Analysis of Commercial Company Displacement in Boston Areas

CS506 Final Project Report

Xiaoyan Su, xiaoyanc@bu.edu

Yirong Zhang, yizh@bu.edu

Jianghong Man, manjh@bu.edu

Background and Introduction:

Businesses always go through a hard time when declaring bankruptcy, whether forced or voluntarily. In order to help the Small Business Unit to generate the development of relevant policy by scrutinizing the closure of commercial displacement in Boston area (specifically focus on the neighborhoods of Roxbury, Mattapan, Dorchester, and East Boston) and identifying underlying causes/drivers of the changes, our team will look at diverse data sets to:

1. Understand where are the closed businesses properties;
2. Seek to understand potential correlations or causal patterns associated with the closure.
3. Try to find patterns that may help the city predict and prevent businesses from closing in the future.

Research Questions:

1. Are there any relationships between commercial displacements and economic information (average home price and price per square foot for commercial property) or other information (crime rates) in particular neighborhoods and areas?
2. What are the common types of businesses that are being closed?
    a. What are the similarities between owners in the businesses? (ethnicity and gender)
    b. Property Value? (rental expenses)
    c. Size in terms of sq ft? (square footage)
    d. Are there geographic trends in the closed? (district wise, neighborhood wise)
3. Can we predict where the next closed business is likely to happen? (prediction)
4. Are there common attributes of property associated with closed businesses at a given location? (determine business type)

Data description:

1. **Data file description**

    This dataset, U.S. Business dataset contains a total of 56 million businesses including 15 million verified and 41 million unverified businesses that are updated weekly. It is the only business database that is enhanced with more than 24 million phone calls per year providing the most accurate data possible.

    We choose the data in the year 2019 and limited the regions of only 4 neighborhoods: Roxbury, Mattapan, Dorchester, and East Boston. After applying the choosing conditions, 5841 numbers of data are finally selected.

2. **Data attribute description (table)**

Field Name	**Field Description**	**Data Type**
Company Name	Name of Company or Professional Name	Object

Executive First Name	First Name of Contact	Object
Executive Last Name	Last Name of Contact	Object
Address	Location Address of the company	Object
City	Location city address of the company	Object
State	Location state address of the company	Object
ZIP Code	Location zip code address of the company	Int
Record Type	Type of record - Verified, Unverified, Closed or Out of Business	Object
Executive Gender	Gender of contact on Executive Detail	Object
Executive Ethnicity	Ethnicity of contact on Executive Detail	Object
County	County name based upon location address zip plus four postal	Object
Neighborhood	Name of the neighborhood where the company is located	Object
SIC Code	Code for an additional line of business the company engages in - up to 10 lines available for output	Int
SIC Code Description	Description for an additional line of business the company engages in - up to 10 descriptions available for output	Object
Primary NAICS	The primary line of business code used by a government agency for the purpose of collecting, analyzing, and publishing statistical data related to the U.S. business economy.	Int
Primary NAICS Description	The primary line of business description used by a government agency for the purpose of collecting, analyzing, and publishing statistical data related to the U.S. business economy.	Object
Location Employee Size Range	A number range of employees at the location address. Data is modeled.	Object
Location of Employee Size Actual	Actual number of employees at the location address, if available.	Int
Location Sales Volume Range	A number range of sales by the company at the location address. Data is modeled.	Object
Location Sales Volume Actual	Actual number of sales by the company at the location address, if available.	Int
Type of Business	Indicates if the company is a public company, private company, or a branch.	Object
Location Type	Indicates if the company is a single location, headquarter, branch or subsidiary.	Object

Square Footage	Indicates the square footage of the company's location.	Float
Credit Score Alpha	Alpha Credit rating of the company.	Object
Credit Score Numeric	Numeric Credit rating of the company.	Int
Rent Expenses	Payments made to other companies for the rental or leasing of land, buildings, offices and related structures. estimation of annual expenses based upon industry and size.	Int

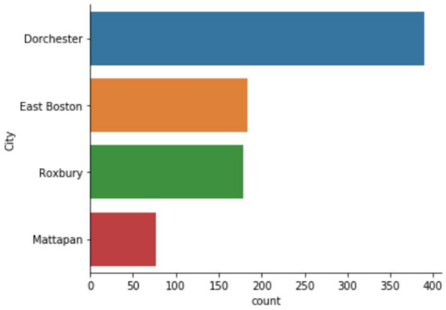
Data cleaning process:

1. **Purpose**: To calculate the rent expenses per sq ft of each business
    a. Methods:
        i. Entries that do not have Rent Expenses and Square Footage are dropped.
        ii. Delete the ‘$’ and ‘,’ in both Square Footage and Rent Expenses
        iii. Rent Expenses column also contains ‘Less than’ and ‘Over’, we dropped both and left with the number provided.
        iv. Since Rent Expenses and Square Footage are ranges, for example ‘$5,000 to $10,000’ and ‘1-1,499’, we split the two numbers and calculated the average.
        v. At last, we created a new column ‘Rent per sq ft’ by dividing Rent Expenses by Square Footage.
    b. Usage:
        i. Later in our project, We hope this attribute will lead to better prediction models for closed businesses since the rent of the company will affect the businesses.

