How Safe is Your (Taxi) Driver?

Rade Stanojevic
Qatar Computing Research Institute, HBKU, Doha, Qatar

ABSTRACT

For an auto insurer, understanding the risk of individual drivers is a critical factor in building a healthy and profitable portfolio. For decades, assessing the risk of drivers has relied on demographic information which allows the insurer to segment the market in several risk groups priced with an appropriate premium. In the recent years, however, some insurers started experimenting with so called Usage-Based Insurance (UBI) in which the insurer monitors a number of additional variables (mostly related to the location) and uses them to better assess the risk of the drivers. While several studies have reported results on the UBI trials these studies keep the studied data confidential (for obvious privacy and business concerns) which inevitably limits their reproducibility and interest by the data-mining community. In this paper we discuss a methodology for studying driver risk assessment using a public dataset of 173M taxi rides in NYC with over 40K drivers. Our approach for risk assessment utilizes not only the location data (which is significantly sparser than what is normally exploited in UBI) but also the revenue, tips and overall activity of the drivers (as proxies of their behavioral traits) and obtain risk scoring accuracy on par with the reported results on non-professional driver cohorts in spite of sparser location data and no demographic information about the drivers.

ACM Reference format:
Rade Stanojevic Qatar Computing Research Institute, HBKU, Doha, Qatar. 2017. How Safe is Your (Taxi) Driver?. In *Proceedings of CIKM'17 , Singapore, Singapore, November 6–10, 2017,* 4 pages.
https://doi.org/10.1145/3132847.3133068

1 INTRODUCTION

Risk assessment using demographic data has been studied in depth and the lessons learned have been widely deployed by underwriters in the car insurance market [1, 6]. However, using solely demographic information to assess the risk of an individual driver limits the potential for differentiating between the safe and risky drivers [10]. With the proliferation of GPS-enabled vehicles and devices, a new opportunity for profiling risk is becoming available with a promise to significantly improve the risk prediction power.

Several studies have examined the impact of mobility [3, 8] (e.g. mileage) or behavioral [1] (e.g. credit rating) features on the car insurance risk. Similarly, actuaries working directly with the underwriters, reported significant predictive power of the location data in the context of risk modeling [10]. However, due to the extreme sensitivity of the location and accident information, the datasets used by these studies are kept confidential which prevents their reproducibility and limits the interest of data mining community.

In this paper we report our ongoing work on understanding the power of location and behavioral data for car insurance. Firstly, we point out that publicly released dataset of 173M NYC Taxi journeys can be used to profile both the driving styles of 40K drivers and also their involvement in (serious) accidents, effectively allowing to examine the relationship between driving patterns and the accident risk. Secondly, we train several models for accident prediction and show that they exhibit similar prediction accuracy to previously reported studies in spite of using much sparser location data. And finally, in contrast with previously studied UBIs which examine personal-car drivers, we here focus on professional drivers and demonstrate that using taxi-related information, such as tip-ratio, or occupancy-rates, remarkably improves risk-assessment quality (measured through decile-lift; see Section 3).

2 DATA

The data we use is collected by New York City Taxi and Limousine Commission (NYC TLC) and was made public in 2014 after a FOIL (The Freedom of Information Law) request by Chris Whong [7]. The data contains information from all journeys operated by the NYC yellow taxi during 2013, in total around 173M journeys. For each journey the following info is available: *medallion_id* (unique vehicle identifier), *driver_id* (unique driver identifier)[1], pickup timestamp, dropoff timestamp, pickup location (latitude/longitude), dropoff location, mileage, fare amount, tip amount, and payment type (cash or card). In total, there are around 13K yellow taxi vehicles shared by around 40K drivers to generate around 500K journeys per day.

