DETECTING TAXI FRAUD IN NEW YORK CITY

Abstract

A Freedom of Information Law request to the New York Taxi & Limousine Commission resulted in the release of troves of valuable information about every taxi ride in New York City during 2013. This paper uses this data and machine learning techniques to identify cases of taxi fraud in the city.

Theo Kulczycki, Santiago Larrain, Natnaell Mammo

Background and Motivation

Anytime you use a car to travel, there exists uncertainty about which route will be best. Myriad conditions--including but not limited to traffic, traffic controls, weather, construction, and the driver's skill or sense of urgency--can affect how long it takes to get to your destination and how far you have to travel to do so. However, some routes are obviously better than others, and even once a route is chosen, some driving styles will result in sooner arrival than others.

Taxi's drivers have a clear incentive to lengthen their trips. The presence of this incentive, along with the uncertainty about the best route, often results in the question that most people who have paid for a cab have probably asked themselves: is this person trying to rip me off? The question is essentially one of fraud detection. The answer requires knowledge of the relevant determinants of an optimal route to a destination, and can also be informed by observation of the past behavior of a particular driver as well as comparisons between drivers along similar routes. We apply machine learning techniques in an attempt to derive a useful metric by which to identify past trips that are potentially fraudulent, as well as drivers that appear to be committing such fraud on a recurring basis.

Related Work

Taxis are an essential form of transportation in metropolitan areas, and the industry both employs and serves a large number of people. Understandably then, issues related to taxis receive a significant amount of media attention, periodically rise to the top of political agendas in cities, and are occasionally explored in academic work. These issues typically fall in one of two general categories: transportation & urban planning, and permitting & regulation. Castro, Zhang, and Li (2012) provide an example of using large-scale taxi GPS data from China to inform urban transportation and environmental policy by modeling traffic flows. The release of similar data from the New York City Taxi & Limousine Commission (NYCT&L) has prompted similar work,

most notably by Donovan and Work (2015), who look at the resilience of urban infrastructure to extreme weather events. Those authors kindly made their data publicly available, which we utilize in this work and describe below.

The recent rise of various ride-sharing services, though, has brought to the forefront questions about passenger satisfaction, which remain relatively unexplored from either an academic or policy-making perspective. Payments for taxi rides represent a sizable portion of an urban economy, and at the same time are subject to almost unavoidable uncertainties and informational asymmetries. Jackson and Schneider (2010) demonstrated the presence and implications of moral hazard between permit owners, who own the right to operate taxis and are generally responsible for insuring the vehicles, and drivers, who typically lease the vehicles that they operate. However, the evidence for how these uncertainties affect transactions between passengers and drivers remains almost entirely anecdotal. Mathew (2005) discusses New York City drivers' resistance to the installation of GPS devices, which indicates that making available greater information about a driver's chosen routes is against their interest. We attempt to bring data-driven evidence to bear on the question of whether or not and how often taxi drivers exploit uncertainties about optimal routes to the detriment of passengers.

Problem Overview

Taxi fares are calculated according to a set combination of fixed fees, incurred once per ride for a flat rate, and variable costs, which depend on the time spent and distance traveled in the vehicle. The current fare breakdown for within city taxi rides in New York City is shown in Table 1.

Table 1: Components of fare for a standard taxi trip in NYC. The variable fees provide incentive for drivers to (fraudulently) lengthen trips.

Amount	**Description**
Fixed Fees	

$2.50	Initial charge (hailing fee)
$0.50	MTA state surcharge (tax)
$0.30	Improvement surcharge
$0.50	Nighttime surcharge (8pm-6pm daily)
$1.00	Evening surcharge (4pm-8pm weekdays)
(toll amount)	Bridge and tunnel tolls
Variable Fees	
$0.50	Each ½ mile (fast moving) or Each 60 seconds (slow moving or stopped)

The income that taxi drivers earn comes directly from the fares that they collect. If a taxi driver is also the owner of the medallion and the vehicle that they operate, then their income is simply the gross fares collected minus expense (primarily insurance, maintenance, and fuel costs). If, as is more common in NYC, the driver and permit owner are separate individuals, then the driver typically operates under a leasing agreement whereby the vehicle is leased for $75-150 per day, and in addition approximately ⅓ of the gross fares collected are paid to the permit owner. Importantly, a taxi driver is obligated to take a passenger to their chosen destination once they have been hailed. Under either arrangement, once a taxi has been hailed and a destination chosen, it is clearly in the driver's interest to choose a route that involves higher variable costs. The other fixed costs depend on the time of day and day of the week on which a trip occurs, as well as the route taken in the case of tolls.