Data visualization:

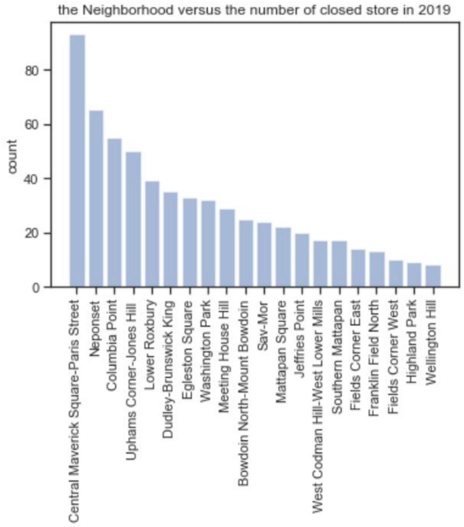
1. **Statistical data visualization Methods**
    a. We used **Seaborn**, a statistical data visualization library based on matplotlib, to construct bar graphs and pie charts. Since Catplot from seaborn is best for the categorical bar graph, we used it to construct all of our bar graphs to give us a sense of the trends and causes of businesses that are closing.
    b. We then used **Folium**, a powerful python library that helps to create several types of leaflet maps, to construct heatmap. It gives an intuitive sense of the average home price, rental price and crime rate in these four districts of interest.
2. **Graphs and analysis (Bar Graph and pie chart)**
    a. **Graph 1**

Number of Closed Businesses in Each District: This graph shows the number of closed businesses located in each district (Dorchester, East Boston, Roxbury, and Mattapan). About 400 of the closed businesses are located in Dorchester and about 175 closed businesses located in East Boston and Roxbury. The lowest amount of closed businesses is located in Mattapan. This only gives an intuitive sense of the distribution of closed businesses.



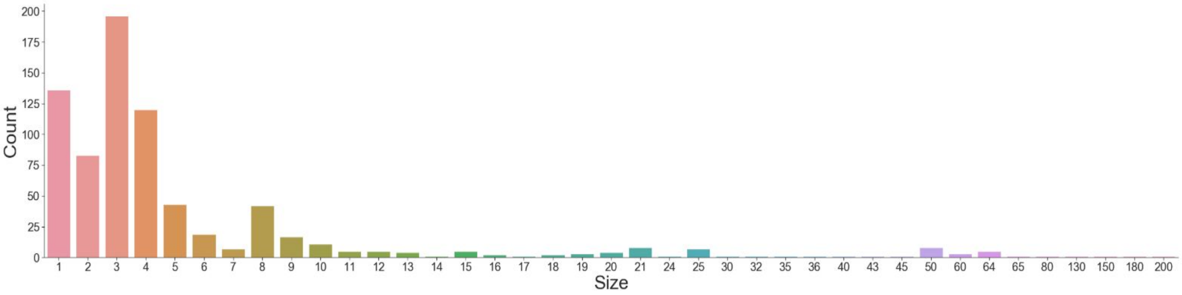
b. **Graph 2**

Number of Closed Businesses in Each Neighborhood: This graph implies the number of closed businesses in the neighborhoods of Roxbury, Dorchester, East Boston, and Mattapan. From our Reference USA database, we generated this graph of 751 closed businesses to identify the number of closed businesses in each neighborhood. Over 80 businesses are in Central Maverick Square and Paris Street in East Boston and about 70 businesses are in the Neponset of Southeast Dorchester.



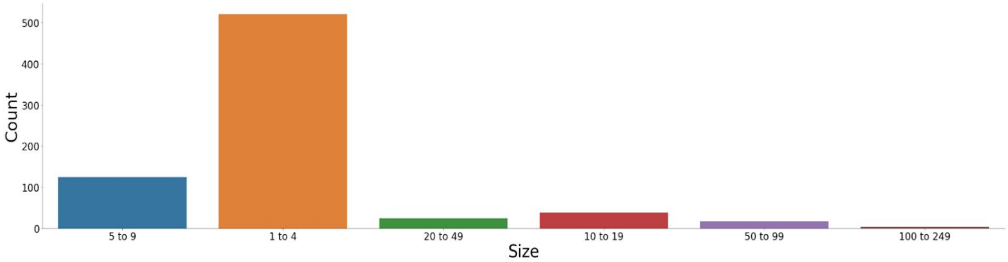
c. **Graph 3**

Actual employee size of closed business: This graph indicates the actual employee size of closed businesses. By deleting rows that do not have employee size attributes, we then build this pie graph to see the size of the businesses to extract small businesses. We can see that most of the businesses are extremely small in size, with an employee size less than 10. This indicates that small businesses with less than 10 employees are more likely to be closed.



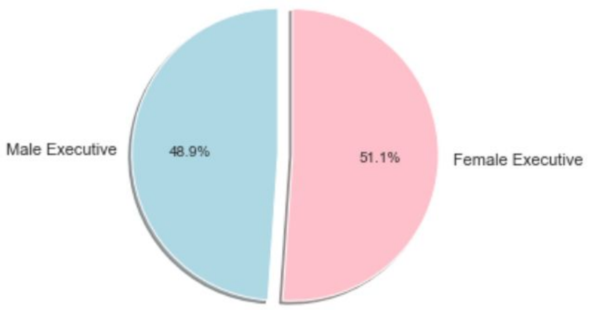
d. **Graph 4**

Employee range of closed business: This graph illustrates the employee size in a range of closed businesses. We can see that a large amount of them are businesses with only 1 to 4 employees. Also, there is a trend of decreasing amount of closed businesses as the employee size increase.



