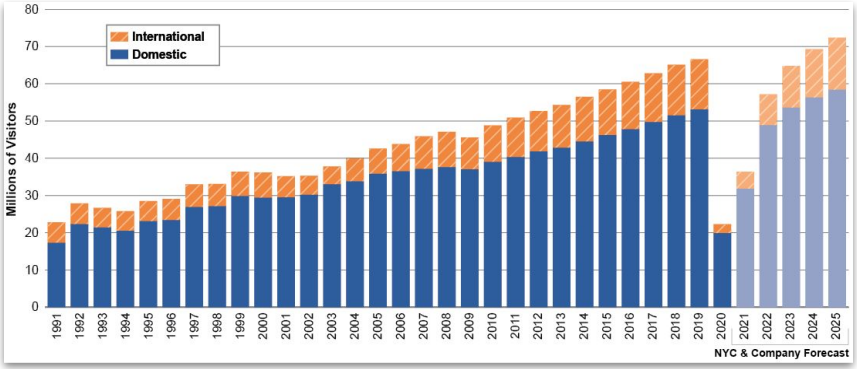
Table of contents

01.

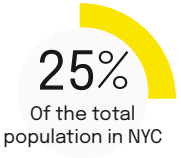
Problem statement

Background

Figure 1. Total visitors to New York City[12]



- New York is one of top 10 global destinations for visitors in 2018 who are interested in its museums, entertainments, restaurants, shows, or even major athletic events. [11, 12]
- The graph has shown us a significant growth of total number of NYC visitors, especially in the last 10 years that the growth has a nearly constant rate.



Approximately, NYC hosted 66.6 million visitors in 2019 with the dominant visitors are domestic and forecasted up to 70 million visitors in 2025 after the drop in 2022 caused by pandemic.

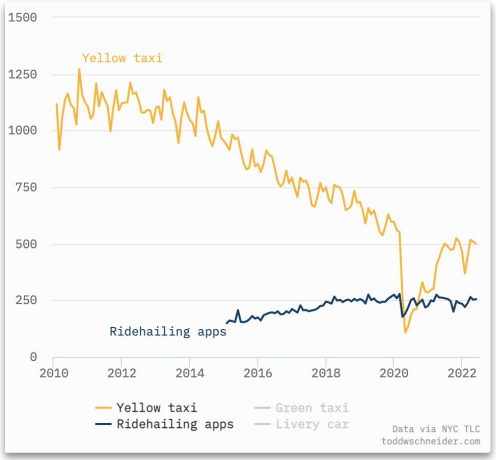
Background

Table 1. Share of visitors by category in 2018-2019[12]

Category	Share of Domestic	Share of International	Share of Total
Leisure	78%	83%	80%
Business	22%	17%	20%

- More than three quarters of total visitors in 2019 visit NYC for leisure purposes.

Figure 2. Monthly taxi trips by vehicle



Even though monthly taxi trips (yellow line) decreased by 33% from 2018 to 2019, NYC yellow taxi trips after pandemic led us to the increasing trend in accordance with the forecasted number of visitors.

Objectives

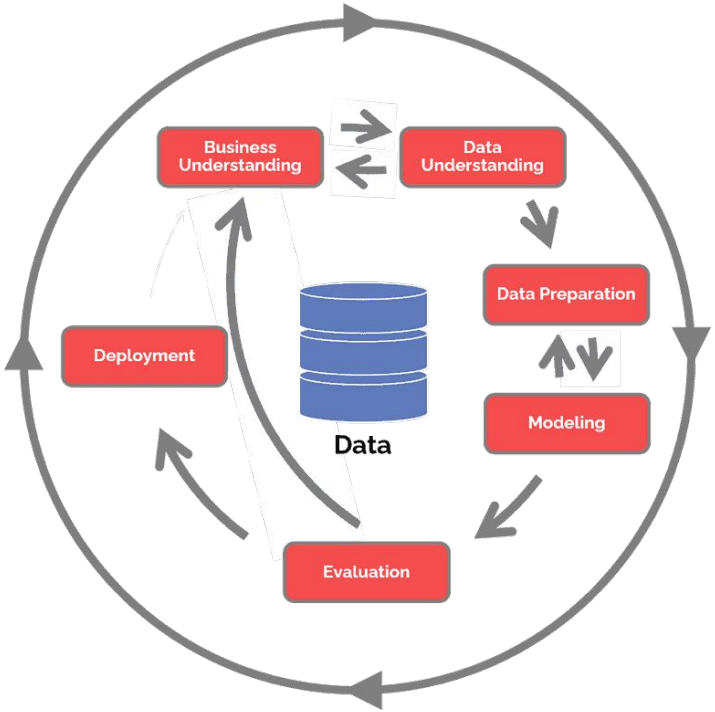
As visitors to NYC take a big part of leisure purposes, they may have a taxi ride for a safe transportation mode. But, bad things may happen like wicked and greedy taxi drivers demand more tips[14] which make them feel uncomfortable, or even the fare amount is too high due to the high/peak season.

Building machine learning model to predict taxi fare amount with given features: pickup and dropoff locations, # of passengers, and time pickup.

The model is aimed to help visitors control their budget based on the predicted taxi fare amount, before deciding to take a taxi rade, travel underground such subways, or even just walk to destinations.

Methodology

Figure 3. Data science methodology cycle[13]

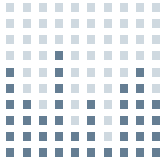


Activities	Tools
Data Understanding and Preparation: Data is obtained through Kaggle repository "New York Taxi Fare Prediction"	kaggle
Data Modeling, EDA: Starting from data cleansing, exploratory data analysis, until two machine learning algorithms are applied to this case and then get evaluated based on the performance.	jupyter seaborn scikit learn

02

Dataset information and modeling

Dataset Information and Modeling



Raw Dataset

- Comprises two datasets: **Trip.csv** and **Fare.csv**, each has +15 millions rows of records.
- Each row in Trip.csv is associated with each row in Fare.csv,

- Due to insufficient memory in CPU, around 10% (~1.5 million rows) of the entire two datasets will be processed for data cleansing, and machine learning modeling.

- Two datasets will be merged and split as follow; **70%** for data training, the residuals for data testing.
- Another 1,000 records from the merged dataset is split as an external data input or test case.