Many things may legitimately increase the total time and distance of a trip. Before a passenger is even picked up, they might choose where to hail the taxi (for example, one side of the street or corner versus the other) and where to be dropped off in part according to what will make the trip cheaper. Once they have been picked up, the passenger may specify a preferred route other than the shortest/quickest one for a variety of legitimate reasons. Even with the best

of intentions, a driver may make excusable mistakes while driving that increase the fare compared to what it should have been. However, the driver also has the opportunity to exploit the uncertainty of the situation for their personal gain. The goal of this work is to develop a methodology to group taxi trips using a measure of similarity based on aspects of the trip that should reasonably affect the fare amount, and then identify individual trips within a group whose fares appear to be outliers among the fares of similar trips. In the next section, we describe the source of the data used in our analysis.

Data Description

The data were initially obtained by a Freedom of Information Law (FOIL) request from the NYCT&L, and thereafter made publicly available. The dataset covers the entire calendar year 2013, totaling roughly 150 million unique trips throughout the year. The data include vehicle and driver identifiers, pickup and dropoff coordinates and timestamps, and a breakdown of the fare for each trip. Approximately 45,000 licensed drivers operated roughly 13,000 permitted vehicles throughout the year. Some summary statistics are shown in Table 2.

Table 2: Summary statistics for all taxi trips in NYC in 2013. Variable fees make up at least half of the “average” trip; more for longer trips.

	Mean	**Std. Dev.**
Fare amount ($)	10	5
Trip distance (mi)	2	1.5
Trip time (s)	680	420

The distribution of fares, distances, and times is quite stable across months of the year, days of the week, and hours of the day--only the total number of trips changes significantly. An example of the frequency and distribution of fares over the course of a single day is shown in Figure 1.

