Analysis of Taxi Demand in NYC with correlation to events and weather

Fiona M Sequeira, Gaurang Gatlewar, Jayesh Mhatre, Shyam Nair
New York University
Computer Science
New York
fms299@nyu.edu, gvg282@nyu.edu, jkm437@nyu.edu, sm387@nyu.edu

***Abstract* — For over a century, Medallion (Yellow) Cabs have been the lifeblood of transportation for New Yorkers. On an average day, there are a significant number of yellow cabs on the streets in comparison to private cars. However, statistics have indicated that more often than usual, people looking for a cab don't get to hail one as the cabs are unavailable at that specific place and time, indicating a certain a disparity between the taxi demand and availability leading to a dearth of cabs in some areas and customers in other areas. Among several factors that affect demand for cabs, scheduling of local events, weather conditions, recurring patterns, population density, etc. plays a significant role.**

This project examines a close analysis on cab, event and weather data of Fall 2017 to determine a correlation with the schedules of local events and/or recurring patterns when observed over a set of days.

Keywords-Yellow Cabs, Machine Learning,

I. INTRODUCTION

A recent survey indicated that there are 14,000 taxi cabs in New York serving 250 million passengers annually, it could be reasonable to question if there indeed is a mismatch between the population density and cab concentration. Statistical data indicates that the population is often concentrated in fewer areas and more frequently cabs when needed are not met with demand. Scenarios such as heavy rains or popular events or both together could exponentially increase the demand to or from a certain location on a certain day. By optimizing their pick-up location, a cab driver could easily increase his sales per day. Not only would this enable them to meet peak demands and reduce the idle time but it would also help them plan their week better.

II. MOTIVATION

Our goal is to build a platform for cab services, such that for the foreseeable future they could better analyze the cab-in-demand density at a given area on a given day. We have prepared a tool where we attempt to provide information to cabs about locations where the chances of getting a customer is very high (more specifically, a place where an event is scheduled to take place on that day, the weather conditions, etc.). We have generated a regression model that predicts a range of cabs required at any place and time, based on the local events and weather conditions. This would in turn provide a cab driver to improve his/her availability by being at the right place and at the right time and thus increase their profits by being at the right time at the right place, thus benefitting one and all.

III. OBJECTIVES

- Classify the cab data and event data to obtain a relation
- Understand the correlation with respect to weather components
- Perform accurate predictions on high cab density locations on future data
- Prepare a tool to help predict potentially high cab density locations in the future.

IV. DATA

We used three datasets for our analysis of taxi demand. The dataset consisted of 3 different CSV files namely-

- **Cab Dataset:** Sourced from NYC Taxi and Limousine data. It has about 10M rows of transactional data, each corresponding to a single cab trip. The cab location was provided in 265 discrete location IDs representing the entire NYC area. The taxi data contains features like pickup and drop-off points, start time, end time, fare, tip, number of passengers, etc.

- **Weather Dataset:** Sourced from weather underground. The dataset contains temperatures and categorical information about the nature of the weather like rain, snow, sunny etc. For our model, we used the information pertaining to NYC. The weather data was collected by scraping the Weather Underground website as it was the most reliable source providing all the historical data including.

- **Event Dataset**: Event data was obtained using multiple APIs like Facebook Graph API, SeatGeek API and Eventful API. The event data included features such as the location, popularity and score of the event, event type, artists and the statistical data.

Since the data was acquired individually from independent sources, the raw data required extensive cleaning and modifications before merging into a single dataset.

V. TECHNOLOGIES USED

To get properly labelled training dataset for our model, we used REST API calls to fetch event data, web scraping to fetch weather data and we extracted cab data manually.

Experiment Setup: The dataset was collected in the form of csv files and JSONs, following which they were aggregated and clubbed into a large dataset to help us analyze the training model. Zeppelin notebook was setup and all three CSV files contained 25 attributes in total were loaded. Data set was cleaned and processed in Spark and visualizations were provided through R and Tableau.

Issues with data and data cleaning: There were a lot of issues with our data such as event data had missing values of event end times, so we replaced them by event start time + 3 hours as it was observed that majority of the events had a duration of 3 hours. In addition, a high percentage of the taxi usage followed a fixed weekly pattern governed by office timings, peak hours, etc. Hence, we segregated the date-time field to get the Month, Day, Weekday and Hour of the data which was used as a major factor in our prediction. The event and cab datasets were merged based on date-time (day, month and hour) and Location ID which was further merged with the weather dataset on the date-time column using MongoDB.