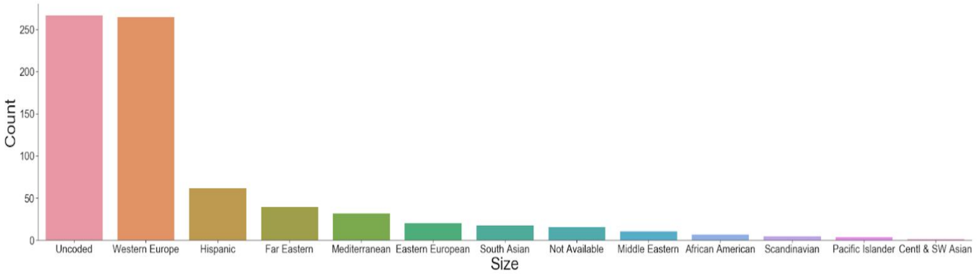
e. **Graph 5**

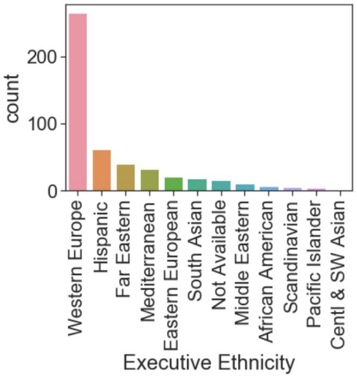
Gender of the executive: This bar graph demonstrated that the executive of these 800 closed businesses is 51% females and 49% males. Generally, there is little gender influence on whether the business is closed or not.



f. **Graph 6**

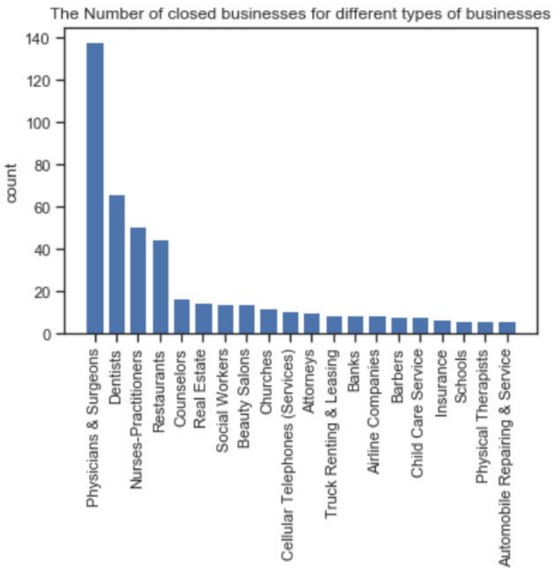
Ethnicity of the executive: According to the data and bar graphs, most of the owner's ethnicities are Western European and Uncoded in a total of about 800 closed businesses. Since there are no Americans and there is Not Available Ethnicity, we guessed that Uncoded should be Americans. Moreover, about 60 of the businesses are run by hispanic owners.





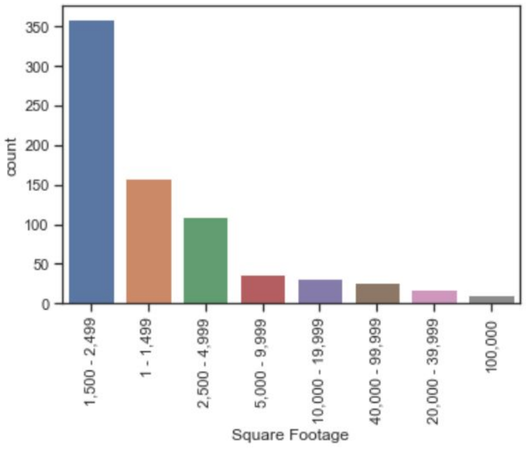
g. **Graph 7**

Business types of closed businesses: This graph indicates the business types of closed businesses. We can see from this bar graph we generated form Reference USA dataset, in a total of 346 closed businesses, there are over 100 businesses classified as physicians and surgeons. 66 of them are dentists and 51 are nurses-practitioners. Therefore, closed businesses are mainly medical-related businesses. This may be due to the high competition of medical businesses in Boston areas. Moreover, there are also about 50 closed restaurants.



