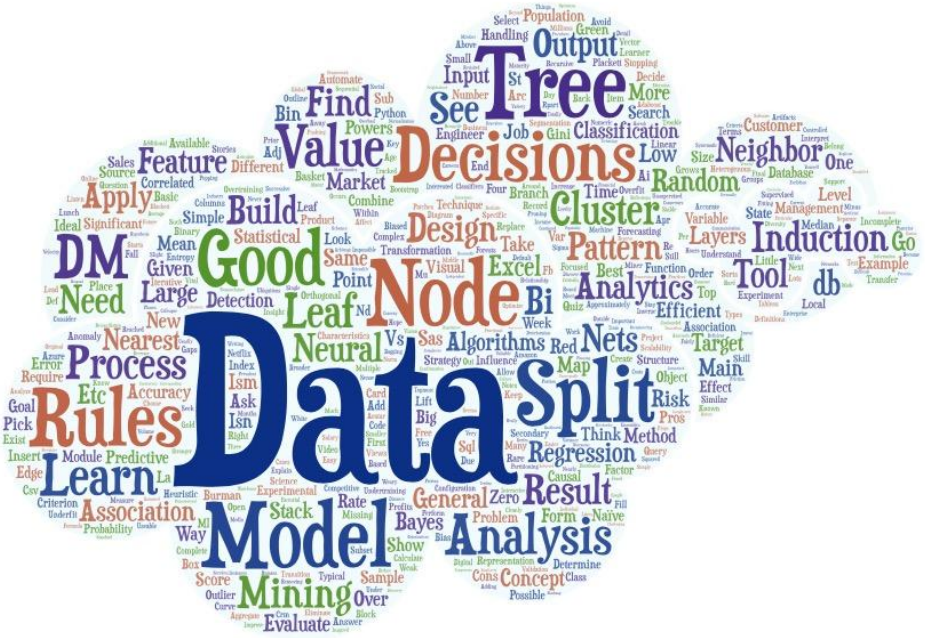


NYC RIDE SHARING STARTUP

Prepared for Dr. Daniel Conway, PhD – Spring 2019

ABSTRACT

A private equity firm's analytics group has been asked to provide data analysis for the strategic planning committee tasked with answering the question: Should we enter the ride sharing market in New York City?

Authors:

Karina Aguiar Goncalves

Chinmayi Karmalkar

Arpita Nayak

Pedro Ortiz Rodriguez

Jon Waggoner

ISM 6136 Data Mining

Table of Contents

Executive Summary

With the advent of ride-sharing aggregators like Uber, Lyft and Juno, traditional taxi operations are taking a major hit. The comfort of using a smartphone equipped with GPS has disrupted the entire taxi and limousine services sector. Simply put, technology is bringing major disruption in transportation industry. In the United States,the ride sharing market consists of various companies which utilize recent technological advances to match drivers with passengers at short notice for one-off shared rides. However ride sharing companies do not fall under regulatory and licensing requirements unlike their traditional counterparts. This gives them a major price advantage as compared to traditional cabs and hence it's being hailed as a 'breakthrough' technology.

According to the Statista 2018 report, Global net revenue of Uber is $11.3 billion calculated from 2013-2018 and 53% of people in US use different ride-sharing apps. Some additional statistics are:

- Revenue in 2019 in US is $18.4 billion in the ride-sharing segment.
- Current annual growth rate is about 9.3%.
- 16.2% of riders used apps in 2019 and usage is expected to be 18.1% by 2023

This said, these are our findings of the ride-share market as we see it.

We have taken NYC taxi data that has around 25K rows consisting of various companies and 9 columns where we have used some feature engineering to select the most important features. The dataset has been sourced from open dataset available from https://opendata.cityofnewyork.us

Along with the significant findings, this report entails recommendations for the company wanting to enter the ride-sharing market. This report focuses more on decisions to be taken rather than just showcasing what the data says.

Business Questions

The taxi dataset that we have has variety of features and many companies operating in NYC offering ride-sharing services ranging from the year 2015-2019.
Problem Statement: Should we enter the ride share market in NYC?
We created our own business questions to be answered by analyzing the dataset. We pondered over many questions but finalized four main questions to answer. They are:

- Primary competitors
  - ➢ There are many companies in the dataset, but we would like to know who would be giving the most competition to our company if we enter into the market.

- Market share segmentation
  - ➢ We would like to know, which company has the biggest market share and how many segments are present in NYC dataset catering to different customers.

