A Predictive Model for Business Failure Based On City and Customer Review Data

By: Allan Ruan, Botao Yao, Julian Romero, and Julia Vida

Abstract

To predict the failure of retail businesses is of high importance to a variety of parties, including financial organizations, governments, and retail businesses themselves. When a business is started, certain *uncontrollable* factors, such as city unemployment and local traffic, present inherent challenges and advantages. This makes it essential to understand how *controllable* factors, such as product quality and price, contribute to the success/failure of businesses. Other studies have investigated how uncontrollable factors correlate to business closure, but have not examined in great detail how controllable factors contribute to a business's longevity. In this study, customer review data, trendiness data, traffic data, and census tract data will comprise the input to the classification algorithm. This allows a thorough investigation of both controllable and uncontrollable factors alike, without giving one precedence over the other prior to running any algorithm. By performing this study we will be able to predict whether a restaurant business will close based on the data describing the relevant controllable and uncontrollable factors. The results will be evaluated against the ground truth data of a business's status (open/closed) as indicated in the customer review dataset.

1: Introduction

Retail businesses have a strong effect on the economics within a city. Growth of retail businesses indicates that a region's economy is strong and a region's economics also have an impact on the success of these businesses. Retail businesses and policy makers have a strong interest in knowing the likelihood and factors associated with business success/failure. With this knowledge, retail businesses can adjust their strategies to become more competitive and policy makers can make decisions to better assist these businesses and enable a thriving economy. There are two types of factors when considering business failures: controllable factors and uncontrollable factors. Controllable factors are ones that the business' decisions can easily change. For example, these are quality of service, quality of product, and price of product. Uncontrollable factors are more difficult to alter. These include traffic to a certain area, business regulations, and competitiveness in the region of operation. One of the challenges that comes with predicting business closure is establishing when a business has closed. In the past, studies have used bankruptcy or a lack of social media check-ins to tag a business as closed. The problem of using bankruptcy to determine business failure is that bankruptcy fails to capture cases where business owners decide to shut down the business. Another study uses a lack of foursquare check-ins as an indicator that a business is closed; however, this is inaccurate because people may just not want to check in to a business.

Previously, there have been studies that utilized urban mobility, and social media data.With other social media and urban mobility based studies, the focus was on uncontrollable factors, such as how many people move to the area. Currently, the studies do not account for controllable factors because it is a challenge to quantify quality of service and customer satisfaction. Our goal is to incorporate controllable factors into the prediction of business failure through the use of the business review platform Google Reviews. In this work, we will use the following datasets: 1) TLC trip record data from OpenNYC, and 2) business review dataset from Google Reviews. The purpose of our study is to predict business failures using both uncontrollable and controllable factors pertaining to restaurant businesses in NYC. The Google Reviews dataset keeps track of whether a business is opened or closed. Therefore, this can serve as the ground truth in the evaluation of the prediction model.

2: Related work

There has always been an interest in uncovering the causes of failure of retail business, and predicting the likelihood that a prospective business will fail. Most other works identify that there are two categories of

A Predictive Model for Business Failure Based On City and Customer Review Data

By: Allan Ruan, Botao Yao, Julian Romero, and Julia Vida

factors that contribute to a business's susceptibility of failure. There are controllable factors and uncontrollable factors. Controllable factors include the quality or price of the store's products, the operating hours, popularity, and its customer satisfaction. Uncontrollable factors include: unemployment rates of the city, overall economic conditions, and urban policies.

It is often difficult to determine what constitutes business failure and this has limited the number of studies on business failure. Prior works have used financial records instead when sampling and used bankruptcy as an indication of failure. However this approach is limited due to not capturing cases where a business decides to shut down. The inherent low frequency of financial reporting also leads to a poor dataset that can only focus on static macro factors. This leads to failure and not being able to recognize that a business is at high risk of closure in the near future.

