Company Longevity Analysis

NYU Center for Data Science

Chuan Chen
cc6580@nyu.edu

Bonnie Deng
yd881@nyu.edu

Yanqi Xu
yx2105@nyu.edu

Yihang Zhang
yz2865@nyu.edu

December 9, 2019

1 Business Understanding

The life expectancy of businesses has long been a center topic of concern in the financial world [4]. Being able to not only predict the lifespan of a business but also understand how different factors impact the lifespan can provide vital information to business owners, employees, and investors. Our project aims to explore these problems by investigating businesses that have opened and closed in New York City since 1977 from available data.

By obtaining information on when an individual business went in and out of business, along with unique feature-values of that business such as its location, its surrounding competitors, and its industry, we can utilize data mining to discover patterns that can help us answer questions such as: Is one location more business-friendly than the other? Does a particular industry have lower risks than another? Moreover, given known features about a particular business, can we predict how long it will stay in business? The ability to answers to such questions using easy-to-obtain and general statistics can provide important information that can help aid potential business owners to make decisions on when and where to open their business, aid investors on deciding whether or not to invest, and provide answers to many other valuable business questions while offering usability.

2 Data Understanding

We obtained the data of businesses in NYC from NYC Open Data: *Legally Operating Business* [5]. This data set contains 200375 instances of the businesses that have existed or are still operating in New York City. Each instance contains location information: City, Street, Zip Code; license information: License Type (Business or Personal), License Creation/Expiration Date; and business information: names, phones and Industry Type. This data is provided by Department of Consumer Affairs (DCA) and is updated on a weekly bases. Hence, we can reasonably assume that the data represents the business population with little to no selection bias. We accessed the data on Nov 22, 2019. Any update afterwards was not considered.

In order to make better prediction on the lifespan of NYC businesses and to take advantage of having a zip code location for each instance, we obtained the state demographic information by zip code from website *Zip Code Demographics By State And County Batch Report*[11]. The demographic data contains all 2153 zip codes of the New York state. Demographic information includes Population, Ethnicity Distribution, Households, Sex Percentage, Income etc. One potential risk of using this data along with the business data in modeling is that all demographic information is specifically for this year only, but the business data contains businesses from 1977 until now. Since the demographic information by zip code for each year between 1977 to 2019 is not easy to obtain, we will be using this data and make the assumption that the demographic distribution according to Zip Address maintains the same over time.

In order to obtain the lifespan of businesses, we are going to make the assumption that businesses without a valid license are no longer operating, and that the time difference between license creation date and license expiration date indicates the lifespan of a business. One problem raises from this assumption, however. For the businesses whose licenses expire after Nov 22, 2019, we have assumed that they will not renew their licenses after expiration, an assumption we can not reasonably make. This is the problem of right censoring in the context of survival analysis, which will be addressed in detail in section 4 Model and Evaluation.

3 Data Preparation

Two data sets altogether contain 78 features, and many of them are irrelevant to our research objective. We deleted all the irrelevant features and were left with 14 variables for feature extraction. For a full list of features we used in the dataset, see Appendix A. NYC business license data contains 70520 missing values, which is around 37% of the overall data. The plot of distribution of variables in the business data set by creation year (Figure 1) indicates that most of the missing values are from the Longitude and Latitude columns. Interpolating the missing Longitude/Latitude with either means or medians would result in the same geographic locations for many of the businesses and might distort this features. Hence, we decided to drop all the missing values. Figure 1 suggests that the missing values have a similar distribution to the population, so we can safely say these values are missing at random and dropping them would not cause serious selection bias.

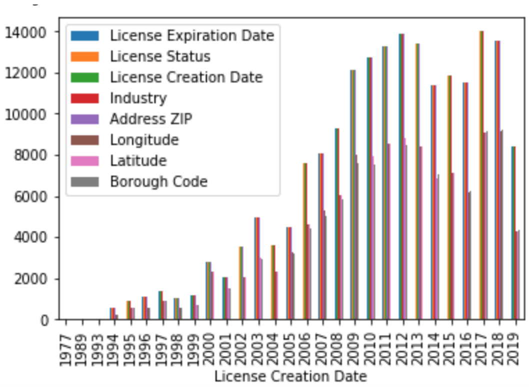


Figure 1: Business counts sorted by creation year

Once we have gathered and cleaned up the two datasets, we merged them together by corresponding zip codes. The merged data has 114719 instances and contains two categorical features: industry type and zip code. We one-hot encoded on the industry type, which generated 47 more binary variables. Instead of one-hot encoding the zip code as well, we dropped this feature due to the fact that the dataset has 155 zip codes, and adding another 155 binary variables to the dataset would largely reduce the accuracy of our model. Another justification for dropping is that our dataset has demographic information by zip code, so we can reasonably assume that the characteristics of each zip code category are sufficiently represented by these demographic features like population, gender ratio, income per household and number of businesses. After that, we modified several feature variables. First, we created a license creation year feature from the license creation date because dates are not numerical but categorical in nature. Also, female ratios and colored ratios by zip code were added that were calculated based on the population, since the ratios to the population rather than the exact numbers better reflect the influence of these factors. Then, we created one dummy variable *right censoring* to indicate whether a business is "dead" nor not. That is, if the expiration date of a business is after Nov 22, 2019, we let *right censoring* $= 1$, and 0 otherwise. This variable will only be used in the survival analysis models. At last, we calculated the target variable, lifespan, that we are going to model throughout this paper. The target variable is created by two cases. For businesses whose licenses expired before Nov 22, 2019, we can safely assume they are no longer operating, and the time difference in days between the license creation date and expiration date can be used their lifespans. For those whose licenses expire after Nov 22, 2019, their expiration date cannot be assumed to be the date they will go out of business, since they may also renew their license and continue operation after then. For these businesses, we can not effectively see their "death" date. Thus, the time difference in days between license creation date and Nov 22, 2019 will be used as their lifespans, which is how long they have been operating until now.