- Features engineering
  - ➢ We have 9 features given in this dataset; but we would like to know if all the features have equal importance or some have more weight than others.

- Which should be the ideal month If we are entering the ride-sharing market?
  - ➢ This question may seem less important, but plays a very significant role in any business. Product's release date generally determines its success and profits for the upcoming weeks.

Data Source and Assumptions

Dataset used: NYC Taxi dataset (2015-2019)
Dataset source: https://opendata.cityofnewyork.us
Features and Instances: There are around 24.5K instances and 9 features.
The features are as follows:

- Base License Number:The TLC License Number
- Base_Name: The official name of the base entity
- DBA: “Doing Business as” name of the base
- Year, Month: Timeline of the trip
- Total Dispatched Trips: The total number of dispatched trips performed in the week, to both affiliated and non-affiliated vehicles
- Total Dispatched Shared Trips: The total number of dispatched trips which were shared trips performed in the week, to both affiliated and non-affiliated vehicles
- Unique dispatched vehicles: The unique number of vehicles the base dispatches to in the week, including both affiliated and non-affiliated vehicles.

Assumptions:

- The variable we are taking to isolate is Total Dispatched Shared Trips.
- We edited metadata and considering the features that are more important for our models that are:
    - ➢ Year, Month
    - ➢ Total Dispatched Trips
    - ➢ Total Dispatched Shared Trips
    - ➢ Unique Dispatched Vehicles

Tools and Data modelling techniques

Tools:

- Microsoft Excel to explore the data set and the results.
- Azure Machine Learning (ML) Studio to build two predictive analytic models to answer the business questions.

Data Modelling Techniques:

- In order to answer the business questions, we prepared two experiments. Each of these was focused on a specific Data Mining concept:

  ➢ **Clusters**: Clustering is the process of making a group of abstract objects into classes of similar objects. This technique was used with the purpose of finding out who are the stronger competitors in the ride share industry in New York City, and for grouping competitors into market segments.

  In the clustering experiment, two main modules were used:

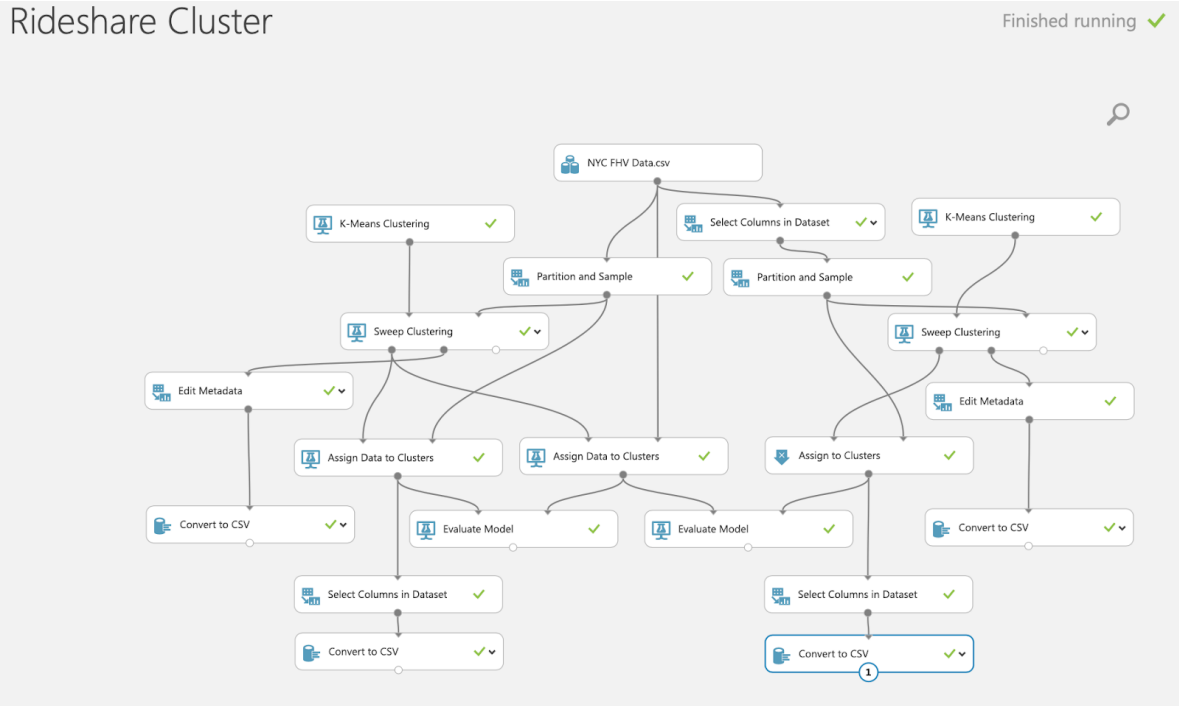
  - **K-Means Clustering**. This data mining algorithm is one of the simplest and best know unsupervised learning algorithms. It aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean.