Recent advances in new urban datasets, in particular datasets related to urban mobility, social media activity, and crowd sourced review forums such as google reviews, offer opportunities for sampling controllable factors. Mobility data has the ability to reveal the dynamics of a particular location. For example, does a certain area like time square attract visitors from other areas like the upper east side? Social media activity has been used in previous studies to examine locations based on social media. This is called location based social network (LBSN). This data has been used to examine consumer interactions for individual businesses. For example this data can tell you how popular a venue is relative to others in its area. Other papers have used this data for business analytics. For example, Wang et al. utilised LBSN to predict the failure of food establishments using a set of over 600 restaurants in New York City over a 6 month period.[3] Other researchers such as Karamshuk et al also used LBSN data to make strategies using LBSN-based features that were able to find optimal locations for new businesses to open.[2] Uncontrollable factors in urban mobility were a big focus in a paper by Krittika D'Silva et al. [1] that focused on establishing three classes of features to predict if a business would be successful in their model. They used features such as: static locality profiles, visiting patterns, and neighbourhood mobility dynamics. Static locality profiles capture the properties of locality in which a business operates. Visiting patterns reflected the volume and spatiotemporal patterns of Foursquare check-ins. Neighbourhood mobility dynamics reflect the visitation patterns across distinct neighbourhoods.

Most papers focus on uncontrollable sources such as the ones mentioned previously. However they fail to mention controllable factors which our paper does. Our paper has a unique feature that other papers do not. This feature is using crowd sourced review forums. We will use crowd sourced review forums because they are a great way to know controllable factors such as popularity, quality, and customer satisfaction. This dataset will be one of the main focuses of our paper as no other paper that we know of has used the google local reviews dataset to reveal data about controllable factors.Most assume or just don't acknowledge that some controllable factors cannot be measured such as popularity. This is not the case however. Our model uses google reviews data to be able to measure these controllable factors that others do not seem to measure. Most also assume that their datasets reflect the actual population at hand. For example the amount of visits to the businesses. Most assume that using a dataset such as foursquare is enough to predict actual visitations. The dataset would need to be more diverse to be actually implemented in deployable systems. Foursquare has only a limited number of users and has only a limited time frame. Our model fixes this issue by using a google reviews dataset that gives us diverse reviews. It also has labels telling us that the business was opened and if the business closed.

3: Motivation

Our motivation for performing this study is that there is currently no study that assesses how controllable

A Predictive Model for Business Failure Based On City and Customer Review Data

By: Allan Ruan, Botao Yao, Julian Romero, and Julia Vida

factors will impact business survivability. Even though uncontrollable factors are important to use in prediction there is little businesses can do to impact them as they are uncontrollable. Being able to identify controllable factors will enable a business to actually do something about their situation if they are likely to fail. In order to show our motivation, we will investigate check-in data that is being used in related studies location based social networks (LBSN) and investigate how using controllable factors may provide a more desirable output.

3.1: Check-in data

As in figure 1, there are a lot of factors that can be extracted from check-in data.[1] Features in this data are all extracted via that use of the number of check-ins to locations from foursquare, hence controllable factors cannot be assessed.

Locality profile from this data such as competition and catchment of locality is based solely on the number of check-ins, but qualities that these businesses have that render them more competitive are absent.

Visit pattern shows how popular a place is and how far people travel between venues. However, this does not show what makes a place popular. Given this feature we can show that perhaps people who go to restaurants on Friday night are more likely to visit a bar afterwards or they will stay at a restaurant longer since they are waiting for a party to start. The aspects of the business that attracted customers cannot be determined with check in data.

Mobility dynamics include information regarding temporal alignment with locality, for example if a business is popular at times when the area has a lot of traffic or whether it is the driving factor of traffic at a time of the day. Also, there is a reachability factor. However, there is little a business can do besides relocate given this data.

Moreover, there are business attributes that are considered, which include price tier which is controllable. However, since we only have price tiers, it is hard to distinguish if price is causing the business to fail or if it is another factor.

Lastly, using check-in data only, it is hard to determine the ground truth, which is whether the business is closed. This is because there is no label of whether or not the business is actually closed. Instead it relies only on a decreased number of check ins.