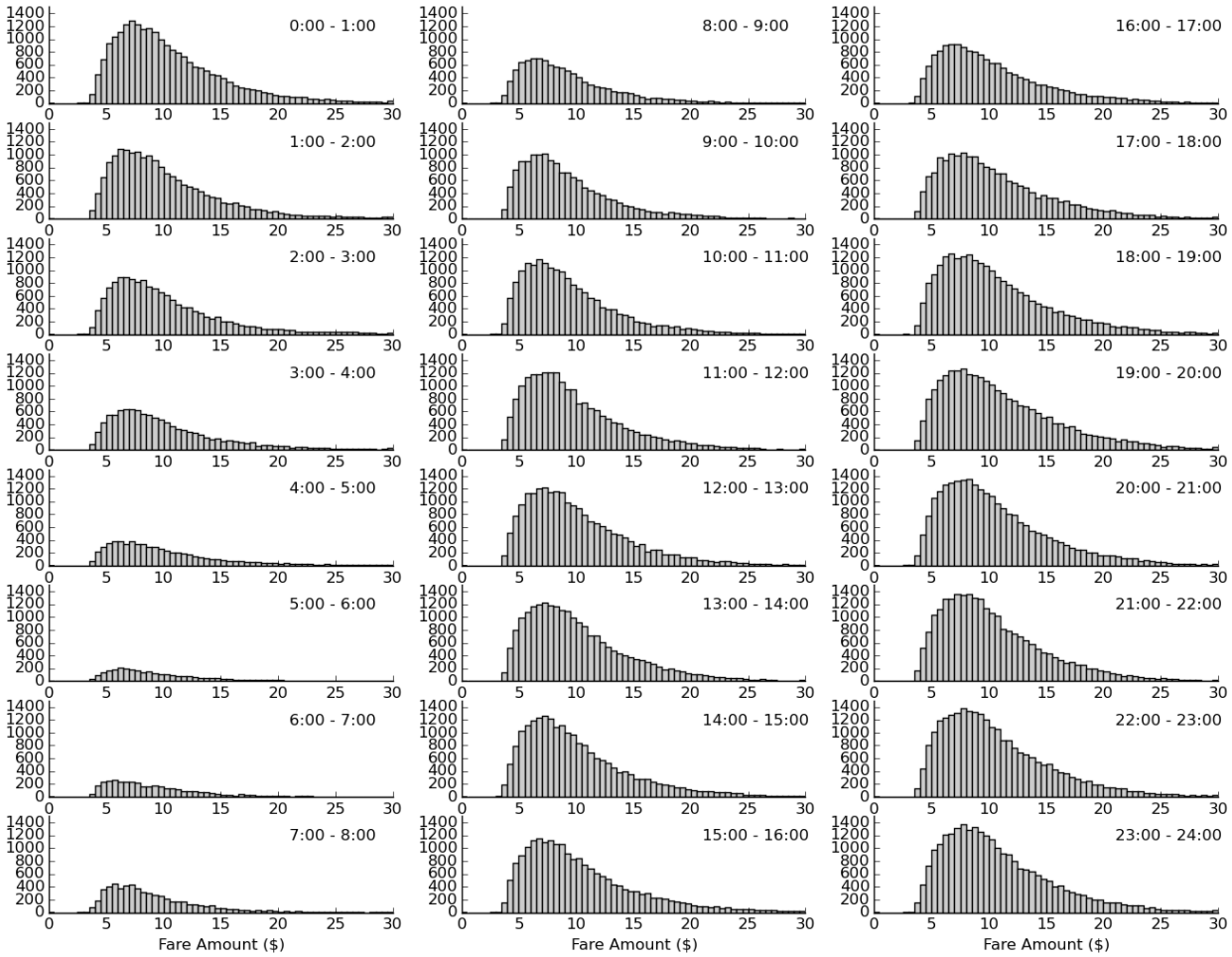

Figure 1: Distribution of fares every hour on a typical Saturday. The shape of the distribution remains relatively unchanged across hours of the day, days of the week, and throughout the year.

Data Cleaning

Since the dataset came directly from a public agency, it understandably required significant inspection and cleaning before it could be used in our analysis. Furthermore, the sheer size of the dataset made a full analysis, using the entire year of data, quite cumbersome. For reasons of practicality, we took several steps to limit the scope of our analysis by considering a smaller subset of the data.

Among the data errors encountered were: duplicate records, missing data, apparent data entry errors, and records that did not represent actual taxi trips. We removed observations

with missing or incorrect locational data--every entry that had either the latitude or the longitude equal to zero for either the pickup or the dropoff location. When observations were duplicated (more than one rows with exactly the same data for each column), we removed the duplicate rows and kept a single row representing that data. Many observations were given an MD5 encoded hack license equivalent to zero--we dropped these observations. We also removed all trips that lasted less than a minute or were less than 0.3 miles long, since many of these represented apparent accidental meter operations on the part of the driver.

Finally, we removed every trip that originated or ended outside of Manhattan, and rounded every GPS coordinate to the 3rd decimal (representing a window of roughly 100 meters). These data operations resulted in about 10% of the observations being dropped from the original dataset. We then selected a single month worth of data (October, 2013) on which to perform the analysis. The result was a set of trips that was small enough to be analyzed in the allotted time, with trips that were similar enough to be compared and also numerous enough to obtain adequate precision in our models.

Solution Methodology

Our approach consisted of examining the results of both supervised and unsupervised learning methods applied to a single month worth of data, in order to detect trips for which the fair appears to be fraudulent and drivers associated with an abnormally high number of such trips, using metrics developed and discussed below. Our definition of fraud was limited, and includes only efforts made by the driver to increase the fare by selecting routes or using driving styles that make the trip longer in terms of time or distance than otherwise necessary.

In the unsupervised context, we applied three clustering algorithms—Mini-batch K-means, Mean Shift, and Density-Based Spatial Clustering of Applications with Noise (DBSCAN)—on a set of features including the pickup latitude and longitude, hour of the day, day of the week (weekday/weekend), and number of passengers. The figure below shows the

various parameter combinations that we used to test the appropriateness of the clusters generated by each algorithm; the parameters were chosen based on the objective of minimizing the total within group sum of squares of all clusters.

Table 3: Set of values used in the unsupervised parameter sweep. Underline indicates parameters that yielded the lowest WGSS.

Mini-batch K-means	Number of Clusters: 100, 200, 500, 1000, <u>2000</u>, 2500
DBSCAN	Neighborhood Size: 50, 100, <u>200</u>, 500, 1000
Mean Shift	Bandwidth: <u>1</u>, .01, .02, .03

To identify potentially fraudulent trips, we examined the distribution of fares within each cluster. Trips are considered fraudulent based on the deviation of the associated fare amount from the within-cluster mean fare. This score is interpreted as an intensity of the belief that a particular trip involved fraud on the part of the driver. A measure of confidence is assigned based on the within-group sum of squares (WGSS) of a trip's associated cluster. We are more confident about identifying outliers in clusters with a smaller WGSS, because the trips within that cluster tend to be more similar than trips within clusters with a higher WGSS.