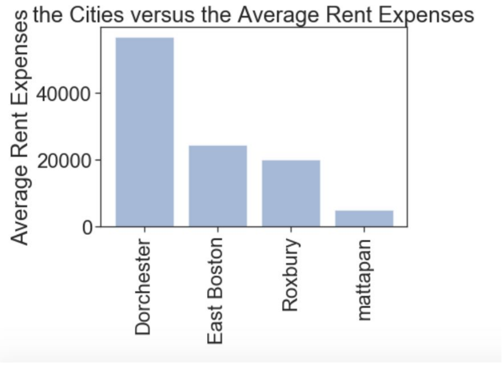
h. **Graph 8**

Physical size of closed business: This graph implies the amount of businesses are closed in different ranges of company size in square foot. It is indicated that most of the closed businesses are 1500 to 2499 square feet and 1 to 1499 square feet in size. Therefore, most of the closed businesses are also small size companies. It is reasonable that small companies are more susceptible to the rental and crime rate around.



i. **Graph 9**

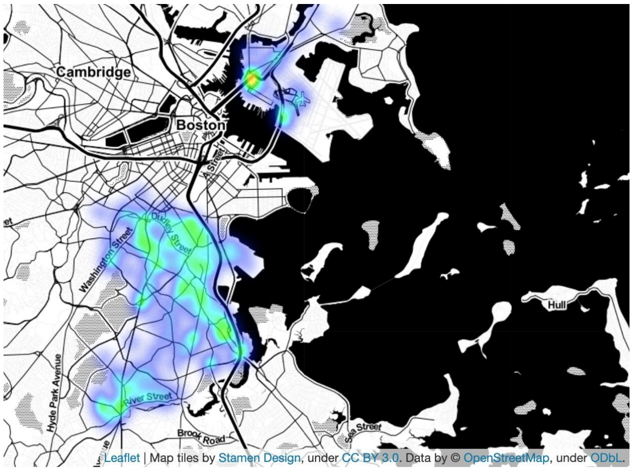
Average rental expenses in four districts: This graph demonstrates that the average rent expenses of the four districts. Among these four districts, the average rent expenses of Dorchester is the highest, at around $60000. The rent expenses in Mattapan is the lowest. This graph corresponds to the number of closed businesses in these four places.



3. Heatmap and analysis

a. Heatmap 1

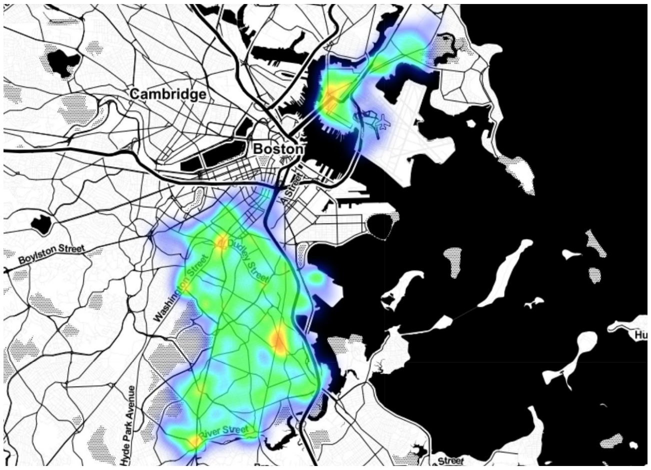
Closed business distribution in four districts: The heat map was obtained from the Reference USA dataset. We selected the closed businesses located in the four districts of interest and Only closed businesses in the four targeted areas were included on this heat map. Most of the closed businesses are aggregated at the most East part of East Boston and the north part of Dorchester and Roxbury.



Leaflet | Map tiles by Stamen Design, under CC BY 3.0. Data by © OpenStreetMap, under ODbL.

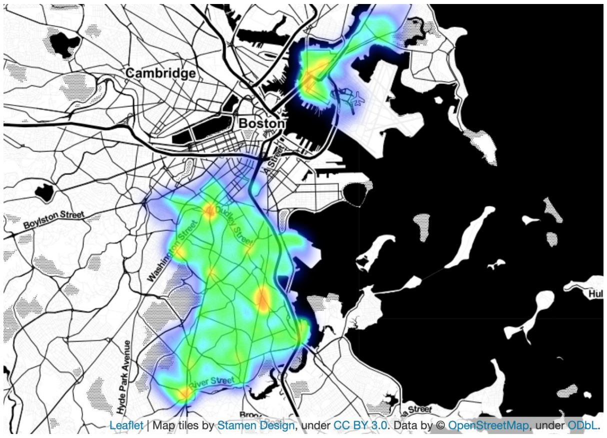
b. Heatmap 2

Average home price: The heat map was obtained from the data set on references USA. Only closed businesses in the four targeted areas were included on this heat map. There are four major areas, South Dorchester, central Dorchester, East Boston, and north Roxbury, all showed high rent expenses for the small business owner. A few minor areas, such as West Roxbury, North Dorchester, and west Dorchester also had relatively high rent expenses.

i. 

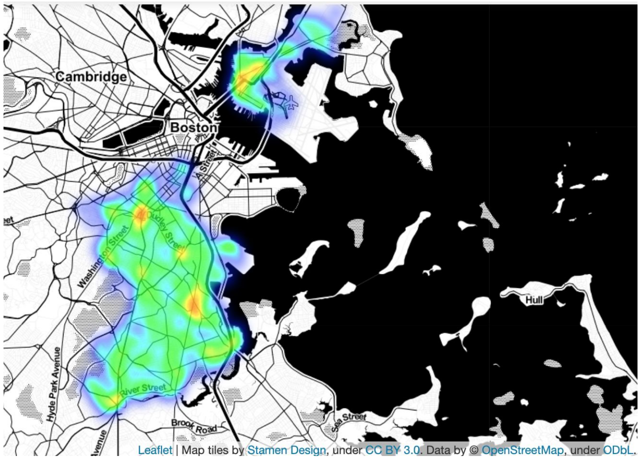
c. Heatmap 3

Price per square foot for all businesses: The heat map was obtained from the data set on references USA. All small businesses with the ones currently closed and the ones remaining opened, in the four targeted areas were included on this heat map. Three areas, which were west of East Boston, central Dorchester and north Roxbury, with a high average of rent per square feet, are shown on the graph.

i. 
Cambridge
Boston
Boylston Street
Washington Street
Hyde Park Avenue
River Street
Hull
Leaflet | Map tiles by Stamen Design, under CC BY 3.0. Data by © OpenStreetMap, under ODbL.

