Capital One Code Challenge Executive Summary

We had the NYC green taxi data set that's available online by the NYC taxi & limousine commission. There was some data errors in the data that needed to be fixed and this will be addressed in the assumptions section. We had about 1.5 million record and we found that the the trip distance follows a lognormal distribution. We found that the median trip distance of the hours is always less than the mean trip distance of the hour. We also find that the highest mean of trip distance is between 5AM to 7AM. We hypothesized that people tend to take a taxi in the morning so they are not late for work. We found that there is 4274 trips that went to JFK & Newark and the average fairs was $53, the peak hour of people going to the airport is at 3PM, also the trips with longest distance occurs at 5AM, the highest tip amount occurs at 5AM.

We try to predict the tip percentage from the fare amount using 9 variables from the original data set and 11 variables derived from the trip data, such as the commute time, hour of the day, day of the month, week of the month and we derived other geospatial data(zip code/county/population/area). Our base model was a linear Regression, we used also Lasso Regression and Ridge regression to prevent overfitting(also know and L1 norm & L2 norm) The measure of performance(MOP) is usually chosen according to the business requirements so i will go with mean square error as my MOP, I also report the r-squared. Our best performance is Ridge regression with mse of 51.6 and r-squared of 0.64. I found out that the model is very biased to the zero's since 59% of the data contains zeros. So I try other regression algorithms, and i find out that Random forest regressor out performs all the algorithms with mse of 31 and r-squared of 0.78. I tried another idea of classifying first if the customer will tip or no(we use only accuracy as the MOP - other metrics can be used like precision, recall, f1 score, AUC, ROC), I applied my own written function of logistic regression, and got 93% accuracy. The best accuracy I got was by Gradient boosting of 94%, After that I ran a regression only on the people who tipped. Random forest outperformed all other algorithms and got mse of 0.04 and r-squared of 0.88. I recommend using the ensemble of classification + Regression

It was interesting to know that 99.9% of the people who tipped used credit card as the type of payment, we also found that 86% of the people who paid with credit cards did tip. Also we find that most of the tip percentage lie in between 18% & 24% of the total amount, which raises of question, is there any percentage of tipping that people are used to like in restaurants? We also found that people tend to tip more on the weekends, is that because they are in a good mood? I haven't performed a test to see if they are different means though.

We performed a test to see if the speeds are different over the week of the month and we found that at least 1 mean is significantly different, we do a tukey's post hoc analysis to see which means are different and it appears that only week 2 and 3 are statistically not different. It's hard to hypothesize something about why are they different because we have a very big sample size so the power of the test is very high so the statistics is detecting the small difference in means

even though the means are between 12.8 mph and 13.2 mph, but one reason might be the weather(in my recommendation at the end I talk about that)

We test to see if the speeds are different over the hour of the day and we find that at least 1 of the means is statistically different, we also perform Tukey's post hoc analysis and we find that most of the hours are statistically different from each hours with the exception of 7 cases that is discussed in the analysis.

Finally,

This data set is very rich with information, and a lot of things can be derived from the geospatial data, I spent a lot of time trying to extract features, google api limits the number of calls to reverse geocode the latitude and longitude to get the zip code and census data, so i had to a GIS platform to grab the information and that was time consuming but I believe the information is valuable after doing that unfortunately while exporting the results of the joined data with the census data the software parsed some of the columns to a new format to read it in QGIS, and I only discovered that after I read the data back in python and I couldn't used the script for cleaning the data. So I decided to leave it at this point.

Please note that I am a full time student and i work full time and I was travelling on the weekend that this challenge was sent to me. I had a very limited time budget because of school midterms. But this is a recommendation of what i would do if i have more time:
1- I would clean the joined data to use the zip code for modeling.

2- I am not expert of how New York city is divided from the socio economic perspective but we can use an expert that has domain knowledge and try to identify the different areas from the zip codes(like Manhattan) and then use that in my model

3- For locating the zip codes I only selected the pick up location, it would be interesting to see the drop off location and compare them, are the trips in the same area? Are the trips going to a specific area has higher or lower tip percentage

4- For modeling, since we have the problem of a lot of zeros, i would be interested to test out a zero inflation model or a hurdle model which assumes that the excess zeros are generated by a separate process from the count values and the excess zeros can be modeled independently(same idea as doing classification they regression but the adjust for the overlap)

4- Due to lack of computation resources, i didn't do any grid search for the hyper parameters of the algorithms since they use brute-force methods and only used the default parameters(this is

a bad practice) I would definitely do a grid search for the hyper parameters, I included some code to do that for boosting but didn't run it.

5- I would also use K-fold cross validation on the algorithms, to make sure that the classifier is consistent across the different folds(number of folds usually between 5 and 10 is the best since it's a bias-variance trade off) i didn't use to save time in computation, but using SKlearn API it's easy to use.

6- There is a reason to believe that this a time series, for example the school schedule, UN meetings (etc..) data but this factor was not taken in consideration in the modeling for simplification, and there was no business requirement regarding time series analysis. But a good practice to include time series if we are trying to predict over the year.

7- I would take in consideration the weather in consideration using the weather api of WSI and map the weather to the trips and see if the weather is a factor in the difference of speed & tip percentage between the days of the months and the days of the week.