This dataset has been studied in the past in diverse domains from traffic monitoring [12] to ride-sharing [18]. Here, we utilize the data in a different way. On one hand we profile the journey data to create an array of features which represent driving and behavioral characteristics of the drivers. On the other hand, we use the same dataset to create the labels for the accidents. Namely, most of the taxi licenses (medallions) in NYC are owned by companies who lease them to the drivers and due to the NYC TLC rules such medallions must be active two shifts per day, 365 days per year, as long as the vehicle is operational [14]. By observing anomalous disappearance of a medallion from the dataset, we can effectively identify the appearance of a major disruption and the driver who operated the vehicle when it happened.

[1]Both medallion and driver identifiers are anonymized. While there are reports which describe de-annomization of medallion and driver id's [17], we do not attempt to de-annonimize the identity of the drivers/medallions at any stage in this work.

CIKM'17 , November 6–10, 2017, Singapore, Singapore

ACM ISBN 978-1-4503-4918-5/17/11…$15.00
https://doi.org/10.1145/3132847.3133068

2.1 Feature extraction

As we explain above, we use the journey data to build a profile of the driver which captures his/her driving style and behavioral profile. In particular, for each driver we extract the features described in Table 1. The features related to location can be (and often are) extracted by the tracking systems used by standard UBI products. However, features related to billing (revenues, tips, cash-rides, etc.) are specific to taxi drivers and are a characteristics of our driver cohort not shared by non-professional drivers.

Note that UBI products normally track much richer data with location sampled frequently (from once per second to once per minute) together with accelerometer and gyroscope sensor readings which allow inference of speeding, harsh events (acceleration and deceleration) as well as harsh turns [6, 8]. The data we study is much sparser, with only two location points per journey (start and end).

In Section 4 we discuss potential additional features which we believe could refine profiling of the driving style.

In Figure 1 we depict the empiric CDFs of four representative features: `mileage per shift`, `shortmiles-ratio`, `revenue-per-shift`, and `tip-ratio`. Mileage is by far the most widely used factor in UBI and is known to have strong prediction power for inferring risk levels [3, 8]. Due to professional nature of our drivers we observe relatively low variability of miles driven by them: 10th and 90th percentile are 40.9 and 71.4 miles per day. In general population, mileage is much more variable: [3] reports 10th and 90th percentile mileage per day (in their dataset) at 19km (11.8 miles) and 76.4km (47.5 miles). Intuitively, with lower variability, the power of the mileage feature for predicting the accidents is likely to be reduced, yet we still see a non-trivial effect of mileage on the accident frequency (see Sec. 3). The `shortmiles-ratio` features measures the fraction of miles which belong to short trips (those <3miles) and we see that more than half of all miles are driven on such short trips which are typically in urban environments which have higher accident risk levels compared to rural and highway roads [8]. The variables `revenue-per-shift`, and `tip-ratio` are proxies of the capability of the driver to generate revenues and could potentially reflect the characteristics of the drivers' personality which correlate with her safety behind the steering wheel.

There are some inconsistencies in the data and we filter out all the journeys with average speed (defined as the ratio between the length and duration of the journey) less than 3mph or greater than 100mph. This filtering eliminates less than 1% of all journeys.

2.2 Accident labels

In this paragraph we describe the process of inferring the anomalous disappearance of the vehicles which we attribute to the accident caused by the driver who drove the vehicle immediately before disappearing.

For given *medallion_id* we construct a bit-vector of length 365, with bit *i* indicating whether it was active on *i*-th day of the year or not. By *maxgap*(*medallion_id*) we denote the length of the longest sequence of zeros in the activity bit-vector: the largest number of consecutive days the vehicle was not in service. With *n_drivers*(*medallion_id*) we denote the number of different drivers which have operated the vehicle *medallion_id* at least once in the year.