Figure 1. Features from D'Silva et al [1]

Feature Class	Feature	Definition	Source
Locality Profile	Competition	$\frac{CN_i}{N}$	Foursquare
	Specific Competition	$\frac{SN_i}{CN_i}$	Foursquare
	Place Entropy	$\sum_{i=1}^{k} p_i * \ln p_i / \ln k$	Foursquare
	Category Counts	$\lvert CN_i \rvert, \quad \lvert SN_i \rvert$	Foursquare
	Attractiveness to the Neighbourhood	$\lvert CN_i \rvert \times \ln\left(\frac{\lvert V \rvert}{\lvert Vc \rvert}\right)$	Foursquare
	Catchment of Locality	$\frac{\lvert D_i \rvert}{\lvert D \rvert}$	Transport
	Temporal Catchment of Locality	$\frac{\lvert D_{w,i} \rvert}{\lvert D_w \rvert}$	Transport
Customer Visit Patterns	Inflow & Outflow	$\frac{\sum_{j=0}^{\lvert V \rvert} t(v_j, vi)}{M}, \quad \frac{\sum_{j=0}^{\lvert V \rvert} t(v_i, v_j)}{M}$	Foursquare
	Distance Travelled to Reach Venue	$\frac{\sum_{j=0}^{N} dist(v_j, v_i)}{N}$	Foursquare
	Speed of Travel to Venue	$\frac{\sum_{j=0}^{N} dist(v_j, v_i) \times t_{i,j}^{-1}}{N}$	Foursquare
	Temporal Popularity Skew	$\sum_{i=1}^{24} h_i * \ln h_i / \ln 24$	Foursquare
	Visit Trend	$\frac{c_t(v_i) - \bar{b}}{t}$	Foursquare
	Temporal Alignment with Competitors	$\sum_{j=1}^{24} (h_i(j) - H_i(j))^2$	Foursquare
Mobility Dynamics	Temporal Alignment with Locality	$\sum_{j=1}^{24} (h_i(j) - h_l(j))^2$	Both
	Reachability	$r_{(a,b)}$	Both
	Distance-weighted Reachability	$dr_{(a,b)}$	Both
Business Attributes	Cuisine Type	Categoric variables	Foursquare
	Price Tier		Foursquare

3.2 User Review Data

Many features that were present with check-in data can also be determined using review data. However, reviews have two aspects: rating and user comments. Due to the nature of user comments, there is a huge range of data that can be extracted.

Competition and specific competition can also be identified with review data. However, instead of just using a number of competitors we can also assess the competitiveness of the competitors. This will lead to a more valuable analysis of competition. With competition levels based solely on the number of competitors, we have no assessment of how good the competition is. With review data we can use the number of reviews, and rating scores reach a better view of competitiveness within the locality.

Different localities may have different preferences, therefore even though price tiers may be controllable,

A Predictive Model for Business Failure Based On City and Customer Review Data

By: Allan Ruan, Botao Yao, Julian Romero, and Julia Vida

we cannot just assume a high price or low price is bad. For example, having a higher price may be associated with a perceived higher level of quality. While having a low price in an expensive area may be a bad business choice that leads to business failures. Even though a higher price might lead to less customers in low income areas, it may be good for higher end places. For instance, a fancy restaurant may benefit from a higher place whereas a comfort foods restaurant might lose traffic by setting prices high. With reviews we can use the age of a business to show whether they have a history of operation in the area. Businesses that are older have the advantage that they do not need to spend as much on advertisement. Also, if a business has been in operation for a long period of time, they will have accumulated more money to survive in difficult times. Lastly, given that a business has a higher age, indicates that its quality was good enough for it to survive to the present date.

3.3 Fusion of Mobility and Review data

In our study we will be using the dataset from Google Local Reviews. This includes reviews of a business, and its ratings. By combining both mobility and review data, we can have features from both types of data. Also, we can fuse them to make them more effective. Given only check-in data we can only see that there may have been an increase to traffic to a business, but we do not know whether the check in indicates a good experience or a bad experience. With review data, we do not know if there is an increase in traffic, even though more reviews indicates that more people are actually trying the service. Using a fusion of both we may now see a connection between reviews and number of visits to the area. This can show how a business with high traffic but low review count or review ratings can be highly related to closure. Areas that have high traffic are generally more expensive to operate in, thus may be more likely to fail if a business fails to capture customers. Therefore, by looking at both reviews and traffic, we are better able to determine whether a business will fail

4: Main Design

The problem considered in this study is the probability of business closure given customer review, traffic, and census data. The constraints imposed on this study include availability of Google reviews for certain businesses and the limited ability to extract the finer nuances of customer satisfaction through the use of ratings. The ability to use review ratings will help us determine if quality of a restaurant is causing business failure. The design goal is to obtain a clear representation of the significance of quality factors to the success or failure of restaurant businesses.

