Promoting Sustained Entrepreneurship in Chicago

Katy Koenig, Eric Langowski & Patrick Lavallee Delgado

Github Link: https://github.com/erhla/ml-chicago-business

I. Background

Chicago is "the city that works." In January 2018, the Illinois Department of Employment Security announced that almost 175,000 jobs had been created in Chicago since 2010.[1] This growth is not indicative of a healthy region, however, as the majority of new jobs were in Chicago's Loop and the Chicago area population has consistently declined over the same time period. Creating healthy businesses in every community is critical for the city's future sustainable growth. An unhealthy business community can lead to negative outcomes for neighborhoods and negative externalities for the city, as declining tax bases force resource to be cut in death spirals. Or as Sociologist Robert J. Smapson wrote in *Great American City,* "bad locations mean bad business."

Alongside long standing narratives of segregation and racial inequity, the current business environment is indicative of structural and systematic under- and disinvestment in Chicago's communities of color. People of color own only 32% of businesses in Chicago while only 32% of Chicago's population identify as non-Hispanic white. The average white-owned business is valued more than 12x Black-owned business in Chicago.[2] We aim to offer actionable analysis and policy recommendations that equip the City to improve the state of entrepreneurship and position traditionally underinvested communities towards sustainable business growth.

II. Related Work

Prediction of business failure and success has historically been viewed through the lens of statistical and financial analysis. Only in roughly the past decade have we seen machine learning tools applied to prediction of business failure. In our review of previous analysis, we found that most applications of machine learning in this field have centered not only around business failure but also around business-specific attributes as indicators of business success/failure. Specifically, we see a focus on the financial statistics of each business, e.g. liquidity ratios, net profit margins, as selected features[3] imputed into machine learning models.

While models centered around monetary attributes of businesses have found success in predicting business failure and expanded the use of machine learning methods in predictions of

---

[1] https://www.chicagobusiness.com/article/20180102/BLOGS02/180109998/chicago-job-growth-hits-highest-peak-in-decades

[2] https://prosperitynow.org/files/resources/Racial_Wealth_Divide_in_Chicago_OptimizedforScreenReaders.pdf

[3] Example: https://www.researchgate.net/publication/286969333_Application_of_machine_learning_algorithms_for_business_failure_prediction

business failure, in our analysis, we seek to expand upon this research by analyzing the "neighborhood effect" (e.g. the effect of location) on business success.

III. Problem Formulation

In this analysis, we seek to obtain the attributes of a successful business for use in subsidizing the initial capital investment of new businesses throughout Chicago. Through analyzing the businesses for which our models deem the most successful, we are able to find the attributes which indicate whether a new business will be successful. In application, we will use these important features to justify the granting of subsidies to new businesses by the City of Chicago. Please note that throughout our analysis a business is deemed "successful" if it in operation for more than two years. We discuss the reasoning for this definition in our Section IV. Data Description below.

We begin by running a variety of machine learning classifiers on our data, as described below, to find the classifier with the highest area under the receiver operating characteristic curve (AUC). We chose AUC as our metric for evaluation as it is a metric that evaluates the models for overall performance (at all thresholds).

We then create a list of features that are strong indicators of success. In practice, this list will be applied when the City of Chicago chooses new business ventures to subsidize initial investment: businesses exhibiting these characteristics will be more likely to receive grants. We use AUC as our evaluation metric because AUC evaluates models with the best fit overall. Because the application of our best model is not an intervention of a certain percentage of storefronts, we do not use precision or recall at a given threshold to evaluate our models.

Furthermore, we also aggregate the predictions of our best model to create an average predicted score of business success for each census tract approximating the "neighborhood effect" that opening a business in that tract has on future success. There are approximately 800 census tracts in Chicago. We then create a list of census tracts with the top ten percent of highest average predicted scores. Regarding our intervention, this census tract list can be used as weeding mechanism: we will not grant subsidies to businesses within these census tracts as businesses within these census tracts are predicted to succeed at higher levels than those in other census tracts and therefore, our grant money would be better spent in areas in which businesses are less successful: our goal is to promote equitable entrepreneurship throughout Chicago, not new businesses in areas in which there are already many structural factors which support business success.

IV. Data Description

We use the following data sources in our analysis:

1. Business License Data from the City of Chicago Data Portal
2. American Community Survey Data as provided by the Census Bureau
3. Reported Crimes Data from the City of Chicago Data Portal

IV.A Business License Data

The business license data is publicly available through the City of Chicago Data Portal. This data is available for licenses issued from 2002 to present. While this data originally is formatted as each row being a license, we converted this data to be formatted as each row being a business storefront[4]. When we make this conversion to each observation becoming a storefront, we simultaneously complete our temporal split, thus creating three sets of training and testing data for the following dates:

Training Set Dates	**Testing Set Dates**
01 Jan 2010 - 31 May 2012	1 June 2014
01 Jan 2010 - 31 May 2013	1 June 2015
01 Jan 2010 - 31 May 2014	1 June 2016
01 Jan 2010 - 31 May 2015	1 June 2017

When pivoting our data from license-focus to storefront-focus, we create rows that denote the type of license each storefront held during the given time period, denote if the business was successful during this time period or during the validation time period in the next two years.

