Springboard Data Analytics Course

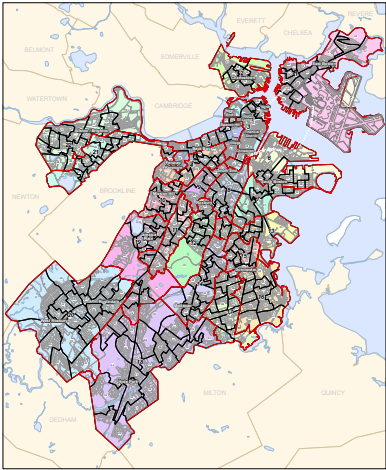
Milestone Report 1B for Capstone II

Comparing Ontime and Overdue Boston 311 Calls

November 2019

John L. Parsons

**Map of the Boston Wards and Precincts in Massachusetts**



Contents

[**Introduction** 3](#_Toc24214776)

[**1.1 Problem Statement** 3](#_Toc24214777)

[**1.2 Dataset Summary** 5](#_Toc24214778)

[**Data Wrangling** 7](#_Toc24214779)

[**2.1 Missing Values** 7](#_Toc24214780)

[**2.2 Transforming and Binning** 8](#_Toc24214781)

[**2.2 Time Analysis** 15](#_Toc24214782)

[**2.3 Geo Pandas** 16](#_Toc24214783)

[**3.0 Correlation Plots** 17](#_Toc24214784)

[**Bivariate and Time-Date Analysis** 19](#_Toc24214785)

[**4.1 Bivariate Analysis** 19](#_Toc24214786)

[**4.2 Time Analysis** 22](#_Toc24214787)

[**Time Analysis in 2019** 25](#_Toc24214788)

[**5.1 Time Results** 26](#_Toc24214789)

[**5.2 Time Analysis for PWDx and BTDT** 28](#_Toc24214790)

[**Machine Learning Models** 30](#_Toc24214791)

[**6.1 Base Line Model** 31](#_Toc24214792)

[**6.2 Model Optimization I** 32](#_Toc24214793)

[**6.3 Model Optimization II** 33](#_Toc24214794)

[**6.4 Hyperparameter Testing** 34](#_Toc24214795)

[**6.5 Champion Model** 35](#_Toc24214796)

[**Geo Pandas** 37](#_Toc24214797)

[**Conclusion** 40](#_Toc24214798)

[**References** 41](#_Toc24214799)

[**Appendix** 42](#_Toc24214800)

# **Introduction**

## **1.1 Problem Statement**

Boston Massachusetts paid the fifth highest property tax rates compared to the other 49 states and their average property tax rate was 6,019 dollars based on a 2018 survey. New Jersey had the highest tax rate at 8,780 dollars per home and Alabama had the lowest tax rate of 788 dollars per home in this study. Boston Residents pay 15.48 dollars per 1,000 of the assessed property values and this amount is up 0.38 cents from the previous year. In a recent poll in Iowa, the majority of Iowans want more accountability and transparency in their tax bills and “believe it is time to tackle the issue of high property tax” as tax rates continue to increase every year for most states. This internal study will be the first of several audits that will be conducted from an independent agency to show the residents a more accountable and transparent government.

This audit will conduct a three-year analysis for all 311 Boston Calls made from October 1, 2017 to October 31, 2019 to determine how many calls were completed ONTIME compared to OVERDUE Calls for this three-year period. 311 Boston is a Government service that connects Boston Residents with highly trained Constituent Service Representatives for all non-emergency City Services and the goal of this study is to determine how effective this agency is in completing non-emergency issues ONTIME and within budget. The four milestones for the Internal Audit are;

1. Analyze the time (days) it takes to complete all 311 service calls for the three-year and the 2019-year period. The close\_open\_diff variable was created to measure the length of time to complete all calls for Boston and will be compared to San Francisco, Chicago and New York City in 2019. This time analysis will determine if Boston’s completion time for 311 calls is similar to other cities or they are significantly longer.
2. Develop a champion machine learning model that will accurately predict which 311 calls made from Boston Residents have a higher probability for not being completed on time. The higher risk calls will be forwarded to the special task force that will be qualified to handle problem cases to minimize the time it takes to complete these incidents.
3. Map specific locations using the latitude and longitude coordinates with Geo Pandas for select overdue 311 Calls in the Boston Area to determine if location may be a factor in length of time for completing these cases.
4. Submit a Final Report that will list possible areas that can be improved on to cut cost and time it takes the city to service 311 calls and be more efficient to the public.

This Internal Audit will contain three phases. The first phase will contain the Milestone Report 1A. This report will be the data exploratory phase and will list the dataset selected for the study, all data wrangling steps used to process the data and the univariate and bivariate analysis of the variables. The second phase will contain the model development section of the project and will provide the champion model that will be used to determine completion time for this study. This will have the comparison of the completion time between Boston and for Chicago, San Francisco and New York City areas and Geo Pandas maps of overdue calls. The Final Report will contain the condensed version of Millstone Reports 1A and 1B, the results from the internal audit and the recommendations to reduce overdue completion time rates.

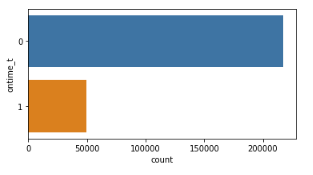
## **1.2 Dataset Summary**

The 311 Boston Open Source Data file can be found from the following website: <https://data.boston.gov/dataset/311-service-requests/resource/2968e2c0-d479-49ba-a884-4ef523ada3c0>. The 311 Service Requests was created by the City Constituent Relationship Management System (CRM) in 2011 and is updated daily. The 311 Boston open source data currently contains a total of 1,629,245 data entries and the data file contain 30 attributes. This study will only review the data from October 1, 2017 until the October 31, 2019 and this data file contains 267,312 entries of data. The Data Quality Report can be seen in Figure 1A, 2A and 3A and shows a list of the original variables and the ones that have been transformed to complete the analysis for this audit. All attributes that end with a “\_t”,” \_t1” or” \_t2” have been transformed and will be discussed in the data wrangling section of this report. The following is a brief description of the data set and the description can be seen in Figure 2.

