Final Project

Zayd Abdalla, Areeya Aksornpan

5/10/2021

Abstract

We sought to understand the tipping behavior of NYC Yellow Taxi riders from last spring, the onset of the COVID-19 pandemic in New York City. We analyzed NYC Yellow Taxi data from the months of March, April, and May of 2020. After managing and cleaning our data, we created multiple visualizations characterizing demand and tipping behavior across multiple factors such as pick-up time or fare amounts. Notably, we found that demand for taxi rides fell by 92% between March and April. Furthermore, we illustrate that high levels of taxi demand occurred between 7 AM and 10 PM, peaking between 5-7 PM. We then built two models (a simple model and an improved one) predicting tip amounts using random forests and k-folds cross validation. Our improved model yielded a roughly 12 percentage point increase in accuracy of predictions. We found that riders will generally tip taxi drivers either nothing or highly, with far fewer occurrences of small tips.

Introduction

Over the past year, the entire world had been struck by COVID-19, leaving many to stay at home in fear of the virus. New York City had become an early epicenter for the virus with over 200,000 cases between March to May. Furthermore, the city experienced historic unemployment rates of 15% and 20%, in April and May respectively. Resultantly, this combination of a public health crisis and the massive displacement of jobs left many taxi drivers with far fewer riders than a typical recession. Recent reforms such as in March 2021 have aimed to help taxi drivers with debt relief using a $65 million federal fund. Our reasoning for pursuing this field was quite simple; we believed the damage inflicted to this market had lacked national news coverage and sought to inform ourselves analytically.

Our research question is simple: what is the tipping behavior of taxi riders during this time period? Tips earned on the job can function as an additional revenue stream for taxi drivers. Given the circumstances of a global pandemic, we would like to see how accurately we can build a predictive model for estimating the tips earned using the available data. Answering this question would be beneficial in understanding consumer behavior for tipping across different factors as well as visualizing their demand for taxi rides.

Methods

Our data set contains over 3.5 million observations from last Spring (i.e. March, April, and May). The information within the data includes charges such as fare rates and surcharges, as well as trip information like number of passengers and distance traveled. We had begun our exploration of the NYC Yellow Taxi Market by managing and cleaning our data set. This process included adjusting many variables such as factorizing categorical variables (i.e. payment type, tip amount category, etc.) and even converting string pick-up/drop-off times to usable datetime formats. Importantly, we define tips by categories: zero, one, two, and high. The first three categories are dollar amounts that are typical of this market according to summary statistics. The last category considers tips greater than 2 dollars, including large tips up to 200 dollars. Following our initial data management and cleaning, we visualize the tipping behavior of consumers across different factors. This process helps us deliberate between relevant variables for our predictive model later on. We trained our data to predict tipping using random forest models and K-fold cross validation. The data set is rather large for our student computer systems, so we will draw a smaller sample of 80,000 observations to build our models from. Our simple random forest model uses 2 predictors and 500 trees. Our simple model's k-folds cross validation uses 10 folds and 5 repeats. Our improved random forest model utilizes 8 predictors and 300 trees (resource and time constrained made running with 500 problematic). Our improved k-folds cross validation uses 10 folds and 5 repeats.