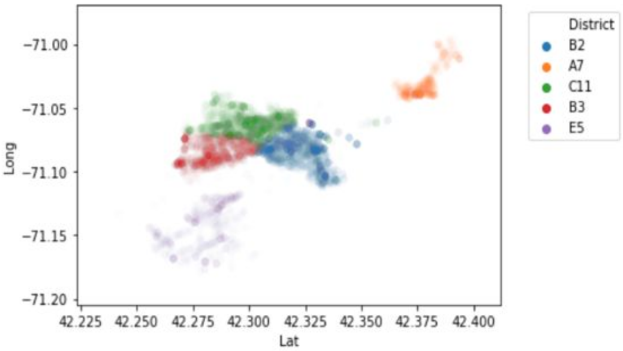
d. Heatmap 4

Price per square foot for closed businesses: The heat map was obtained from the data set on references USA. Only closed businesses in the four targeted areas were included on this heat map. Three areas, which were west of East Boston, the center of Dorchester and north Roxbury, with a high average of rent per square feet, are shown on the graph. A few more spots, such as South Dorchester, Lower Roxbury, and northwest of Dorchester showed a high price of square footage as well.

i. 
Cambridge
Boston
Washington Street
Hyde Park Avenue
River Street
Brook Road
Hull
Leaflet | Map tiles by Stamen Design, under CC BY 3.0. Data by © OpenStreetMap, under ODbL.

e. Heatmap 5

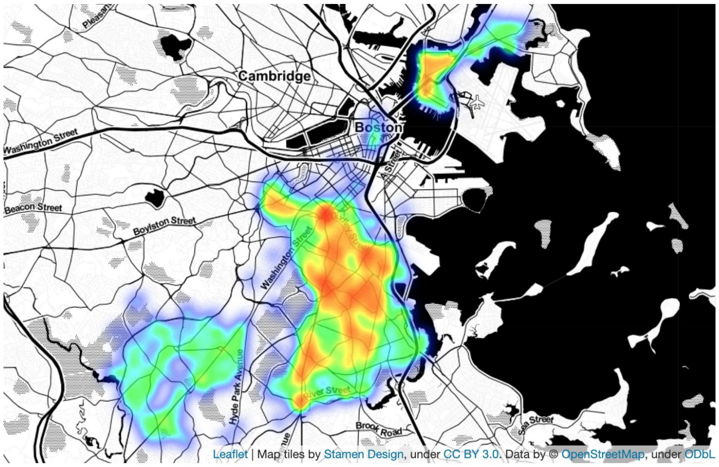
Crime rate heat map 1: This graph shows the distribution of crime incidence in these four places (B2 = Roxbury, A7 = East Boston, C11 = Dorchester, B3 = Mattapan, and E5 = West Roxbury). According to this heatmap, crime incidents mostly take place in the middle, which is B2, B3, and C11. The high crime incident in Dorchester and Roxbury is coherent with the number of closed businesses there. Therefore, there is a positive correlation between the crime rate and the number of closed businesses.



i.

f. **Heatmap 6**

Crime rate heatmap 2: The heat map was constructed from the Boston Crime incident in the 2019 data set. Only specific areas: Mattapan, Roxbury, East Boston, and Dorchester were included on the graph. Most crime incidents happened in North Dorchester, central Mattapan, and Roxbury. Few streets, such as Bowdoin St, Warren St, Washington St, and Blue Hill St.

i. 

Prediction Using Models:

1. Feature selection and generation

a. **Feature selection**: Initially, we have 28 attributes representing the characters of a specific business. We delete some attributes which can give us repetitive information of a company (such as “City”, “State”, “County” , “Location Employee Size Range”, “Location Sales Volume Range”, “Credit Score Alpha”), some attributes which are unique and can cause overfitting in the model (such as “Company Name”, “Executive Name”, “Address”, “SIC Code”, “Primary NAICS”, “Latitude”, “Longitude”), and some missing values which are the same for all businesses (such as “Type of Business”, “Location Type”). We finalize the attributes to 11 numbers which can be put in our training model.

b. **Feature generation**: We drop the NaN attributes in “Rent Expense” column and then use “Rent Expense” divide “Square Footage” to calculate “Rent per square foot” in order to predict whether this attribute correlates to the closed businesses directly. By deleting the “Rent Expense” and “Square Footage” and adding the new column called “Rent per square foot” in our model, we want to get a more precise prediction of whether the business will close or not.

i. Problems: However, after we drop NaN attributes, we observe that there are no sufficient closed business for us to evaluate. Although we will probably get a better result using only one number instead of two columns, the overall dataset will be biased and the result will be inaccurate. Thus, we decided to use our original data (columns included "Rent Expense" and "Square Footage" instead of one column named "Rent per square foot").