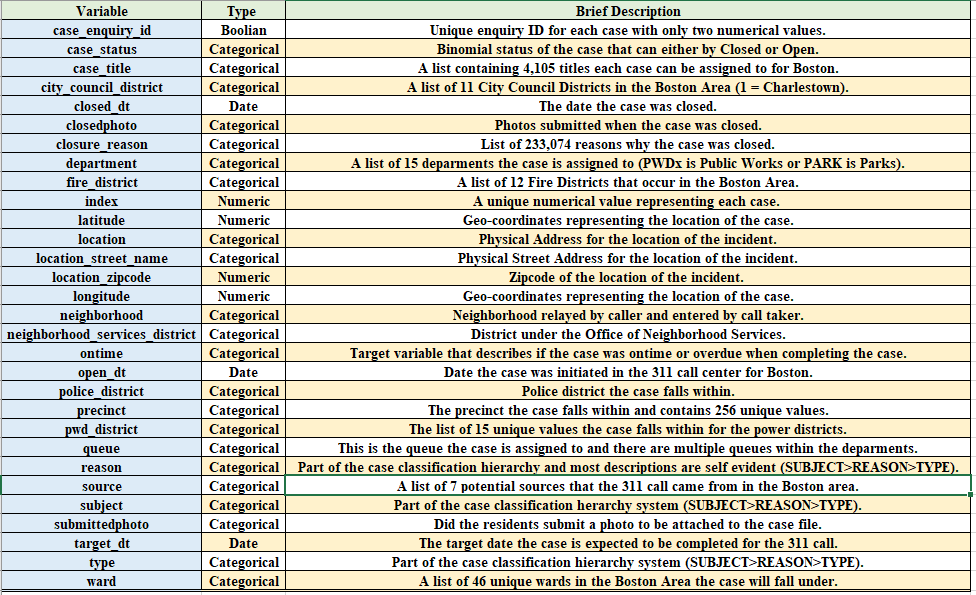
The ontime variable is a categorical attribute and will be the target variable for this study. This is a binary target variable and has been transformed to ontime\_t to convert the text to a binary value. The ontime\_t value will be represented with a level of 0 and the overdue value will be represented with a value of 1. This is an imbalanced dataset and level 0 contains 217,605 rows of data and level 1 only contains 49,705 rows of data. The remaining 29 attributes contains a total of 21 categorical variables ranging from a binary to a multifactorial variable, four numerical attributes, three dates and one Boolean attribute for the data set.

The 29 input variables can be divided into location, time, description and unique identifiers. The location input variables had many different levels of descriptions such as the physical location of the event or the district the event took place.

**Figure 1: Bar plot of the Target ontime\_t variable.**



**Figure 2: Boston 311 Description of the 30 Attributes.**



The specific location was recorded as the latitude, longitude, location, location\_street\_name, and location\_zipcode for the 311 events. The district location was divided up into a total of five districts found in the Boston Area. These five districts are; city\_council\_distict, fire\_district, neighborhood\_services\_distict, police\_district and pwd\_district. The last type of description for the 311 calls was placed into a case hierarchy system. This hierarchy system was subject (denotes which department the case is assigned to), the reason of the incident and then the type of incident that occurred. The time attributes for the Boston 311 dataset were given as the open\_dt, target\_dt and the closed\_dt of the event for the dataset. The open\_dt is the time the 311 case was opened and then the Boston Government gave a specific time the case was expected to be completed and the closed\_dt is the time the case actually took to complete. The unique identifiers for the dataset was the index, case\_enquiry\_id and case status. The dataset also contained photos of the event if submitted by Boston residents and government officials. These input variables were submittedphoto and closedphoto. The complete description of the attributes can be seen in Figure 2 and Figures 1A, 2A and 3A in the Appendix.

# **Data Wrangling**

## **2.1 Missing Values**

The Boston 311 dataset had a total of 16 attributes with missing values and these are; ontime, fire\_district, city\_council\_district, police\_district, pwd\_district, neighborhood\_services\_district, neighborhood, location\_zipcode, ward, precinct, case\_title, closedphoto, location\_street\_name, submittedphoto, target\_dt and closed\_dt. The list of the missing values and percent missing can be seen in Figure 4A in the Appendix. The missing values that had counts less than 35 were removed from the data set and these are; ontime, city\_council\_district, neighborhood\_services\_distict, ward and case\_title. The date and time variables target\_dt and closed\_dt had a total of 33,600 and 31,983 missing variables and were included in the dataset for descriptive, bivariate and model development phase. The missing values from the target\_dt and closed\_dt variables were removed for the time analysis section of this report to determine the average number of days it took to complete the case.

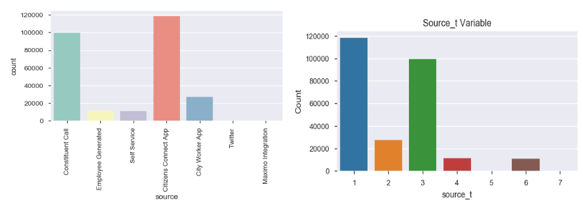
There are five input variables that had missing values and were not removed from the dataset. These five input variables with the number of missing values in parenthesis are; fire\_district (615), police\_district (125), pwd\_district (130), neighborhood (325) and precinct (194). These attributes with missing values had the missing value replaced with 0 and then were binned into a specific value called 33. The 33 bin will be used for most of the attributes selected for model development.

## **2.2 Transforming and Binning**

The input categorical variables in the dataset had descriptive names to explain the category like “Traffic Division” for subject or 10A for the pwd\_district that needed to be replaced with an integer for analysis. The input variables had a lot of missing values represented with ‘ ‘ or NaN in the dataset that needed to be removed (discussed earlier) or binned into a distinct value. The categorical variables also had a lot of sparse categories that will lower model performance and will be binned into a group to help increase model performance. The input variables that were modified with these approaches will be discussed next.

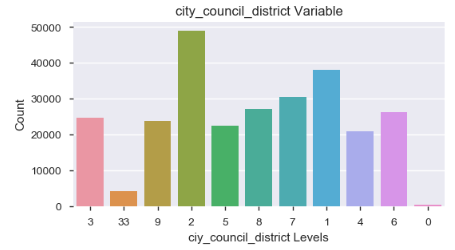
The source variable had a total seven levels ranging from “Self Service” to “City Worker App”. The count for all seven levels can be seen in Figure 3 below. The levels were arranged in alphabetical order and then assigned a number from 1 to 7 and placed in a new attribute called source\_t and can be seen in Figure 3. Levels 5 (Maximo Integration) and 7 (Twitter) were very sparse and another variable was created called source\_t2 and level 5 and 7 were binned into level six.