    To find the optimal parameters, we used a Parameters Range configuration, also a Euclidean function was used to measure the distance between cluster vectors.

  - **Sweep Clustering**. This module was used to train a model using a parameter sweep. A parameter sweep is a way of finding the best hyperparameters for a model, given a set of data. This module receives as input a clustering model (K-Means Clustering) and a dataset (a 75% sampling of the data). The module iterates over the set of parameters building and testing models with different parameters until it finds the model with the best set of clusters. It automatically computes the best configuration and then trains the model using that configuration (3).

    The following parameters were used for the Sweep Clustering:
    - Total Dispatched Shared Trips
    - Unique Dispatched Vehicles

- Base Name
- Year,
- Month
- Base License Number



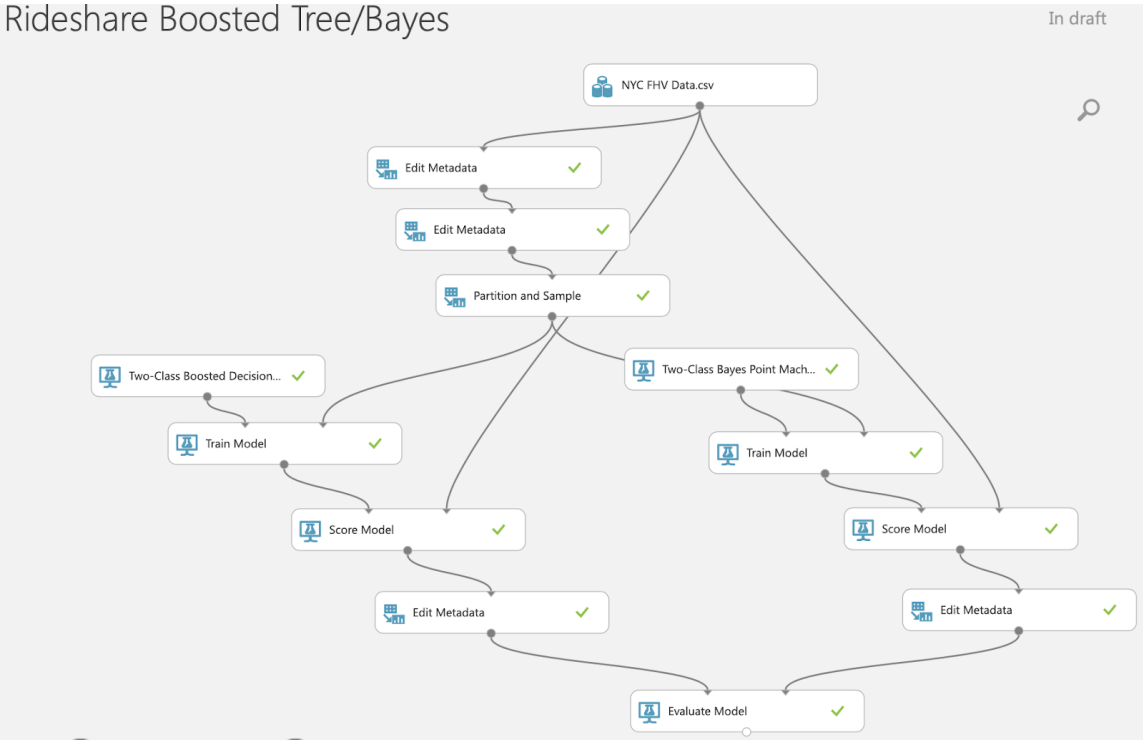
➢ **Decisions Trees**: A Decision Tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. This technique essentially was used to answer the business question when should we enter the market.

In the decision trees experiment, two main modules were used for comparison:

- **Two-Class Boosted Decision Tree.** A boosted decision tree is an ensemble learning method in which the second tree corrects for the errors of the first tree, the third tree corrects for the errors of the first and second trees, and so forth. Predictions are based on the entire ensemble of trees together that makes the prediction (4).