4 Modeling and Evaluation

We split the entire data set into 75% for training and 25% for testing. Observation on the data set renders that many of the features are numeric variables and have different ranges. We normalized all the numeric variables on the training set and transformed the test set to the same scale with the means and variances from normalizing training set. After that, we first fitted a simple linear regression with default parameter settings as our baseline. This simple baseline has a R-score of 0.836 and a RMSE of 0.40490 on the validation set with a 5-fold cross validation.

The poor performance of baseline suggests that linear regression models might not be a good solution to our business problem because it does not deal with the problem of right censoring. Our target variable, lifespan, does not indicate the real lifespan of each business because many of the business have not yet closed and we have no information of their genuine life span. One potential solution is to drop all the businesses that have not closed and only fit a regression model on the businesses that have an exact lifespan. However, this would produce selection bias and underestimate the overall lifespan of the businesses in NYC. In the next two sections, we propose two solutions to address the problem of right censoring.

4.1 Multiclass Classification Models

Since regression models do not work very well on our data due to the bias of the target variable, we first decided to combat this problem by using multiclass classification models. The target variable, lifespan, can be converted into a categorical variable which has 9 classes: <3 years, $3 \sim 6$ years, $3 \sim 9$ years... >21 years, and open, where each category represents the lifespan intervals of the businesses.The category "open" is assigned to these businesses whose "death" has not yet been specified. Figure 2 demonstrates the distribution of 9 classes.

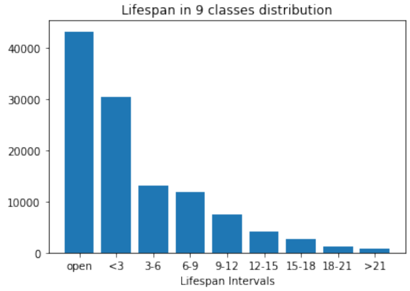


Figure 2: Class Distribution

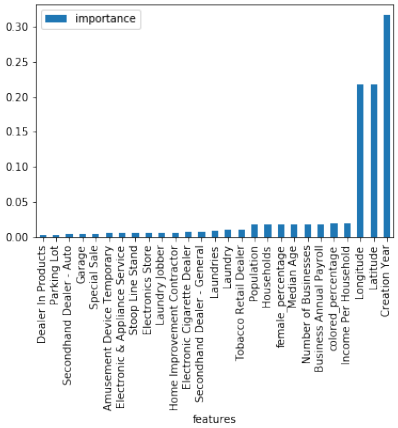


Figure 3: Largest 34 Feature Importance Distribution

By classifying the lifespan, the original target variable can be completely captured by the new target variable bypassing the problem of right censoring. One drawback of this method is its low precision. The prediction results of this type of models would only give a rough lifespan interval that the businesses belong to, while a regression can predict the exact days of their lifespan. However, a classification model can produce more meaningful results. We can sacrifice some precision for more significant results. Besides, a rough interval of their survival time might already be sufficient in some contexts.

We first tried five different multiclass classification algorithms: decision tree, logistic regression, K-nearest neighbors classification, random forest, and support vector machine. We evaluated their performances on 5-fold cross-evaluation considering that our training set does not have enough instances compared to the amount of features we have. While running the code, we discovered that the Support Vector classifier did not seem to converge and took forever to finish, so we took it out of our algorithm list. We used accuracy as our evaluation metric. Accuracy in the multi-class classification is the proportion of the instances that are correctly classified among all testing instances. Figure 4 and Figure 5 demonstrate the results of 5-fold cross validation performed on four different algorithms before and after oversampling, respectively. We concluded that oversampling largely improved the performance of decision tree and K-NN while the random forest classifier and logistic regression did not change much. Additionally, we found that the variance of the cross validation results are smaller on the over-sampled training set. We should be alerted of the possibility of over-fitting the model.