**Figure 3:** **Count plot for the source and transformed source\_t variables.**



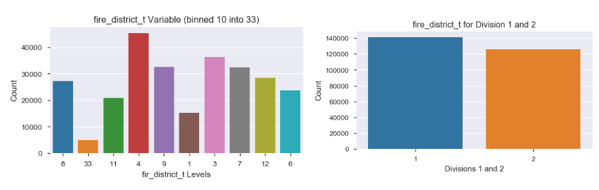
The city\_council\_district had a total of 10 levels and missing values listed as ‘ ‘ in the data column. The missing value ‘ ‘ was replaced with the number 33. The results from the replacement method and binning of empty values to 33 can be seen in Figure 4. The level 0 for the dataset is also very sparse and was binned into the level 33 and this can be found in the new attribute called city\_council\_district\_t.

**Figure 4. Count plot for the city\_council\_district variable.**



The fire\_district had a total of 10 levels and missing values listed as ‘ ‘ and nan in the data column. The missing values ‘ ‘ and nan was replaced with the number 33. The results from the replacement method and binning of empty values to 33 can be seen in Figure 5. The level 10 for the dataset is also very sparse and was binned into the level 33 and this can be found in the new attribute called fire\_district\_t. The city of Boston has two Fire District Divisions containing nine Districts and these were placed in new attribute called fire\_district\_t2. Districts 1, 3, 4, 6 and 11 were placed in Division 1 and 7, 8, 9, 10, 12 and 33 were placed in Division 2 (Figure 5). District 10 is not originally from the two Divisions and 33 is the binned empty values for data set. The names of the districts can be seen in Figure 5A in the Appendix.

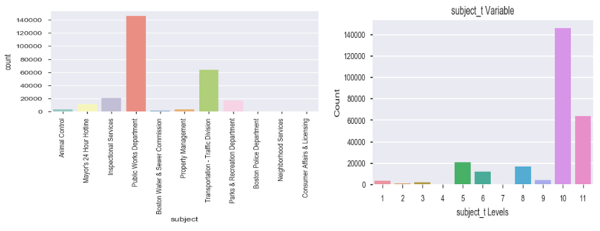
**Figure 5: Count plot for the fire\_district\_t and transformed fire\_distirct\_t2 variables.**

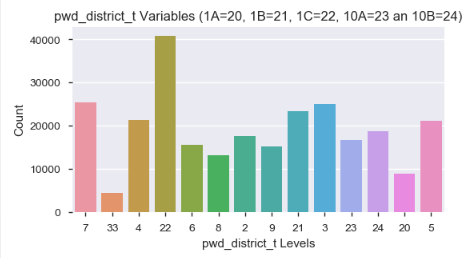


The subject attribute had a total of 11 levels for the data and did not contain any missing values. Figure 6 is a count plot with the original subject names and the distribution of these levels. The subjects were arranged in alphabetical order and then replaced with a value of 1 thru 11 and placed in the new variable called subject\_t. The levels 1 to 4 were very sparse and then another attribute called subject\_t2 was created and all four of these levels were binned into level 7 to increase model performance.

The pwd\_district had a total of 13 unique levels in the dataset and missing values that were binned into level 33 and can be seen in Figure 7. The pwd\_district had a mixed set of names which mostly used numbers to represent the district but also contain letters A or B for 1 and 10 pwd\_districts. The pwd\_district\_t was created and the letters from these pwd\_districts were replaced with a unique number starting with 20 for 1A to 24 for 10B.

**Figure 6: Count plot for the subject and transformed subject variables.**

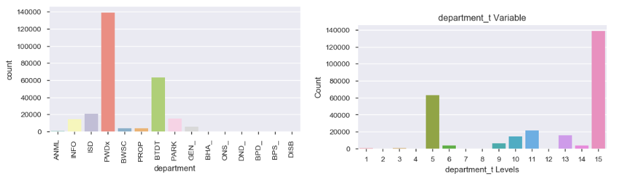


**Figure 7: Count plot for the transformed source pwd\_district\_t.**

The department attribute had a total of 15 levels and did not contain any missing values for the dataset. The subject was arranged in alphabetical order and replaced with a number 1 for “ANML” up to 15 for “PWDx”. The count plot for the original and replaced values can be seen in Figure 8. The department\_t2 attribute was created to bin the sparse levels in the dataset. Levels 1, 2, 3, 4 and 7 were all binned into level 8 for this variable. An explanation of these values can be seen in Figure 6A.

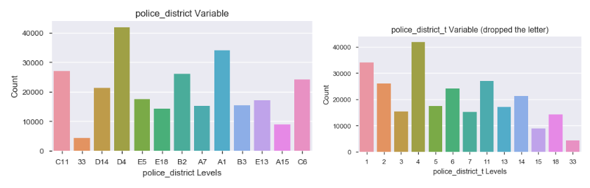
The neighborhood attribute contains the name of all 17 neighborhoods (including missing values) relayed by the caller and entered into the data base. This attribute also contained missing values that were replaced with number 33. All values in this variable were integers and will be used for the analysis. Level 0 was very sparse and will be binned into 33 only if this variable is used in the final model.

**Figure 8: Count plot for the department and transformed department\_t variables.**



The police district is the location where the case falls within for the Boston residents. There is a total of 12 districts and are labeled as a letter attached to the number. The missing variables were binned into level 33 and then all letters were removed and placed in a new attribute called police\_district\_t. This can be seen in Figure 9.

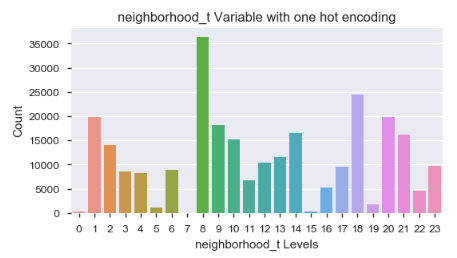
**Figure 9: Count plot for the police\_district and transformed police\_district\_t variables.**



The neighborhood attribute contains the name of all 24 neighborhoods (including missing values) relayed by the caller and entered into the data base. This attribute also contained missing values that were replaced with the name “UNKNOWN”. The attribute neighborhood\_t was created and then One Hot Encoding was used to assign specific values to the neighborhoods. The results can be seen in Figure 10 and the names of all the neighborhoods can be seen in Figure 7A in the Appendix. The number of calls for levels 0, 5, 7, 15 and 19 are sparse compared to the other neighborhood levels and will only be binned if this is used as the variable for the final model.

