

NYC TAXI - TIP PREDICTION

MUKUNTH RAJENDRAN & RAPHAEL PRESBERG

MULTIVARIATE DATA ANALYSIS

AMIRHOSSEIN GANDOMI

PROJECT REPORT

Project Report

ABSTRACT

The purpose of this project is to solve the difficulties faced by many people in tipping the taxi drivers. This study was done by using data modelling techniques and by grouping the independent variables and finding their influence on the dependent variable. This study uses Dimension Reduction, Regression, and Classification techniques to find the prediction. The data used in the prediction had many variables, we used different techniques to find only those variables which have an impact on our dependent variable. This finding can be used to predict how much the customer should pay the driver by using several factors involved in the travel. In the future we'll be working on the various other situations using this dataset and will be predicting lot more.

Table of Contents

INTRODUCTION:	4
PROBLEM DESCRIPTION:.....	5
EVALUATION OF THE DATABASE:	5
DATA PROCESSING AND PREPARATION:	6
OUTLIERS:	7
USING DUMMY VARIABLE:	11
PRINCIPAL COMPONENT ANALYSIS:	12
METHODS:	14
MULTIPLE REGRESSION:	14
CLASSIFICATION:	16
KNN MODEL:.....	16
CONCLUSION:.....	18
FUTURE RESEARCH:	19
THANKYOU	20

INTRODUCTION:

Initially, to start with there are many predictions out there, but do they really solve all the problems, or the difficulties that we people are facing? The NYC taxi says that “Please tip your driver for safety and good service” in their official site. But we don’t know how many people are willing to pay.

Tipping practices may vary among different factors depending upon the location in US. Moreover, the published guidance will vary depending upon the source.

Here, we worked on finding the prediction to solve real life problem using a real data. We wanted to find the tip amount that must be paid to the customer and categorize them based on various scales. People are confused or cheated when they deal with such a situation, not knowing exactly how much must be paid. We wanted to work on this prediction as we faced the same difficulties as the general people. We were demanded by the Driver to pay the amount that we weren’t willing to pay. So, we went to search for real data on taxi services and got this data from Wharton Research Data Services (WRDS). We didn’t believe that we will be able to predict the tip amount when we were initially worked on it, the knowledge that we acquired from understanding and dealing with the data was vast.

This prediction is just a start to our journey to find more predictions with the same dataset. Furthermore, we’ll be focusing on the Customers, Drivers, Company and Society to find all the possible prediction. The predictions that we acquired from these machine learning algorithms can be used by the driver and customer to know what should be the actual tip amount that has to be paid to the driver.





















Project Report

PROBLEM DESCRIPTION:

When an international traveler arrives in Newark Airport, he takes a cab to go to his hotel or apartment. In the end of the trip, the taxi driver demands to pay a certain tip amount and he isn't satisfied with the amount given to him. Here we are to predict the tip amount that must be paid to the driver using machine learning algorithms. We have done a model to predict the tip amount as well as to scale the customer based on different categories based on the amount paid by him.

EVALUATION OF THE DATABASE:

The database consisted of 20 variables and 727310 records. Each record had their own predefined datatype and all the required details regarding the variable name, position of the variable, the datatype, the length and the description.

Position	Variable Name	Type	Length	Description
1	pickup_date 	Num	8	Pickup Date
2	pickup_time 	Num	8	Pickup Time
3	dropoff_date 	Num	8	Dropoff Date
4	dropoff_time 	Num	8	Dropoff Time
5	vendor_id 	Char	3	Vendor ID
6	passenger_count 	Num	8	Passenger Count
7	trip_distance 	Num	8	Trip Distance (in miles)
8	pickup_longitude 	Num	8	Pickup Longitude
9	pickup_latitude 	Num	8	Pickup Latitude
10	rate_code 	Num	8	Rate Code
11	store_and_fwd_flag 	Char	1	Store and Forward Flag
12	dropoff_longitude 	Num	8	Dropoff Longitude
13	dropoff_latitude 	Num	8	Dropoff Latitude
14	payment_type 	Char	3	Payment Type
15	fare_amount 	Num	8	Fare Amount
16	surcharge 	Num	8	Surcharge
17	mta_tax 	Num	8	MTA Tax
18	tip_amount 	Num	8	Tip Amount
19	tolls_amount 	Num	8	Tolls Amount
20	total_amount 	Num	8	Total Amount

Project Report

Data Source: Wharton Research Data Services (WRDS) – NYC Yellow Taxi Trips – 2015.