Dataset Information

This dataset includes trip records from all trips completed in yellow taxis in NYC, in April 2013

hack_license	Taxi medallion
vendor_id	Taxicab technology service provider. **1: CMT Group; 2: Verifone, Inc**
payment_type	Taxicab technology service provider. **1: Credit card; 2: Cash; 3: No charge; 4: Dispute; 5: Unknown; 6: Voided trip**
pickup_datetime	The date and time when the meter was engaged
dropoff_datetime	The date and time when the meter was disengaged
pickup_latitude	Latitude at pick-up point
pickup_longitude	Longitude at pick-up point

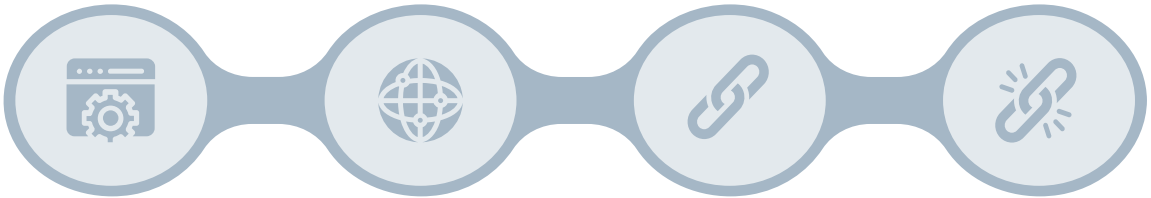
dropoff_latitude	Latitude at drop-off point
dropoff_longitude	Longitude at drop-off point
passenger_count	# of passengers in taxi
fare_amount	The time-and-distance fare calculated by the meter
tolls_amount	Total amount of tolls paid in trip
tip_amount	Tip amount given to the driver, by credit card only. Cash tips are not included.
total_amount	Total amount of fare, tolls, and tips

Notes: not all features are displayed in this slide due to the removal.

03.

Data cleansing and EDA

Dataset Cleansing



Data transformation

Converting raw pickup/dropoff datetime column into datetime type, extracting **hour** and **weekday** number from datetime column, converting categorical features into numerical (encoding)

Geolocation data filtering

Setting New York City boundaries and filter out the data beyond the boundaries

Univariate analysis and outliers removal

Identifying the distribution of all numerical features of the dataset, outliers, and doing the outliers removal

Bivariate analysis

Exploring the correlation of numerical features, gaining insights from categorical features

Data Transformation

Converting pickup/dropoff datetime object data type column into proper data type (datetime), and extracting the hour, weekday, and the name of day

pickup_datetime	dropoff_datetime	Hour-dropoff	Weekday-dropoff	Dayname-pickedup	Dayname-dropoff
2013-04-04 18:47:45	2013-04-04 19:00:25	19	3	Thursday	Thursday
2013-04-05 07:08:34	2013-04-05 07:17:34	7	4	Friday	Friday
2013-04-04 17:59:50	2013-04-04 18:21:48	18	3	Thursday	Thursday
2013-04-04 18:12:01	2013-04-04 18:25:24	18	3	Thursday	Thursday
2013-04-04 20:12:57	2013-04-04 20:29:55	20	3	Thursday	Thursday

Notes:

Monday=0 → Sunday=6

Encoding the categorical features

Vendor_id → CMT = 1 and VMF = 2.

Payment_type → CRD = 1, CSH = 2, UNK = 3, NOC = 4, else 5.

Geolocation data filtering

The coordinates of New York City is shown on the table below. Since there are few points in the dataset that lie outside the boundaries, these points have to be removed.

Table 5. New York City coordinates[6]

Longitude	**-74.006**
Latitude	**40.7142**
West Longitude	-74.257159
East Longitude	-73.699215
South Latitude	40.495992
North Latitude	40.915568

Feature creation: **The trip distance, which is one of the most defining factors for predicting taxi fare, can be calculated by using the Haversine Formula. The Haversine (great circle) distance is the angular distance between two points on the surface of a sphere. As the earth is nearly spherical, calculating this is a good approximation.**

Figure 4. Illustration of calculating shortest distance in sphere[15]

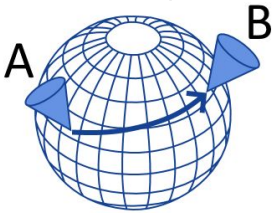


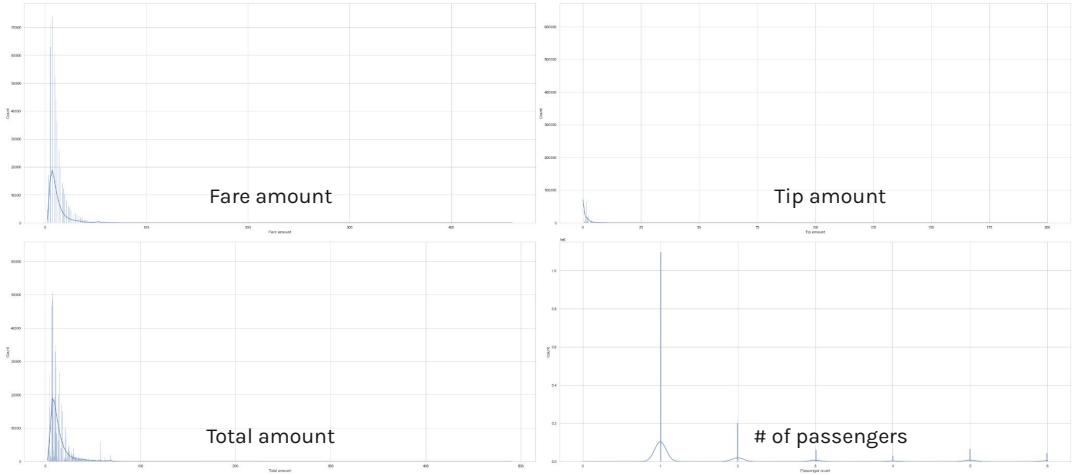
Figure 5. Equation of calculating Haversine distance (d)

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

Where Phi = latitude and Lambda = longitude

Univariate Analysis

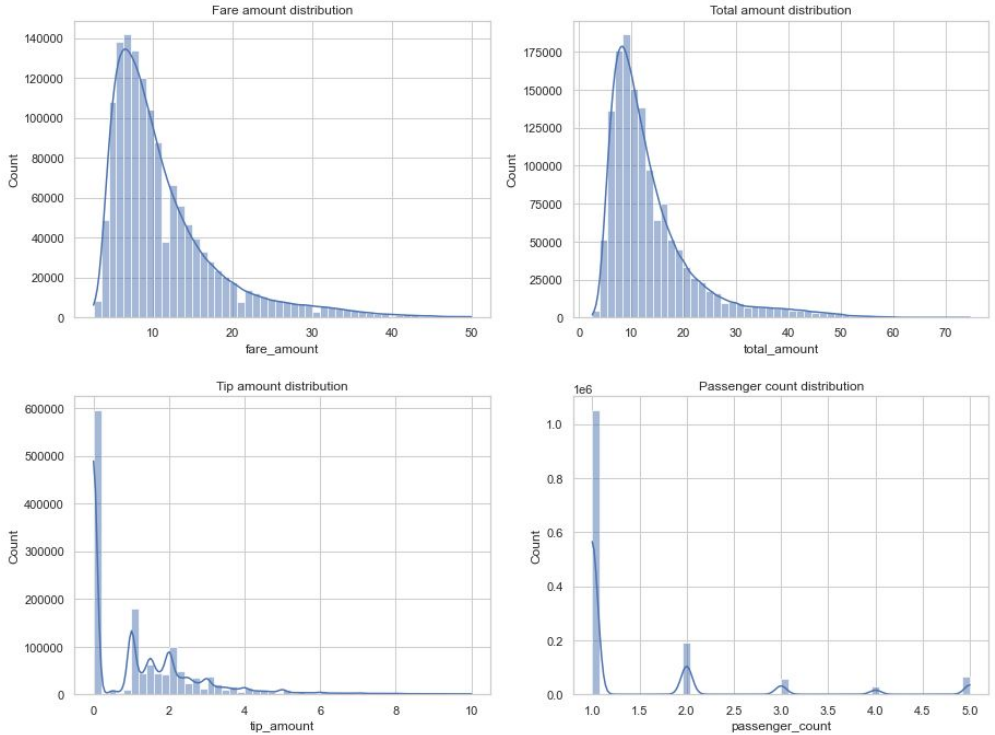
Figure 6. Univariate analysis of numerical features before outlier removal



- As we can see in Figure 6, the distribution of fare amount, tip amount, total amount and number of passenger in taxi, they are **heavily distributed to the left**, which means there are few points indicated as outliers.
- We can limit these features with this conditions:
  1. Fare amount <= 50 USD (with an initial charge is USD 2.5)
  2. Tip amount <= 10 USD
  3. Passenger count between 1 and 5
- Also, we filter out non-logical things like travel time and coordinates which equal to zero.