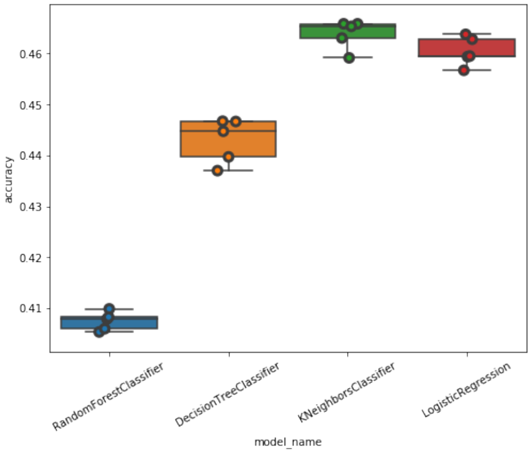


Figure 4: Algorithms Comparison

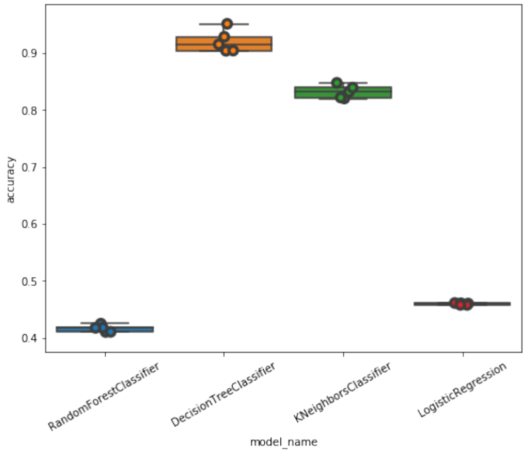


Figure 5: Algorithms Comparison After Oversampling

We chose to use the decision tree algorithm and random forest algorithm for further parameter tuning and feature selection after testing the four algorithms on the test set, shown in Table 1. They outperformed K-NN and logistic regression probably due to the fact that tree algorithms are good at dealing with categorical variables, and our data contains many binary features. For decision tree models, we first fitted a decision tree with default parameters and used the feature importance to perform feature selection. We discarded all features that are below 0.001 and are left with 34 features in total. Then, we tried to decide on what value to use as the max-depth of the model. The learning curve demonstrates that the model achieves its optimal performance on validation set when max-depth is at 40, and the accuracy stays the same for max-depths larger than 40. We fixed max-depth to be 40 and fitted the model on the selected set of features to get an optimal decision tree model. The same procedure was performed on the random forest models, but this time we fixed the max-depth to be 40 and tuned another parameter, the n-estimator. The results of all the models on the test set are summarized in the Table 1.We used three metrics for final evaluation: accuracy, macro average F1-score, and weighted average F1-score[7]. F1-score combines the information of precision and recall. The major difference between macro average and weighted average is that macro average averages the

F1-scores of all classes while weighted average weights the score by number of samples in each class. We concluded that, after tuning, random forest supersedes other models in all three evaluation metrics. For more detailed results, see Appendix B.

Model	Accuracy	Macro Av F1	Weighted Av F1
Decision Tree	0.44	0.32	0.45
Knn	0.40	0.30	0.42
Random Forest	0.45	0.33	0.46
Logistic Regression	0.36	0.32	0.38
Decision Tree after tuning	0.44	0.33	0.45
Random Forest after tuning	0.45	0.34	0.46

Table 1: Multiclass Classification Results Comparison

In the process of performing multiclass classification on the data, we discovered that Figure 2 reflects a pattern of survival curves and Figure 3 indicates that creation year is the most important feature that influences a business' lifespan. These information discloses the influence of time in our dataset and led us towards a survival analysis solution to the problem.

4.2 Survival Analysis

To combat both the problem of right censoring and lack of precision, we proposed another class of models to be used on our data set, survival analysis models. Survival analysis is a branch of statistics that not only predict whether or not an event will happen but the time it takes for this event to take place. Such techniques were originally developed by medical professionals interested in finding out the expected lifetime of patients, but its uses can be further applied to many other problems where the time-to-event is of interest [10]. In our business problem, the lifespan of a business is our target variable. We can view the license creation date as the time of origin for each business, the license expiration date as the time of the event, and lifespan can be viewed as the time-to-event. Thus, we can see why survival analysis will be a great model for our problem.

Survival analysis methods include a powerful consideration of our data, right censorship. For businesses whose licenses have not expired yet, we have no information on when their "death" will occur. Yet, we are still using those instances to estimate the lifespan of a business. When we chose to represent their lifespan as how long they have been operating till the current, our estimates are highly biased and under-estimated. Survival analysis corrects this bias by the inclusion of censored data when calculating the estimates, which makes it stand out compared to our previous models. In our model, all businesses whose licenses have yet to expire are assigned a 0 in the censorship column and are censored when the model was fitted.