We judge the success of a business as a business surviving more than two years. This definition was chosen as length of business license vary by type and by year, with the longest license issued being two years. Therefore, by choosing just over two years, any given business license would have the opportunity for renewal.

Regarding missingness, we confirmed imputation strategies on storefront characteristics that we do not observe elsewhere in the data. For business licenses with either an empty issue or expiry date, we will deduce the date from the average term of that license type. However, we cannot impute dates on licenses missing data in both fields, which could affect the accuracy with which we measure how long a business has existed. Also, we cannot impute census blocks on storefronts for which the City has withheld data on its location on each of its licenses upon

[4] A storefront is defined as each unique account number plus site number from the business license data: if a business has multiple locations, each location has its own storefront id.

request of the owner. Without further recourse in the confines of this project by which to recover these missing data, we must drop these observations.

In addition to features regarding the type of license owned by each storefront, we also include the month the first license for each storefront was issued as an indicator to see if time of year of opening impacted the success of a given storefront. Additionally, we also include the number of open storefronts in a given storefront's block[5]. Below we include summary statistics for our largest training set (1 Jan 2010 to 31 May 2015):

Attribute	**Count**
Total Storefronts	112,729
No. of Successful Storefronts	91,907
No. of Non-successful Storefronts	20,822
No. of License Types	127
Avg. No. of Storefronts on Block	39
Max. No. of Storefronts on Block	852

From the table above, we can conclude that roughly 82 percent of storefronts in our training set survive for more than two years. We also see a large variation from block to block. Specifically, with respect to the number of storefronts on a block, we can conclude that location is a factor in business success.

IV.B American Community Survey Data

The American Community Survey (ACS) five-year data is publicly available through the U.S. Census Bureau website. We used the ACS data representing the average of time window from 2009 to 2013 to aggregate demographic data by block group. While, there is data available for the time frame 2013 to 2017, this data was not utilized as it does not reflect the data that will be available in realistic machine learning scenarios: The five-year ACS data is not available until the following year, i.e. ACS five-year data from 2013 to 2017 is not available until 2018. For our models to predict success for all storefronts open on 1 June 2017 in realtime, i.e. make predictions on 1 June 2017, we would only have the ACS data from 2013 to 2017. Thus, by using the ACS data from 2013 to 2017, we are mimicking a realistic application of our process.

Similarly, we sought to train our models on only data that would be available during the training periods. Unfortunately, due to the availability on API for the U.S. Census Bureau, we were unable to request data for 2004 to 2009, so our analysis does suffer from training our models being on demographic data that was currently being collected for some of our training sets.

[5] Please note that a block is a subset of a block group. On average there are 39 blocks per block group.

Additionally Census data before 2010 does not include block level data and uses 2000 census tracts, making it generally incompatible with our analysis.

Demographic attributes of a block group are included as percentages as high raw numbers in all demographics are inherently correlated with large population, which was included as a feature in and unto itself. Below, we provide summary statistics of the demographic features across block groups included in our analysis for our largest training set (1 Jan 2010 to 31 May 2015)[6]:

Attribute	**Avg. Count/Percentage[7]**	**Range**
Population	1245.20	0.00 - 8,758
Pct. Male Children	23.63	0.00 - 80.30
Pct. Male Working Age	66.40	0.00 - 100.00
Pct. Male Elderly	9.65	0.00 - 80.10
Pct. Female Children	20.95	0.00 - 52.69
Pct. Female Working Age	65.94	0.00 - 100.00
Pct. Female Elderly	12.83	0.00 - 86.92
Pct. White	31.34	0.00 - 100.00
Pct. Black	36.23	0.00 - 100.00
Pct. Asian	4.96	0.00 - 90.85
Pct. Hispanic	25.56	0.00 - 100.00
Pct. Below Poverty Line	30.24	0.00 - 100.00
Pct. Below Median Income	31.48	0.00 - 100.00
Pct. Above Median Income	19.29	0.00 - 58.11
Pct. High income	14.42	0.00 - 72.77
Pct. Low Travel Time	35.51	0.00 - 96.41
Pct. Medium Travel Time	48.27	0.00 - 100.00
Pct. High Travel Time	16.01	0.00 - 70.97
No. of Bachelor's Degrees	307.47	0.00 - 3752

[6] We include a description of each column in Table 1 of our Appendix.
[7] All percentages have been rounded to the nearest second decimal place.

IV.C Reported Crimes

Our reported crimes data is publicly available through the City of Chicago data portal. For our training sets, we aggregate the reported crimes data by block group, and then find the total reported crime data for the length of our training set and divide by the number of days in the training set for an outcome of features depicting average crime statistics for a day. For our testing set, as it is a snapshot of storefronts alive on a particular day, we aggregate the reported crimes data for the year prior to this snapshot and divide by 365 days.