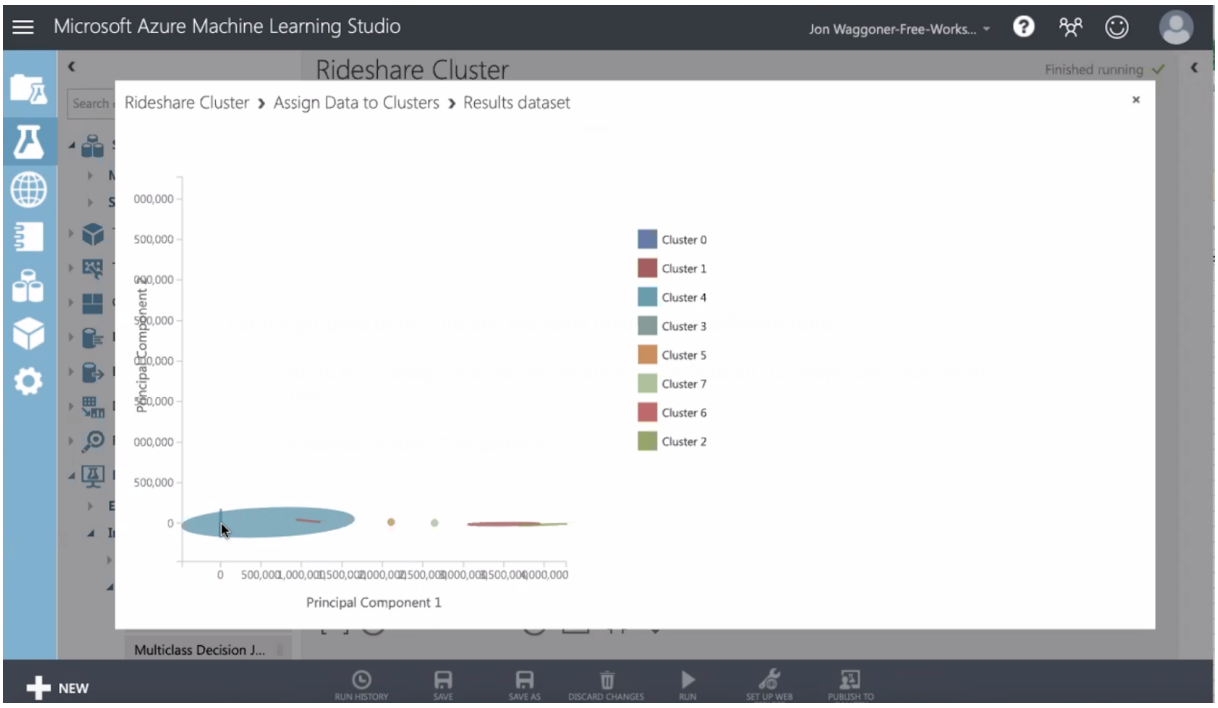
A single parameter setup, with a minimum of 10 and a maximum of 20 leaves, and a learning rate of 20% was used for this experiment. Total Dispatched Shared Trips was the target parameter, defined on metadata connected to the Train Model module.

- **Two-Class Bayes Point Machine.** This technique usually is used to create an untrained binary classification model. The algorithm uses a Bayesian approach to linear classification. This module was used for experimental purposes but didn't add value when coming to answer our business questions.



Results and Interpretation

➢ **The Cluster experiment**:



The cluster experiment led to the above results. The experiment generated eight clusters, all of which are represented in the graphic above. The more elliptical the cluster's shape the more significant or relevant the cluster is. From results, it can be highlighted how cluster number four 4 is the most representative of the dataset.

To interpret the results a CSV file was generated.

I	J	K	L	M	N	O	P	Q
Assignment	Company(s)	Base Count	Shared Trips	Unique Vehix	Std Dev Shared	Std Dev Vehix	Shared Trips per Vehicle	sharede trips per month
0	All Taxis and JUNO	18340	36283	2241163	77	1460	0.02	740
1	LYFT	4	14008777	282613	110648	4432	49.57	285893
2	UBER	3	11931598	224444	84330	1784	53.16	243502
3	UBER	2	3035256	121624	67985	801	24.96	61944
4	VIA or LYFT	28	16426092	663497	100138	16536	24.76	335226
5	UBER	1	2106689	62549	0	0	33.68	42994
6	UBER or LYFT	3	3241044	161094	40675	5360	20.12	66144
7	UBER	1	2643389	65375	0	0	40.43	53947

VIA or Lyft is the ride sharing company cluster most representative of the dataset, it has about 16 million shared trips, about 660,00 unique vehicles, and 330,000 share trips per month.