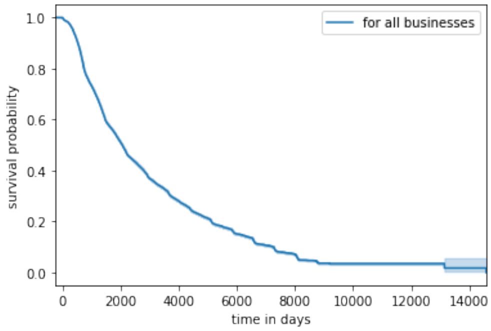


Figure 6: Survival probability over time for all businesses

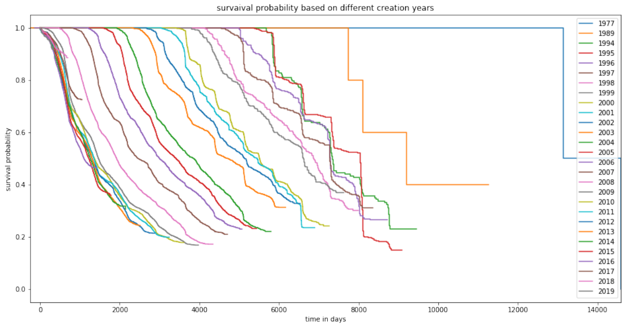


Figure 7: Survival probability over time for different cohorts

4.2.1 Kaplan-Meier Estimate

A simple yet powerful survival analysis model is a non-parametric method called the Kaplan-Meier estimator. This model estimates the survival probability from the observed survival times. It uses a survival probability function, which is a function of time, whose calculation depends only on the number of instances alive before time t, the number of events at time t, and the initial survival probability, which is 1 [3]. From this model, we can plot a K-M survival curve, which plots the survival probability against time [6]. Figure 6 plots the survival probability calculated by the Kaplan-Meier model for all businesses in our data, and we can see that Figure 6 demonstrates strong similarity with Figure 2.

We can see that the lifespan of businesses shows exponential decay over time. From the model, the median survival time for all business is 2058 days. That is, all businesses have a 50% chance of staying open past this 2058 days.

From the classification models, we can see that one of the features which has the strongest influence on the lifespan of a business is the creation year of that business. Thus, we have divided businesses into different cohorts based on their different creation years to get a clearer picture on their survival probabilities. Figure 7 shows their survival probabilities side-by-side. We can calculate the median survival time for each cohort accordingly.

The median survival time is calculated as the smallest survival time for which the survival probability function is less than or equal to 0.5 [6]. For businesses who were created in 2017, 2018, and 2019, they do not get this far, thus their median survival times were not calculated.

From Figure 7 and Table 2, we can clearly observe a strong relationship between the creation year of the business and its survival time, the earlier the business was created, the longer its survival time.

creation year	1977	1989	1994	1995	1996	1997	1998
median survival time	13146.0	9198.0	7346.0	8045.0	7293.0	7281.0	6865.0
creation year	1999	2000	2001	2002	2003	2004	2005
median survival time	6607.0	5510.0	5448.0	5134.0	4720.0	3852.0	3419.0
creation year	2006	2007	2008	2009	2010	2011	2012
median survival time	2998.0	2462.0	1923.0	1463.0	1364.0	1353.0	1303.0
creation year	2013	2014	2015	2016	2017	2018	2019
median survival time	1309.0	1262.0	1221.0	1102.0	inf	inf	inf