Location features:	
`mileage`	total number of miles
`#trips`	total number of trips
`#shifts`	total number of shifts
`trip-duration`	total duration of all trips
`shortmiles`	mileage on short trips (<3miles)
`shortmiles-ratio`	ratio: `shortmiles/mileage`
`miles-per-trip`	`mileage/#trips`
`miles-per-shift`	`mileage/#shifts`
`duration-per-shift`	`trip-duration/#shifts`
`avgspeed`	`mileage/trip-duration`
`first-last-day`	# days between first and last ride
`old-driver`	binary (1: active in Jan'13, 0: otherwise)
Billing features:	
`revenues`	total revenues (fare+tips)
`tips`	total tips
`cashrides`	fraction of rides paid in cash
`shifts-per-week`	avg number of shift per week
`tip-per-trip`	`tips/#trips`
`revenue-per-shift`	`revenues/#shifts`
`tip-ratio`	`tips/revenues`

Table 1: Features extracted from the data.

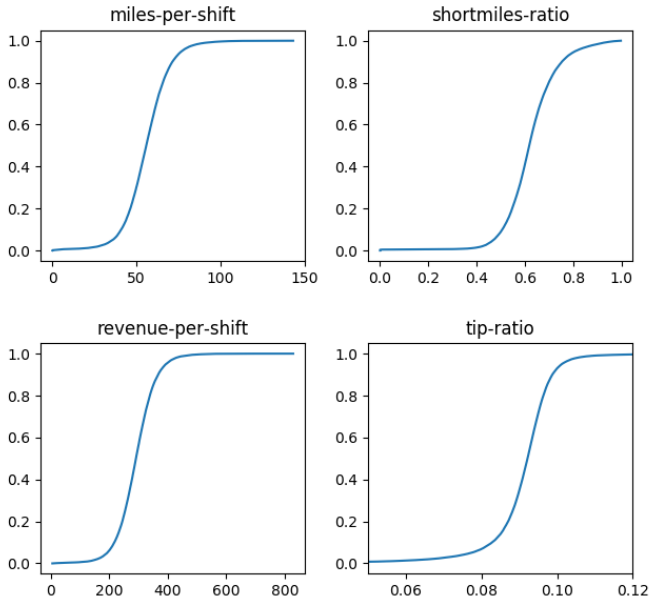

Figure 1: Epiric CDF for four representative features.

Given the high medallion value (leased for 200-250USD per day [15]), it is in the best interest of the medallion owner to keep it active as long as the vehicle is operational. Additionally, corporate medallions (also known as mini-fleet medallions and which represent a dominant majority of all medallions in NYC) *must* be in function 365 days per year and are otherwise penalized [14]. We label an accident in the event a vehicle is inactive for several days. An important consideration here is the fact that while most medallions are owned by professionally managed taxi companies (mini-fleet), there is a fraction of medallions who are privately owned and operated by a taxi driver and typically have small *n_drivers*(*medallion_id*). We notice that *medallion_id*'s with very small *n_drivers*(*medallion_id*),

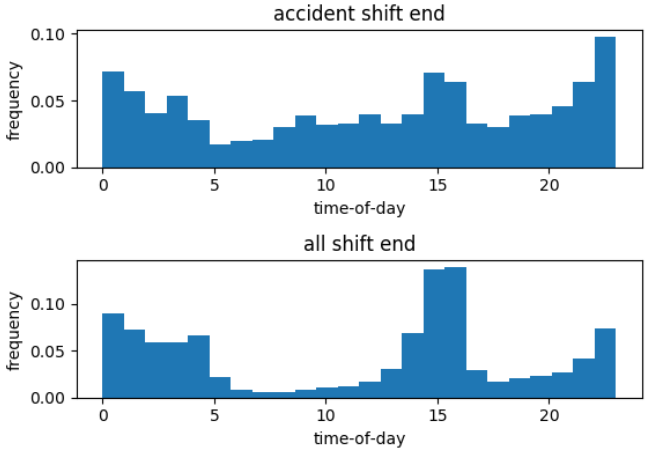

Figure 2: Histogram of time-of-day shift ending: accidents (top) and all (bottom). Most shifts end in two time intervals: 14-17h and 23-05h. Accident shift ends are spread more uniformly throughout the day.

