**Emergence of Location Services Has Created New Capabilities**

Bernard Savarimuthu

March 11,2019

# Introduction

## Background

The emergence of location services has created new capabilities and existing businesses are finding new threats every day. Start-ups leveraging new technologies and innovative business models have continued to disrupt and influence how customers choose their travel destinations, activities to engage in, places to stay and eat, and the number of travellers suitable for a specific season. To remain competitive, businesses have embraced use of location data to identify consumer patterns, tastes, and trends. Location data is available from a number of social media platforms including Google maps, Facebook check-in data, location reviews on sites like TripAdvisor and Booking, Foursquare among other location data providers.

After learning data science, two friends who have a young start-up travel business decide to change the business model by applying data science skills and techniques. It is their believe that data science will improve their customer experience, grow customer loyalty, promote the company brand and set the company out as a choice travel agent.

## Problem

The start-up faces a lot of competition from established businesses and their services are not differentiated from those of their competitors. Their recommendation on travel itinerary to their customers has not yielded the much needed growth in business. The company might close down if this trend continues.

## Interest

Obviously, to remain competitive, businesses have embraced use of location data to identify consumer patterns, tastes, and trends. Inaccurate data is plaguing the marketer’s capacity to enrich experiences using location, and in many cases, raising some of the aforementioned questions around privacy.This is especially prevalent for advertising, which is not always reliant on real-time tracking from smartphone signals. According to data from xAd and Foursquare, up to 60% of ad requests contain some form of location data. Yet of these requests, less than half are accurate within 50-100 meters of the stated location.

# Data acquisition and cleaning

## Data sources

Location data is available from a number of social media platforms including Google maps, Facebook check-in data, location reviews on sites like TripAdvisor and Booking, Foursquare among other location data providers.

## Data cleaning

Data downloaded or scraped from multiple sources were combined into one table. There were a lot of missing values from earlier seasons, because of lack of record keeping. I decided to only use data up intends to collect location data from Foursquare and apply data science techniques and tools. The data collected will involve comparison of two locations to determine which the best location to recommend to a customer is. However, there were reviews on sites like TripAdvisor and Booking, Foursquare among other location data providers. This cause the start-up faces a lot of competition from established businesses and their services that are not differentiated from those of their competitors. Secondly, their recommendation on travel itinerary to their customers has not yielded the much needed growth in business. The company might close down if this trend continues. After fixing these problems, I checked for outliers in the data. I found there were some extreme outliers, mostly caused by some types of small sample size problem. For example, some Location data providers offer identical data coverage or quality. Indeed, there’s evidence to suggest that a majority of programmatic bid stream location data, which some providers use as their primary source, is either fraudulent or of questionable quality. Brands, retailers and agencies working on their behalf should take advantage of location-based insights and attribution capabilities. Those that aren’t doing so are missing a significant opportunity.