Univariate Analysis

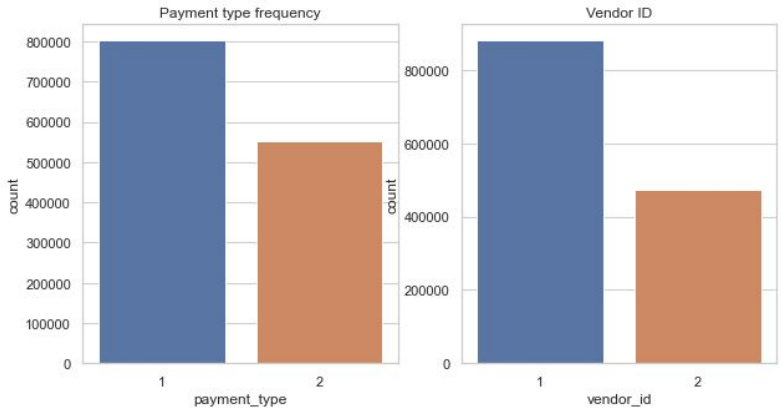
Figure 7. Univariate analysis of numerical features after outlier removal



- The graphics are now more reasonable to be analyzed after the boundaries get filtered out.
- **Key highlights**:
  1. Average taxi fare is around USD 7.5 - USD 10.
  2. The vast majority of trips have only one passenger on board.
  3. Most of passengers do not give tips to the driver (tips = 0).

Univariate Analysis

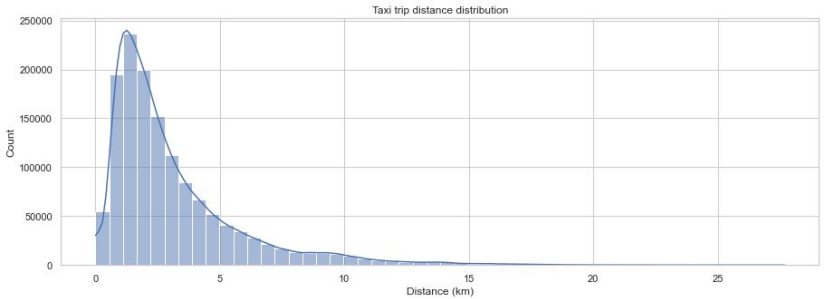
Figure 8. Univariate analysis for categorical features (payment type and vendor)



- Two most dominant payment types; cash and credit. Only few thousand trips have their payment registered as Disputed, No Charge, or even Unknown, and they do not really contribute much to the whole payment time. So, the other three types are neglectable or able to be taken out.
- **Key highlights**:
  1. It is NYC citizens' habit to pay trips with credit card rather than cash
  2. Nearly 75% of the total taxi trips using vendor 1 (CMT Group) for their service provider.

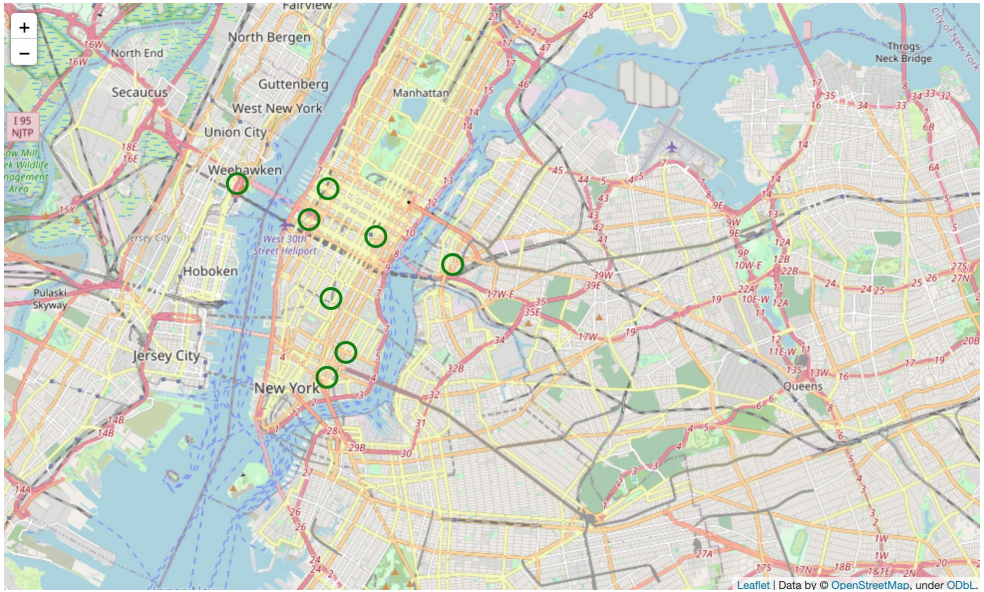
Univariate Analysis

Figure 9. Taxi trip distance distribution



- **Key highlights**:
  1. Average taxi trip distance in NYC is 3.1 km.
  2. The most popular pickup points in NYC based on the data is around the apartment (5th Avenue), train station (Pennsylvania Station), and airport (La Guardia Airport).

Figure 10. Ten most popular pickup points



Univariate Analysis

Figure 11. Five busiest hour in a day (pickup)