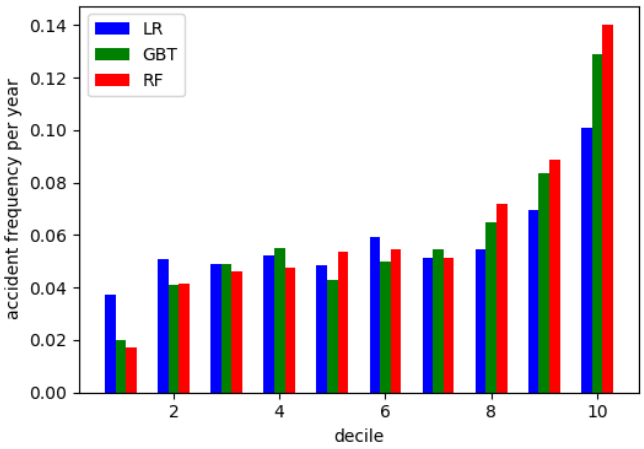

Figure 3: Per-decile accident frequency per year. In all three methods the decile predicted to be the riskiest (#10) generates significantly more accidents than the others.

tend to have more (and longer) gaps. Hence, we declare an accident of *medallion_id* if:

$$maxgap(medallion\_id) \geq \delta_1 \text{ , and, } n\_drivers(medallion\_id) \geq \delta_2,$$

where δ_1 and δ_2 and parameters which determine the overall number of caught accidents. The first condition aims to filter out the medallions owned by individuals which normally are shared by a small pool of drivers. The second condition eliminates minor issues which may not be caused by serious accidents. For most of the analysis in this paper we chose $\delta_1 = 4$ and $\delta_2 = 4$ which results in 1931 medallions with an accident. We identify the driver responsible for the accident simply as the last driver associated with the medallion prior to the inactivity gap.

Since we do not have the ground truth accident data (and it is unlikely that such data is made publicly available due to privacy implications of the involved drivers) it is difficult to say how many of the labeled accidents are true accidents and what fraction is caused by other factors (e.g. mechanical issues, expired license, etc.). To demonstrate that most of the labeled accidents are caused by irregular events we plot the distribution of accident shift ending throughout the day and compare it with the distribution of the shift ends for all shifts, see Figure 2. While most shifts end in two time intervals 14-17h and 23-05h, the accident shifts ends are spread more evenly throughout the day suggesting the premature end of most shifts we labeled as accident.

While we certainly have errors (both false negatives and false positives) in inferring the actual accidents using the simple technique described above, we believe that we do capture most of serious accidents and that such (noisy) labels still allow for detecting relevant factors associated with the risk.

3 SUPERVISED LEARNING

To understand the power of the factors discussed in Section 2.1 on predicting the accidents, we pose the binary classification task of predicting whether the driver was involved in one of the labeled accidents. For both training and testing the model we use the features extracted for the whole year. We experimented with three classification methods to evaluate which one performs best in our specific problem: L2-regularized logistic regression (LR), random forest (RF), and gradient boosted trees (GBT). For each of the studied methods we select the hyper-parameters automatically using a simple grid-search over a validation set sampled from the training data (sample of 20% of the training cohort). We use fivefold cross-validation to predict the likelihood of having an accident label, and the tested fold is not used in the training phase at any instance.

To evaluate the impact of different features we use two sets of the user features. **Location** refers to set of the features which could be extracted from the location alone, while **Location+Billing** refers to the whole feature set including both location as well as billing features which are likely to carry certain information on the user behavior which may correlate with the risk factors; see Table 1.

We evaluate the *decile-lift*, a metric often used by actuaries to measure the quality of the risk assessment process. For given risk scores, we rank the drivers according to the risk scores, and split them in ten deciles: top-10% of least-risky predictions go in decile 1, the following 10% drivers go to decile 2, and so on. Within each decile we evaluate the frequency of accidents and report the average frequency per decile. Then, *decile-lift* is the ratio between the accident frequencies od the riskiest and the safest deciles.