The variables that are examined in this study can be divided into two main categories: 1) product/service quality, and 2) traffic. Variables in the model are: Average Rating, Number of Ratings, Competition, Number of Drop Offs, Number of Passengers, and Age.

Variable descriptions:

1. Product/ service

   A. Average rating: takes the average rating across all ratings of a particular restaurant.
   B. Number of ratings: This gives the number of ratings a particular restaurant has.
   C. Age: This gives the age of the individual who gave the review.

2. Traffic

   A. Number of drop offs: This gives the number of drop offs to a particular location.
   B. Number of Passengers: This gives the number of passengers that were dropped off.

One assumption made in this study is that customer reviews are an accurate measure of controllable factors such as customer service and product quality. This assumption is valid on a few accounts. Firstly, it is the quality of a business's product from the point

A Predictive Model for Business Failure Based On City and Customer Review Data

By: Allan Ruan, Botao Yao, Julian Romero, and Julia Vida

of view of the customers that is of interest; that is, there is no need to establish an "objective" measure of quality as the "subjective" measure of customers is the crucial metric from a business perspective. Secondly, it can be asserted with a high degree of confidence that Google reviews are in fact indicative of the customer point of view. While Google does not enact a review system that is designed to weed out illegitimate reviews like other review sites such as Yelp, the fairly high degree of specificity of most feedback suggests legitimacy.

5: Evaluation

5.1 Dataset description

The data is obtained online from TLC trip record data from OpenNYC and from a Google reviews dataset. The Google Local Reviews dataset consists of a business dataset and a reviews dataset. There were originally 10 million reviews and 3 million businesses across the world. The reviews are dated from 1990 to 2014. Our study is focused on restaurant failures after year 2012 and utilizes all reviews prior to that. We first processed the businesses data by mapping the businesses to the NYC census tract 2010, which consist of 2164 distinct tracts. Then we joined the businesses data to the reviews data and filtered for those which were categorized as restaurants. This resulted in a total of 9,161 restaurant businesses and 70,486 business reviews. We also obtained the 2012 NYC yellow taxi cab data, consisting of 167 million trip records, and also mapped them to the census tracts.

5.2 Feature Extraction and Model

In order to perform predictions, we have chosen features which are indicative of a business' quality, the businesses environment, and the trendability of the business. To represent a businesses quality we use the age of a business and a businesses average review rating. Age was calculated as the difference between the last review of the business and the first review of the business. This is a proxy for how long the business has been in operation and known to the public. The mean of all business reviews for the business is used to represent the overall quality and customer satisfaction of a business.

Next, we have features for the business' environment which are the traffic in the area and the competition in the area. The traffic feature is based on yellow cab drop offs in the business' tract. Two features are used for traffic: number of drop offs, and passengers dropped off. The number of drop offs indicate how popular the area is. Number of passengers dropped off also gives a baseline of how many people have actually gone to the area. For competition, we utilized a count of all restaurants in the same tract. This shows how many other restaurants are in the vicinity of the restaurant and the choices the customers have. Lastly, we use the amount of reviews a business has to represent how trendy a given business is.

We modeled the features using a logistic regression model because it assigns a probability score for classifications. If a probability is over a certain threshold the model will predict it to be closed. Therefore, logistic regression is suitable for this type of prediction. The results are evaluated against the Google review dataset's ground truth data about the open/closed status of each business. However, with only an open or close label, we cannot determine when the business is closed. In order to approximate the close date of a business we utilize the date of last review. If a business is labeled as closed and the last review for the business is prior to half a year after our testing period, we will consider that the business closed during that time. Logistic regression results will be assessed using the Area Under the ROC Curve as well as Area Under the Precision/Recall Curve.