Table 2: Median Survival time in days for each creation year cohort

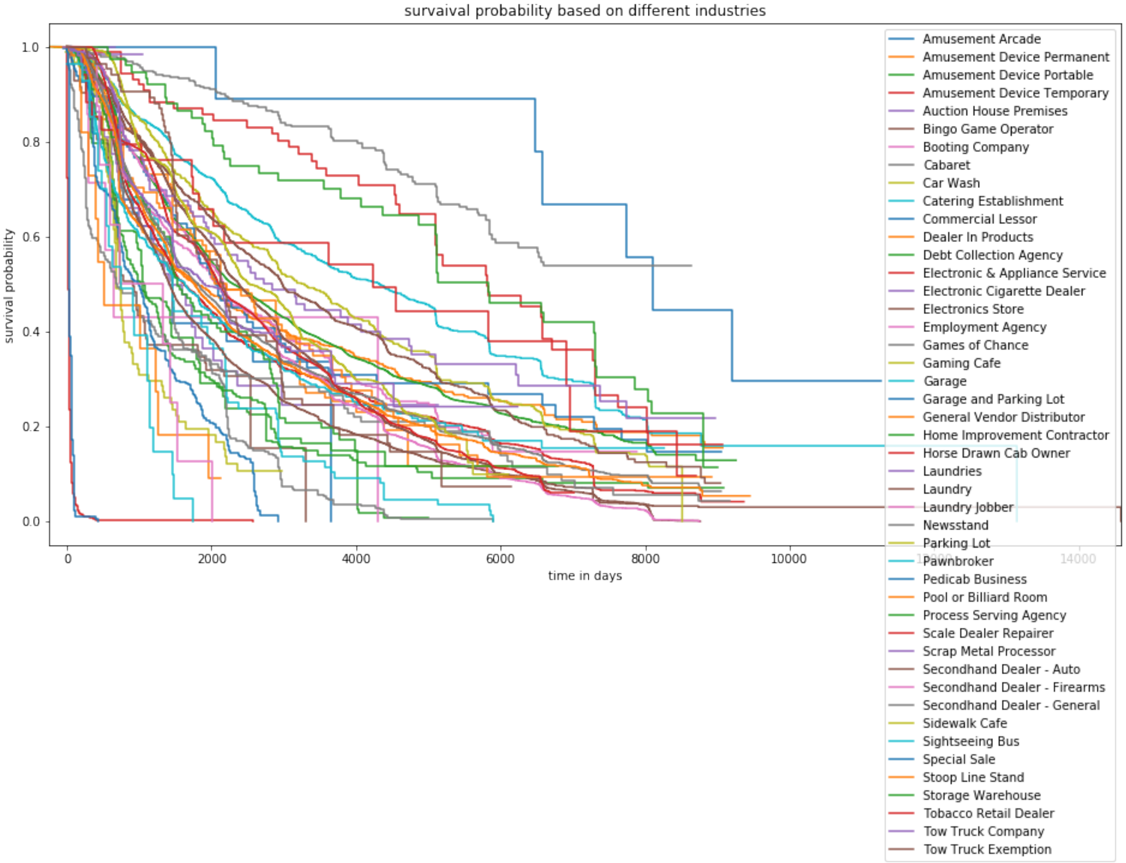


Figure 8: Survival probability for different industries

In order to understand the effect of industry type on the lifespan of businesses, we have also separated cohorts based on industry type. The survival probability over time for each industry is plotted in figure 8. After calculating the median survival time for each cohort, we find that the 3 industries with the highest median life span are Commercial Lessor, Storage Warehouse, and Horse Drawn Cab Owner, as shown in table 3. The 3 industries with the lowest median life span are General Vendor Distributor, Special Sale, and Amusement Device Temporary, as shown in table 4.

Industry	median survival time
Commercial Lessor	8102
Storage Warehouse	5829
Horse Drawn Cab Owner	5814

Table 3: median survival time for the top 3 industries

Industry	median survival time
General Vendor Distributor	509
Special Sale	32
Amusement Device Temporary	23

Table 4: median survival time for the bottom 3 industries

We can do more such cohort analysis from the survival curves of cohorts separated on different features, such as zip code, population, and etc. However, cohort analysis is only a limited use case of survival analysis since we can only use it for the aggregated level of the data. Also, we are limited to separating cohorts based on one feature variable at a time. To overcome these limitations, we decided to use another survival analysis method, the Cox Proportional Hazard Model.

4.2.2 Cox Proportional Hazard Regression

The lifespan for businesses in a particular population can provide important insights to the data at the aggregate level. However, in real life situations, each individual business also has its own set of covariates (features). It is very important to learn the impact of those covariates on the survival probability so we can predict survival time on the individual level.

The Cox Proportional Hazard Regression is a popular model that can combine covariates with the survival function. It models the hazard function which is defined as the rate at which the event is taking place, out of the surviving population at any given time t, conditioned on certain features. The hazard function can then be used to calculate the survival probability for each individual instance [2]. Using the variance inflation factor, we found that the feature, 'Population', causes high collinearity in the dataset. Thus, we removed this feature to help with model convergence.

As a first attempt, the cox regression model was fitted with the training set with a penalizer of 0.5 and a step size of 0.005. Despite the large increase in computation time, we had to increase the penalizer value and decrease the step size values in order for the model to converge. Convergence was difficult since some of our industry dummy features had very low variance when conditioned on the right censoring column [6]. Despite the obstacles, our initial fit gave us a concordance value of 0.72. The three covariates with the highest absolute value of coefficients are the industries Amusement Device Temporary, Special Sale, and Electronic Cigarette Dealer. A visual representation of the feature coefficients can be found in Appendix C.

Due to the fact that censoring is present, we can not obtain the true survival time for censored instances. Thus, for survival analysis models, it is not appropriate use a loss function like mean-squared-error or mean-absolute-loss for model evaluation. Instead, measures such as the concordance index can be used. The concordance index is a value between 0 and 1 that evaluates the accuracy of the rankings of predicted survival time [8]. Simply put, if observations with the higher survival times also has the higher probability of survival as predicted by the model, the better the concordance index will be. The concordance index is a generalization of AUC, and can be interpreted similarly where

- 0.5 is the expected result from random predictions,
- 1.0 is perfect concordance, and
- 0.0 is perfect anti-concordance.

According to literature, fitted survival models typically have a concordance index between 0.55 and 0.75, due to the fact that even perfect models have a lot of noise which can make a high score impossible [6]. Under this context, our model concordance of 0.72 indicates a pretty good performance. However, upon testing the proportional hazard assumptions, we find serious problems.

The Cox proportional hazards model is built entirely on the assumption that the covariates have a linear multiplication effect on the hazard function, and the effect stays the same across time [9]. Thus, if our fitted model violates this assumption, we may experience a serious loss in prediction power. We have tested our covariates with a null hypothesis of no violation and a p-value threshold of 0.01 [6]. Under this test, we have found that 45 features violates the hazard assumption. Out of those 45, 44 were industry dummy columns. Thus, we have fitted the model on the training set again, but this time stratifying on the 44 columns in an attempt to reduce violations. Our second cox model has a concordance index of 0.69, with only 3 features that violates the hazard assumption. Although our concordance has experienced a 0.04% decrease, we have successfully reduced 93.33% of violations in our model. Thus, overall, the second model with stratification is the better model of the two.

After the model has been trained, we can utilize the powerful prediction ability of the cox model on individual business given a set of covariates. For illustration, consider two business that are still operating with the following criteria shown in Table 5:

We can predict their estimated survival probability shown in Figure 9.