We include the raw crime numbers, as opposed to those scaled by population, because while we expect that more crimes happen in more populous block groups, we believe that amount specific crimes in a block group more directly influences the success of a storefront.

As there are 33 types of reported crime, each of which was made into a dummy column and thus, too multidimensional to easily aggregate below, we include summary statistics for each block group of the total reported crimes for our largest training set (1 Jan 2010 to 31 May 2015), the number of reported crimes that resulted in arrest and the number of domestic crimes below. We include the number of reported crimes that resulted in arrest as it serves as an indicator of successful policing (the reported crime led is closer to a “solved” crime outcome), and it also tempers the issue of using reported crime data: reported crimes may include crimes that did not in fact occur; while crimes that resulted in arrest seem more likely to include crimes that actually took place in a block group. We include number of domestic crimes as crimes that happen outside of the home and therefore in public seem more likely to affect storefronts as they are public areas.

Attribute	**Count/Percentage**
No. of Different Types of Reported Crime	33
Avg. Total Reported Crimes by Block Group	0.416
Range of Total Reported Crimes	0.00 - 7.63
Avg. Total Arrests by Block Group	0.12
Range of Total Arrests	0.00 - 2.14
Avg. No. of Domestic Reported Crimes by Block Group	0.06
Range of Total Domestic Reported Crimes	0.00 - 0.44

V. Methodology

We link our storefront data to our reported crimes, which has been grouped by block group, and ACS demographic data through a spatial join of the storefronts’ latitude and longitude. We run the following machine learning classifiers on four sets of training/testing data sets (dates of each

set are noted in the data section above): random forest, logistic regression, decision tree, k nearest neighbors, Adaboost and bagging. We use a variety of parameters for each model: the details of which can be viewed in our Github repository. We ran our models both with and without business license types as features, but found that the policy applications of license types being deemed our most important features to be limited and ultimately removed license type from our final models.[8] For each model we created a small decision tree to extract feature importance for successful businesses.

From our results, we choose the model with the highest AUC score as our best model. Once we identify our best model, we aggregate the predicted scores for each storefront by census tract: for each census tract, we find the average predicted score of business success. This average predicted score is then used as a proxy for neighborhood effect and the census tracts are ranked into deciles. We also create a list of the top ten percent of census tracts (roughly 80 tracts total).

VI. Evaluation & Interpretation of Results

Overall, we found minimal performance differences across different testing-training splits. Below, we include a chart detailing the models and the parameters with the best AUC scores for each testing and training period.[9]

Model	**Parameters**	**Testing Dates**	**Training Dates**	**AUC**	**Accuracy at 5%**
KNN	Weights: distance Neighbors: 50	01 Jan 2010 - 31 May 2012	1 June 2014	0.525877	0.331733
KNN	Weights: distance Neighbors: 50	01 Jan 2010 - 31 May 2013	1 June 2015	0.522634	0.330156
Random Forest	Max. Depth: 1, N. Estimators: 10, Min. Samples Split: 2	01 Jan 2010 - 31 May 2014	1 June 2016	0.526554	0.350147
Random Forest	Max. Depth: 1, N. Estimators: 10, Min. Samples Split: 10	01 Jan 2010 - 31 May 2015	1 June 2017	0.528473	0.41287

While our AUC score dips slightly from our smallest training set to our medium training sets, our model trained on the largest dataset does ultimately have the highest AUC. Because the best model for our largest and second largest dataset is a random forest which averages ten decision stumps, our results suggest that our available features offer minimal predictive power of business success above the “random” baseline for AUC (0.50). Still, we believe that our

[8] For example, diversity of business type is important for a thriving community and if the “tavern” license type was most indicative of business success, do we only want to grant subsidies to storefronts that are applying to be “taverns?”

[9] The models in the table were not trained with license type.

model has predictive power on aggregated geographic levels (the census tract) and the poor performance of our models are indicative of random (and uncaptured) traits of individual businesses.

Furthermore, when we view the performance of the accuracy of our best models at the arbitrary threshold of five percent, we see that our models performed worse on accuracy than if each observation had been classified as successful: as noted in the data section above, roughly 82 percent of our observations were successful. Since our data is imbalanced (about 82% of businesses were successful) accuracy is not the best metric.

While our models themselves may not be the most useful in classifying businesses as successful using all of the features we provided, they may provide us with a list of features that are exhibited by successful business. Below, we provide the top ten most important features for our largest training set when business license type are not included. For a more visual comprehension, we also include a decision tree of these features in our appendix.