For each tested fold, and each driver in the tested fold we evaluate the probability p_i of having the accident label using the respective trained model. Since not all of the drivers have been active during the whole year we rank the drivers not on p_i but rather on the probability of an accident per day during the period the driver is active: p_i/T_i, where T_i is the number of days between the first and last journey of the driver. Similarly, when evaluating per-decile accident frequencies we evaluate the number of accidents normalized per active year. Thus, for a decile with x accident labels and average active period of $\bar{T}$ days, the accident frequency per year is $x \cdot \frac{365}{\bar{T}}$.

In Figure 3 we report the accident frequency per year within each risk decile averaged over all 10 folds. Decile 10 (the riskiest decile),

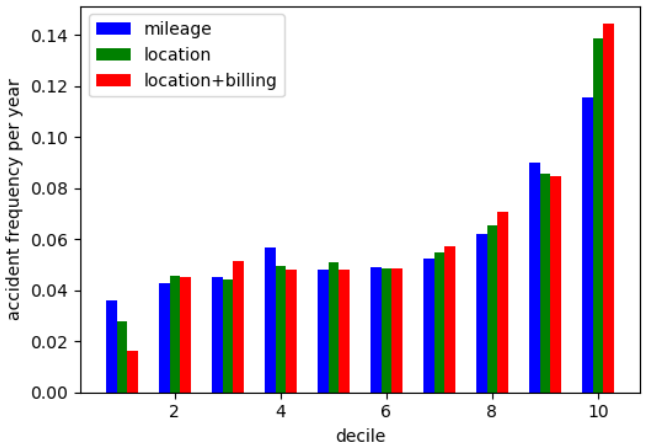

Figure 4: The impact of mileage-only, location and billing features on decile lift.

expectedly, contains drivers which have more accidents that drivers from other deciles. However, the classification method used for ranking the drivers' risk results in different levels of differentiation. Decile-lift (ratio of accident frequencies between the top and bottom deciles) in LR-ranking is only 2.5, while the decile lift using random forests ranking is around 8.5.

Here we would like to stress that traditional risk scoring which relies on demographic data alone has decile lift in the range between 2.5 and 3 [1, 10]. However, when augmented with location data [2] albeit without billing information and on a non-professional driver cohort actuaries have reported decile lift of around 10 [10], not very far from RF-based scoring which achieves 8.5 with significantly less location/speeding/acceleration data.

To evaluate the impact of billing-related features on the ability to differentiate the risky from the safe drivers we run the RF classification using three different sets of features: *mileage only* which ranks driver using only overall mileage, **Location** and **Location+Billing** as the whole feature set (see above). Mileage only results in the decile lift of around 3, the **Location** features result in the decile lift of around 4.9 while the the whole feature set has decile-lift of 8.5 indicating that billing features carry non-trivial power for differentiating between the risky and safe taxi drivers.

4 DISCUSSION

With more than a billion vehicles globally, motor insurance is a large industry generating around USD 500B of revenues globally. However, due to heavy competition most underwriters have very low profit margins, well under 10% [11]. As claim/accident costs dominate the costs of the underwriters (around 60-80% of the revenues go into servicing claims [11, 13]), understanding and predicting the risky drivers is of paramount importance for structuring a profitable portfolio.

In this paper we examine the potential of using location as well as billing information to predict the potential risk of professional taxi drivers. While professional drivers constitute relative minority of drivers, they do drive significantly more than non-professionals, and consequently pay more for insurance. For example, in New York the average personal car premium is around USD 1200, while average taxi driver premium is in the range of USD 5000-10000 [16].

With the proliferation of ride-sharing services such as Uber or Lyft, an increasing number of people start driving (semi-)professionally. We believe that our study can help design appropriate insurance packages for them either offered by the ride-sharing platform itself, or with cooperation between the platform and the underwriters.