A web application using HTML, PHP was developed so that cab drivers can use this prediction model to aid their business.

VI. ARCHITECURE DESIGN:

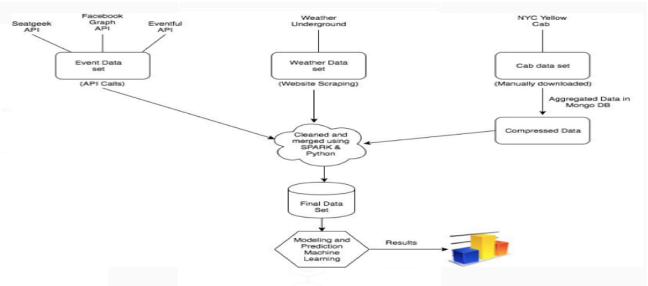

Figure: Architecture flow chart

The architecture design consists of two modules out of which one is data extraction(python) and merging(spark) and the other is prediction(python and spark) of future cab densities

Modeling and Evaluation:

We used an approach based on classification where the finite number (positive integer values) of taxi availability was trained using the NYC Taxi frequency. Our observation from the historical data is that this would be a countably small number that can be estimated using a classifier. Given the nature of the prediction problem and the availability of pre-existing Taxi data, Supervised algorithms offer an estimation approach which is easily understood and can be improved upon, iteratively.

We trained our datasets on three Supervised Classification approaches to pick the best performing algorithm -

We have a supervised learning problem with a continuous target variable and hence, we used a Regression model instead of a classification model. For this problem, Logistic Regression, Random forest, Gradient Boosting and Support Vector Machines usually perform well.

Performance Analysis
Our analysis of the accuracy indicates that, among the alternatives considered, Random Forest model has the highest accuracy and least Mean squared error.

We tuned each of the three classification models using GridSearchCV to optimise the parameters. On the optimal model, we evaluated the performance of each parameter and their contribution.

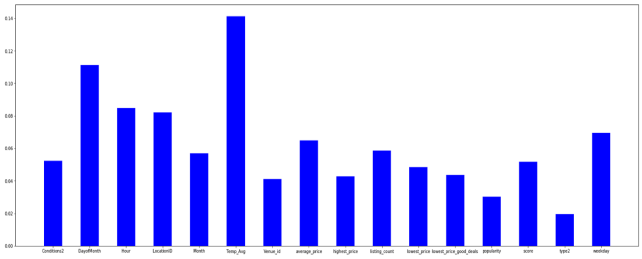

Figure: The most important parameter for demand prediction is 'Temp_Avg' followed by 'DayOfMonth'.

VII. CODE

The code for the final report has been attached along with the report and uploaded in the NYU Classes.
Git: https://github.com/ShyamNair9/Taxi_Demand_Predictor

VIII. VISUALIZATIONS

The visualizations were done using R and Tableau are shown below. The web app was built using HTML, CSS and PHP and the related images of web application are as shown below

IX. RESULT AND EVALUATION FRAMEWORK

In estimating the Classification's usefulness, we considered the following priorities –

• value numerical closeness of the predicted to the actual value. This would directly translate into usefulness as a directional pointer to the taxi driver.

• minor differences in the number of cars predicted to the actual does not carry a huge penalty

As we can see from the table, Random Forest has the highest precision of all the 3 models.

Classifier	Original Precision	Cross-Validated precision
Logistic Regression	0.6978	0.7967
Decision tree	0.9176	0.9065 (maybe overfitting)
Random Forest	0.9286	0.9341

The table shown above describes the number of times the model correctly predicts the count of taxis for the test instances as well as the cross-validated score of the model generated by choosing the optimal parameters determined by Grid Search CV. This also ensures prevents the model from overfitting.