WRDS collected this public data from various sources in the public domain, the raw data is converted into a consistent format and updated on a regular basis. This dataset is organized into the content area - New York

<https://wrds->

[web.wharton.upenn.edu/wrds/ds/publicdata/taxis_yellow_2015.cfm?navId=321](https://wrds-web.wharton.upenn.edu/wrds/ds/publicdata/taxis_yellow_2015.cfm?navId=321)

DATA PROCESSING AND PREPARATION:

To start with the data preparation process, we selected some initial variable to perform our analysis and they are as follows passenger count, payment type, vendor ID, total amount, trip distance and surcharge. We have two missions to accomplish and they are as follows:

1. How much tip should the customer give the driver?
2. Which scale of generosity does the customer belong to?

While, focusing on our former mission of tip prediction, there are three stages and they are pre-processing of the data, using machine learning algorithm and finally finding the tip amount.

The later mission focuses on predicting the scale of generosity of the customer from the variable 'tip amount' which was used in our former mission. We end up finding whether the customer is generous or ungenerous.

To start with we focus on removing the outliers to make a better prediction.

OUTLIERS:

When an observation lies in an abnormal distance from the other values in the random sample of a population it is called as outliers. We need to remove these outliers for a better prediction of result. Initially we focused on individual variables and remove the outliers and here below there is a difference in the statistical measure of the same variable with and without the outliers. As we can notice that the mean, median, range, SD and variance have reduced by a large extent, making it easier to work on our prediction. Based on the outliers' visualization for with outliers we can see that there are 2 huge strikes which can destroy our prediction accuracy. We used Inter Quartile Range (IQR) method to remove the outliers.

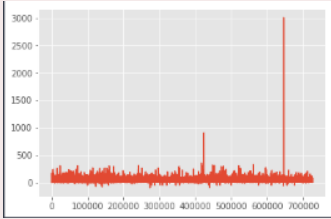
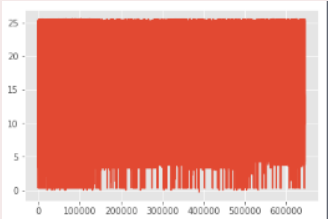
IQR method can be used as a measure to find out the spread out.

$$\text{IQR} = Q3 - Q2$$

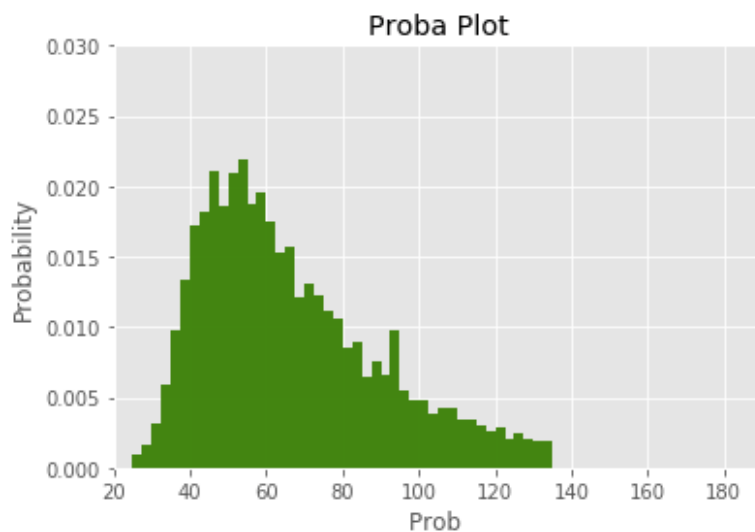
After the outliers were removed the visualization plot becomes much better for analysis. Here, we have 2 examples to show how the statistical measure got reduced.