Ten Most Important Features Without Business License Type Excluded	
Feature	**Gini Importance**
Month Issued	0.27010217178311974
Storefronts on Block	0.1547413330639465
Total Arrests	0.08637195735852941
Total Bachelor's Degrees	0.0859666875481833
Pct. Low Travel Time	0.05185265668933306
Pct. Female Work	0.0424295487988854
Pct. Male Working	0.04211129963741711
Police District	0.0406427375423391
Pct. Below Median Income	0.038303627746717785
Theft	0.03728184580067102
Total Population	0.03145075498338729
Pct. White	0.031134557170196427

Below, we provide the top five most important features for when business license type is included. Please note that we only included five features in the list below as only five features exhibited a gini importance above 0.01.

Five Most Important Features Without Business License Type Included	
Feature	**Gini Importance**
Limited Business License	0.6176142104505442
Regulated Business License	0.1549213150622894
Retail Food Establishment	0.12521408731145942
Home Occupation	0.09182518166756577
Home Repair	0.010020747651007265

As we can see above, models trained with business license type information deemed some license types as most indicative of business success. While this may be valid, its future application seems problematic: do only grant subsidies for new businesses to businesses that applied for a limited business license? Moreover, a limited business license represents an incredibly large variety of businesses as this license is required for "general retail sales and services."[10]

More applicable are the features deemed important when our models are not run with license type information. As shown above, the number of other successful storefronts in the block group is a barometer of a given storefront's success, implying that there exists a positive feedback mechanism at work.

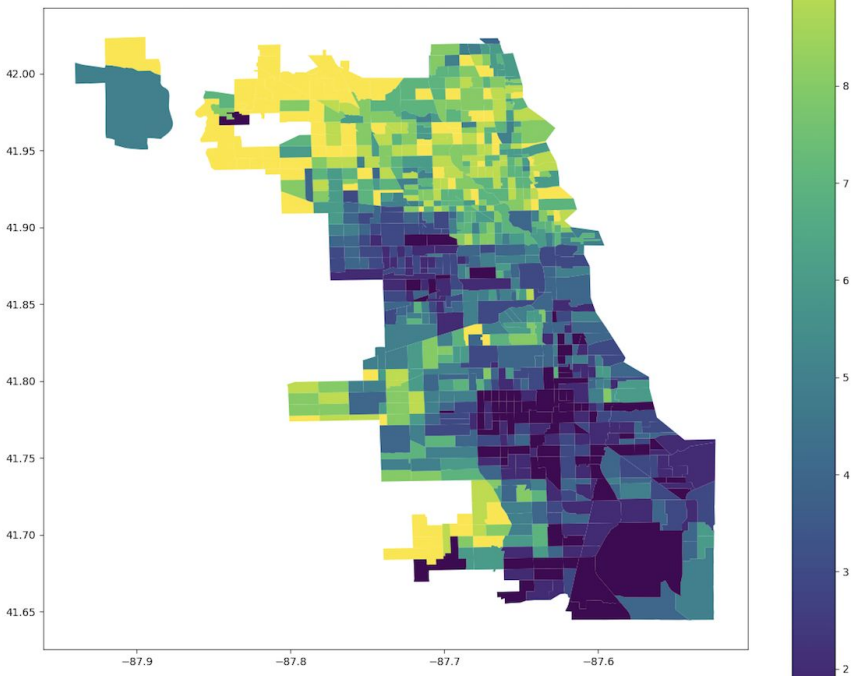
While our important features list has total reported crime and theft as important features for the success of a storefront, because we did not scale these numbers by population, it is likely that their high ranking in importance is due to the correlation between population and crimes, i.e. generally, when a population is high, raw number of crimes in all categories are higher than in less populated areas[11]. It is also important to note that police district, which was included as a location marker, is an important feature, signifying that location in and of itself is a factor is business success.

In the map below, the census tracts are plotted by deciles by deciles, ranked on average predicted scores of our best classifier for our largest training dataset. Please note that ten, or yellow, represents the tracts with the highest average prediction scores of success while one, or

[10] https://www.chicago.gov/city/en/depts/bacp/sbc/business_licensing.html

[11] While most block groups contain 600 - 2,000 residents, in our analysis, we noticed that there are occasional block groups that have populations much larger and much smaller, e.g. one block group in which a business was located had a zero population as it only contained part of N. Lake Shore Dr. and a beach.

the blue, represent the census tracts with the lowest predicted scores of success. We also provide a list of the census tracts in the highest decile in our appendix. We ultimately aggregate up to the census tract level as opposed to the block group level for ease of policy application: while there are more than 10,000 block groups, there are roughly 866 census tracts[12] in Chicago, so aggregating a list of census tracts within the top decile that is easily viewable and usable. It is worth noting the uneven distribution of census tracts within the top decile: we see census tracts with the highest average predicted success in mostly in the Northside while the tracts in the lowest decile are focused in the Southside, further demonstrating the necessity of promoting equitable entrepreneurship throughout the City.