In the supervised context, we tried different regression models: linear regression, K-Nearest Neighbor, Decision Trees, Random Forests, Bagging and Boosting to predict the fare for each trip (the "fare_amount" column, not the total value that included taxes and tips). Using accuracy as the evaluation metric (precision and recall doesn't make much sense here since is not a classification problem) Boosting scored the highest. From there, we iterated through a few different parameters to determine the best ones to use, as shown in Table 4.

Table 4: Set of parameters used in the parameter sweep for boosting (9 combinations). Underline indicates parameters that yielded the highest accuracy.

Number of Estimators	10, <u>100</u>, 1000
Maximum Depth	3, 5, <u>7</u>

We initially used the trip distance and the trip time to predict its fare, but we changed the approach and used the same features that were used in the unsupervised algorithms. This resulted in an accuracy score of 72.7%. Since the data still included potentially mislabeled information, we removed cases where the difference between the predicted and the real fare was more than $100, since it is highly likely that was due to a data entry error.

With the fare prediction for each trip, we calculated the difference between real and predicted fare, then scaled it, normalized, and finally aggregated it by driver. Once all trips have been scored using the various methods, we examined the frequency and intensity of fraudulent trips associated with a particular driver and compare the scores given to drivers by different trip scoring methods. Drivers that are ranked above the 95th percentile (in terms of the weighted average of their associated trip scores) by multiple methods are identified as suspicious drivers.

Results and Discussion

The below table contains the correlation between the four scores assigned to each driver by the models. As evidenced in the table, Mini-batch K-means, Mean Shift, and Boosting have highly correlated scores. DBSCAN discards trips that it assigns as noisy, meaning that there are not enough similar trips with similar characteristics. For this reason, it has a lower correlation.

Table 5: Correlation matrix for the driver scores generated by each method.

	K-Means	*Mean Shift*	*DBSCAN*	*Boosting*
K-Means				
Mean Shift	0.943717			
DBSCAN	0.704623	0.498043		
Boosting	0.958703	0.941167	0.682418	

Since each driver's score represents the strength of the belief that they are engaging in repeated fraudulent activity, we look for suspicious drivers in the group with the highest scores. However, since we do not ever know the truth about which trips included fraud, and therefore since neither the unsupervised or supervised scores can be considered definitive in their detection of fraud, we must compare the results of the methods to each other in order to detect "agreement" on the likelihood of fraudulent activity for each driver. Figures 2, 3, and 4 show, for drivers whose score is above a certain percentile according to one method, the proportion of drivers whose score is above the same percentile according to a different method. We achieve the highest agreement between the scores produced by K-Means clustering and Boosting. Focusing on drivers with scores above the 99th percentile according to these two methods, we identified 252 drivers that appear to be the most suspicious in terms of the amount of potentially fraudulent activity. The appropriate next step would be to investigate the details of these drivers' behaviors, but this is beyond the scope of the current project.

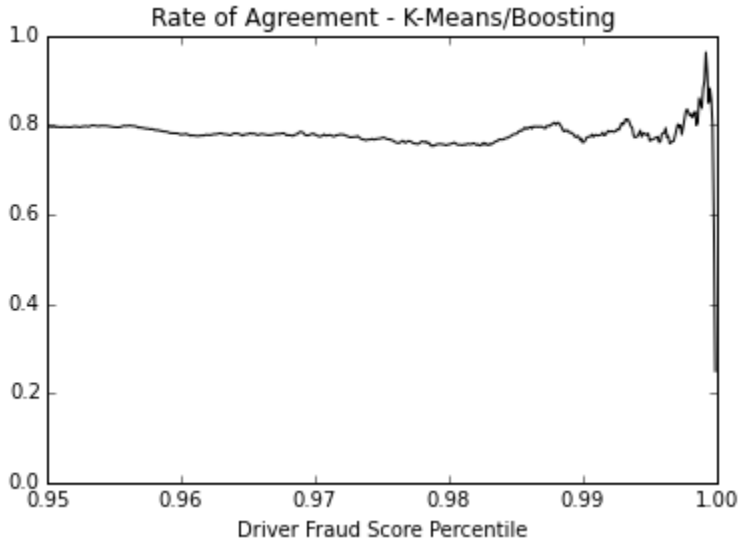

Figure 2: Over 90% of drivers with a K-Means score above the 99th percentile also had a Boosting score above the 99th percentile.