OUTLIERS IMPACT ON VARIABLE 'TOTAL AMOUNT':

Project Report

	WITH OUTLIERS	WITHOUT OUTLIERS
MEAN	15.1	11.5
MEDIAN	11.15	10.3
RANGE	3106.65	26
STANDARD DEVIATION	13	4.8
VARIANCE	173	23
OUTLIERS VISUALIZATION		

Probability plot after removing outliers:

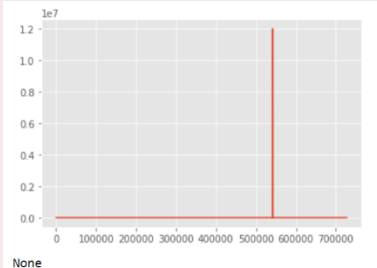
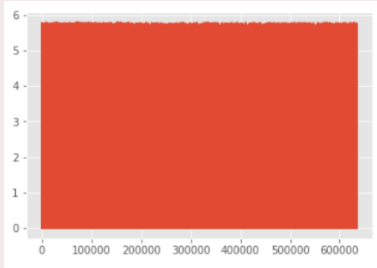


Project Report

We can see that it forms a normal distribution, having the highest probability value of .020 and X-axis ranging from 20 to 130.

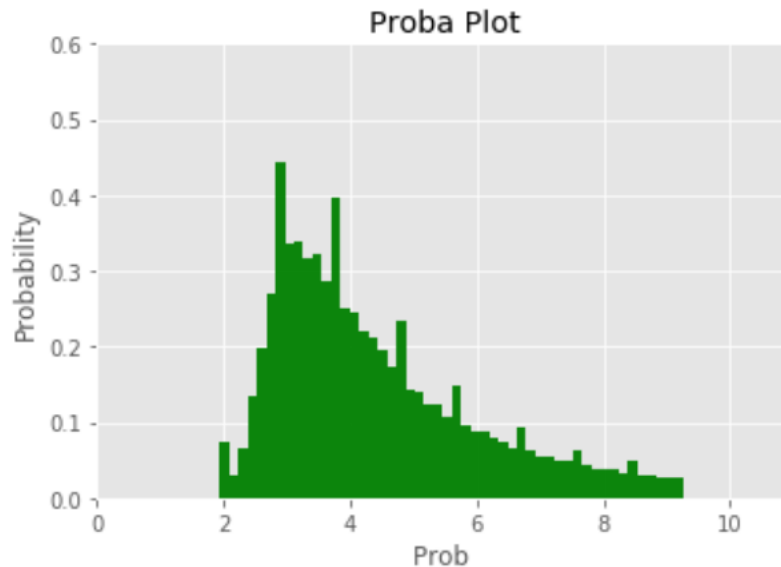
After the outliers were removed the probability plot becomes normal, making it a perfect data to predict our finding.

OUTLIERS IMPACT ON VARIABLE 'TRIP DISTANCE':

	WITH OUTLIERS	WITHOUT OUTLIERS
MEAN	19.6	1.94
MEDIAN	1.83	1.6
RANGE	12000000.0	5.8
STANDARD DEVIATION	14070.8	1.3
VARIANCE	197989493.4	1.6
OUTLIERS VISUALIZATION		

Project Report

Probability plot after removing outliers:



We can see that it forms a normal distribution, having the highest probability value of 0.40 and X-axis ranging from 2 to 10.

OUTLIERS IMPACT ON ALL THE DATABASE:

In the previous section, we removed outliers from one variable. Now, we are doing it for all the variables that we are going to work with: 'Total Amount', 'Passenger Count', 'Trip distance', 'Surcharge', 'Tip Amount'. By using the function `describe()` of a pandas data frame, and we obtained the mathematical properties of our data frame.