**Figure 10: Count plot for the neighborhood and transformed**

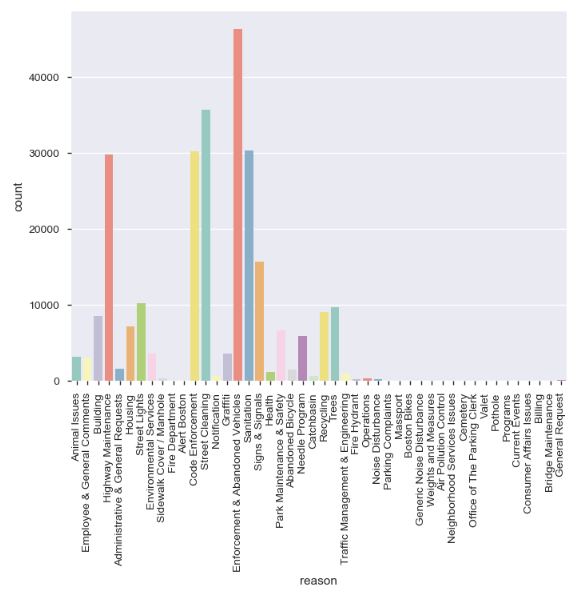
**neighborhood\_t variables.**



The ward attribute had a total of 44 levels and is the election ward a case will fall into for this variable. The ward category had an integer designated to the ward or a letter and integer designated each ward. A new attribute called ward\_t was created and One Hot Encoding was used to create a unique integer for each ward.

The reason variable was a text describing what the problem was and there is a total of 46 different reasons from the raw dataset. This attribute did not have any missing values and another attribute called reason\_t was created and One Hot Encoding was used to create unique numbers for this variable. The list of all reasons can be seen in Figure 11 and 8A (Appendix).

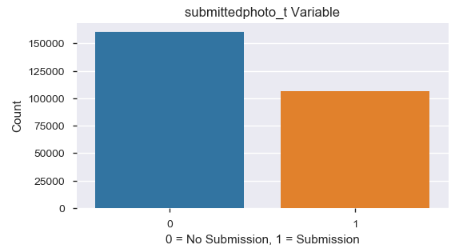
**Figure 11: Count plot for the 46 different levels listed as a reason.**



The last two attributes are the type and precinct for this study. The type of complaint had a total of 134 unique categories ranging from “Animal Lost” to “Walk-In Service Inquiry”. There were no missing values for this input variable. A new attribute type\_t was created and One Hot Encoding was used to generate unique numbers for analysis. A word cloud will be used in the final report to see which types were used the most for this study. The precinct variable had a total 255 levels and is the Election precinct a case falls within for the 311 calls. There were missing values for this variable and these were replaced with the number 33 and precinct 0502A was replaced with 0502001. A new attribute called precinct\_t was created and One Hot Encoding was used to generate unique numbers for this variable.

The last variable of interest was the submittedphoto category and this had a hyperlink attached to a picture which could not be uploaded or it was empty with no link to a photo. A new attribute called submittedphoto\_t was created and all missing rows of data was given a level of 0 (no photos) and all rows of data with a submitted photo was given a level of 1. Figure 12 shows the distribution of no photos to submitted photos.

**Figure 12: Bar plot of the binary submittedphoto\_t variable.**



The latitude and longitude attributes had no missing values for the dataset and will be used in the Geo Pandas to map out specific locations from areas that had higher overdue cases and will be compared to ontime cases.

## **2.2 Time Analysis**

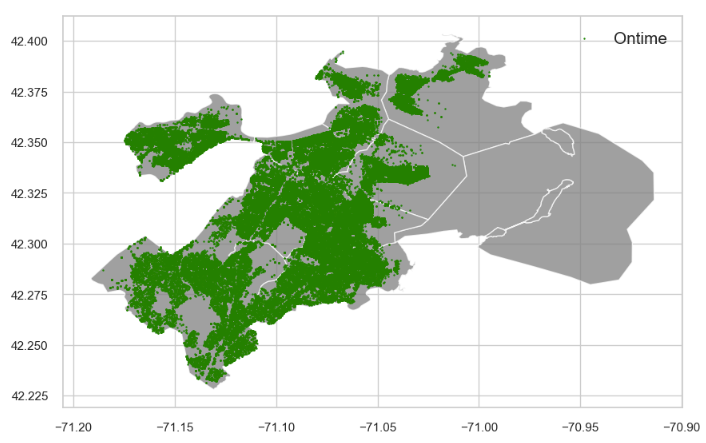
The three attributes that were listed as time and date the call was made is classified as an object in python. These three attributes are open\_dt, target\_dt and closed\_dt. The target\_dt had a total of 33,600 missing values and the closed\_dt had a total for 31,933 missing values. The raw data set was converted to the datetime function and the dataframe was filtered on all rows with an open\_dt greater or equal to October 1, 2017 for analysis. The two “time” attributes with missing values were removed and placed in another file called time. There was a total of 207,386 cases and only 28,168 of these were classified as overdue. The time dataset was separated into the on\_time and the over\_time data files and contained either all level 0 (ontime) or level 1 (overdue) attributes for the time analysis section. The comparison of the completion time (number a days to complete a project) will be discussed further in the Time Analysis Section.

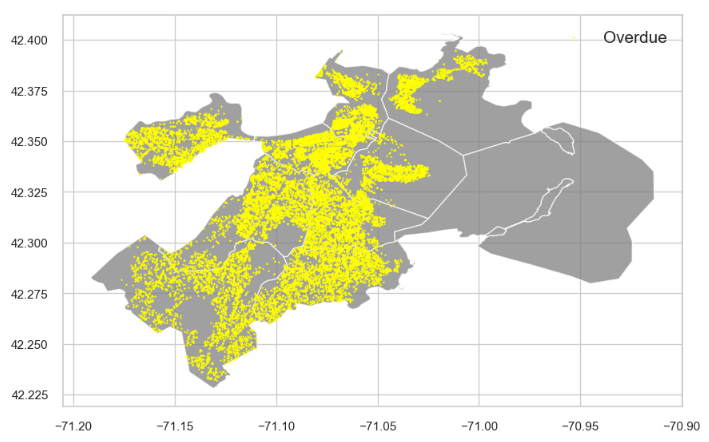
## **2.3 Geo Pandas**

The attributes used for the Geo Pandas was the latitude and longitude for location and the target variable ontime\_t. The latitude and longitude attributes did not have any missing values and every call was assigned a specific coordinate. A base map had to be downloaded from the

Boston Open Data web site and the two base maps that will be used is the street map and the Fire District map. The data was separated into ontime and overdue cases and the coordinates were plotted for all data points in this file and can be seen in Figures 13 and 14.