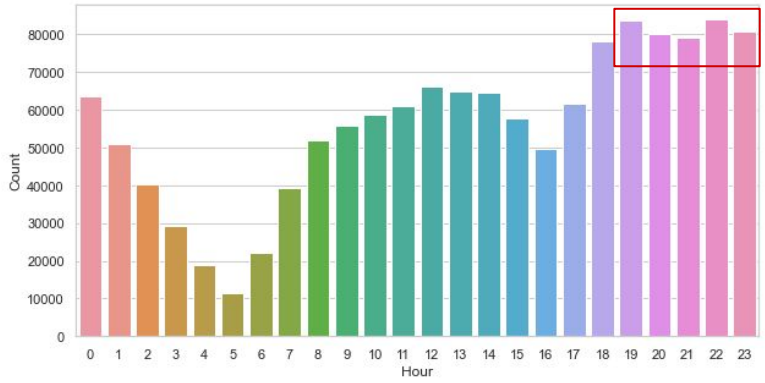


Figure 13. Average fare amount by # of passengers in taxi

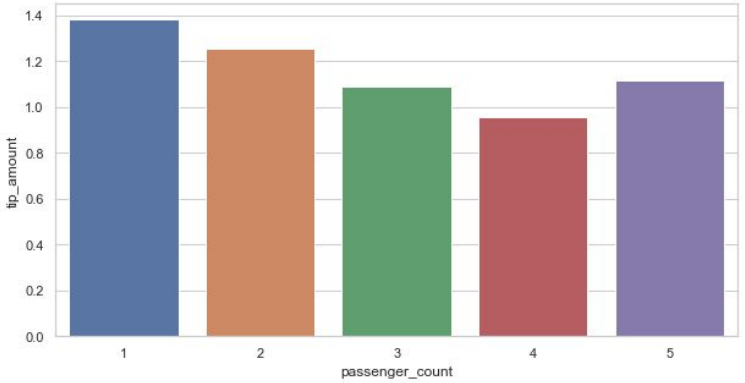
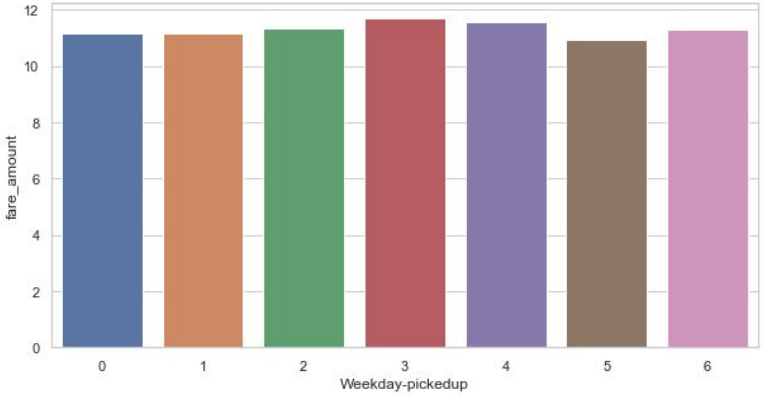


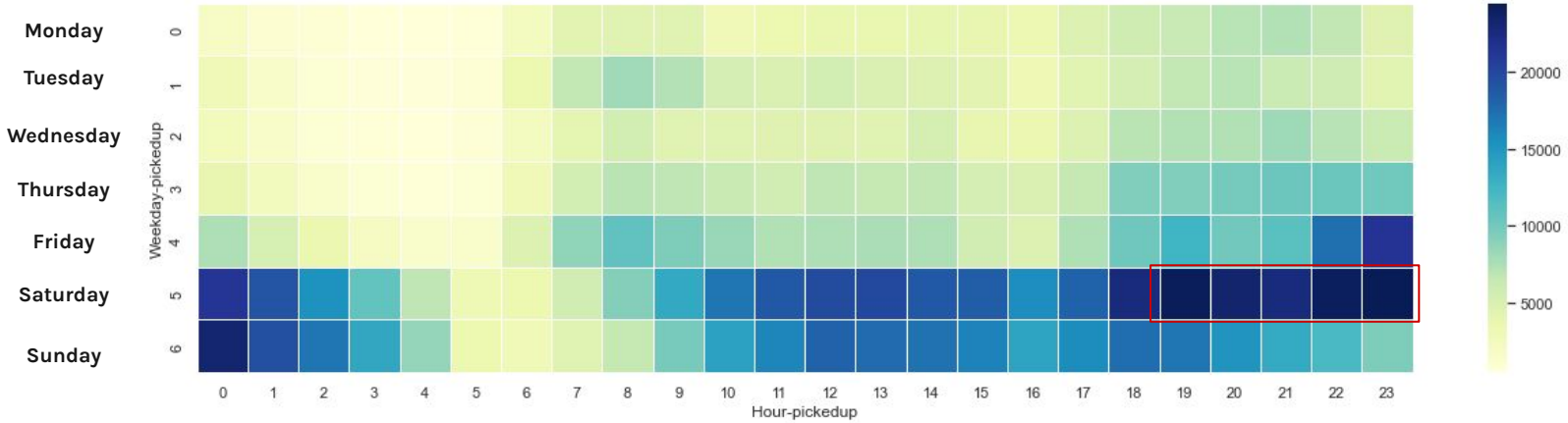
Figure 12. Average fare amount by day (Monday=0; Sunday=6)



- **Key highlights**:
  1. We see that busiest hours where taxis are being hailed are from 7 pm - 11 pm which make sense as this is the time when people return from the office or have dinner
  2. There is not any significant correlation between pickup day to taxi fare, the average fare amount per day remains constant, and so is the # of passengers correlation.

Bivariate Analysis

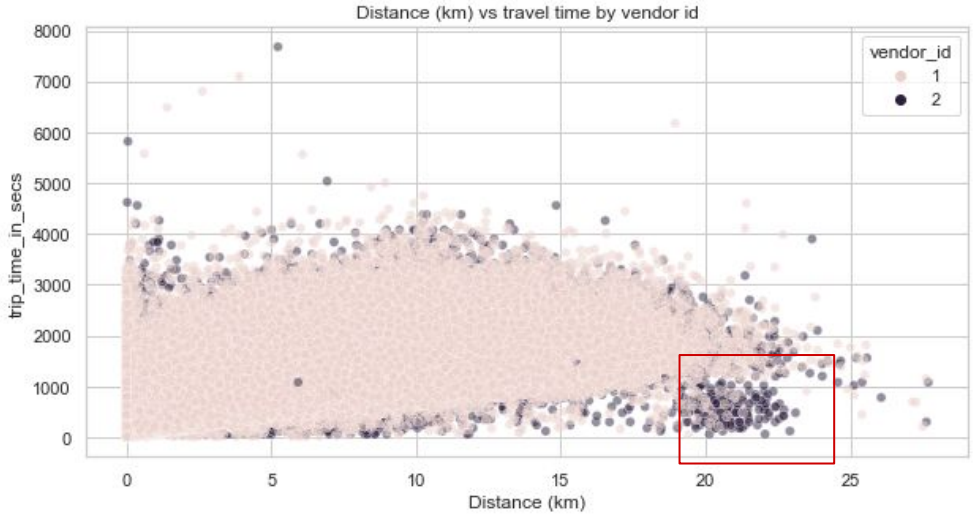
Figure 14. Heatmap visualization of Hour of pickup and Day of Week



- It is obviously clear that from the heatmap above, the busiest time where people order and drop off from taxi in NYC is around the Saturday Night (19.00 - 23.00) which is the perfect time for leisure or entertainment purposes.
- Time for picking up a taxi; hour and day (weekend or weekday) is one of the other defining factors for predicting taxi fares.

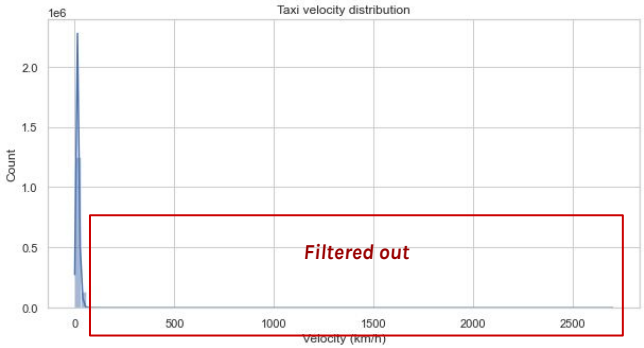
Bivariate Analysis

Figure 15. Distance (km) vs trip time in sec Scatterplot by vendor id



- See Figure 15 and the red box. The relation between distance and time from the equation of velocity is supposed to be proportional. The red box shows us something odd which does not make any sense.
- The **longer** the trip takes, the **longer** time it takes.