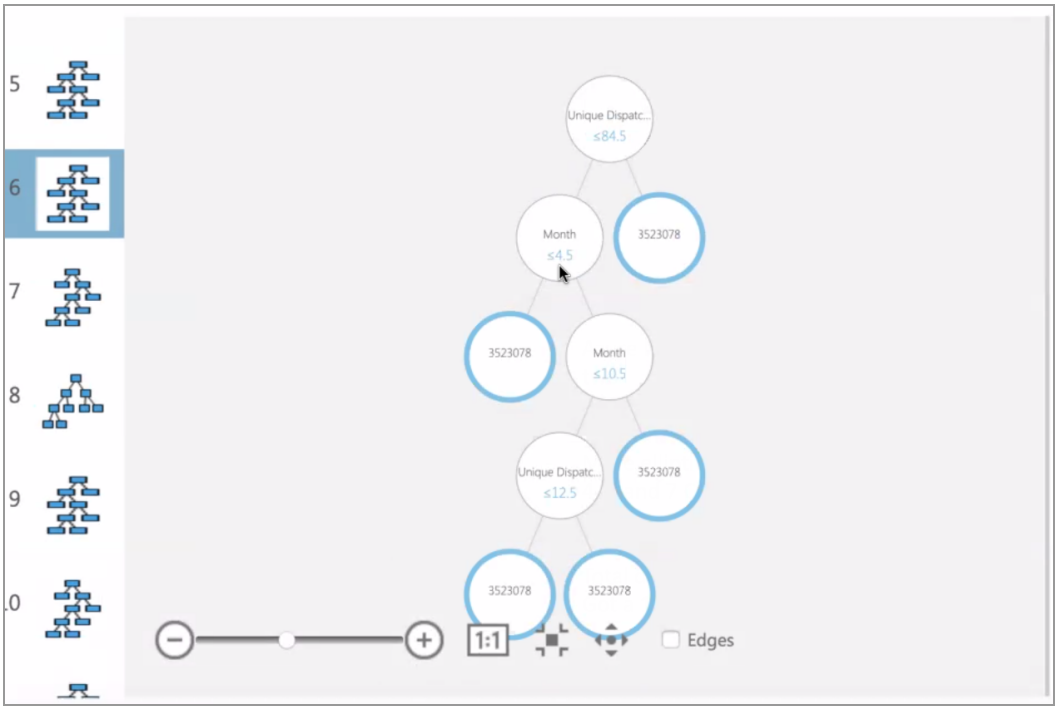
The interpretation of these results led to answers for the following business questions:

1) Who is our competition? Uber, Lyft and VIA.
   Regular Taxis and Juno, another share riding company, are not direct competition.

2) How is the NYC market share segmented?
   - 8 clusters (assignments 0-7) based on total dispatched trips.
   - 1 & 3 Uber - shared trips
   - 2 Lyft - shared trips
   - 4 Lyft and VIA - shared trips
   - 5 Uber - non-shared cluster 1
   - 6 Uber and Lyft
   - 7 Uber - non-shared cluster 2
   - 0 Taxis and Juno - not direct competition

➢ **The Decision Tree experiment**:

The decision tree experiment based on the Two-Class Boosted Decision Tree technique, generated 100 trees, however, the first 9 trees came out with very high accuracy of 98.5%. The analysis and interpretation of the results was focused on Tree #6.