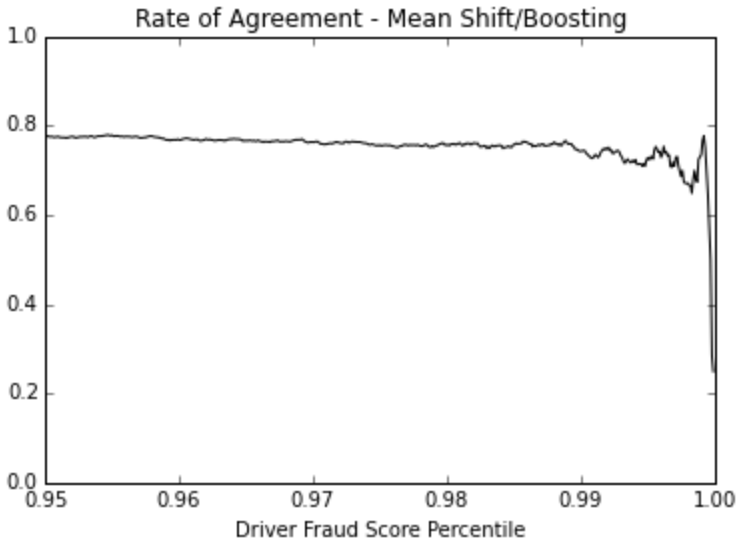

Figure 3: Less than 80% of drivers with a Mean Shift score above the 99th percentile also had a Boosting score above the 99th percentile.

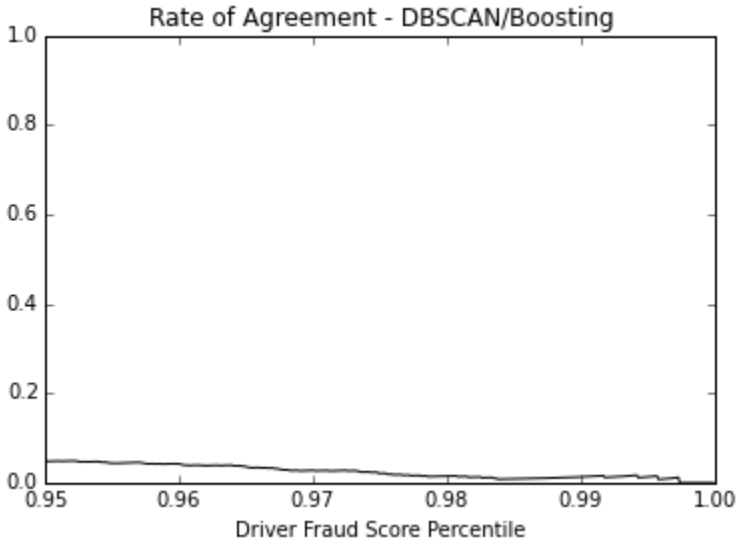

Figure 4: Very low agreement in driver percentiles between DBSCAN and Boosting.

Limitations, Recommendations, and Future Work

The fundamental problem with the question we are trying to answer is that the answer is unknown. We have no way of verifying whether or not a particular trip was truly fraudulent, and therefore whether I driver is being flagged as suspicious for good reason. This difficulty primarily

arises because we cannot, using the data available, identify a (reasonably defined) optimal route for each trip to be compared to. Data describing the trips taken by livery cabs in NYC would be particularly useful toward this end, since the livery cabs charge a flat rate for getting from one neighborhood to another and therefore do not face the same incentive to elongate their trips. One possible evaluation would be to replicate this analysis on years of data where there are substantial known cases of taxi fraud and observe whether this analysis would flag those drivers. In order to consolidate all of this information into one easily digestible deliverable, we have prioritized identifying the drivers that scored in the top 1% of displaying fraudulent behavior; this list of 131 drivers along with the code to produce this report has been provided on github at NYC-Taxi-Fraud. We hope that these results and more importantly, this type of analysis will become used in the Taxi & Limousine Commission and throughout the government of New York City.