1. Identify the speed of the trip distribution by creating a new feature
2. Eliminate the irrelevant data. Discover the maximum speed of taxi which is allowed in US/NYC.



Bivariate Analysis

- Eliminate the trips with speed more than 100 kmh.
- The maximum speed allowed in NYC has been reduced to 25 miles per hour, from 30 mph, which is approximately **40.2** km per hour.
- Average trip time in NYC : 705 seconds (11.75 mins)

Figure 16. Taxi velocity distribution after irrelevant data being eliminated

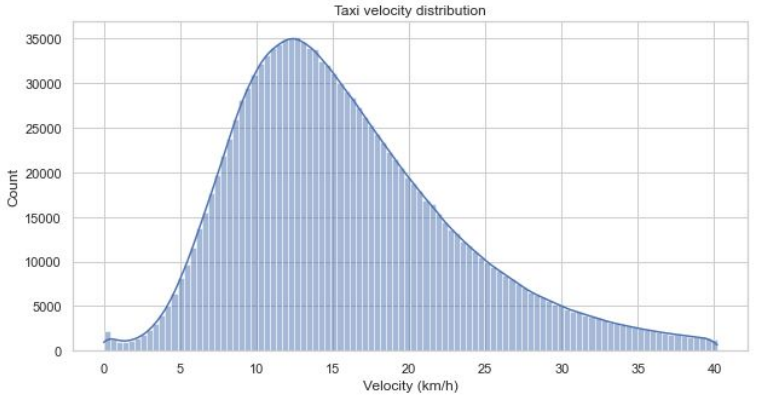
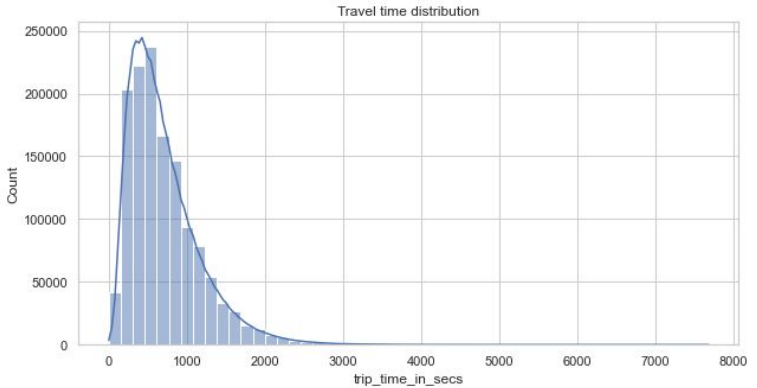
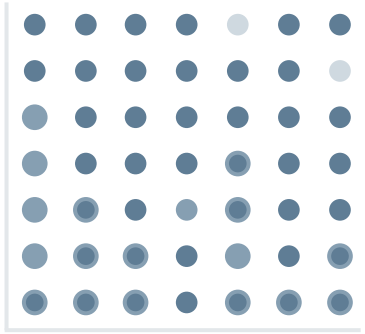