**Figure 13: Geo Mapping for all ontime cases within the 9 Fire Districts.**

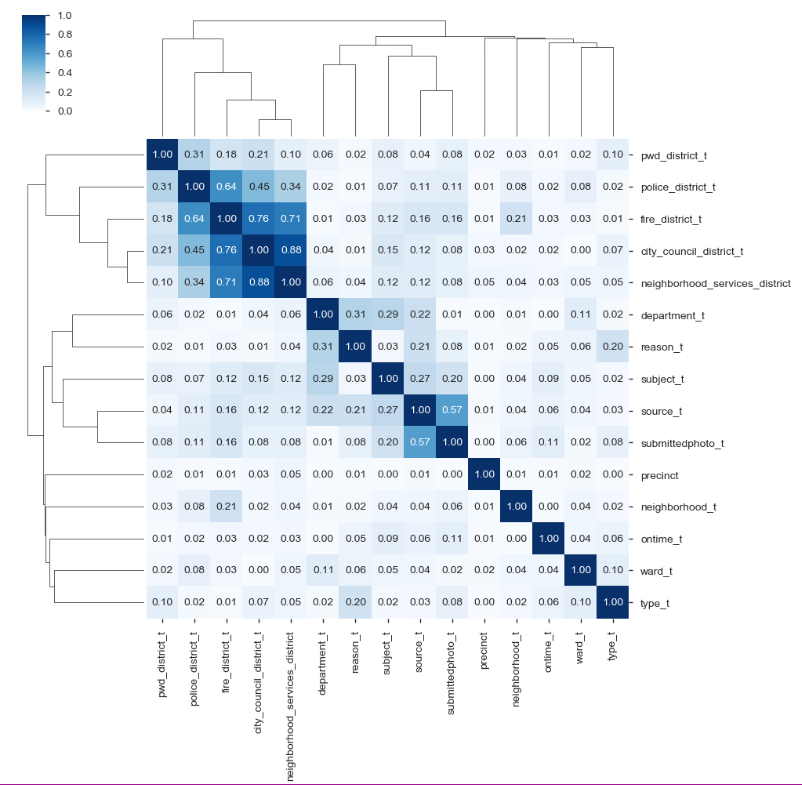


**Figure 14: Geo Mapping for all overdue cases within the 9 Fire Districts.**

## **3.0 Correlation Plots**

The pdf file for Boston 311 discusses how several cases could be placed under different attributes such as districts. The correlation cluster map in Python was created to determine which attributes were closely related to each other and will not be used in the final model. The Cluster map can be seen in Figure 15 and the five districts are grouped together and have a higher correlation than the other attributes except for source\_t and submittedphoto\_t (57%). The city\_council\_district\_t and neighborhood\_district\_t had a correlation value of 88% and the next highest value was between fire\_district\_t and city\_council\_district\_t with a value of 76%. The last variable with a high correlation value is fire\_district\_t and neighborhood\_district\_t with a value of 71%. These will not be used together in the final champion model to reduce multicollinearity from the dataset.

**Figure 15: Seaborn Cluster Map for 15 variables.**



The correlation values were generated between the target and the input variables in Python and the LogWorth values in JMP SAS to help in the attribute selection process for the champion model. The target variable ontime\_t was compared to the input variables that were transformed in the data set and the number of days (time differential) it took to close the case between all three time-date input variables was only included in the first correlation matrix. The open\_dt, closed\_dt and target\_dt time-date stamps was only included in the LogWorth values for JMP SAS. The time-date input values will only be used as the metric to be compared between the three metropolitan cities for time to completion for 311 cases and the Geo Pandas. The results can be seen in Figure 9A in the Appendix. These results show the top six correlated values (excluding time differential attributes) between the input and target variables are; submittedphoto\_t, subject\_t2, source\_t, type\_t, reason\_t and fire\_district\_t2.

# **Bivariate and Time-Date Analysis**

## **4.1 Bivariate Analysis**

A total of 17 attributes will be used to create the best champion model to predict which cases have a higher probability of becoming overdue. These are; ontime\_t, source\_t, city\_council\_district\_t, fire\_district\_t, subject\_t, pwd\_district\_t, department\_t, neighborhood\_services\_district, police\_district\_t, neighborhood\_t, ward\_t, reason\_t, type\_t, precinct and submittedphoto\_t. The bivariate analysis section contained the OLS Regression, Tukey Comparison of Means, histogram and normal probability plot, seaborn catplot with ontime\_t as the target variable and the panda’s cross tabulation table. Graphs were generated in the data wrangling section for all of these 17 variables. The three attributes that will be discussed in this section are the department, subject and reason variables.

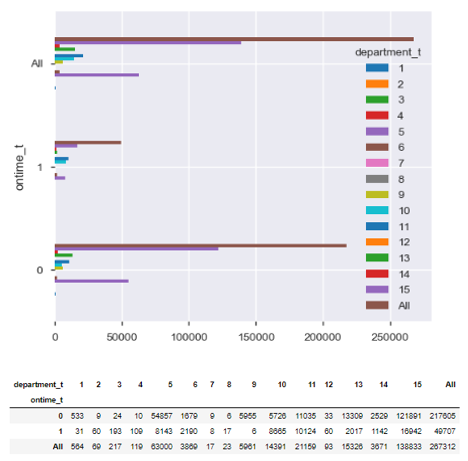
The cross tab was graphed and the table was included for the department\_t variable and the results can be seen in Figure 16 below. Department\_t 10 and 11 had very interesting results

and these departments are the Information Channel (10) and the Inspection Services (11). There were more overdue cases (8,665) than ontime cases (5,726) for the Information Channel and

48% of the cases were overdue for the Inspection Services (10,124) compared to the ontime

**Figure 16: Catplot and Cross tabulation results for department\_t**

**variable and ontime\_t.**

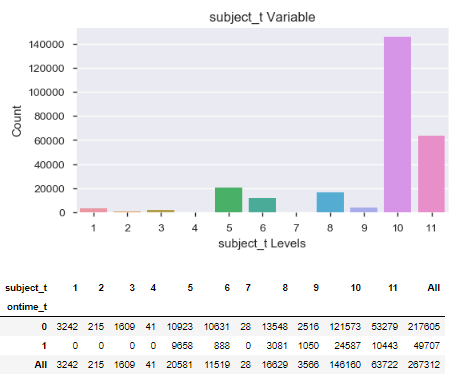


cases (11,035). There was a total of seven out of nine levels that had significant p values of 0.0 for the OLS Regression Model and both of these were included in this group with t value of 27 and 57. These two attributes contain 18,789 overdue cases out of 49,707 and is close to 38% of the total number of overdue cases and warrants further investigation.