Business ID	Annual Payroll	Households	Income Per Household	competitors
52758	920912	25158	119203	1819
	female percentage	colored percentage	industry	Creation Year
	0.557808	0.104894	Sidewalk Cafe	2019
	Annual Payroll	Households	Income Per Household	competitors
639	8251251	12096	106056	5229
	female percentage	colored percentage	industry	Creation Year
	0.496354	0.421552	Home Improvement Contractor	2017

Table 5: features for the two businesses

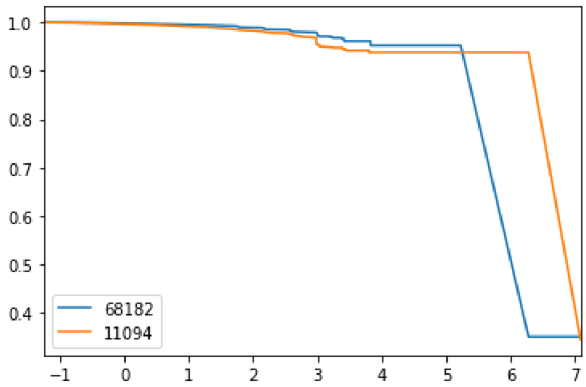


Figure 9: survival probability for the two businesses

Accordingly, we can calculate their expected lifetime, shown in Table 6.

business ID	expected survival time
52758	5553.50
639	3521.63

Table 6: Expected survival time of the 2 businesses

5 Deployment

Our fitted Kaplan-Meier model can be deployed and utilized to generate business answers on the aggregate level. According to our results, we have listed the 3 best industries that result in the longest expected life span for businesses and the 3 worst industries correspondingly. Investors should take this into consideration before they decide to invest in businesses of a particular industry. As time progresses, more business data can be added to help further train the model, and updates on industry rankings should be monitored in case of any changes. From our multiclass classification models, we have given a ranking of feature importance in affecting the business lifespan. Potential business owners can use this result to help make business plans and shift business focuses accordingly to help prolong their operation time.

Our models also offer powerful prediction capabilities if the lifespans of individual businesses are of interest. Due to its efficiency, if only a rough estimate of lifespan is sufficient as a prediction, our random forest model can be used to predict a rough estimate of lifespan within an error of 3 years given specific statistics about an individual business. If a more precise prediction is needed, our cox regression model should be used to give a survival probability along with each time point in the future and be used to predict an expected lifetime of said business. As new data are added to the model, hazard assumptions should be checked, and if serious violations occur, updates to feature processing should be made accordingly.

Deploying our models on any business outside the NYC area should be used with caution as results might be largely inaccurate. Besides careful model evaluation and the limitation of deployment location, there are no other apparent ethical considerations or risks associated with model deployment.

References

[1] Lan Davis. *Reflections on corporate longevity*. 2014. URL: https://www.mckinsey.com/business-functions/strategy-and-corporate-finance/our-insights/reflections-on-corporate-longevity. (accessed: 12.06.2019).

[2] Cox DR. “Regression models and life tables”. In: 34 (1972), pp. 187–220.

[3] Kaplan EL and Meier P. “Nonparametric estimation from incomplete observations”. In: 53 (1958), pp. 457–481.

[4] Mark Goodburn. *What is the life expectancy of your company?* 2015. URL: https://www.weforum.org/agenda/2015/01/what-is-the-life-expectancy-of-your-company/. (accessed: 12.07.2019).

[5] *Legally Operating Businesses*. 2019. URL: https://data.cityofnewyork.us/Business/Legally-Operating-Businesses/w7w3-xahh. (accessed: 11.20.2019).

[6] *Lifelines Python Documentation*. 2019. URL: https://lifelines.readthedocs.io/en/latest/Quickstart.html. (accessed: 11.29.2019).

[7] *Multi-Class Metrics Made Simple, Part II: the F1-scores*. 2019. URL: https://towardsdatascience.com/multi-class-metrics-made-simple-part-ii-the-f1-score-ebe8b2c2ca1. (accessed: 12.7.2019).

[8] Sebastian. *How to interpret the output for calculating concordance index (c-index)?* 2013. URL: https://stats.stackexchange.com/questions/29815/how-to-interpret-the-output-for-calculating-concordance-index-c-index. (accessed: 12.07.2019).

[9] Steve Simon. *The Proportional Hazard Assumption in Cox Regression*. URL: https://www.theanalysisfactor.com/assumptions-cox-regression/. (accessed: 12.07.2019).

[10] Clark TG et al. “Survival Analysis Part I: Basic concepts and first analyses”. In: 89 (2003), pp. 232–238.

[11] *Zip Code Demographics By State And County Batch Report*. URL: https://www.cdxtech.com/tools/bulk/demographics/state-and-county/?from=singlemessage&isappinstalled=0. (accessed: 11.22.2019).