VII. Policy Recommendations

Our analysis allows us to make the following recommendations:

Regarding our list of features which are indicative of successful businesses, we recommend applying only some of the features generated through our models trained not using license data. Specifically, we would prioritize new businesses in areas where there is a high working population and high percentage of residents below the median income. Moreover, while the importance of our features such as police district and percent white indicate successful but also inequitable growth, we do not recommend applying either demographic feature as benchmark to be reached in granting subsidies to new businesses.

Our aggregated predictions allow us to create neighborhood effect scores for each census tract. A policy maker could use these scores to identify tracts with high potential for business success or areas in need of greater investment. Moreover, in our grant application process, we would

[12] https://www.lib.uchicago.edu/e/collections/maps/censusinfo.html

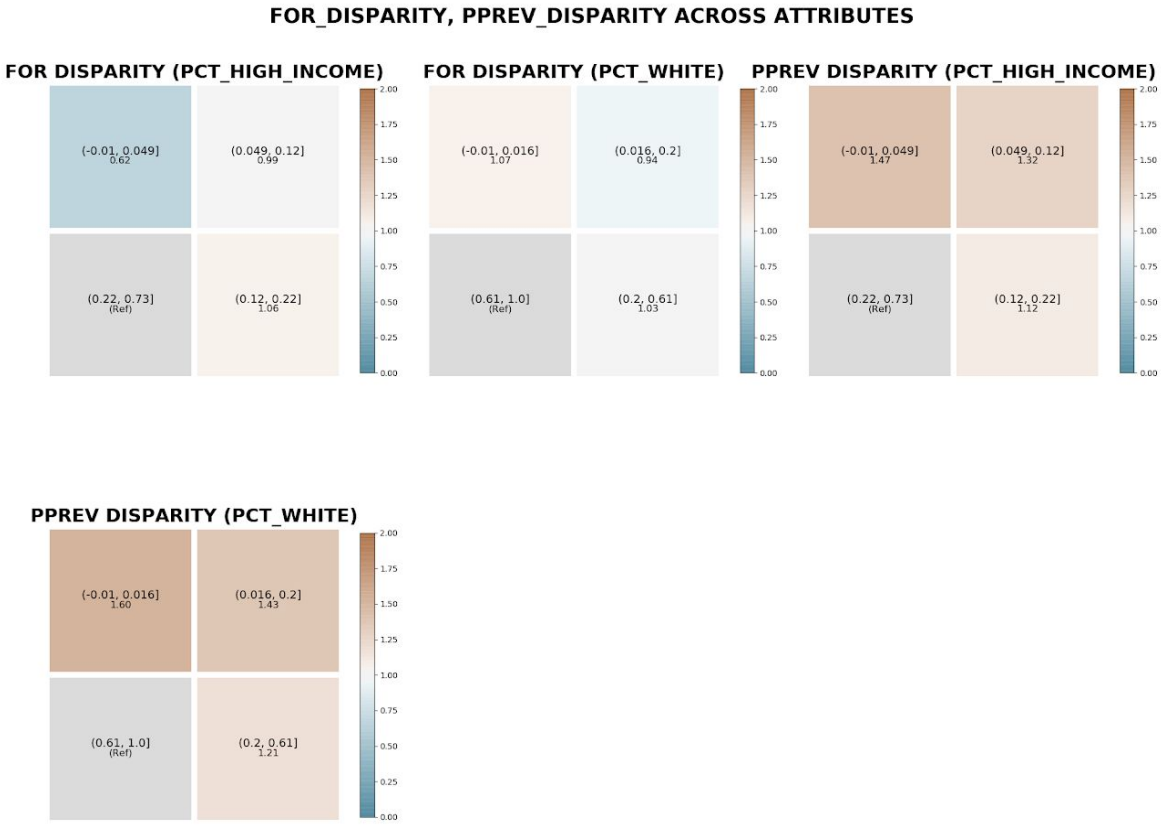
deprioritize businesses in the highest deciles of our neighborhood effect scores and our prioritize new ventures located in the lower deciles to foster equitable growth throughout Chicago.

VIII. Bias & Fairness Analysis

Just as our methodology considers block group as a feature, we evaluate bias and fairness in our results on the same unit of analysis. Our model assigns a score to each storefront in the validation set, so we aggregate to the block group level per the average score of storefronts in each block group. The top decile of census tracts realized an average score above 0.8207; we use the same threshold to assign block groups to the positive class. This allows us to audit demographic representation in the collection of census blocks that our analysis suggests exhibit characteristics favorable to business.

Aequitas is a bias and fairness analysis toolkit developed by the Center for Data Science and Public Policy at the University of Chicago. Aequitas discretizes our validation set to assign block groups with shared characteristics to mutually exclusive buckets. For this analysis, we consider quartiles on the proportion of white residents and quartiles on the proportion of high income residents in each block group. We compare representation of block groups among these quartiles with respect to the most white and most wealthy block groups. We must note that while other parts of our study consider census tracts, we use block groups here to complement the resolution in demographic data in the American Community Survey.