We would like to conclude with a number of questions which remain open. First, we have a fairly primitive way of inferring accidents which most certainly has plenty false positives/negatives. Having a more sophisticated anomaly detection approach for inferring accidents would significantly improve the confidence of our methodology. Second, we strongly believe that richer features either derived from richer data (e.g. sampled more frequently than twice per journey) or inferred from the existing data (e.g. speeding behavior using expected vs. actual journey time) could add additional power to the proposed risk-scoring methodology. And finally, in this work we use off-the-shelf ML methods, which may be sub-optimal for the insurer. It would be very interesting to formally capture the metric of interest which the insurer may want to optimize (such as decile-lift) and then develop ML techniques which explicitly optimize it.

[2] Which is normally much richer than origin-destination (lat, lon) pairs available to us and includes fine-grained location, speed, accelerometer readings for in-depth profiling of the driving style.

REFERENCES

[1] Federal Trade Commission. "Credit-based insurance scores: Impacts on consumers of automobile insurance". A Report To Congress by the FTC, July 2007.
[2] S. Lee, A. Katrien. "Why High Dimensional Modeling in Actuarial Science?". IACA Colloquia 2015.
[3] J. Lemaire, S.C. Park, K.C. Wang. "The use of annual mileage as a rating variable". ASTIN Bulletin vol. 46(1) 2016.
[4] SAS Institute. "Using Data Mining for Rate Making in the Insurance Industry Written". June 2003
[5] Y. Kahane, N. Levin, R. Meiri, J. Zahavi. "Applying Data Mining Technology for Insurance Rate Making: An Example of Automobile Insurance". Asia-Pacific Journal of Risk and Insurance, vol. 2(1), 2007.
[6] J. Paefgen, T. Staake, F. Thiesse. "Evaluation and aggregation of pay-as-you-drive insurance rate factors: A classification analysis approach". Decision Support Systems, vol. 56, 2013.
[7] C. Whong. "Foil NYC Taxi". Available Online. http://chriswhong.com/open-data/foil_nyc_taxi/
[8] M. Ayuso, M. Guillen, A.M. Perez-Marin. "Time and distance to first accident and driving patterns of young drivers with pay-as-you-drive insurance". Accident Analysis and Prevention, vol. 73, 2014.
[9] Available online. https://uberpeople.net/threads/what-do-you-pay-for-commercial-insurance-in-nyc.9665/
[10] B. Rooney. "Why Car Insurance Has Yet To Be Disrupted". http://www.informilo.com/2015/06/why-car-insurance-has-yet-to-be-disrupted/
[11] P. Desyllas, M. Sako. "Profiting from business model innovation: Evidence from Pay-As-You-Drive auto insurance". Research Policy, vol. 42(1), 2013.
[12] B. Donovan, D.B. Work. "Using coarse GPS data to quantify city-scale transportation system resilience to extreme events". arXiv:1507.06011, 2015.
[13] T. Stormer. "Optimizing insurance pricing by incorporating consumersâĂŹ perceptions of risk classification". Zeitschrift fÃijr die gesamte Versicherungswissenschaft, 104(1), pp.11-37.
[14] NYC Taxi and Limousine Commission. "2014 Taxicab Factbook". 2014.
[15] NYC Taxi and Limousine Commission. "Notice of Promulgation of Rules". 2012. http://www.nyc.gov/html/tlc/downloads/pdf/lease_cap_rules_passed.pdf
[16] Independent insurance agents. "How Much Does Commercial Vehicle Insurance Cost?". https://www.trustedchoice.com/commercial-vehicle-insurance/compare-coverage/rate-cost/
[17] V. Pandurangan. "On Taxis and Rainbows". https://tech.vijayp.ca/of-taxis-and-rainbows-f6bc289679a1
[18] M. Ota et al. "A scalable approach for data-driven taxi ride-sharing simulation". IEEE International Conference on Big Data, 2015.