WITH OUTLIERS:

	Total Amount	Passenger Count	Trip Distance (in miles)	Surcharge	Tip Amount
count	727310.000000	727310.000000	7.273100e+05	727310.000000	727310.000000
mean	15.081973	1.783453	1.959882e+01	0.329787	1.280769
std	13.160194	1.361950	1.407088e+04	0.344260	2.473669
min	-100.300000	0.000000	0.000000e+00	-1.000000	-92.420000
25%	8.000000	1.000000	1.090000e+00	0.000000	0.000000
50%	11.150000	1.000000	1.830000e+00	0.500000	0.000000
75%	16.800000	2.000000	3.440000e+00	0.500000	1.950000
max	3006.350000	9.000000	1.200000e+07	9.000000	850.000000

Project Report

WITHOUT OUTLIERS:

	Total Amount	Passenger Count	Trip Distance (in miles)	Surcharge	Tip Amount
count	638764.000000	638764.000000	638764.000000	638764.000000	638764.000000
mean	11.372894	1.776689	1.994112	0.338608	0.935600
std	4.838188	1.360779	1.323293	0.347733	1.198002
min	-6.800000	0.000000	0.000000	-1.000000	-2.700000
25%	7.800000	1.000000	1.000000	0.000000	0.000000
50%	10.300000	1.000000	1.610000	0.500000	0.000000
75%	14.000000	2.000000	2.700000	0.500000	1.700000
max	34.400000	9.000000	6.120000	9.000000	6.120000

Analysis:

Initially, we can see that from 727310 observations, we deleted around 10 % of outliers. We now have 638724 rows in our data frame. Then, we observe that few properties significantly decreased when we deleted outliers. For example, the range, but even more important the standard deviation.

Indeed, when the standard deviation decreases, the variance decreases too, and we can have a better use from the data when the variance is not distorted due to outliers.

To conclude, we checked our variables distribution by plotting their density plot. All the variables we are using are now normally distributed.

USING DUMMY VARIABLE:

The Discrete variables are transformed into Dummy variables to make the analysis. All the machine learning concepts are expected to be in numerics to make the prediction easier and faster for working on the large datasets. As our dataset is large and being a real data we decided to apply dummy variables to the string variables that we used in the prediction process.

Payment Type

Initial Factors	Dummy
CSH (= Cash)	0
CRD (= Card)	1

We created two columns of dummy variables using 0 and 1 as rows, stating that 0 means it doesn't correspond to the variable and 1 means it corresponds to the variable.

Payment type has 2 outcomes: either cash or card. Card takes the value 0 and card will take the value 1.

PRINCIPAL COMPONENT ANALYSIS:

PCA is a feature extraction method, the proportion of variance is explained by the Principal components.

We use PCA to select few independent variables out of the many variables in the database. The variance is used to select the components. We want 90% variance and we decided to select the first 5 components.

COMPONENTS	VARIANCE
0	0.352
1	0.189
2	0.174
3	0.141
4	0.109
5	0.024
6	0.008

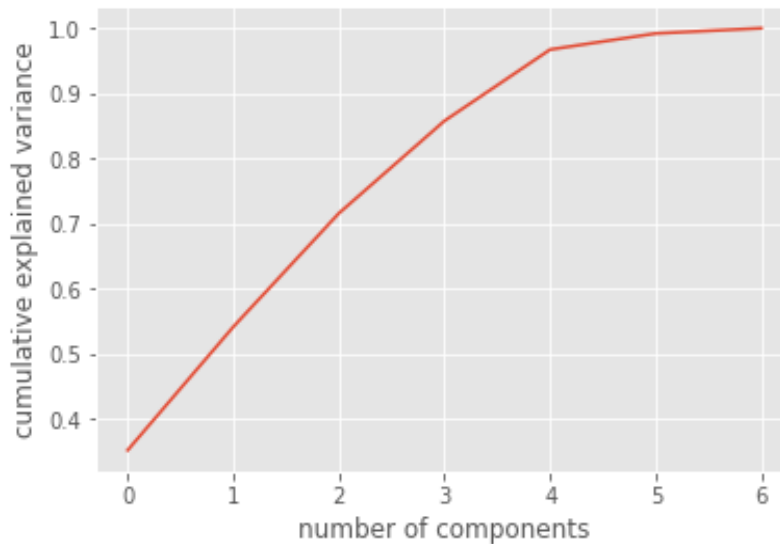
These 5 components are selected having more than 90% variance.