5.3 Model Prediction and Metrics

The three most statistically significant variables in the logistic regression model are Age, Competition, and Average Rating. Age is negatively correlated with closure, meaning that the longer a business has been around (and subject to reviews), the less likely it is to experience closure. Competition is positively

A Predictive Model for Business Failure Based On City and Customer Review Data

By: Allan Ruan, Botao Yao, Julian Romero, and Julia Vida

correlated with closure, suggesting that businesses that are not top tier will get beat out by their competition. Since competition is determined based on the number of businesses near a given business (in the same Census tract), one might expect that the increased local activity would drive more business and therefore reduce the probability of closure. The results suggest otherwise--that a business must be even more exceptional to stand out in a busy area--but more variables would be needed to confidently draw that conclusion. Nonetheless, the finding contributes to a deeper understanding of competition and its relationship to closure. The third most significant variable is Average Rating. Rating magnitude negatively correlates to closure, which is intuitive as one would expect a higher rating to mean better business prospects.

A further examination of the results reveals that Number of Passengers and Number of Drop Offs near a given business are minimally statistically significant. These variables are less predictive by several decimal places compared to other variables. In an investigation of the signage of the feature importance, it can be seen that like Competition, Number of Passengers is positively correlated with closure. This suggests that as the number of people circulating around a business increases, so does the demand for a high quality product.

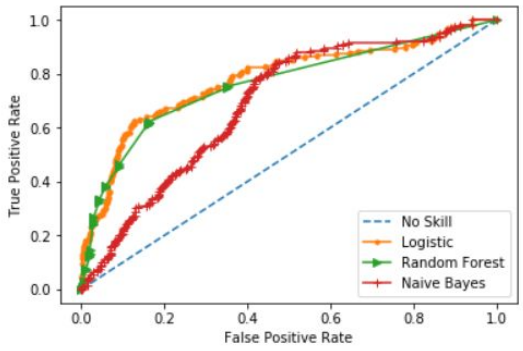
The Average Rating variable is somewhat predictive of business closure, but does not provide as clear a link to closure as one might expect. It is intuitive that a lower average rating would signify a higher probability of closure, and this is in fact the result that is obtained in the model. However, it seems that other "meta" qualities of reviews, such as Age (amount of time between oldest and newest review) shroud the significance of the actual rating values. A possible way to correct for this is to further test for variable interference and to include the number of low, average, and high ratings for a given business, respectively.

Figure 2. Complete feature significance in Logistic Regression model

Feature	Feature Importance
rating_mean	-4.133428e-05
rating_count	-1.144163e-05
competition_first	3.759759e-04
count_dropoffs	-1.982718e-07
passenger_count	5.626287e-08
age	-2.780276e-03

Furthermore, we evaluated our logistic regression model against other models, such as Random Forest Classifier, Naive Bayes, and No Skill. The strongest model was logistic regression which produced an Area Under the ROC curve value of .818. When a random forest algorithm was used, the AUC was .763 and with Naive Bayes the AUC was .691.

Figure 3. A comparative analysis of models predicting business closure by AUC



An analysis of the area under the precision-recall curve reveals that the logistic regression model is ~90% accurate in total predictions and is only ~30% accurate in predicting business closures specifically. This result may be due in part to the fact that the classes "closed" and "not closed" were vastly imbalanced in the data set, with many more "not closed" businesses comparatively.

A Predictive Model for Business Failure Based On City and Customer Review Data

By: Allan Ruan, Botao Yao, Julian Romero, and Julia Vida

Figure 4a. Precision-recall results in full

	precision	recall	f1-score	support
False	0.93	0.94	0.94	1658
True	0.36	0.30	0.32	175

Figure 4b. Confusion Matrix of predictions

		Predicted	
		FALSE	TRUE
Actual	FALSE	1565	93
	TRUE	123	52

While there are corrective measures available for such class imbalances, the predictive results obtained from the models are strong enough to obviate the need for such corrections. Studies have shown that a 30% predictive accuracy can reasonably be translated to a 70% accuracy by shifting the focus to the complement of the originally studied outcome [4]. The model results are sufficiently far from random (i.e. 50%), and therefore so-called “rare events” corrections are not needed.

6: Future Work

In order to introduce greater specificity into the study, it would be useful to take a closer look at a given business’s competitors. Competitors could be categorized by both location and type of business (e.g. fast food, French cafe, Italian restaurant). Then key characteristics for each business are determined, such as average Google Reviews rating. A finer-tuned analysis might include an investigation of factors such as daytime vs. nighttime competition.