Measuring bias and fairness, we are interested in *false omission rate* to identify demographic groups who reside in census blocks conducive to business that our model classifies as otherwise. We are also interested in *predicted prevalence* to measure the fraction of census blocks correspondent to a demographic group that our model classifies as likely to support businesses to reach our definition of success. The results are not too surprising:



The treemaps above demonstrate the Aequitas bias assessment. We do not see meaningful disparity in false omission rate, which may be because the majority of businesses satisfy our definition of success for having existed for at least two years. We do see disparity in predicted prevalence, which proves that fewer successful businesses exist in block groups with smaller proportions of high income and white residents. The disparity widens as block groups become less affluent and less white with respect to the most white and most wealthy block groups.

These treemaps of the Aequitas fairness assessment reconfirm the previous bias analysis. The false omission rate suggests that block groups with any proportion of white and high income residents realize parity in this respect, except for block groups with the least affluence. The predicted prevalence, however, shows that our data do not have parity of representation with respect to the most white and most wealthy block groups. We believe that this also results from the unequal concentration of storefronts in Chicago around whiter and wealthier communities.

IX. Limitations/Caveats

Lastly, our analysis will not be without its limitations. The City's business license has low dimensionality. The data does not have demographic information of its licensees or financial information of the business itself, so while we can suggest blocks in socioeconomically diverse neighborhoods, there is no indication of whether the entrepreneurs who benefit from intervention are themselves representative of those communities. Other factors contribute to the success of a business, particularly its business model and management; these may correlate with the level of education of the business owner, and other identifying characteristics that do not exist in the dataset. Our models may improve drastically if information regarding the business owner and the businesses themselves were available for our analysis as has been included in previous machine learning work regarding business failure prediction. Also, our

recommendations would only assist storefront registered in Chicago, which is complicated by telecommuting organizations.

Regarding our crime data, we must note that this data includes only reported crime. Therefore, it lacks both crimes that occurred and were never reported and includes crimes that never occurred but were reported, introducing errors into our training data. Similarly, as we used a scaled aggregate of crimes over our training periods with each increase in length of our training periods, we lose precision in average crimes and crime types reported per day: our analysis disregards changes in reported crime data for each block group over time.

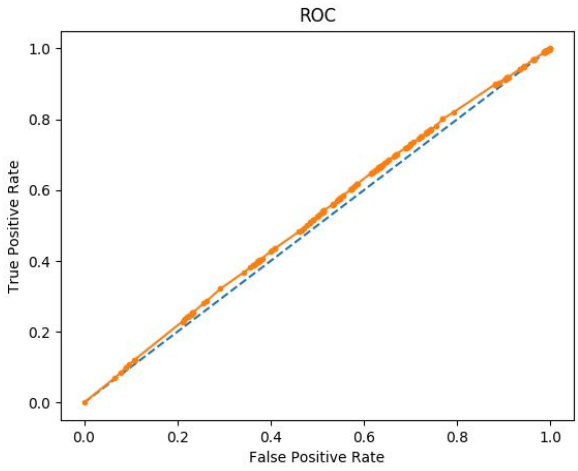
It is important to note that while we seek to optimize our results via training our models on the longest time frame, the poor performance of our models and the variation of the model deemed most successful based on our evaluation metric suggests that there may be underlying trends in that data. Specifically, we suspect that our storefronts data may have been impacted by the 2008 recession while our later data was not, thus violating a common trends over time assumption. Moreover, because our data varies greatly in the time period of January 2010 (the start of our training data) to May 2017 (the last testing set), models may see increased performance if trained on datasets beginning after the Great Recession.

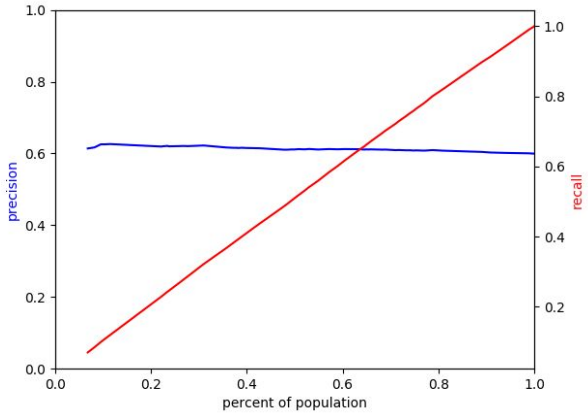
X. Appendix

Table 1: Description of Demographic Feature Breakdown	
Column	**Definition**
Child	Residents Aged 0 through 19
Working Age	Residents Aged 20 through 64
Elderly	Residents Aged 65 and Older
Below Poverty Line	Annual Household Income below $25,000
Below Median Income	Annual Household Income between $25,000 and $60,000
Above Median Income	Annual Household Income between $60,000 and $100,000
High Income	Annual Household Income Above $100,000
Low Travel Time	Travel Time to Work Below 25 Minutes
Medium Travel Time	Travel Time to Work between 25 and 59 Minutes
High Travel Time	Travel Time to Work 60 minutes and above

We made the age division as it seem to storefronts to identify working age population as a proxy to those with income to spend at a given storefront. We divided income in the above way as the federal poverty line for a household of four is $25,750 and the median household income in the U.S. in 2017 was $59,039. We divided travel time to work at 25 minutes as the average travel time to work in the U.S. is 26.6 minutes. We arbitrarily chose 60 minutes as our high travel time as it was more than twice the cutoff for our low travel time.