Figure 17. Trip time distribution



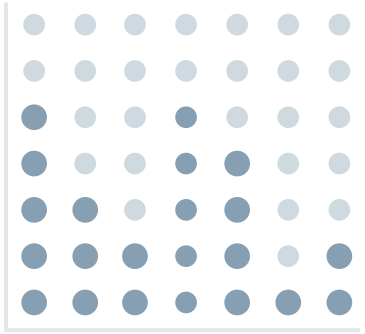
Dataset Final Information



1.500.000 million rows

-9.7%

Data cleansing and EDA



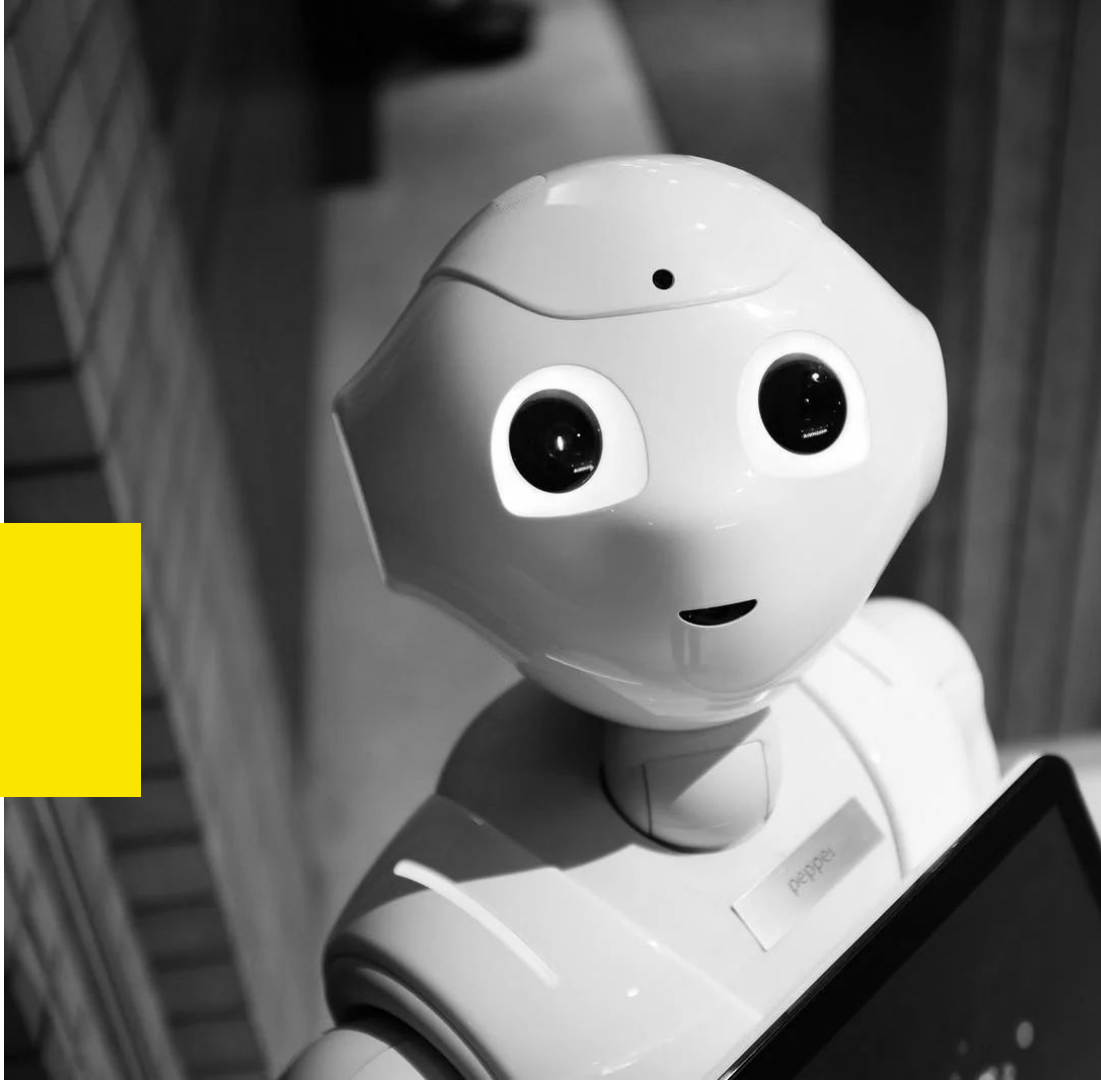
1.354.694 million rows

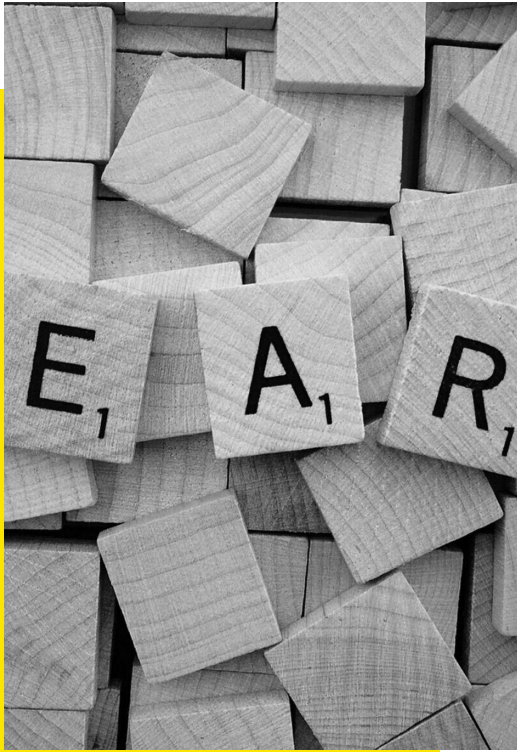
(cleansed dataset)

04

Machine learning model

You can enter a subtitle here if you need it

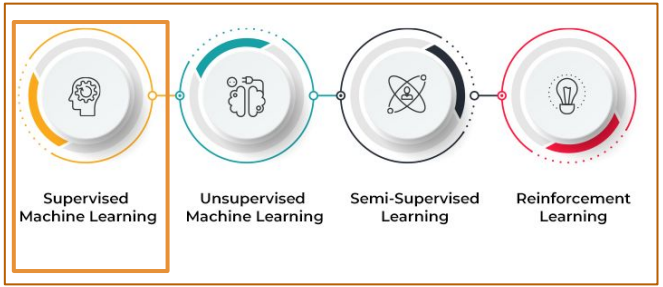




Machine Learning

A discipline of artificial intelligence (AI) that provides machines the ability to automatically learn from data and past experiences to identify patterns and make predictions with minimal human intervention.

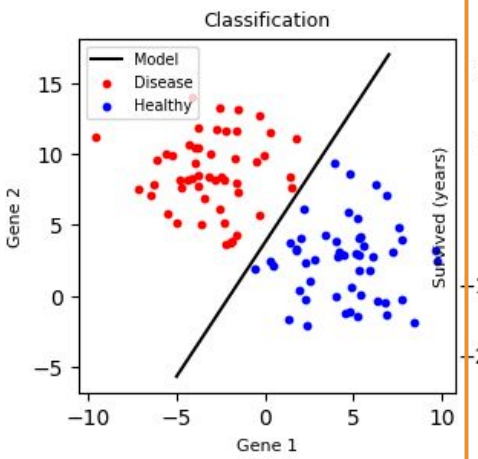
Figure 17. Types of machine learning

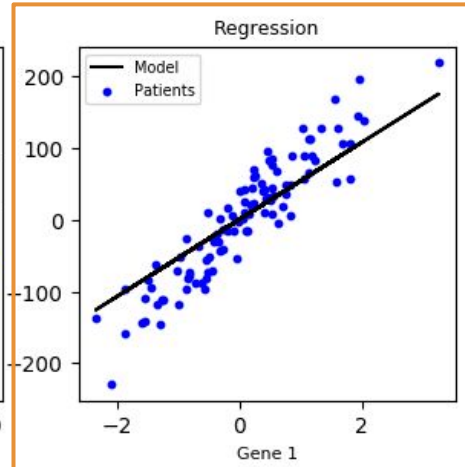


Machines are trained on labeled datasets and enabled to predict outputs based on the provided training.

Supervised Machine Learning

Figure 18. Types of supervised machine learning (classification and regression)





Refers to algorithms that address classification problems where the output variable is categorical; for example, yes or no, true or false, male or female. Algorithms: *Random Forest Algorithm, Decision Tree Algorithm, Logistic Regression Algorithm, and Support Vector Machine Algorithm*

Handle regression problems where input and output variables have a linear relationship. These are known to predict continuous output variables. Examples: *Simple Linear Regression Algorithm, Multivariate Regression Algorithm, Decision Tree Algorithm, and Lasso Regression.*

In this case, the algorithm we are going to apply is Linear Regression since we must predict the taxi fare amount based on several predictors.

Linear Regression

The most common and used type of predictive analysis. **Defined as a statistical method that helps us identify and the relationship between two or more variables.**

Simple Linear Regression	$y = mx + c$
Multiple Linear Regression	$y = mx_1 + nx_2 + \ldots + c$
Polynomial Regression	$y = m_0 + m_1x^1 + m_2x^2 + \ldots + m_nx^n$

Since we have many variables that will predict the outcome (predicted taxi fare amount), we are going to use Multiple Linear Regression.

Linear Regression Evaluation Metrics

R^2 (R squared)

Also known as coefficient of determination, It measures the strength of the relationship between the model and the dependent variables (goodness of fit).

$$R^2 = 1 - \frac{RSS}{TSS}$$

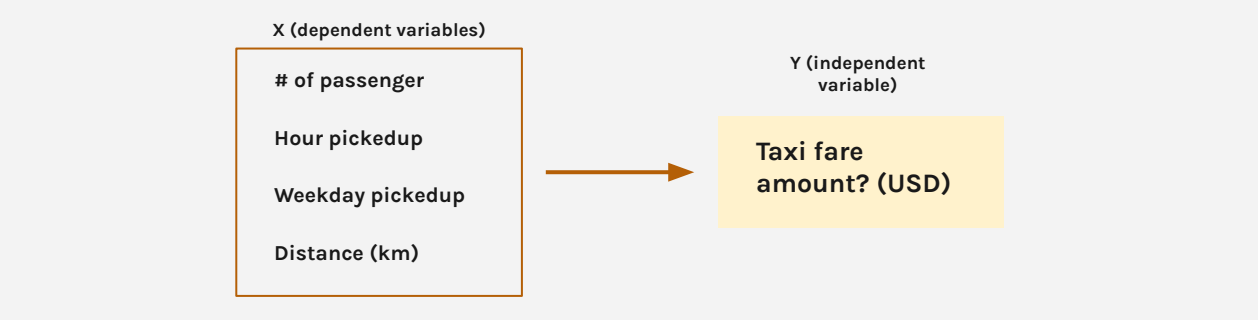
Where:
RSS = Residual Sum Square
TSS = Total Sum Square

$$SE_{Line} = RSS = (y_1 - (\theta_0 + \theta_1 x_1))^2 + (y_2 - (\theta_0 + \theta_1 x_2))^2 + \ldots + (y_n - (\theta_0 + \theta_1 x_n))^2$$

