Heinz 95-845: Project Proposal

Na Su nsu1/nsu1@andrew.cmu.edu
David Mitre Becerril dmitreb/dmitrebe@andrew.cmu.edu
Raphael Wild rwild/rwild@andrew.cmu.edu
Heinz College of Information Systems and Public Policy
Carnegie Mellon University
Pittsburgh, PA, United States

Our analysis will predict the expected revenue per driver per week in the City of Chicago based on an open dataset of timestamp taxi trips. We expect that driving more time would be positively correlated with higher profits. Also, driving in specific areas of the city or during weekends may lead to higher profits. Finally, we will complement our dataset with weather conditions, which we expect to be highly relevant features to predict the expected revenue.

The taxi business in Chicago generates more than $400 million yearly from 27 million rides of Chicago (2016). Due to the rising of online and peer-to-peer ridesharing platforms, taxi drivers face more competition. This situation lead them to look for better strategies to increase their revenues. However, achieving such goal is not an easy task given the multiple dynamic factors involved. For instance, which day and hour of the week are the most profitable? Which areas are the most profitable? To leverage the use of machine learning for taxi drivers, we will build a predictive model to recommend which strategies could help taxi drivers achieve their revenue goals.

Different cities have published taxi trips datasets, which usually contain geo-referenced, timestamp information, along with fare costs, distance, and time traveled for each trip. The granularity of the data allows making predictions that are useful for businesses. For instance, based on data from Porto, Portugal, the Taxi Trip Time Prediction Challenge run by Kaggle aimed to predicted the final destination along with the total traveling time of taxi trips based on their initial trajectories. As most drivers do not report their final destination, the prediction helped the taxi central to optimize the efficiency of their electronic dispatch system, see Hoch (2015). Also, a taxi app dataset in China was used to predict the number of taxi-calling requirements submitted per time and region. This prediction would allow an online taxicab platform to predict its demands better, and hence efficiently dispatch idle taxis, see Tong et al. (2017). Similar approaches have been conducted to predict the density of taxi pickups and drop-off locations in New York City on an hourly basis, see Daulton et al.. Our analysis will contribute to existing work by building a machine learning model that will predict the drivers' revenue allowing to distinguish the best behavior they could follow to increase their profits.

We will use the Chicago Taxi Trips dataset, which has approx. 90 million rows, each representing a single taxi ride in Chicago, see Portal (2017). However, to allow reasonable computation times, we will only use observations from 2016. The available information for each ride is the ID of the taxi, time data, distance, the company, pickup and dropoff location, and payment details (fare, tips, tolls, extras, total, payment type). We will complement the dataset with the hourly weather information at the county level based on the National Oceanic and Atmospheric Administration's Local Climatological Data, see NOAA. For the purpose of this analysis, the dataset will be aggregated to represent the driving behavior for each driver in each week. Our analysis will consider:

Outcome: Revenue (fare aggregated by driver and week)
Treatment: Number of rides on each weekday, rides per time of day, rides per pickup location, hours worked per day
Covariates: Taxi company, weather conditions per day of the week (i.e., precipitation and temperature)
Population: Taxi trips in Chicago in 2016

Due to the numeric outcome variable, evaluation measures, including R square, Mean Absolute Error, Mean Square Error, Root Square Error and Mean Square Prediction Error are appropriate for the analysis. Given the range of our dataset, we consider the Mean Square Prediction Error (MSPE) the most reasonable measure. We will test the models by computing their MSPE for different input variables against our outcome variable. In the end, we will compare the model's MSPE to evaluate which model performs best. The analysis will use the training set to learn methods and through the test set, we can get the performance estimate for the learned models. A single training set doesn't tell us how sensitive accuracy is to a training sample so we will use cross-validation to include multiple training/test partitions.

The initial data, which shows data per ride in each row, will be aggregated so that rows contain data about each driver's behavior for each week (approx. 400,000 rows in total) including the outcome, treatment, and covariates as stated above. The dataset will be split into a test and training set using random sampling. A first approach will employ a lasso regression to try a model that is easy to interpret, which hopefully provides insights about decisive factors. In a different approach, a neural network will be used as it is likely to show better performance for prediction. We will tune our hyper-parameters. There are several modifications to the data to be done for pre-processing and several hyper-parameters that can be optimized, e.g., the number of hidden layers and neurons. Furthermore, feature engineering and selection will be done to improve the model's performance. Therefore, this analysis is expected to come with appropriate effort.

Our analysis is limited by the variables the dataset contains. Even though we may test these variables extensively and believe they are useful to us, it is not to say there aren't other better predictors available. There is missing information that is likely to have an impact, e.g., information about special events and holidays. Another possible limitation is that due to the data reporting process, not all trips are reported even though the City of Chicago believes that most are, which may result in bias in prediction. Also, since the study will predict revenue rather than profit, users of this analysis need to consider cost elements, such as gas and taxi rental fees. Lastly, this analysis will use weather data measured at a single weather station near the airport, but the weather might vary throughout different locations.

The target users of this analytic pipeline could be drivers, taxicab companies and other related industry researchers. For drivers, they can use the model to predict revenues based on their estimated behavior including number of rides on each weekday, rides per time of day and rides per pickup location.

References

Samuel Daulton, Sethu Raman, and Tijl Kindt. Nyc taxi data prediction. URL `https://sdaulton.github.io/TaxiPrediction`.

Thomas Hoch. An ensemble learning approach for the kaggle taxi travel time prediction challenge, 2015.

NOAA. Local climatological data. URL `https://www.ncdc.noaa.gov/data-access/land-based-station-data/land-based-datasets/quality-controlled-local-climatological-data-qclcd`.

City of Chicago. Chicago taxi data released. *Chicago Digital*, 2016. URL `https://digital.cityofchicago.org/index.php/chicago-taxi-data-released`.

Chicago Data Portal. Taxi trips, 2017. URL `https://data.cityofchicago.org/Transportation/Taxi-Trips/wrvz-psew`.

Yongxin Tong, Yuqian Chen, Zimu Zhou, Lei Chen, Jie Wang, Qiang Yang, Jieping Ye, and Weifeng Lv. The simpler the better: A unified approach to predicting original taxi demands based on large-scale online platforms. In *23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17)*, pages 1653–1662, New York, NY, USA, 2017. ACM.