Results

Figure 1. Number of Yellow Taxi Rides in NYC du

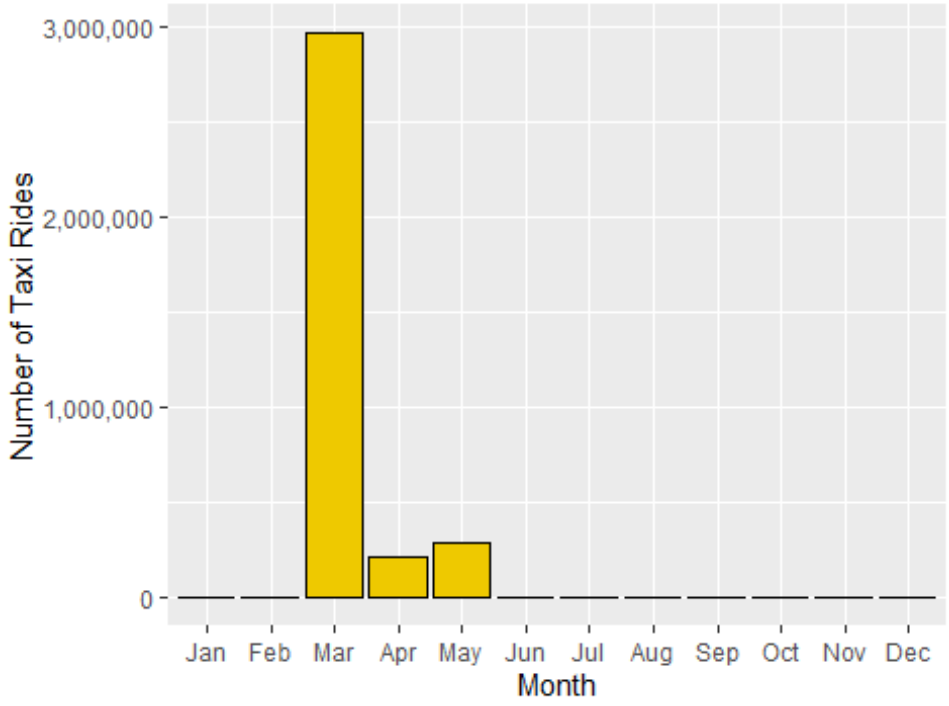
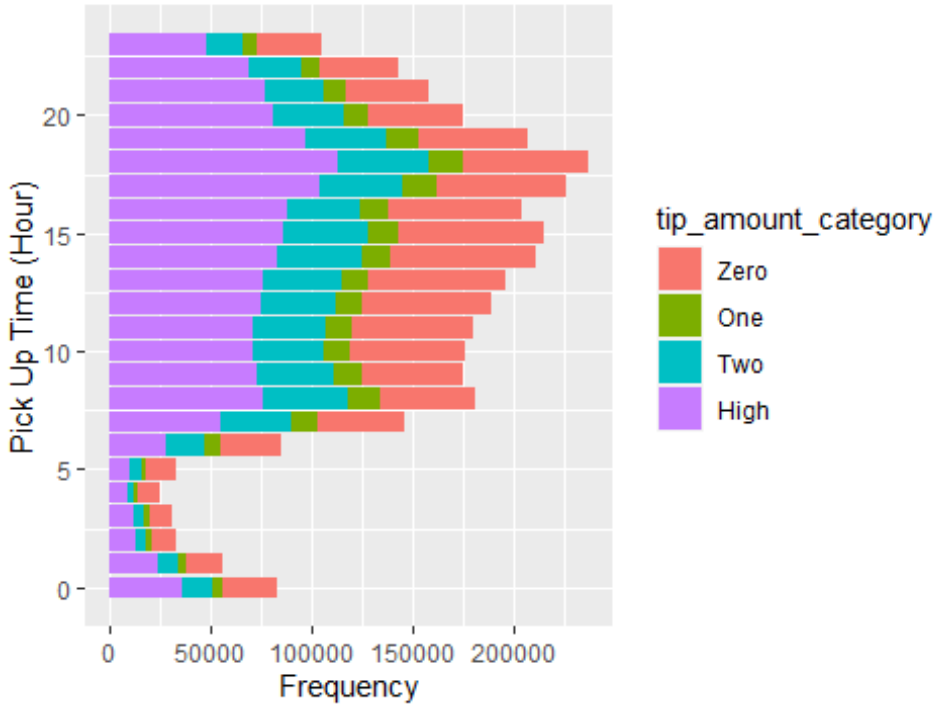
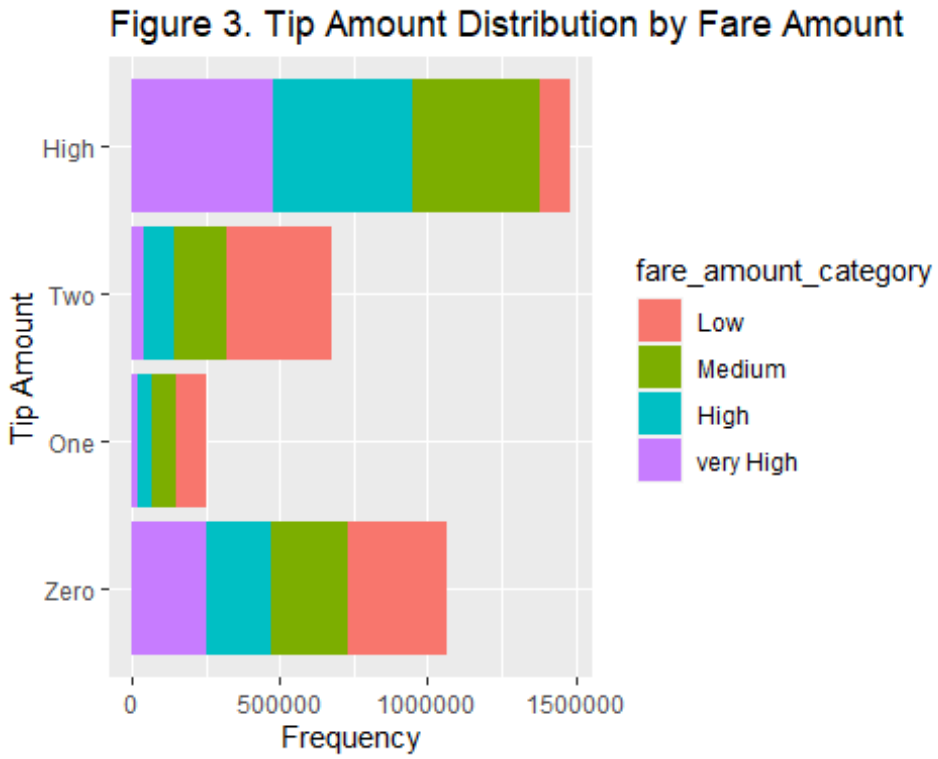