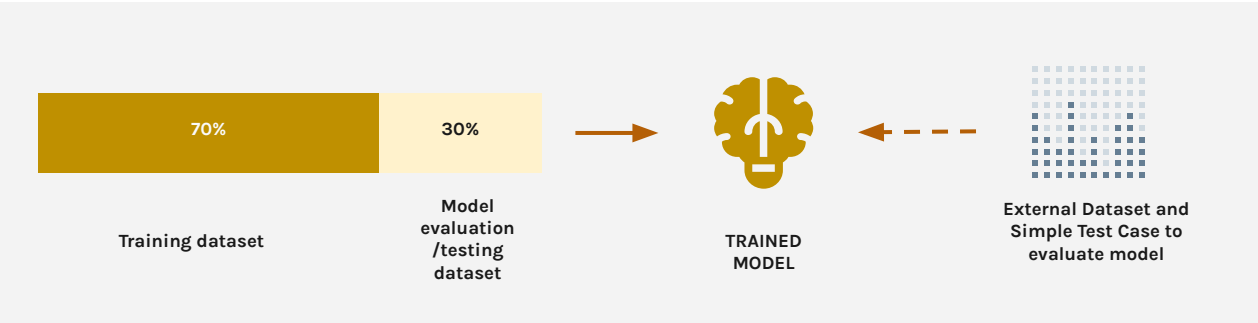
$$SE_{\bar{y}} = TSS = (y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \ldots + (y_n - \bar{y})^2$$

Machine Learning Model Illustration

Model illustration



Dataset separation



Linear Regression Model Algorithm

A. Splitting dataset

```
y = BigData['fare_amount']
X = BigData[['passenger_count', 'Hour-pickedup', 'Weekday-pickedup', 'Distance (km)']]

# for external data

y_ext = ExternalData['fare_amount']
X_ext = ExternalData[['passenger_count', 'Hour-pickedup', 'Weekday-pickedup', 'Distance (km)']]

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=50)

print("Train shape :  ", X_train.shape, " - ", y_train.shape)
print("Test shape :  ", X_test.shape, " - ", y_test.shape)
```

```
Train shape :   (948285, 4)  -  (948285,)
Test shape :   (406409, 4)  -  (406409,)
```

Linear Regression Model

B. Building machine learning model

```
from sklearn import linear_model

reg1 = linear_model.LinearRegression()
reg1.fit(X_train, y_train)

X_linear = pd.DataFrame(X_test)
Y_linear = reg1.predict(X_linear)

r2s = r2_score(y_test, Y_linear)
print("R2_score: ", round(r2s, 2))
```

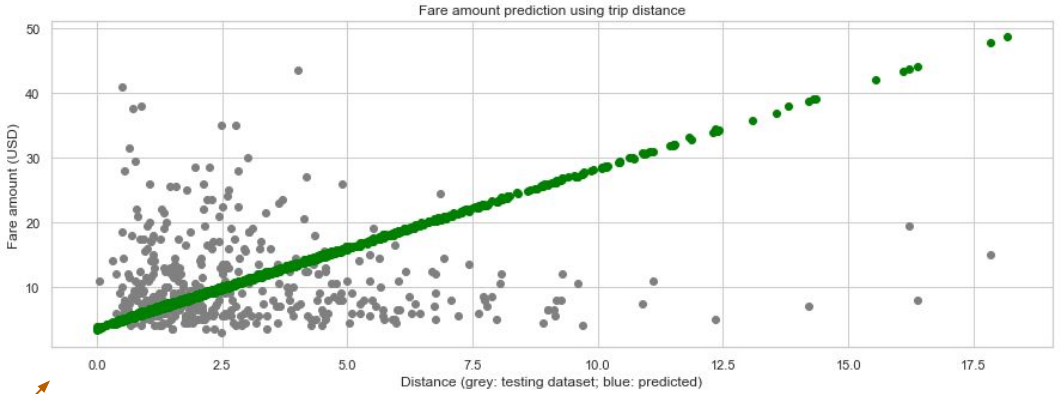
```
R2_score:  0.84
```

X_test (# of passenger, hour pickedup, weekday pickedup, distance)

y_test

y_pred

Figure 19. Scatter plot of predicted, actual fare amount and distance of trip (in km)



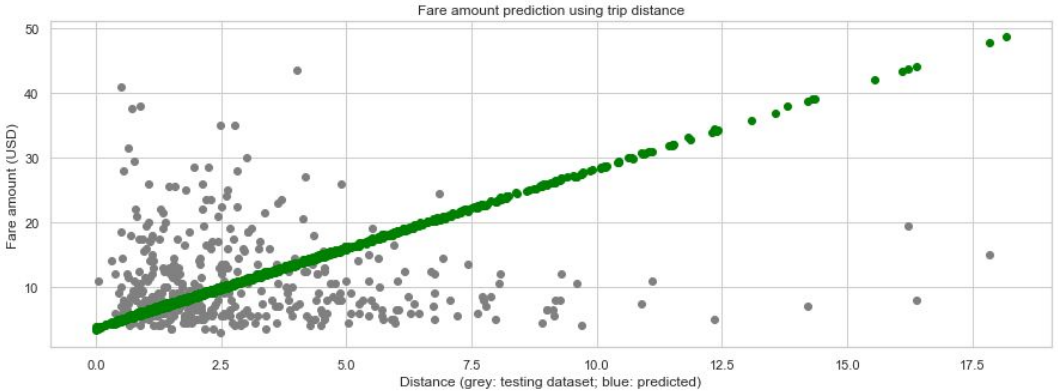
Note: The scatter plot below only shows the first 500 records in testing dataset

Above is the scatter plot of the predicted fare amount and the actual fare amount with distance as the dependent variable, from the X_test dataset. It is clear that the model gives us a **positive correlation** between distance and fare amount of taxi.

The model has given an R2 score of 84% which in Machine Learning is a very good model (good fit).