The subject\_t has a total of 11 levels containing all 267,312 calls made to the 311-service call center (Figure 17). The two areas that contain the majority of the calls are number 10 which is Public Works Department and number 11 and this is the Transportation – Traffic Division. These two levels contain 70.5% of all overdue cases (35,030). The Boston 311 data set has values that are highly related as show in the correlation matrix cluster plot and have high multicollinearity. The level 10 for subject\_t variable is the Public Works Department and is

**Figure 17: Count plot and Cross tabulation results for subject\_t**

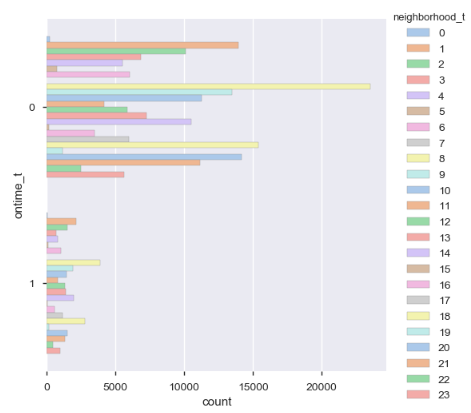
**variable and ontime\_t.**



identical to the department\_t variable 15. Level 5 is Inspectional Services and is similar to 11 in the department\_t category from the previous figure. These variables will not be used together for the final model and will be addressed in the final analysis.

The last variable in this analysis is neighborhood\_t to determine if certain locations had more overdue cases. The OLS Regression Results had four levels with a p-value of 0.001 or less in the data set and these are neighborhood\_t levels 0, 3, 20 and 21 (Figure 10A). These corresponding neighborhoods are Allston (0), Allston/Brighton (3), South Boston Waterfront (20) and South End Boston (21) (Figure 7A). There were three more levels with values between 0.001 and 0.05 and were not included in this section. The Seaborn catplot shows the distribution of these attributes and this can be seen in Figure 18. The mean time taken for the ontime and overtime will be calculated and discussed in date-time section of this report.

**Figure 18: Count plot for neighborhood\_t and ontime\_t.**

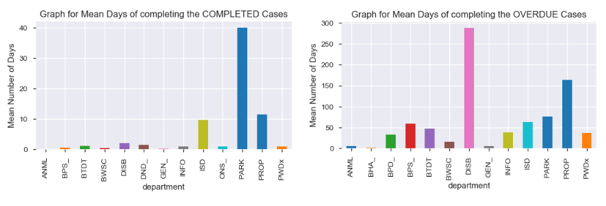


## **4.2 Time Analysis**

The time attributes had the missing values removed and the data set only contains a total of 207,386 rows of data. The total number of ontime cases in this data set is 179,218 and there were only 28,168 cases classified as overdue. There were 21,539 cases that were not closed (missing) and had to be removed to complete the analysis. The mean number of days was calculated from the time the project began and the time it was closed in the dataset. The ontime cases on the average only took 4.1 days to complete compared to 50.6 days for the overtime cases. The ontime variable minimal mean value to complete a case was 0 days compared the longest time to compete the case was 593 days. The overdue variable also had the minimal mean value of 0 days to complete the project and 738 days for the longest time to complete the project. The time-data analysis will discuss the mean number of days to complete for the department\_t and neighborhood variables and the box and whisker plot for the subject variable.

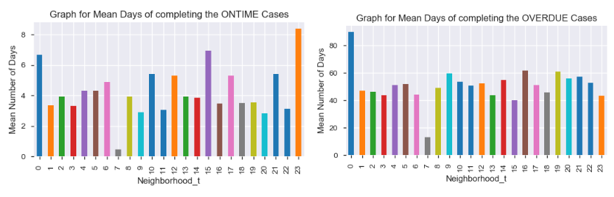
The time dataset was separated into two groups and the on\_time contained all the ontime variables and the over\_time contained all the overdue cases. The number of days it took to close a project was calculated and placed in the close\_open\_diff column and the number of days it took to complete an overdue project was placed in the target\_close\_diff column for analysis. The groupby function and graphs for the department and neighborhood can be seen in Figures 19 and 20. Figure 19 shows the PARK department had the longest mean time to complete the ontime case which was 40.03 days and the ANML had the quickest time of 0.0 mean days. The longest mean days to complete the overdue cases was DISB with 288.25 days and the BHA\_ had the lowest level with only 2.0 mean days to complete. The DISB is the Disability Commission and only contained a total of 10 cases, but had the longest time for the dataset.

**Figure 19: Graphs for mean number of days of completed and overdue cases.**



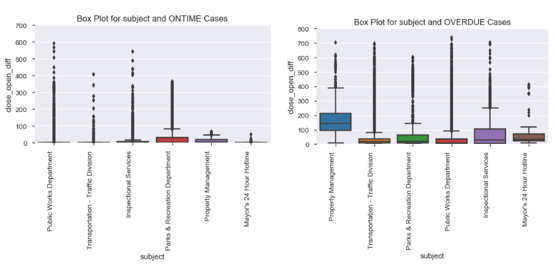
The OLS Regression model showed that neighborhood\_t levels 0, 3, 20 and 21 had significant p values of 0.001 or less and the mean time to complete the case was graphed and can be seen in figure 20. The one level that stood out the most was level 0 and this had a total of 225 cases that took on the average of 6.8 days to complete ontime, compared to 43 cases that took on the average 89 days to complete for overdue cases. The other three neighborhoods had an average ontime completion time between 3 to 5 days compared to 43 to 57 days for the overdue cases. Level 0 is the Alston Neighborhood and will be plotted in Geo Panda to determine if specific locations are more difficult to complete ontime.

**Figure 20: Mean number of days for completed and overdue cases for department.**



The Box plot was created for the subject attribute with the ontime and overdue cases and this can be seen in Figure 21. The ontime case mean for the Parks and Recreation was 39.9 days and the remaining four levels ranged from 1 to 11.6. The overdue case average for the property management was 169 days and the remaining five levels were ranging between 36 to 88 days for the dataset. All six overdue subjects had a lot of outliers in the dataset. The outliers for both distributions will be further analyzed in the results section to try and determine the possible cause of these data points that take more than 200 days to complete for the Boston area. The two areas of concern are the Transportation-Traffic Division and public works Department which contain around 70% of the overdue cases.