From the above table we have selected 0 to 4 components which has contributed for more than 90% variance.

Moreover, we used the cumulative variance plot to understand the components and select the components using the elbow method.

	Payment Type Dummy	Vendor ID Dummy	Total Amount	Passenger Count	Trip Distance	Surcharge	Tip Amount
0	0.399713	0.006327	0.563607	-0.002426	0.483532	0.095125	0.528847
1	-0.585443	-0.253372	0.347317	0.286191	0.488125	0.074535	-0.383018
2	0.251672	-0.660308	-0.141764	0.644800	-0.202416	0.025065	0.152284

Project Report



By using elbow method, we can see that the number of component equals 5 (from 0 to 4). The point at which the line starts to fall is used to predict the number of component. We can see that there are 5 components, which consist of most of the variance towards our dependent variable.

METHODS:

We used several methods like regression and classification to find our prediction and they are as follows:

MULTIPLE REGRESSION:

We use multiple regression for our analysis as we have many independent variables to find their dependency over the dependent variable.

For the prediction process we use the data after the processing steps removing the outliers and using dummy variable. If the processing of the data isn't done the accuracy would have been just 33%, which is less than flipping a coin.

Project Report

Then, we use the Multiple regression algorithms to predict our tip amount. Now, we get an accuracy rate of 77%, which is more than twice the previous prediction using the same method.

For example:

For a single passenger, who will be paying by cash 15\$, for a 2 miles trip, without surcharge or any special system, we are 77% confident that he will give the driver 1.64\$ of tip

R-Squared

R-squared:	0.596
Adj. R-squared:	0.596
F-statistic:	2.351e+05
Prob (F-statistic):	0.00

Equation and significance

	coef	std err	t	P> t	[0.025	0.975]
Total Amount	0.2290	0.000	499.133	0.000	0.228	0.230
Passenger Count	-0.1242	0.001	-151.977	0.000	-0.126	-0.123
Trip Distance (in miles)	-0.5873	0.002	-291.048	0.000	-0.591	-0.583
Surcharge	-0.4891	0.004	-134.835	0.000	-0.496	-0.482

Using adjusted R-square method and p-value to predict the dependency:

- Adjusted R-square is what is more important as it gives us an understanding of how close the data are to the fitted regression line
- Before any data processing, we had a R^2 equal to 53%; now, we have $R^2=60\%$.
- Significance: as we can see on the top, for each variable, we have $p - \text{value} < \alpha$ (with a 95% confidence interval)

The equation of our multiple regression using the coefficients:

$$y = -1.15 + 0.23(\text{Total Amount}) - 0.12(\text{Passenger Count}) \\ -0.59(\text{Trip Distance}) - 0.49(\text{Surcharge})$$

CLASSIFICATION:

To answer our problematic for the drivers, to classify customers into groups of generosity (Generous, Ungenerous), we used two classification algorithms: KNN & Logistic Regression.

We first split our data into Train & Test, 80% of the initial dataset went into the train and the 20 other % went into the test dataset. We then train our KNN model on the train, and we tested its accuracy on the test dataset.

KNN Model:

KNN stands for k Nearest Neighbors, this model classifies points based on the distance with the other closest points.

Here, we proceed to a KNN without précising the number of neighbors to take in the analysis. We let the scikit learn library choose for us.