## Feature selection

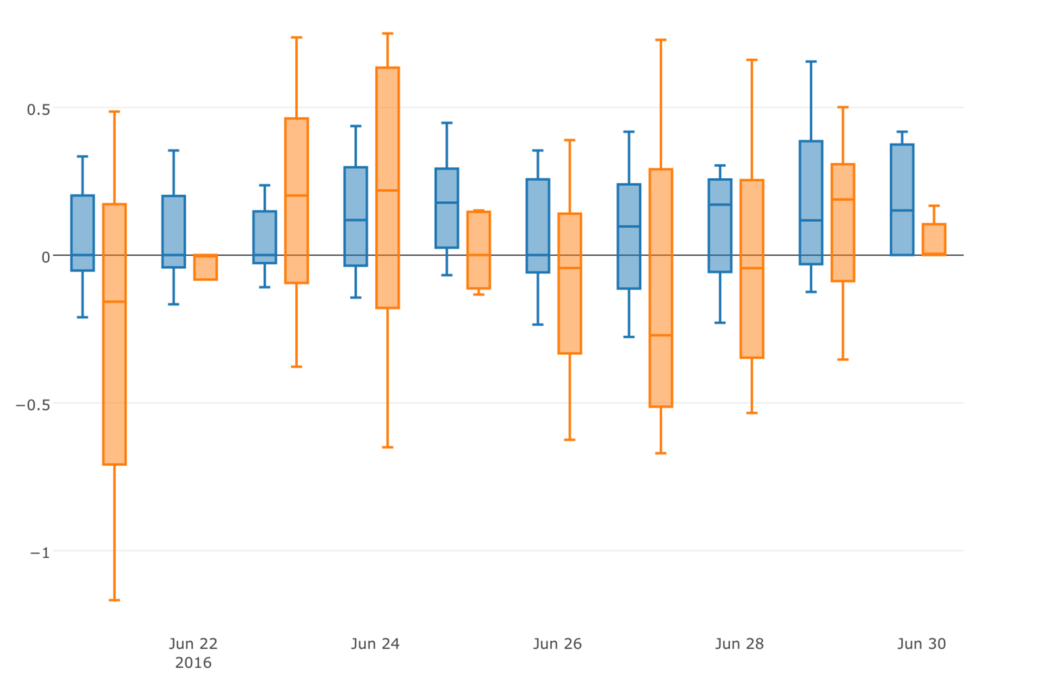
After data cleaning, there were 260 samples and 86 features in the data. Upon examining the meaning of each feature, it was clear that there was some redundancy in the features. For example, there was a feature of the number of rebounds a location data provider collected, and another feature of the rate of rebounds it collected. These two features contained very similar information (a location data provider ability to rebound), with the difference being that the former feature increased with Coverage, Accuracy, Regency, Contextual metadata and Transparency, while the latter feature did not. Such total vs. ratio of first-party to third-party data also existed between other features. These features are problematic for two reasons: (1) How many of the total venues/visits are represented? How detailed is the metadata about a given set of venues. (2) A mix of signals and sensors (i.e., GPS and cell tower signals, SSIDs, Bluetooth or beacon signatures) can help when distinguishing between places in densely populated urban areas. Marketers should look for multi-sensor inputs. How often is the data updated? What is the methodology for ensuring that any fluctuations at business locations are logged and verified in a timely manner? Datasets can be classified into three groups: first-party sources, third-party data aggregations and a combination of the two. Foursquare throws out more than 80 percent of third-party data that doesn’t meet data quality standards. Sources are important because there may be larger privacy and regulatory implications in certain regions. Contextual metadata indicates how much additional information is provided for a given location beyond the most basic attributes of name and address. This can be useful in providing a more complete understanding of a place, why people go there, or what’s unique about it. In order to fix this, I decided to keep all features that were rates in nature, and drop their cumulative counterparts.

# Exploratory Data Analysis

## Calculation of target variable

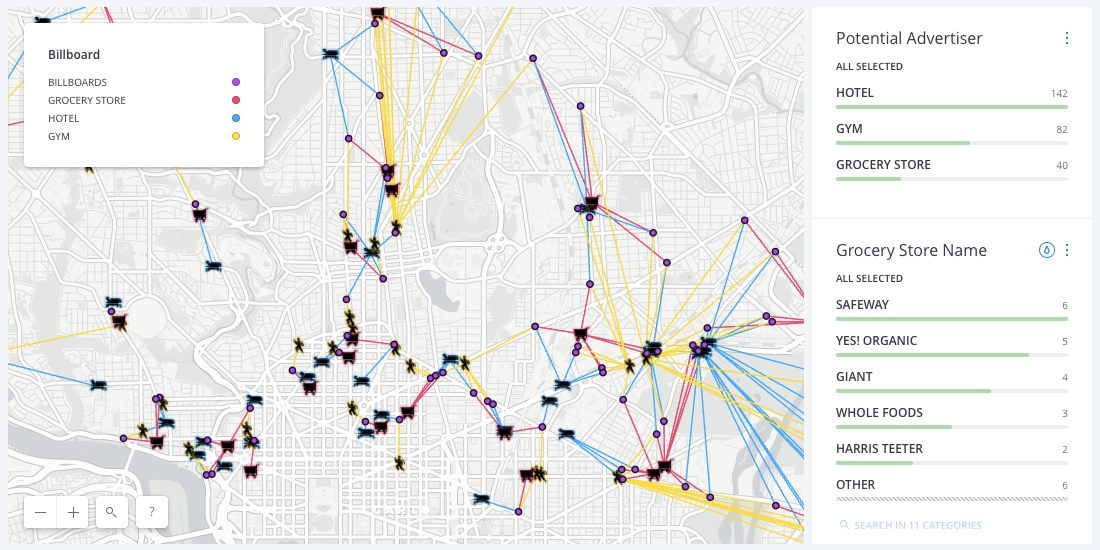
Understanding sources of location data: How do you source location data? What is the breakdown of first-party data and third-party data? What percentage is bid stream vs non-bid stream? Making sure you are using data with strong density Choosing a provider whose data has been third-party verified Using a provider with a broad scope of applications. Having maximum Omni channel flexibility to make sure offerings go across more than mobile to be effective for all marketing channels. User (s) data ready-made and location data providers have integrations in place to plug into the existing marketing stack (i.e., DSPs, DMPs, social platforms, TV platforms)