**Figure 21: Box plot for ontime and overdue cases for the subject variable.**



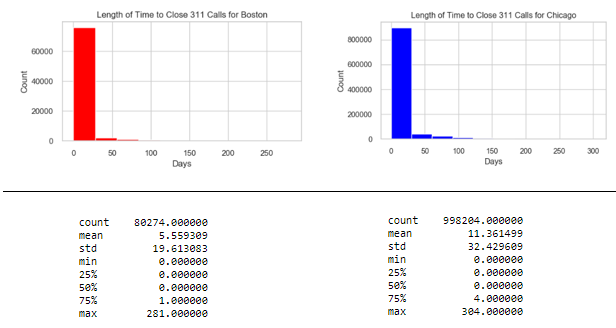
# **Time Analysis in 2019**

New York City, San Francisco and Chicago time to complete a case for the 2019 year was compared to Boston to determine if Boston’s completion rates were within range of other metropolitan cities. The Open Source Data Portal was used to select only the data for the 2019 year from all three cities and then was imported into Python. The date-time columns were converted to a datetime function in python and were checked to make sure all rows had a creation date greater than or equal to 01/01/2019. The missing values for both the time a case was created and the time the case was completed were dropped from the dataset and then the closed\_open\_diff column was generated which measured time in days to complete the case. The descriptive statistics, histogram and students T test for independent samples was run on all cities for completion time of the 311 calls for the 2019 year.

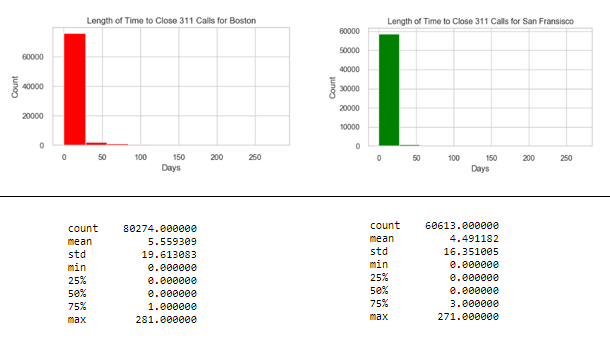
## **5.1 Time Results**

The histogram and descriptive statistics can be seen in Figure 22, 23 and 24. The mean number of days to complete the case for Boston was 5.6 compared to 11.4 for Chicago, 4.5 days for San Francisco residents and 8.0 days for New York City Residents. The minimum time to complete the case was 0 days for all three cities and the maximum time to complete a case was 304 days for Chicago. There were a few cities that actually had -1 days for the minimum value and these values were dropped from the analysis. Chicago had the greatest dispersion around the mean with the standard deviation of 32.4 days and San Francisco had the lowest standard deviation value of 16.4. Boston had the second lowest standard deviation with a value of 19.6. The maximum values for the 2019 year was between 271 days for San Francisco and 304 days for Chicago. The students T test was used to compare the means for all three cities and all three cv (critical values) and p-values were significant with a p-value less than 0.0 (Figure 25).

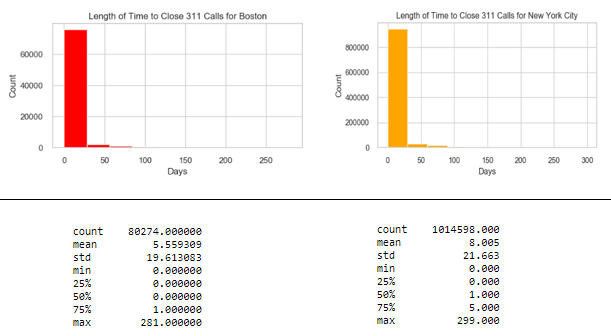
**Figure 22: Histogram and Descriptive Statistics for 311 calls for Boston and Chicago.**



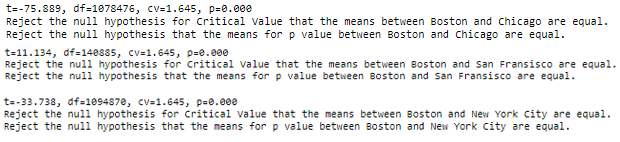
**Figure 23: Histogram and Descriptive Statistics for 311 calls for Boston and San Francisco.**



**Figure 24: Histogram and Descriptive Statistics for 311 calls for Boston and New York.**



**Figure 25:** **T Test for Independent Samples Comparing the mean Completion Time for 311 Calls.**

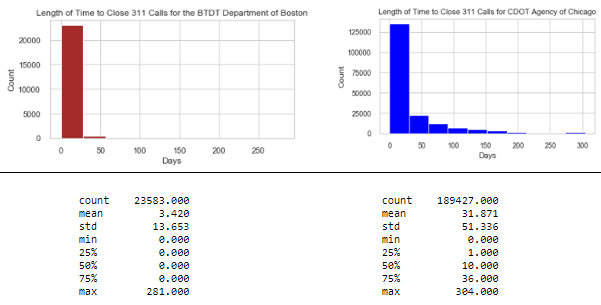


## **5.2 Time Analysis for PWDx and BTDT**

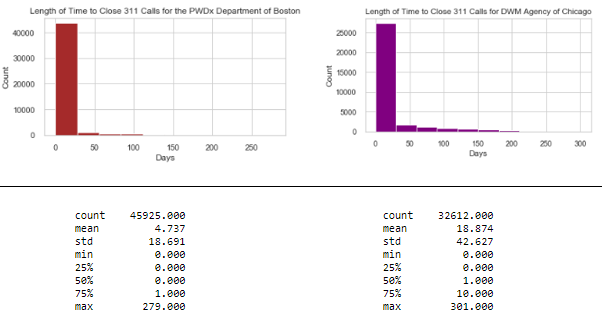
The average time to complete a 311 call was analyzed previously and Boston average was 5.6 and is very reasonable compared to the other three cities in this study. The last section for the time analysis is to look at specific departments and the Transportation (BTDT) and Public Works Department (PWDx) were selected for this analysis. The Public Works Department had the most cases for the Boston area and was 121,891 out of 217,605 calls for the three-year study. The Department of Transportation had the second highest number of cases and this was 54,857 out of 217,605 calls. This was difficult to compare to other cities because every city has different categories that are relevant to their specific location and they never were completely similar. The Chicago City was chosen to compare its Department of Water Quality (dpm) and Department of Transportation (CDOT) because they were more similar in agencies than other cities.