Moreover, we took the same variables as the multiple regression, and we transform the 'Tip Amount' variable into a two factors (Generous Ungenerous) variable.

When we tested our model, we obtained 98% accuracy. This is a very high score; we are satisfied about it. For the comparison, when we ran a KNN model without preprocessing the data, for the same problematic and with the same variables, we only got around 70% accuracy. This underlie the importance of the preprocessing.

To have an idea of comparison, we also computed a Logistic regression with the exact same data (same random train and test). Our accuracy with the exact same problem is 97.7%. We conclude

Project Report

that when the problematic is easy, we can use several classification algorithm and we will have good results with it.

Let's now make things a little bit more complicated. We decided to create a scale of generosity and we tried our model to classify customers into 6 different categories.

Here is our (100% created by us, this is not real) scale of generosity:

Categories	Meaning	Range (% of total amount)
0	Very Bad Customer	0
1	Not Generous	>0% & <=2%
2	OK	>2% & <= 5%
3	Generous	>5% & <= 8%
4	Very Generous	>7% & <= 12%
5	Saint	>12%

For this problem, our accuracy is less high. Indeed, we got 93% for the KNN algorithm and only 87% for the Logistic Regression.

In this case, we will choose to use the KNN since the results with this algorithm are better. Nevertheless, when the problematic is becoming more difficult, we can see that accuracy is dropping.

To go further, we tried to improve our accuracy on this model by trying a neuronal network. The work is still in process and we are still trying out to have a better accuracy than 93%, but here are few combinations of layers we tried with their accuracy level.

Project Report

Input	Hidden Layer 1	Output	Accuracy
5	x	1	86%
6	x	1	53%
12	x	1	87%
12	2	1	53%
12	12	1	51%

Unfortunately, we didn't succeed to improve our accuracy yet but as we said before, this is a work in progress.

CONCLUSION:

We learnt a lot from this project and we have been able to notice and try many concepts covered in class.

Importance of data processing:

By comparing our results before and after the data processing phase, we underlined the importance of this phase. Either in regression and classification, our results are way less accurate when we skip this step. By data processing, there are several methods and analysis to make. We have to verify and transform the missing values, by chance, we didn't have any in our project; we have to delete outliers; this we had a lot of them to deal with. We deleted 12% of our data who have been marked as outliers by the z-score method. Moreover, we used dummy variables to make our analysis more accurate and we proceed to a PCA (Dimension Reduction technique).

Meaning of variance:

By working a lot on the outliers and on the PCA, we notified the real importance, and meaning of variance. Indeed, in the outliers management part, we were happy to notify that, with outliers, the variables had a very high variance. And when we deleted the outliers, the variance became normal.

Project Report

Concerning the PCA, we only choose 5 components instead of 7 because these 5 components were representing more than 90% of the variance.

Necessity of choosing the good machine learning model:

After all the data process part, and with a full, good dataset, we implemented models. When the problematic was easy, all the different techniques were having the same results. Nevertheless, when the problematic was more complicated, the choice of the technique was important. For example, here, KNN was better than Logistic Regression

FUTURE RESEARCH:

- Improve our current model: adding more variables to the analysis by processing all the string values, for example, the geographic information (Longitude, Latitude) or the Time & Date
- Extend our needs, answer new problem: where the driver should be at any time to increase his chances of getting “good” customers, which location a customer should avoid if he wants to find a cab, etc...
- Create other models, with other machine learning learning techniques, for example a Neural Network, or try to find a problem with unsupervised learning solution of using clustering.

Project Report

THANKYOU



A special thanks to our professor Amir H Gandomi, He gave us so many opportunities to challenge ourselves and showcase our work to him and guided us through this prediction and made our contributions recognizable. We are the luckiest students to have him as our mentor. He inspired us and gave many suggestions, which motivated us and made us to achieve such a high prediction level.

Right from the beginning he has been an incredible support for pursuing this project with the real data. Now that we are here with great prediction, it is possible only because of him.