Figure 2. Tip Amount Distribution by Pick Up Time





Simple model

```
##
## Call:
##  randomForest(x = rf.train.1, y = rf.label, ntree = 500, importance =
TRUE)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 1
##
##         OOB estimate of  error rate: 32.84%
## Confusion matrix:
##       Zero One Two  High class.error
## Zero 21021   0   0  2320   0.0993959
## One      2   0   0  6526   1.0000000
## Two      4   0   0 17412   1.0000000
## High     9   0   0 32706   0.0002751

## Random Forest
##
## 80000 samples
##     2 predictor
##     4 classes: 'Zero', 'One', 'Two', 'High'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 5 times)
## Summary of sample sizes: 71999, 71999, 72000, 71999, 72000, 72001, ...
```

```
## Resampling results:
##
##   Accuracy  Kappa
##   0.6715    0.4718
##
## Tuning parameter 'mtry' was held constant at a value of 2
```

Improved Prediction Model for Tip Amount

```
##
## Call:
##  randomForest(x = rf.train.2, y = rf.label2, ntree = 300, importance =
TRUE)
##                Type of random forest: classification
##                      Number of trees: 300
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 21.34%
## Confusion matrix:
##       Zero One   Two  High class.error
## Zero 21152   3   582  1604     0.09378
## One      6   5  2916  3601     0.99923
## Two      6   1 10760  6649     0.38218
## High    38   1  1665 31011     0.05209

## Random Forest
##
## 80000 samples
##     8 predictor
##     4 classes: 'Zero', 'One', 'Two', 'High'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 5 times)
## Summary of sample sizes: 71999, 71999, 72000, 71999, 72000, 72001, ...
## Resampling results across tuning parameters:
##
##   mtry  Accuracy  Kappa
##   2     0.7865    0.6767
##   5     0.7755    0.6626
##   8     0.7335    0.6069
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was mtry = 2.
```

Conclusion

Figure 1. illustrates the number of NYC Yellow Taxi Rides by month from last spring. Clearly, there is a massive drop in rides from March to April (a roughly 92% decrease). While the graphic is fairly simplistic, it may serve as a starting point to understand

consumer and driver perspectives. From the consumer side, partaking in taxi rides is understandably hazardous given the onset of COVID-19 within New York. Contrastingly, the nature of some in-person jobs, such as construction, can not function remotely and still requires transportation. From the driver perspective, revenues—a function of quantity—are heavily decreased during this time of little demand.

Figure 2. illustrates the tip amounts a rider receives over each hour of the day. We found that the propensity for each tip amount was fairly consistent across the time of day. From our observation, it appears that most riders seem to either tip highly or not at all. Furthermore, it also serves as a visualization of rider demand by time of day. From observing the graph, we see that yellow taxi demand is quite high from roughly 7 AM to 10 PM, with peak hours occurring roughly between 5-7 PM.

Figure 3. illustrates the relationship between tip amounts and fare amounts. Interestingly, we find that tips of zero are distributed very evenly across the different levels of fare amounts. Furthermore, high tips are distributed fairly evenly for all the fare amounts except for low fares. A plausible explanation for this trend may be that short rides yield lower fares, so kinder tips are far less likely due to the simplicity of the rider's job fulfillment.

We begun our model building by starting with a very simple random forest model for tips using only 2 predictors. Our confusion matrix for this model reported an error rate of roughly 33%, which is not good at all. Our prediction of 1 and 2 dollar tip amounts seems to have a high error, with 1 dollar tip amounts having a high error regardless of the model specification we used. A plausible explanation could be because we have a low number of 1 dollar tip amounts which makes it more difficult for the model to accurately predict. Intuitively, this makes sense since 1 dollar is such a low amount, riders might opt to either just tip nothing at all or more than a few dollars generally. This explanation seems to be plausible for 2 dollar tip amounts as well, but to a far lesser extent due to the larger number of 2 dollar tip amount observations in our data. To continue our analysis, we also performed a k-folds cross validation using 10 folds and 5 repeats. Our results with this method had attained nearly identical results as before with the random forest model.