2. **Outcomes and Model Evaluation**

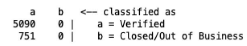
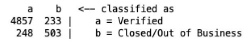
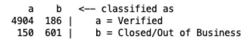
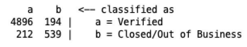
   a. **Introduction**: In this project, we tried various models including NaiveBays, Logistic Regression, K-Nearest Neighbors, OneR, Decision Trees, Random Forest, and AdaBoost. Among these models, Decision Tree works efficiently and gave us the best performance on predicting whether the business with specific attributes will close or not.

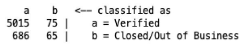
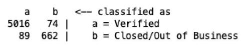
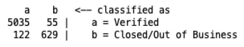
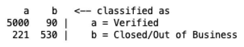
   b. **Model Validation and General Settings**: In order to evaluate different models based on the same standard, we decided to use the accuracy score calculated by the confusion matrix of testing set to compare the performance between different models. We use ZeroR model as a baseline to determine which model can give more precise results and perform better than ZeroR. We used K-Fold for cross-validation with random shuffle (set random seed to 10) to split the training set and testing set randomly. By setting the parameter k = 10, the imputed dataset will be separated into 10 pieces and the model will run 11 times (the final run will be training the whole dataset without splitting). For each iteration, one piece will be chosen as the testing set and the rest will be the training set.

   c. **Results**: We use Weka to predict the results for each model. For each algorithm, the independent variable (or the class) is "Record Type". The goal for our model is to predict whether the business will likely close purely based on factors shown below:

```
ZIP Code                          object
Executive Ethnicity               object
Executive Gender                  object
Neighborhood                      object
Location Employee Size Actual      int64
Location Sales Volume Actual     float64
Location Type                     object
Credit Score Numeric               int64
Square Footage                    object
Rent Expenses                     object
Record Type                       object
```

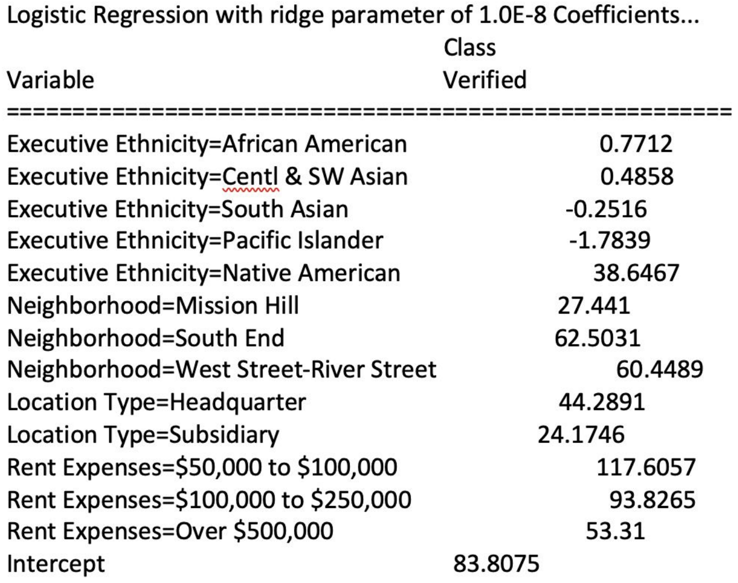
The following is the Classified Instance Accuracy Table for each algorithm

Algorithm Name	**Accuracy**	**Confusion Matrix**	**Mean Absolute Error**	**Recommend?**
Zero R	87.1426%	 a b <-- classified as 5090 0 | a = Verified 751 0 | b = Closed/Out of Business	0.2242	Baseline
NaiveBays	91.7651%	 a b <-- classified as 4857 233 | a = Verified 248 503 | b = Closed/Out of Business	0.1049	
Logistic Regression	94.2476%	 a b <-- classified as 4904 186 | a = Verified 150 601 | b = Closed/Out of Business	0.0813	*
KNN (K=5)	93.0491%	 a b <-- classified as 4896 194 | a = Verified 212 539 | b = Closed/Out of Business	0.0939	

One R	86.9714%	 a b <-- classified as 5015 75 | a = Verified 686 65 | b = Closed/Out of Business	0.1303	
Decision Trees (Pruned)	97.2094%	 a b <-- classified as 5016 74 | a = Verified 89 662 | b = Closed/Out of Business	0.035	*
Random Forest	96.9697%	 a b <-- classified as 5035 55 | a = Verified 122 629 | b = Closed/Out of Business	0.0694	*
AdaBoost	94.6756%	 a b <-- classified as 5000 90 | a = Verified 221 530 | b = Closed/Out of Business	0.2154	

d. **Model Evaluation**: Based on the accuracy score, we choose the top three recommended algorithms to analysis as follows:

i. **Linear Regression**: From the linear regression model, the factors that play a major role in determining whether the business will be closed are as follows:

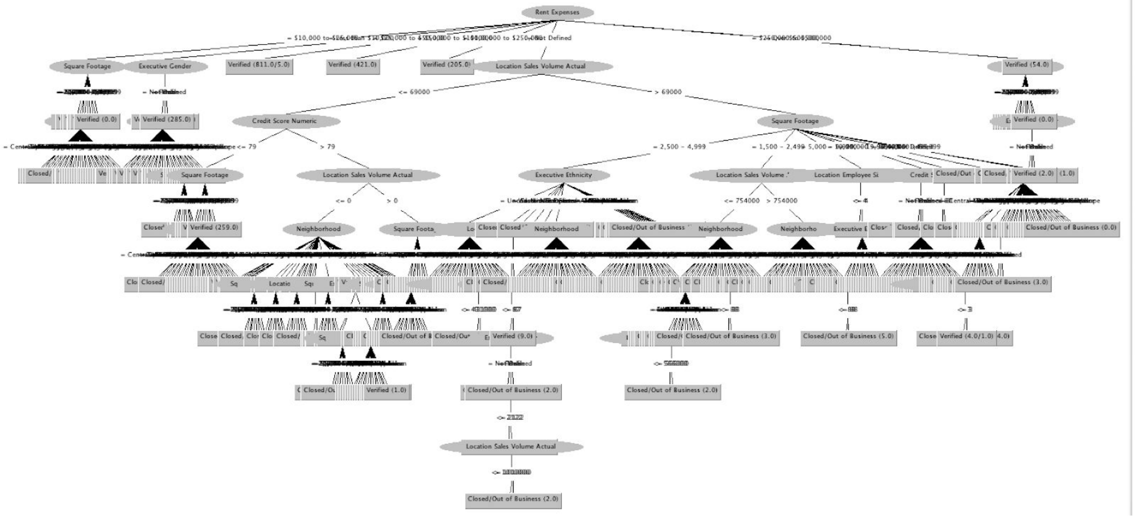


```
Logistic Regression with ridge parameter of 1.0E-8 Coefficients...
                                          Class
Variable                                  Verified
======================================================
Executive Ethnicity=African American            0.7712
Executive Ethnicity=Centl & SW Asian            0.4858
Executive Ethnicity=South Asian              -0.2516
Executive Ethnicity=Pacific Islander         -1.7839
Executive Ethnicity=Native American            38.6467
Neighborhood=Mission Hill                    27.441
Neighborhood=South End                       62.5031
Neighborhood=West Street-River Street            60.4489
Location Type=Headquarter                    44.2891
Location Type=Subsidiary                   24.1746
Rent Expenses=$50,000 to $100,000              117.6057
Rent Expenses=$100,000 to $250,000              93.8265
Rent Expenses=Over $500,000                    53.31
Intercept                                  83.8075
```

Among those variables, we observe that the higher rent expense is, the more likely a small company will go out of business. Also, an executive with the ethnicity of South Asian or Pacific Island is not likely to close their businesses. Beside most native American executives of the company will likely fail (which makes sense because most data represent the business in the native U.S.), the African American and Central & SW Asian executives will likely close the companies compared to other executive ethnicities according to our findings. Maybe, further regulations need to be generated in order to protect those ethnicities to successfully retain the company.

The companies in neighborhoods of Mission Hill, South End, and West Street-River Street are more likely to be closed according to the prediction. The high crime rate and rental expenses can explain why this might happen.

ii. **Decision Tree**: From the Decision Tree model, we get the tree as follows:



As the picture shown above, we can clearly see that "Rent Expense" is a strong predictor, followed by Square Footage, Location Sales Volume, Credit Score etc. This model has the highest accuracy score and can be used for further prediction.

iii. **Random Forest**: Since the dataset is a relatively high-dimensional dataset and from data visualization, linearity is not guaranteed, we decided to use random forest to further improve the model's predictive power. Compared to Logistic Regression, although the accuracy is a bit lower, the MAE(Mean Absolute Error) of Random Forest is also lower, which also indicates a better performance of the model.

Problems Encountered:

- **Problems with First Dataset**:
  - First, we use the 'property assessment FY2019' datasets. We tried to limit the business type to 'RC' and 'C' which is mix used and commercial parcels. We then tried to limit down our interest in four specific areas: 'ROXBURY MA', 'MATTAPAN MA', 'DORCHESTER MA', 'EAST BOSTON MA'. However, the database only provided the complete address for the business billing address, which is often different from the actual address of the business.
  - In order to retrieve the information that whether the business is closed or not in this parcel, we need to plug in the specific location to the yelp API and Google places API. These two APIs will return information that whether this business and the businesses around are "closed". However, this dataset did not provide a specific city and state for the parcel, we are unable to plug the address into the APIs. Moreover, the street addresses provided by this dataset are quite confusing. It often has addresses like '254 256 Bennington St' and empty number with only the street name, so we are not sure which is the parcel of interest.
- **API limitations**
  - Yelp API
    - The first attempt was the yelp API. The goal was to send in addresses and return surrounding businesses in order to check the status of the results. However, yelp API only includes businesses with reviews, such as restaurants, and other service businesses. The businesses which don't have a review would not be recorded in the API, so those would be excluded from our data set. This limitation leads to an incomplete dataset, which would skew the result of our measurement.