Linear Regression Model

Figure 19. Scatter plot of predicted, actual fare amount and distance of trip (in km)



Note: The scatter plot below only shows the first 500 records in testing dataset

	features	coefficient
0	passenger_count	0.042472
1	Hour-pickedup	0.009645
2	Weekday-pickedup	-0.069012
3	Distance (km)	2.463567

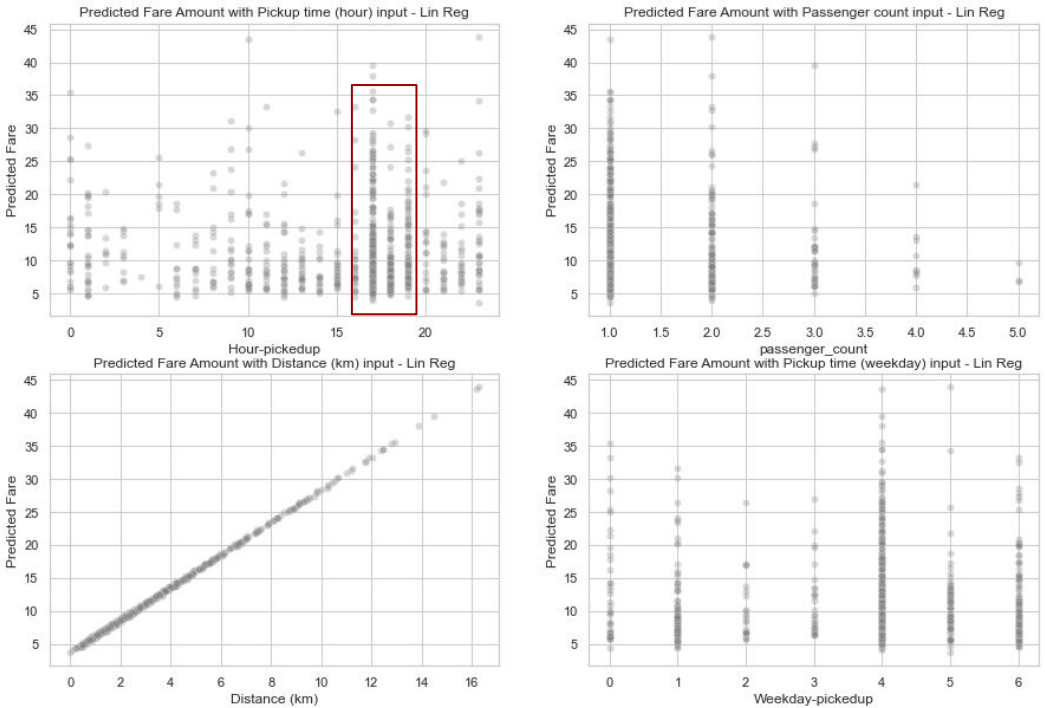
Linear Regression Model

C. Training machine learning with an external dataset

The external dataset is actually part of the whole dataset which records 1,000 rows only. It is aimed to validate and train the model more.

- Based on the findings, the highest traffic of taxi order occurs at around 18 - 20 pm on Weekday 4 (Friday) which really makes sense that people tend to go out after work or get dinner by using transportation.
- There is a positive correlation between distance of trip (in km) to the taxi fare amount.
- Trips with only 1 passenger are predicted to have a higher fare.

Figure 20. Dependent variables correlation to fare amount from external dataset



Linear Regression Model

D. Training machine learning with a simple test case

To make this model more applicable, let's try a new simple case:

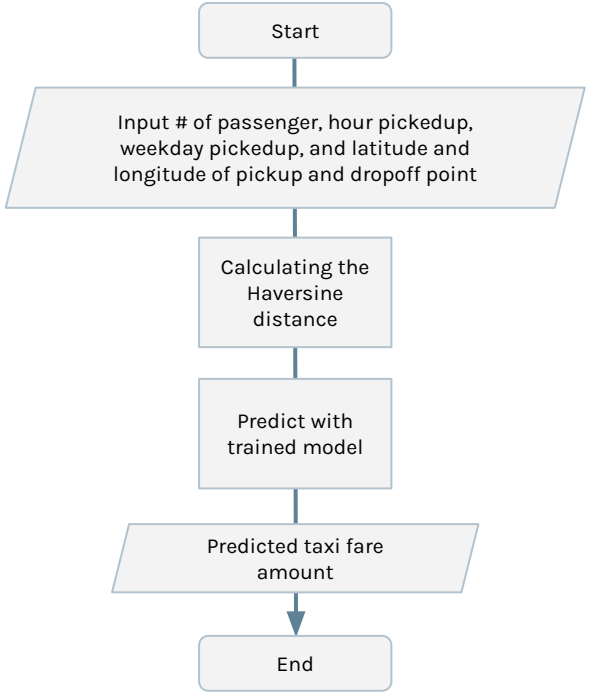
A man with his wife wants to order a taxi from 14th Union Square to Roosevelt Island Tramway at 19.00 pm on Friday.

1. # of passengers : 2
2. Hour pickedup : 19
3. Weekday pickedup : 4
4. Latitude Longitude pu and do: 40.7345° N, 73.9903° W 14th Union Square , 40.7616° N, 73.9646° W Roosevelt Island Tramway

Predicted Taxi Fare at those conditions:

12.91 USD

Figure 21. Program flowchart with manual input (simple test case)



05.

Future Works

Future Works

Department	Jobs to be done
Data Science and Analytics	- Collect more complete data back to 5 years ago and deep dive the EDA to find pattern in a time-series viz - Identify and consider the major events or public holidays in USA to a taxi fare amount prediction - Identify another potential predictor variables; weather, cab size, public holiday or not. - Train and improve model with a specific case of pickup and drop-off point. For example: Determine a flat fare to airport in NYC, and forecast the toll amount charged in the trip.
Tech/Product/ Design	- Build and design a simple web that requests input from users and result a predicted taxi fare. - Determine the success metrics of the product (click rate, bounce rate, and any other traffic metrics)

References

1. Kamali, Emmanuel. 2018. *Predicting NYC Yellow Cab Taxi Fare.* https://nycdatascience.com/
2. Salomonsson, Jacob. 2022. *Predicting the Fare on a Billion Taxi Trips with BigQuery.* https://blog.devgenius.io/
3. Insight Software. 2022. *Top 5 Predictive Analytics Models and Algorithms.* https://insightsoftware.com/
4. Patel, Brij. 2019. *New York City Taxi Fare Prediction.* https://medium.com/
5. Mitchel, Mackenzie. *Selecting the Correct Predictive Modeling Technique.* https://towardsdatascience.com/
6. Archie, Christie Natashia., Velu, Subashini. 2020. *DATA EXPLORATORY ON TAXI DATA IN NEW YORK CITY.* School of Computing, Asia Pacific, Kuala Lumpur, Malaysia.
7. Anand, Sejal. 2021. *Exploratory Data Analysis on NYC Taxi Trip Duration Dataset.* https://www.analyticsvidhya.com/
8. Ramachandran, Aiswarya. 2018. *Machine Learning to Predict Taxi Fare — Part One : Exploratory Analysis.* https://medium.com/
9. Fitria, Laila. 2022. *Regresi.* www.linkedin.com/
10. Das, Abhishek. 2017. *Exploring NYC Taxi Data (Updated).* https://factorwonk.github.io/NYCTaxi/
11. Mastercard. 2019. *Bangkok tops Mastercard's Global Destination Cities Index for the fourth consecutive year.* https://www.mastercard.com/news/
12. State Comptroller NYC. 2021. *The Tourism Industry in New York City.* https://www.osc.state.ny.us/reports/osdc/
13. Hotz, Nick. 2022. *What is CRISP DM? - Data Science.* https://www.datascience-pm.com/crisp-dm-2/
14. NYC 311. 2021. *Taxi Complaint.* https://portal.311.nyc.gov/article/?kanumber=KA-01241
15. Luca, Gabriele. 2021. *Haversine Distance.* https://www.baeldung.com/cs/haversine-formula

Thank You

adolfbordeaux99@gmail.com
+62 822 8402 7104

https://github.com/adolfmrbn
https://www.linkedin.com/in/adolfmrbn/

Copy the link to see the full repository (code)