Having been informed by our exploratory analysis, we built a more improved random forest model with 8 predictors now. This time, our confusion matrix for the model reported an error rate of roughly 21%, which is far better than where we had started. Again, we found that predicting one dollar tip amounts had a high error, but the error in predicting two dollar tip amounts fell significantly. Next, we performed a k-folds cross validation using 10 folds and 5 repeats. Again, our results with this method had attained nearly identical results as before with the random forest model. In both the simple and improved model, we found that the best mtry value was 2.

We improved upon the simple model quite significantly, though adjusting for factors such as better computer resources, data involving other variables such as weather, and conducting this model in a year not marred by COVID-19 might yield more accurate results for the taxi market. Nonetheless, we can see that our model has a solid foundation for prediciting tip amounts, which are typically either zero or high. The few dollar tips (i.e. the

1 and 2 dollar tip amounts) tend to be relatively less common, which may suggest that riders typically behave by either tipping nothing or tipping highly.

Appendix

Figure 4. illustrates tip amounts by the rider's option of payment, which are Credit, Cash, No Charge, and Dispute. We separated the tip amounts into four different categories, namely Zero, One, Two, and High. We see that passengers who use their credit card to pay will also tip highly roughly 50% of the time. Alternatively, when they pay by cash, the data claims that there is no tip (zero). However, the code-book for the data claims that the data set does not record cash tips, hence we see this visual. It is likely that riders who pay in cash are including tips, but these tips simply do not appear in our data.

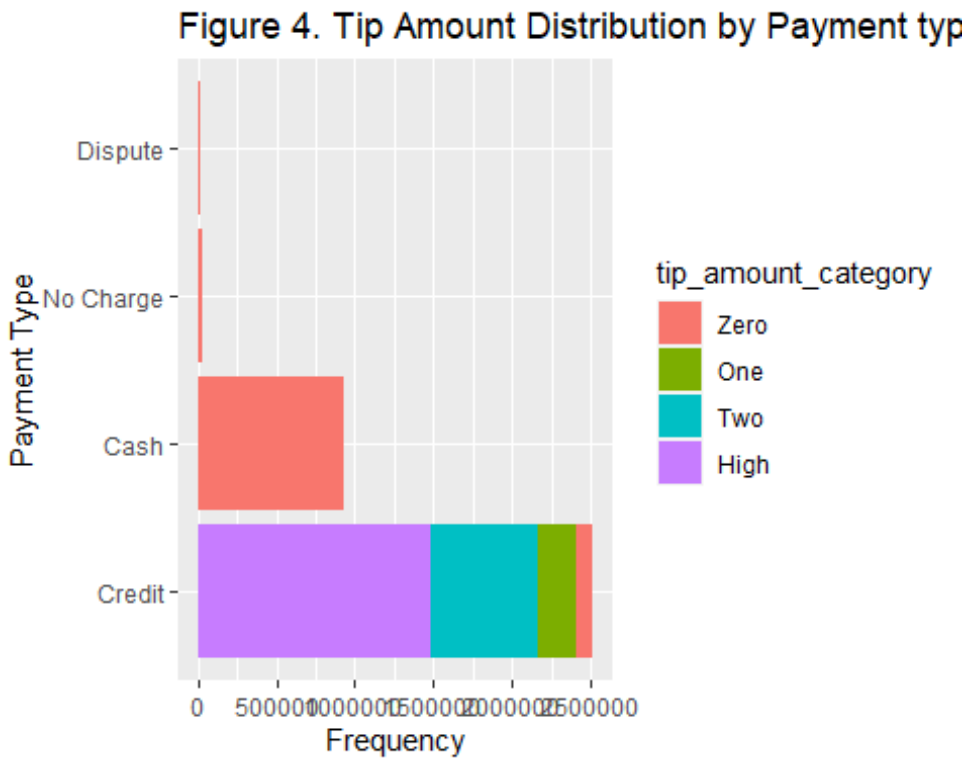


Figure 5. illustrates the tip amount by the rider's option of vendor, Type 1 and Type 2. Type 1 vendor is Creative Mobile Technologies, LLC and Type 2 vendor is VeriFone Inc. We see that riders utilize services from Type 2 vendors roughly twice as much more than Type 1 vendors. However, the type of vendor does not seem to affect the tip amounts passengers provide. We observed that most riders tip high amounts, followed by not tipping at all, then 2 dollars, and lastly 1 dollar. It appears that about 40% of riders tip high amounts, no matter the vendor type.

Figure 5. Tip Amount Distribution by Vendor