- Google places API
    - Since we do not have a formal address to plug into the google places API, we are unable to get the closed businesses around these parcels. Moreover, after we tried to enter our Reference USA data into the API, it returns not only businesses but also other things such as ‘Kenmore’ and ‘BU Sparks!’. So, the returned information by google places API are including non-business type information.

Answers to Research Questions:

1. Are there any relationships between commercial displacements and economic information (average home price and price per square foot for commercial property) or other information (crime rates) in particular neighborhoods and areas?
    a. The bar graphs 1 and 9 for the number of closed businesses and average home prices in four districts (Dorchester, East Boston, Roxbury, and Mattapan), indicated that there is a positive correlation between the rental price and closed businesses. For example, Dorchester has both the highest rental price and the number of closed businesses.
    b. The heatmaps for crime incidents and the identified closed businesses indicate that there are some relationships between the crime rate and closed businesses. As shown on the graph of identified closed businesses, several spots with green color, which means more businesses are closed near this area, such as the one near the center of Dorchester, the one at the west of East Boston and the, have corresponding highlighted spots in the graph of crime incidents. Since the crime data only includes ones that happened in 2019, if a larger dataset could be used, the relationship between the crime rate and closed business might be more obvious.
2. What are the common types of businesses that are being closed?
    a. What are the similarities between owners in the businesses? (ethnicity and gender)
        i. From the bar graph (Graph 6) and pie chart (Graph 5) showing the gender and ethnicity of the executives of the closed businesses, we concluded that most of the business owners are Western European and uncoded (possibly Americans). However, the gender of the owners is half female and half male, so gender has no effect on the conditions of small businesses.
    b. Property Value? (rental expenses)
        i. From Graph 9, the average rental expenses in Dorchester are the highest, at around $60,000 and Mattapan has the lowest average rental expenses, at around $500. This also corresponds to the number of closed businesses in these four places that higher the rental expenses, more of the closed businesses.
        ii. From the heatmap of rent expenses, small business owners near western Roxbury, central Dorchester and the western East Boston face relatively high rent expenses compared to other districts. Compared with the closing rate of this area, there is a clear relation between the high rent expenses and the closing rate.
    c. Size in terms of sq ft? (square footage)
        i. Graph 8, showing the physical size of closed businesses, indicates that most of the closed businesses are 1500 to 2499 square feet and 1 to 1499 in size. This gives a sense that most of the closed businesses are also small size companies. It is reasonable that small companies are more susceptible to the rental and crime rate around.
    d. Are there geographic trends in the closed? (district wise, neighborhood wise)
        i. From Graph 1 and 2, we can see that there are more businesses closing in Dorchester than other Districts. About 400 of the closed businesses are located in Dorchester and about

175 closed businesses located in East Boston and Roxbury. The lowest amount of closed businesses is located in Mattapan. Moreover, over 80 businesses are in Central Maverick Square and Paris Street in East Boston and about 70 businesses are in the Neponset of Southeast Dorchester.

3. Can we predict where the next closed business is likely to happen? (prediction)
    a. From our prediction analysis, there are many attributes combined that increase the likelihood of whether a business will close. Generally, if the business of the following attributes: its executive's ethnicity is African American or Central & SW Asian, its neighborhood in Mission Hill, South End, and West Street-River Street, and its rent expense ranges from $50,000 to $100,000 or from $100,000 to $250,000, the company is more likely to close in the future.
4. Are there common attributes of property associated with closed businesses at a given location? (determine business type)
    a. From our data analysis, we found that most of the closing businesses are medical-related businesses. Graph 7 gives a perceptual intuition that in a total of 346 closed businesses, there are over 100 businesses classified as physicians and surgeons. 66 of them are dentists and 51 are nurses-practitioners. This may be due to the high competition of medical businesses in Boston areas. Moreover, there are also about 50 closed restaurants.

Overall conclusion and suggestions:

Overall, we used a data set that includes all the necessary information about small businesses from reference USA in specific areas (Dorchester, Mattapan, Roxbury, East Boston), and a crime incidents data set from the Boston government. Based on the analysis of these data, we found positive correlations between the closing rate of small businesses and many factors, such as the general rent expenses and crime rate in different districts. The areas with a high business closing rate overlap with the areas with high rental prices and with frequent crime incidents. The data from reference USA does not include Americans, so the investigation about the ethnicities of small business owners indicates that more than half of the closed businesses were owned by Western Europeans. Another big part is uncoded, which might represent Americans. We also found that there is no relationship between closing businesses and gender since it was almost half male owners and half female owners.

Future improvements:

With better data sets from property assessment, more information about these parcels can be analyzed by using google places API, which returns all the business and other types of properties within a set radius. Then, the vacant rate can be calculated by using the returned information from google places API. We can use the building permits dataset and compare it with the places where have a high vacancy to find the relationship. In addition, the closed year of the business and address can help to determine the types of companies that fill the vacant places. Further analysis can be done over the whole Boston area to include more parameters and complexity in our data set.

Reference:

"Reference USA" InfoGroup.

http://www.referenceusa.com/Home/Home

"Boston Government" Crime Incident Report

https://data.boston.gov/dataset/crime-incident-reports-august-2015-to-date-source-new-system