Furthermore, the model can be expanded to include natural language processing on reviews. The results of NLP algorithms can be collected into several different variables, such as Number of Positive Words, Number of Negative Words, and Specificity Ranking. Perhaps even grammatical correctness of reviews can be examined to determine review legitimacy and how persuasive a given review is to potential customers.

The model could also be tailored to solve time-related problems, such as analysis on the economic impact caused by COVID-19. Due to the quarantine regulation in many areas globally, some environmental features, such as traffic volume, are not activated and significantly different from the regular time, which might cause unexpected prediction errors. Therefore, to better predict business results and give feedback to the government to make economic surviving decisions, the model is expected to include some new features--including the number of instant delivery orders. Based on the instant delivery data, which falls into the category of controllable factors, the model can better predict the future number of orders for each individual restaurant.

The model could also be likely modified to predict whether it would be safe to dine in the restaurant within a particular district, given the pandemic context. By collecting the COVID-19 patients’ statistical data per partitioned region, the model should also predict the probability of being infected. By this means, the prediction model can also avoid infection cases in modern cities.

7. Conclusion

In this study, we have presented an approach to predict restaurant closure which focuses on the controllable factors of a business. Our model utilizes features that account for a business’s popularity and quality, as well as urban mobility. In this work we were able to achieve a high prediction accuracy with an AUC of approximately 0.81. The results indicate that controllable factors are indeed of high significance in the prediction of a business closure and can be further combined with modeling of

A Predictive Model for Business Failure Based On City and Customer Review Data

By: Allan Ruan, Botao Yao, Julian Romero, and Julia Vida

uncontrollable factors to improve prediction accuracy.

Citation

1. D'SILVA, Krittika; JAYARAJAH, Kasthuri; NOULAS, Anastasios; MASCOLO, Cecilia; and MISRA, Archan. The role of urban mobility in retail business survival. (2018). Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies. 2, (3), 100: 1-22. Research Collection School Of Information Systems
2. Dmytro Karamshuk, Anastasios Noulas, Salvatore Scellato, Vincenzo Nicosia, and Cecilia Mascolo. 2013.Geo-spotting: Mining Online Location-based Services for Optimal Retail Store Placement. In Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '13).
3. Lei Wang, Ram Gopal, Ramesh Shankar, and Joseph Pancras. 2015. On the brink: Predicting business failure with mobile location-based checkins. Decision Support Systems 76 (2015), 3 – 13. Analyzing the Impacts of Advanced Information Technologies on Business Operations.
4. Bagchi, Rajesh, and Elise Chandon Ince. Is A 70% Forecast More Accurate than a 30% Forecast? How Level of a Forecast Affects Inferences about Forecasts and Forecasters. Journal of Marketing Research.

Predicting Business Failure with Human Mobility Data and Crowd Sourced Review Forums

Authors: Allan Ruan, Botao Yao, Julia Vida, Julian Romero

What is our model?

- A business prediction model
- A logistic regression model (Binary classification)
- Data sources are from businesses in New York City neighborhoods
- We will evaluate our results by using a ground truth.

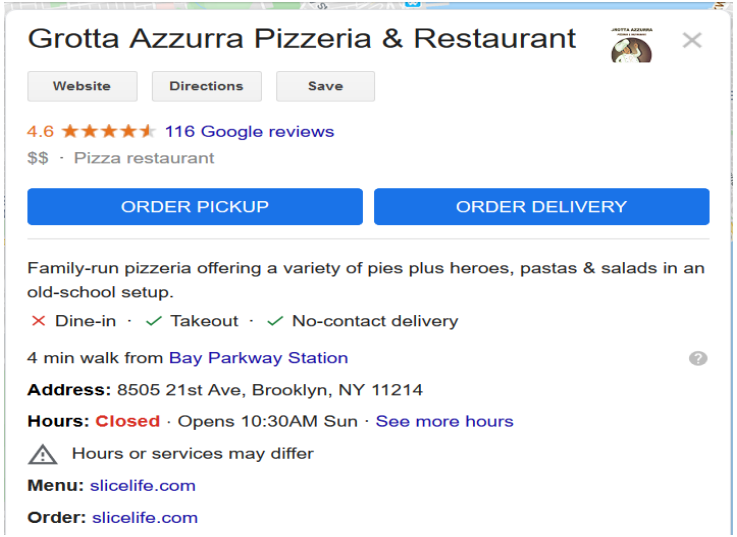
How do we classify a business's susceptibility to failure?