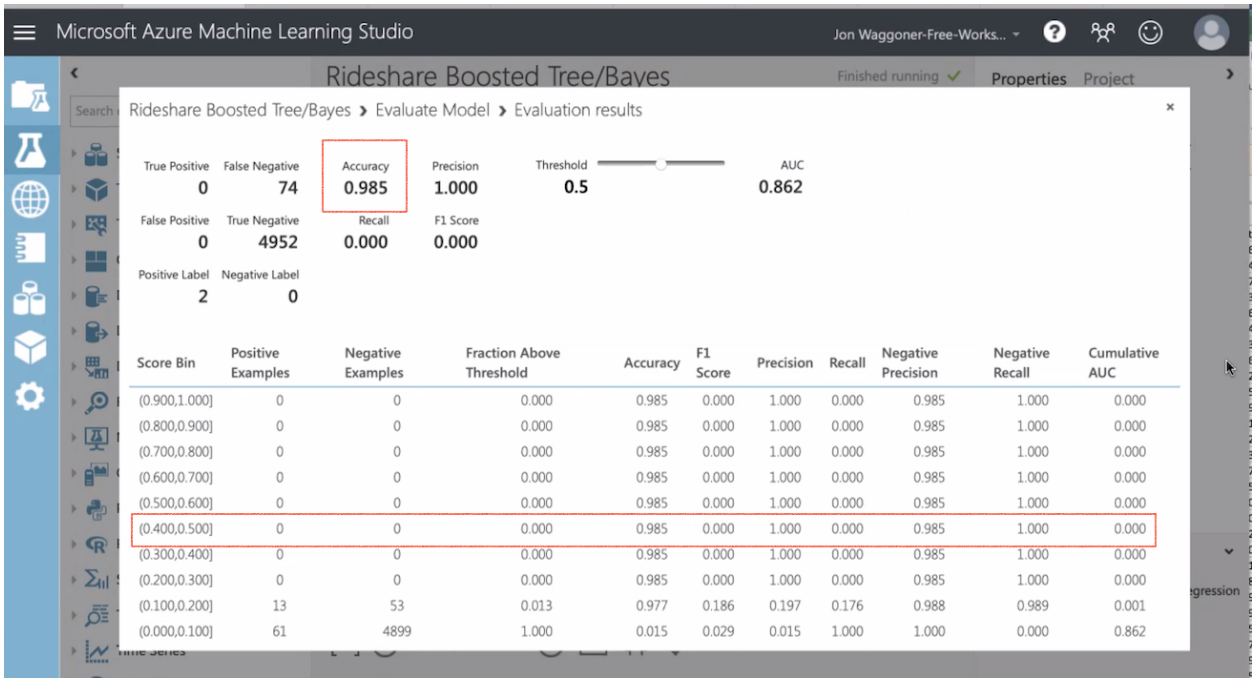


From the results, two business questions were answered:

1) Which variables are most important in determining market viability?
    - Unique Dispatched Vehicles
    - Month

2) When should we enter the market?
   From the decision tree it can be interpreted that if the number of vehicles is a restriction for the financials of the business, in other words, if the company has capacity for less than 85 vehicles, then entering before mid April is an option, however, since dates have already passed for 2019, the best option for the company is to roll out the new ride share service after mid-October.

Accuracy of the model:



Conclusions and Recommendations

From a business point of view, all the data mining experiments lead us to conclude that there are big competitors and important market segments in the NYC ride sharing industry. Three competitors to highlight are Uber, Lyft and VIA, and one cluster, Lyft and VIA combined, is the most important.

In order to enter the market and be competitive, the number of unique dispatched vehicles, and the month to enter the market are key for defining a business strategy.

As a company looking to operate in NYC offering ride-sharing services, we can conclude there is an opportunity to enter the market, however, competition is very high. Global competitors such as Uber and Lyft take a big part of the market share, however, it is interesting to look at VIA a real-time ridesharing company based on New York City only. This competitor plays an important role together with Lyft. This competitor should be evaluated in detail and analyzed to find out which differentiators and added value have led them to be a strong competitor in the NYC market. Our business strategies should be similar to the ones applied by VIA. From research (5) it was found that VIA service works more like a dynamic bus line rather than ride sharing service, and charges users a flat rate for a ride. Moreover, the VIA service accepts commuter benefit cards.

On the other hand, for financial support and strategic release, it is important to consider the number of unique dispatched vehicles. Less than 85, but more than 13 appears to be a very good range for starting the business. Operations before mid-April or after mid-October are recommended from the experiment results. Basically, it would be a very good strategy to start a business during Q1 or Q4, or in other words in a season very close to the beginning or the end of winter, and of course during the US holiday season itself.

All in all, and from an academic point of view, this project was a great opportunity to look at a dataset, analyze it and come up with answers for specific business questions, with the purpose of decision making. Data mining is the science to explore data, build and evaluate models using different algorithms and techniques resulting in conclusions and recommendations that help businesses to make decisions. We learned a lot during the project including the importance of using a variety of tools such as Azure ML to support the data mining process.

References

1. https://www.statista.com/topics/4610/ridesharing-services-in-the-us/
2. https://www.statista.com/outlook/368/ride-hailing
3. https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/sweep-clustering
4. https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/two-class-boosted-decision-tree
5. https://en.wikipedia.org/wiki/Via_(company)