Appendix A Data Features

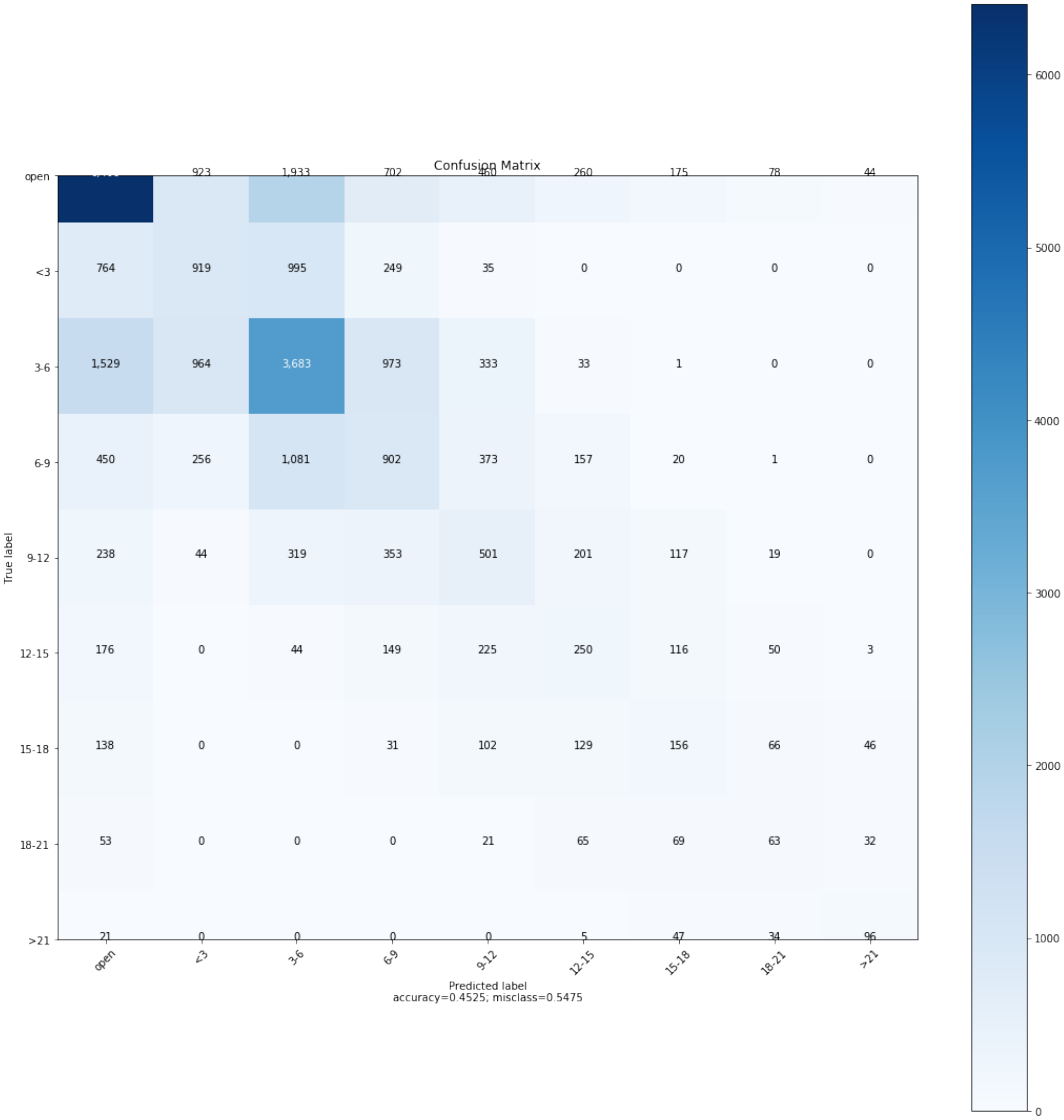
Data Type	Feature Names	Number of Class/ Range
Categorical	Industry	47
	Address ZIP	155
Numerical	License Creation Date	01/01/1994-12/31/2018
	License Expiration Date	01/01/2015-12/31/2021
	Longitude	[-77.51958437167269, -73.70092927652794]
	Latitude	[40.112385336598024, 40.9120628714794]
	Population	[2,109931]
	Business Annual Payroll	[0,23065864]
	Median Age	[20.3, 80.9]
	Households	[0,44432]
	Income Per Household	[0,250000]
	Number of Businesses	[5,7296]
	Female Population	[0,56119]
	Colored Population	[1,90166]

Table 7: All features used in data modeling

Appendix B Results on the Test Set of Random Forest Model after Tuning

	Precision	Rcall	F1-score	Support
open	0.66	0.59	0.62	10983
<3	0.30	0.31	0.30	2962
3-6	0.64	0.49	0.47	7516
6-9	0.27	0.28	0.27	3240
9-12	0.24	0.28	0.26	1792
12-15	0.23	0.25	0.24	1013
15-18	0.22	0.23	0.23	668
18-21	0.20	0.21	0.21	303
>21	0.43	0.47	0.45	203
Accuracy			0.45	28680
Macro Avg	0.33	0.34	0.34	28680
Micro Avg	0.47	0.45	0.46	28680