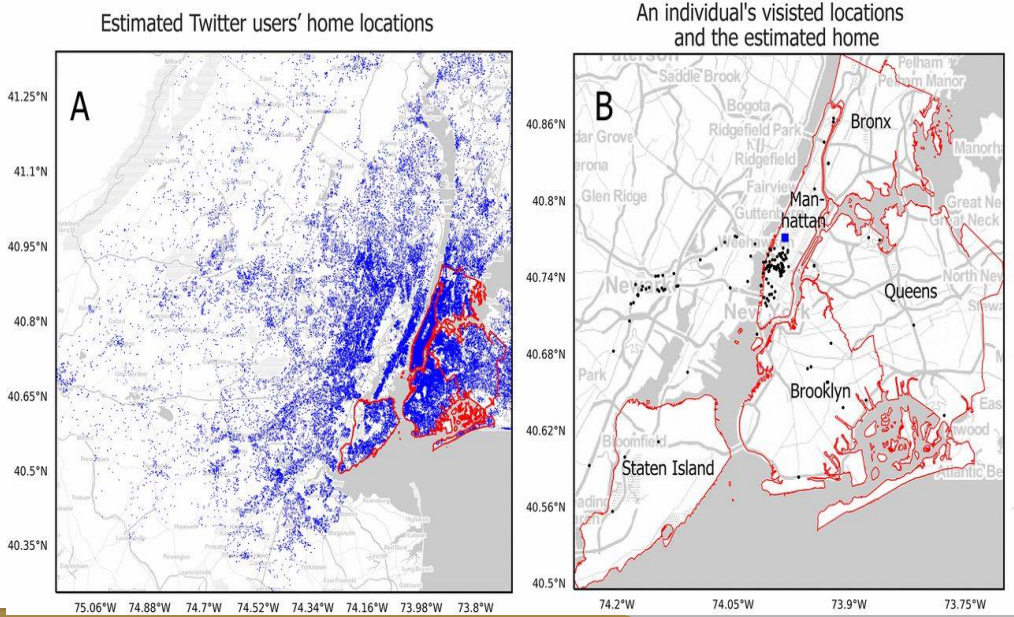
controllable factors:

- **Trendability**
- **customer satisfaction.**

uncontrollable factors:

- **Human mobility**



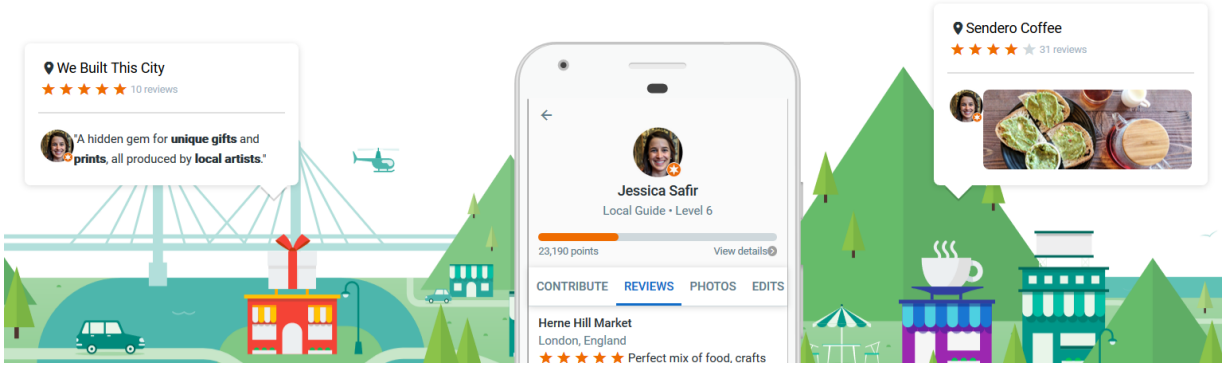


Key contributions

- **Trendable factor:** Measures trendability of a business.
- **Crowd source review forums:** Measure controllable factors.
- **Human Mobility:** Measure uncontrollable factors.