Figure 1. Relationship between improvement and overall ability



## Relationship between improvement

## Well, location data measurement providers take varying approaches to arriving at that exposed visits number that marketers are so interested in when gauging the ROI of mobile and digital ad campaigns.There are big distinctions between reported numbers that reflect the “raw value” of exposed visits and those reporting “extrapolated exposed visits.” Let’s start with raw value. When measurement providers report exposed visits using “raw value,” they are measuring the actual count of devices exposed to an ad that have been tracked in campaign locations.In contrast, when they report exposed visits by gauging extrapolated exposed visits, they are estimating the physical foot traffic using a mathematical model factoring in things like GPS usage, the frequency of ad requests, dwell time and other parameters. The idea behind extrapolating as opposed to reporting raw numbers is to reflect a more accurate estimate of the total number of exposed visits seen in a real-world location.



## Relationship between improvement and last year’s improvement

I hypothesized that a location data provider improvement might be correlated with it’s previous improvement, because their competitors might improve continuously for a few years, and older location data provider (s) might decline for a few years straight. It turned out that the relationship between improvement and prior improvement was negative. In other words, more often than not, a location data provider company will “regress to the mean” rather than continuously improve or decline.

Figure 2. Scatter plot of location data provider (s) improvement and that of last season

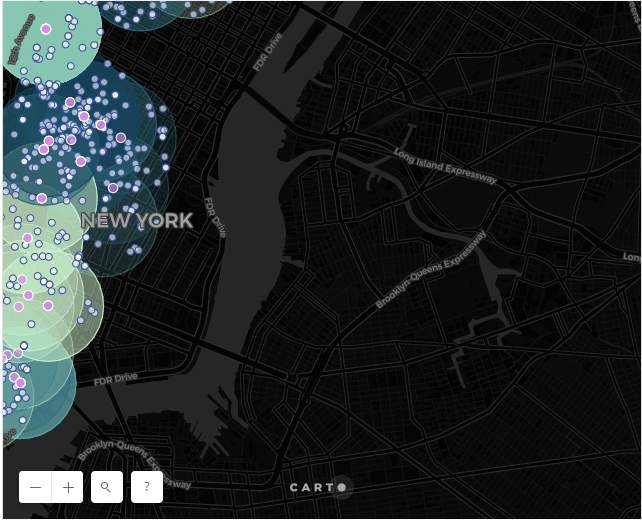
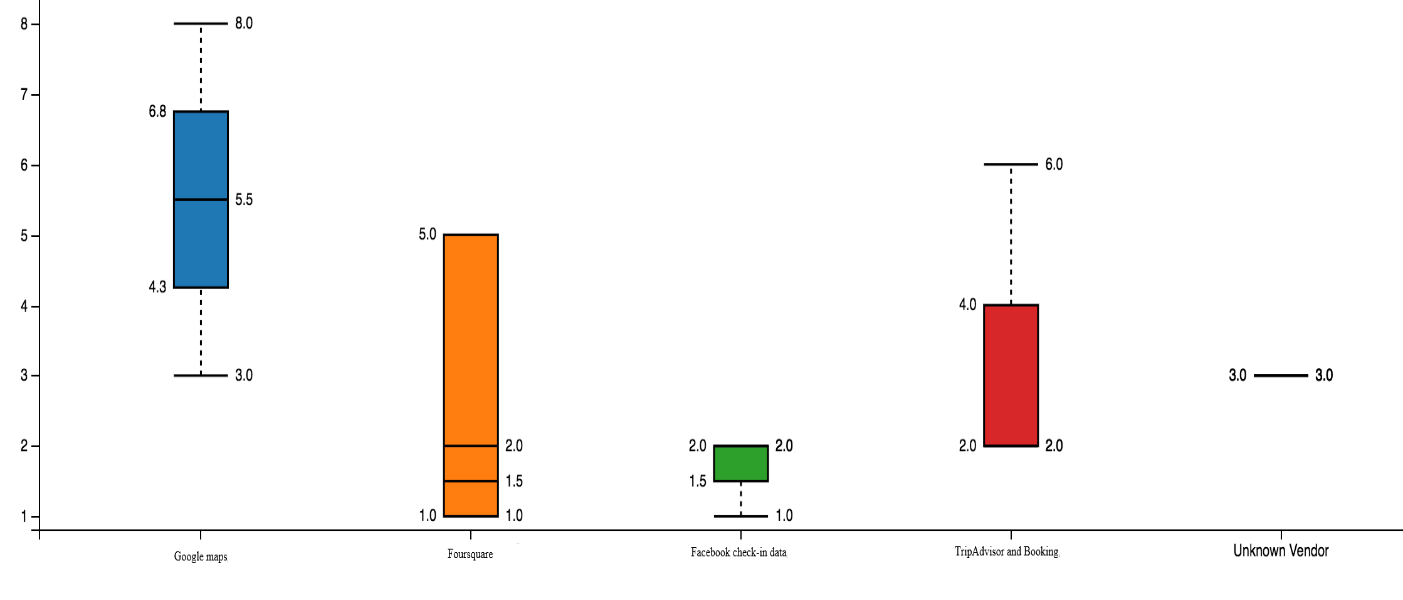