Table 8: Evaluation Summary



	open	<3	3-6	6-9	9-12	12-15	15-18	18-21	>21
open	[illegible]	923	1,933	702	460	260	175	78	44
<3	764	919	995	249	35	0	0	0	0
3-6	1,529	964	3,683	973	333	33	1	0	0
6-9	450	256	1,081	902	373	157	20	1	0
9-12	238	44	319	353	501	201	117	19	0
12-15	176	0	44	149	225	250	116	50	3
15-18	138	0	0	31	102	129	156	66	46
18-21	53	0	0	0	21	65	69	63	32
>21	21	0	0	0	0	5	47	34	96

Figure 10: Confusion Matrix

Appendix C Visual Representation of the Coefficients

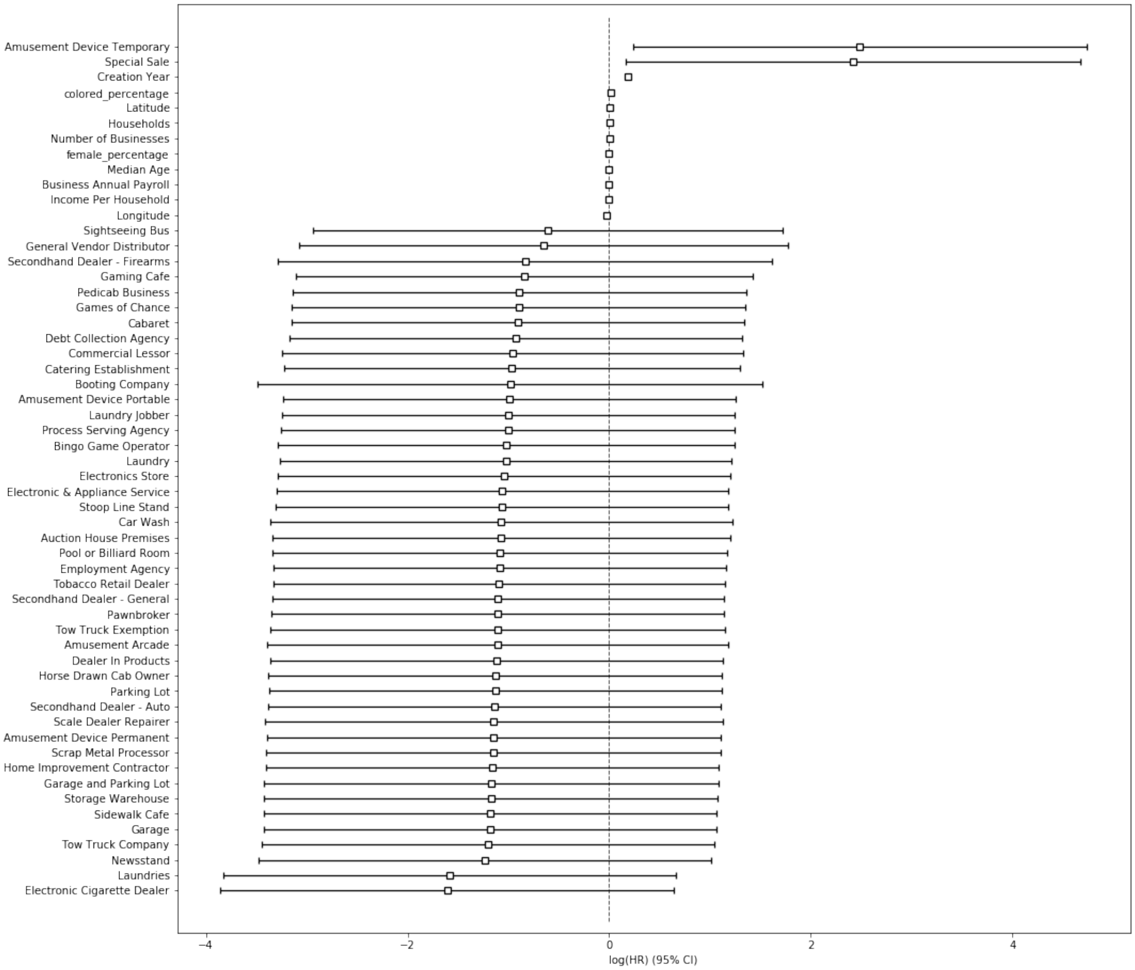


Figure 11: a visual representation of the coefficients (i.e. log hazard ratios), including their standard errors and magnitude

Appendix D Contributions

- **Chuan Chen**: Data Cleanup, Feature Engineering, Kaplan-Meier Estimate Cohort comparison, Cox Proportional Hazard Regression data fitting and model improvement, Write-up
- **Yadi Deng**: Data Cleanup, Feature Engineering, research on relevant literature, data exploratory analysis, Survival analysis as a classification problem (This part was dropped due to limited time), Write-up
- **Yanqi Xu**: Data Cleanup, Feature Engineering, Cox Proportional Hazard Regression data fitting, Multiclass Classification data fitting and model improvement, Write-up
- **Yihang Zhang**: Data Cleanup, Kaplan-Meier data fitting, VIF, Linear Regression Model comparison and backward feature selection (This part is not included in the final write-up due to limited space), Write-up

Appendix E Codes

All of the relevant codes used in our paper can be found in our Github repository