Plot I: ROC and Precision-Recall Curves for Highest AUC Model: Random Forest Model (Parameters: Max. Depth: 1, N. Estimators: 10, Min. Samples Split: 10) Trained on Largest Set of Data not including License Type Included





Plot 2: Decision Tree of Most Important Features

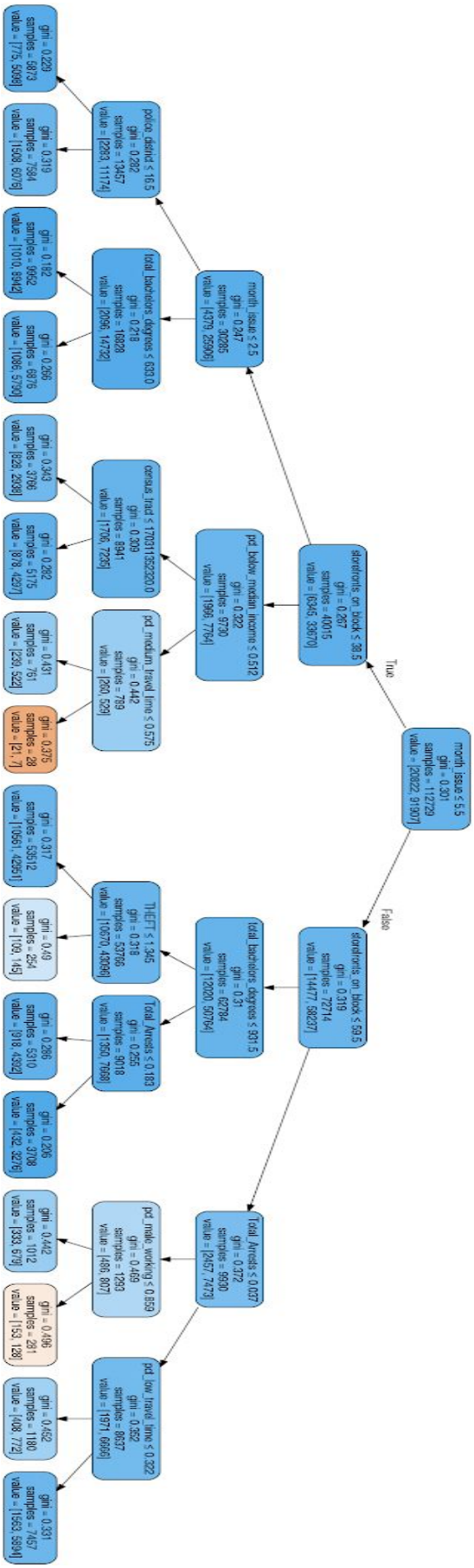


Table 2: Census Tracts with Highest Average Predicted Probability of Successful Businesses					
Census Tract	**Avg. Predicted Probability of Success**	**Census Tract**	**Avg. Predicted Probability of Success**	**Census Tract**	**Avg. Predicted Probability of Success**
17031050300	0.8224877375254688	**17031170800**	0.8216143179389678	**17031030400**	0.8208405831026988
17031760802	0.822487737525468	**17031180100**	0.8213719120298909	**17031020801**	0.820833515850894
17031080100	0.8224877375254676	**17031740300**	0.8213592634309688	**17031191000**	0.8208178255949773
17031170600	0.8224877375254674	**17031220300**	0.8213223725981961	**17031051200**	0.8208077919277742
17031720500	0.8224877375254674	**17031150600**	0.8213122813808992	**17031040800**	0.8208051597202952
17031090200	0.8224877375254674	**17031200300**	0.8213085784314363	**17031770602**	0.8207725970250576
17031170900	0.8224877375254673	**17031770700**	0.8213003325636451	**17031230300**	0.8207725970250571
17031071600	0.8224877375254673	**17031171000**	0.8212999393784298	**17031720200**	0.8207669193156331
17031060800	0.8224877375254673	**17031832200**	0.8212725424439545	**17031640800**	0.8207269751405387
17031100700	0.8224877375254672	**17031150502**	0.8212663773891331	**17031840300**	0.8206790431476224
17031750200	0.8224877375254672	**17031160502**	0.8212515579135048	**17031590600**	0.8206574717176852
17031090300	0.8224877375254672	**17031030604**	0.8212272381044667	**17031160800**	0.8206433733928772
17031030702	0.8224877375254672	**17031070103**	0.8212258041715885	**17031150300**	0.8206297245198205
17031060100	0.8224877375254672	**17031060400**	0.8212258041715881	**17031160100**	0.8206200761617262
17031120100	0.8224877375254672	**17031031000**	0.8212204551001303	**17031130300**	0.8206118577744185
17031150501	0.8224877375254672	**17031100500**	0.8211901924277019	**17031100100**	0.8205668708576456
17031720400	0.8224877375254672	**17031250500**	0.8211751252917434		
17031100400	0.8224877375254672	**17031190200**	0.8211374343484228		
17031810400	0.8224877375254672	**17031560300**	0.8211086374808578		
17031040700	0.8224877375254663	**17031640500**	0.8211086374808577		
17031120300	0.822428584399504	**17031120200**	0.8211024893420904		
17031150700	0.822379599862377	**17031061901**	0.8210798848073557		
17031170400	0.8222203641826636	**17031090100**	0.8210427843179283		
17031190702	0.8220965225623851	**17031210700**	0.8210358094933575		
17031170700	0.8220453561764617	**17031740200**	0.8209976141391975		
17031110300	0.8219470492100159	**17031071400**	0.8209686729127411		
17031071700	0.8219469089452334	**17031020400**	0.8209525636449998		
17031740400	0.821810917208153	**17031110100**	0.820909406698919		
17031050100	0.8217343444783757	**17031170500**	0.8208805168454232		
17031100200	0.8217273469390811	**17031140800**	0.8208666603618916		
17031040402	0.8217264098006875	**17031170200**	0.8208656725791128		
17031100300	0.8216237410332836	**17031063302**	0.82085828505739		