Google Reviews

Main Design - Datasets



Google

NYC Taxi and Limousine Commission (TLC) Trip Records

2019

January	July
• Yellow Taxi Trip Records (CSV) • Green Taxi Trip Records (CSV) • For-Hire Vehicle Trip Records (CSV)	• Yellow Taxi Trip Records (CSV) • Green Taxi Trip Records (CSV) • FHV Trip Records (CSV) • High Volume For-Hire Vehicle Trip Records (CSV)
February	

Main Design

Variables:

- Continuous (how many reviews for a business)
- Scale-Based (how high-traffic is a region on a scale)
- Business-specific (business age and rating)

Assumptions:

- Google reviews are an accurate measure of controllable factors such as product quality

Logistic regression

- Data Date range - 2000 -2013
- 80% for training and 20% for testing

Evaluation

- Data Details
  - 70,486 NYC Restaurants reviews
  - 9,161 Distinct Restaurant Businesses
  - 2012 Yellow Taxi data - 167 million trip records
  - 2164 NYC Census Tracts
- Ground Truth
  - Google Reviews closed label
  - Last Review Date before June 2013

Evaluation - Features

Quality

- Average Review Rating
- Age

Environment

- Competition
- Traffic

Trendability

- Number of Reviews

Feature	Feature Importance
rating_mean	-4.133428e-05
rating_count	-1.144163e-05
competition_first	3.759759e-04
count_dropoffs	-1.982718e-07
passenger_count	5.626287e-08
age	-2.780276e-03

Evaluation - Predictions

- Accuracy
  - ~90% accuracy in total predictions
  - ~30% accuracy in predicting business failures

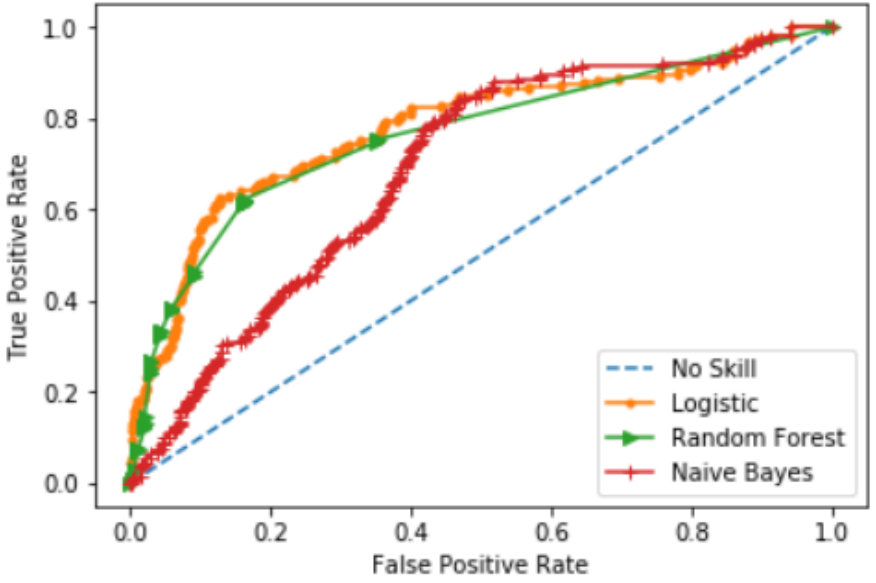
- Prediction Scores

		Predicted	
		FALSE	TRUE
Actual	FALSE	1565	93
	TRUE	123	52

	precision	recall	f1-score	support
False	0.93	0.94	0.94	1658
True	0.36	0.30	0.32	175

Evaluation - Models

- Baseline - No skill
- ROC Curves
  - No Skill
    - Auc = .5
  - AUC
    - Logistic = .818
    - Random Forest = .763
    - Naive Bayes = .691



Conclusion

- Business Failure Prediction
  - Business popularity
  - Controllable factors
  - Urban Mobility
- Further improvements
  - Average google rating of competitors
  - Increase/decrease of traffic over time