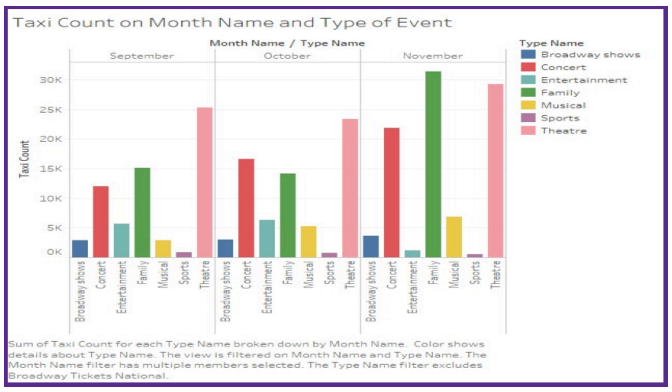

Figure: Comparison of Taxi Count for a given month vs event

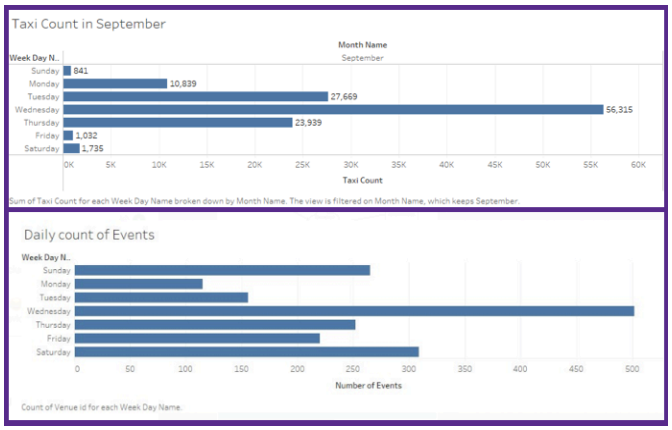

Figure: Weekly distribution Taxi Count in September vs Event Count in September

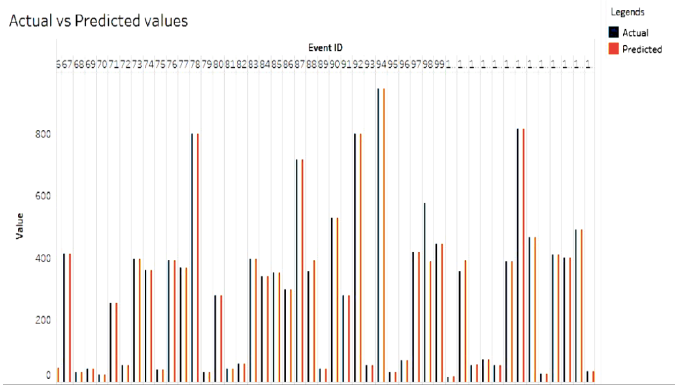

Figure: Actual vs Predicted analysis of cab densities

Figure: web application tool, where the green arrows indicate event and yellow block signify the actual spread of cabs in July 2017

X. CONCLUSION AND FUTURE WORK

The primary insight from the model is the number of taxis available at a particular venue conditional on parameters pertaining to the event, location and weather. There are two scenarios where this can be used -

➢ WHOLESALE: Allow a taxi operator to plan the deployment of the fleet optimally on days when events are scheduled

➢ RETAIL: Allow an individual driver to position himself so that he improves his chances of being hailed, in person or on apps like Uber and Lyft

We have set up a mode of delivery through a web application for the wholesale clients.

For Future work, integration with analytics like fueling locations, shift times, maintenance costs are important to these clients.

For the retail users, a mobile app with tight integration for traffic and resting stops would be more important. Also, a real

time feedback of how quickly the demand is being met at a particular venue would help drivers avoid crowding at venues where demand has already been fulfilled.

Future work also includes dealing with the following risks: One of the major risks of this model is that, the cabs would overcrowd the high demand areas and desert the low demand areas further aggravating the problem that this solution intended to mitigate. This problem can be solved by updating the location of the cabs in real time and managing the fleet centrally.

REFERENCES

[1] http://platform.seatgeek.com/

[2] http://www.nyc.gov/html/tlc/html/about/trip_record_data.html

[3] https://www.wunderground.com/weather/us/ny/new-york

[4] http://scikit-learn.org/stable/

[5] https://pandas.pydata.org/pandas-docs/stable/

[6] https://developers.facebook.com/docs/graph-api/

[7] http://api.eventful.com/

[8] https://matplotlib.org/contents.html

[9] https://stackoverflow.com/

[10] https://docs.mongodb.com/v3.4/