Data Splits:

Training Set Dates	Validation Gap	Testing Set Dates	Validation End Date
01 Jan 2010 - 31 May 2012	01 June 2012 - 31 May 2014	1 June 2014	31 May 2016
01 Jan 2010 - 31 May 2013	01 June 2013 - 31 May 2015	1 June 2015	31 May 2017
01 Jan 2010 - 31 May 2014	01 June 2014 - 31 May 2016	1 June 2016	31 May 2018
01 Jan 2010 - 31 May 2015	01 June 2015 - 31 May 2017	1 June 2017	31 May 2019

Feature List:

1. storefronts_on_block
2. police_district
3. total_bachelors_degrees
4. total_population
5. pct_male_children
6. pct_male_working',
7. pct_male_elderly
8. pct_female_children
9. pct_female_working
10. pct_female_elderly
11. pct_low_travel_time
12. pct_medium_travel_time
13. pct_high_travel_time
14. pct_below_poverty
15. pct_below_median_income
16. pct_above_median_income
17. pct_high_income', 'pct_white', 'pct_black',
18. pct_asian
19. pct_hispanic
20. Census_tract
21. HOMICIDE
22. OTHER OFFENSE
23. ROBBERY
24. THEFT
25. NARCOTICS
26. BATTERY
27. ASSAULT
28. CRIMINAL DAMAGE
29. CRIMINAL TRESPASS
30. PUBLIC PEACE VIOLATION
31. MOTOR VEHICLE THEFT
32. DECEPTIVE PRACTICE
33. WEAPONS VIOLATION

34. INTERFERENCE WITH PUBLIC OFFICER
35. BURGLARY
36. CRIM SEXUAL ASSAULT
37. OFFENSE INVOLVING CHILDREN
38. PUBLIC INDECENCY
39. SEX OFFENSE
40. KIDNAPPING
41. PROSTITUTION
42. INTIMIDATION
43. ARSON
44. LIQUOR LAW VIOLATION
45. CONCEALED CARRY LICENSE VIOLATION
46. GAMBLING
47. OTHER NARCOTIC VIOLATION
48. STALKING
49. OBSCENITY
50. HUMAN TRAFFICKING
51. NON-CRIMINAL
52. NON-CRIMINAL (SUBJECT SPECIFIED)
53. Total_Crimes
54. Total_Arrests
55. Total_Domestic
56. Month_issue
57. License type (as a dummy column for each available license type)[13]

Additional Feature Wishlist:

1. Transit information: number of public transit stops within a one mile radius of the business
2. Total number of crimes divided by the population
3. 311 Data[14]
4. Time between application and issue date of licenses for new storefronts
5. Percentage of businesses still open after two years in block group
6. Length of time a business has existed (while also avoiding perfect multicollinearity with our outcome variable)
7. Count of non-profits per block group
8. Population density of block group
9. Number of storefronts with same license/similar products in block group
10. Rent paid by storefront
11. Reported crime statistics scaled by population of each block group
12. Whether storefront belongs to a franchise or is a unique business
13. Financial information, i.e. profit margin, for each storefront
14. Number of employees for each storefront
15. Walkability scores of block groups

[13] We ran models with and without license type as specified previously in our report.
[14] Initially, we sought to use 311 data instead of reported crimes data but the City of Chicago Data Portal only provides 311 information from 13 Dec. 2018 onward.