Figure 7. Box plot of location data providers’ improvement among different draft groups



# Predictive Modeling

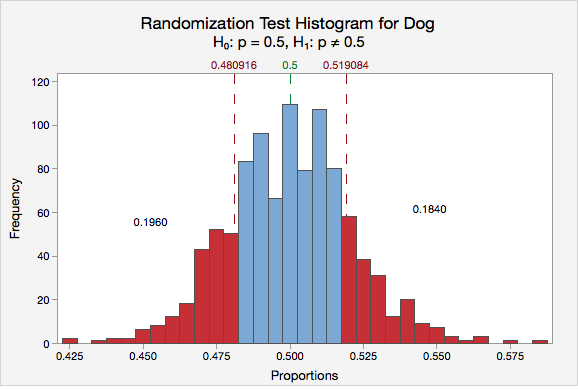
There are two types of models, regression and classification, that can be used to predict player improvement. Regression models can provide additional information on the amount of improvement, while classification models focus on the probabilities of a location data provider (s) might improve. The underlying algorithms are similar between regression and classification models, but different location data user (s) might prefer one over the other.

### Applying standard algorithms and their problems

I applied linear models (linear regression, Ridge regression, and Lasso regression), support vector machines (SVM), random forest, and gradient boost models to the dataset, using root mean squared error (RMSE) as the tuning and evaluation metric.

### Solution to the problems

The reason behind these problems were the uneven Segmentation: Repeated visitation can then be used to build audience segments. With little improvement an audience, they could collect the device IDs of people who demonstrate a particular mind-set based on the places they visit. Therefore, the models tried to prioritize minimizing errors on players with little improvement/decline when RMSE was used as the evaluation metric. My solution to this problem was to assign weights to samples based on the inverse of the abundances of target values. In other words, other types of data, like time of day, day of the week and weather, need to be taken into consideration before sending out messages to users within a geofence.



I also evaluated the models using their ROC curves. In this particular problem, lower false positive rate is more important than higher true positive rate. In other words, it is more important to be sure that a location data provider (s) will improve as predicted.

# Conclusions

In this study, I analysed the relationship between different location data providers improvement/decline and their performance and biographic data. I identified applying big data methodologies to filter out inaccurate locations, and by analysing historical location data as well as first and third party data sets, advertisers can further refine the information about a given device to provide useful insights and determine patterns or place usage. This data helps to build a picture of how any given location is used in real life (e.g. whether a place is residential or commercial) and provides an opportunity to target audiences based on their context and not just their physical location,”

# Future directions

I was able to achieve ~30% improvement from the benchmark model in the regression problem, and ~52% accuracy in the classification problem. However, there was still significant variance that could not be predicted by the models in this study. I think the models could use more improvements on capturing the user (s) traits. “Brands must fully understand what data sources their chosen location provider has access to and how they ensure incoming data is cleaned for errors. At an industry level, overcoming the challenges of location-based marketing also relies on greater standardisation to keep data reliable and accommodate the ever increasing number of data providers in the ecosystem. More data, especially data of different types, would help improve model performances significantly. These interactions data are obviously more difficult to extract and quantify, but if optimized, could bring significant improvements to the models. “These days, we’re all becoming a lot more connected. Whether it’s customers, staff, the store itself or physical products on the shelf, the Internet of Things is connecting us all and is the driving force behind the smart store. By 2020 it’s estimated that we’ll see somewhere in the range of 20–30 billion connected devices across the globe.

Signing Off : Bernard Savarimuthu