The results for these comparisons can be seen in Figures 26, 27 and 28. Boston PWDx mean number of days to complete a call was 3.4 compared to 31.9 for Chicago and the maximum number of days to complete a project was 281 compared to 304 days. The mean number of days to complete the 311 calls for the PWDx was 4.77 compared to 18.8 for the Chicago DPM and the longest time to complete the case ranged from 279 to 301 days. The Students T Test for independent samples compared the means for the PWDx and DWM and were both significant for the cv and p-value of mean days to complete the project. The results show that Boston does a good job in completing projects for these two departments when compared to Chicago

**Figure 26: Histogram and Descriptive Statistics for 311 calls for Boston and Chicago for between the BDTD and CDOT.**



**Figure 27: Histogram and Descriptive Statistics for 311 calls for Boston and Chicago for between the PWDx and DWM.**



**Figure 28: T Test for Independent Samples Comparing the Completion Time for 311 Calls**

**Between Boston and Chicago for the Water Departments.**



# **Machine Learning Models**

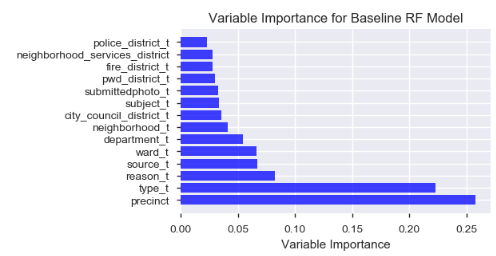
A base model will be generated from the Random Forest Model to give the base line metric for f1 score for model development. The base line model and first stage of model optimization will be using 14 attributes from the dataset. These 14 input attributes and target variable are; ontime\_t, source\_t, city\_council\_district\_t, fire\_district\_t, subject\_t, pwd\_district\_t, department\_t, neighborhood\_services\_district, police\_district\_t, neighborhood\_t, ward\_t, reason\_t, type\_t, precinct\_t, submittedphoto\_t and ontime\_t. The variable importance, confusion matrix and classification report will be generated and a reference point for model development. The Model optimization steps are divided into three stages for the Boston 311 dataset. The first and second stage will contain a group of seven models to determine the best f1 score for the ontime\_t variable and variable importance from the Random Forest Model. These seven models are the Random Forest, KNN Nearest Neighbor, Extra Tree Classifier, Support Vector Machines, Ada Boost and Gradient Boosting and multilayer perceptron. The second iteration will contain the same seven models but will only have select attributes from the previous feature importance values, correlation matrix and logworth values. The third iteration will contain the Ada Boost, Extra Tree Classifier, Random Forest Model and Gradient Boosting for the hyperparameter testing to find the best model. The final two models will be selected with the best f1 scores and tuned parameters to generate the champion model for detecting ontime\_t cases in the Boston 311 Dataset.

## **6.1 Base Line Model**

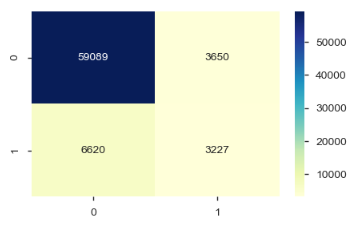
The Random Forest Model results can be seen in Figures 29, 30 and 31. The entire

267,375 rows of data were used in the analysis and the criteria was ‘gini index’ and the class weight = None. The top seven attributes were precinct, type\_t, reason\_t, source\_t, ward\_t, department\_t and neighborhood\_t. The Confusion matrix and Classification report shows the model does a good job in classifying the True Negative values (TN) with an accuracy score of 86%, but has a hard time classifying the True Positive Value (TP). The data set is imbalanced and the target variable that was ontime had a total of 217,605 rows and only 49,707 rows for overdue cases in the dataset. The f1 score will be used in the model development phase and then the balanced approach for the final set of models will be added to increase model performance.

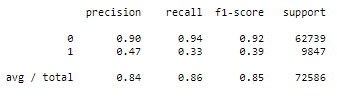
**Figure 29: Base line Variable importance for the Random Forest Model.**



**Figure 30: Confusion Matrix for the Baseline Random Forest Model.**



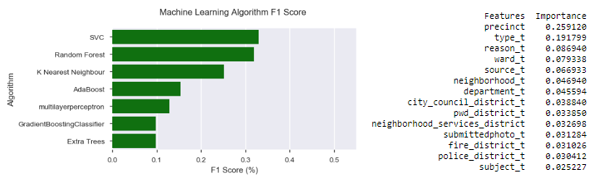
**Figure 31: Classification Report for the Baseline Random Forest Model.**



## **6.2 Model Optimization I**

The first stage of model development contained all 14 input variables with seven models using the f1 score to give a more balanced approach between recall and precision. A total of 41,477 samples was used for the first two runs and this is 20% of the data points. There was a total of 26,960 datapoints for the training set and 14,517 datapoints for the test dataset. The f1 scores from all seven models and the variable importance value from the Random Forest can be seen in Figure 32. The f1 scores ranged from 9.8% for the Extra Trees and 33% for the SVC Classifier. The first eight attributes are the same and precinct is the most important value and subject\_t is the least important attribute.

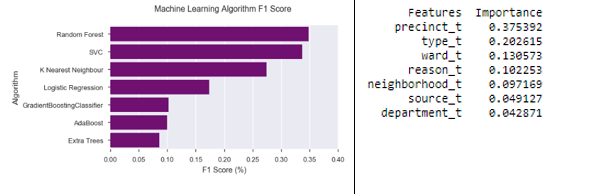
**Figure 32: f1 score and Feature Importance for Model Optimization I.**



## **6.3 Model Optimization II**

A total of seven input variables were selected for the second run based on the variable importance and correlation values in the dataset and all settings are the same as the previous model. These seven attributes are; ontime\_t, precinct\_t, type\_t, reason\_t, ward\_t, source\_t, neighborhood\_t, department\_t. The results from the second run can be seen in Figure 33 and the Random Forest had the highest f1 score with 34.9% and the SVC was slightly better from the previous run with a score of 33.7%. The precinct\_t is still the most important variable and department\_t is the least important for this run.

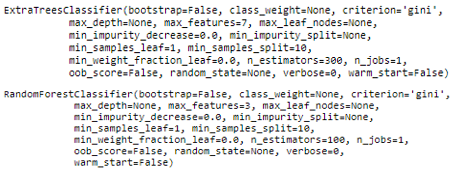
**Figure 33: f1 scores and variable importance for the second model optimization.**



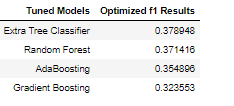
## **6.4 Hyperparameter Testing**

The previous seven input variables were used in four models for hyperparameter tuning and the city\_council\_district was included to increase model performance. The four machine learning models are; Ada Boosting, Extra Trees, Random Forest and Gradient Boosting. The data set only had 25% of the data randomly sampled and 33,699 samples were used for the training dataset and 18,147 samples were used for the test dataset. A five kfold cross fold validation was used to tune the models and f1 was scoring parameter. The best model parameters for the top models and f1 results can be seen in Figures 34 and 35. The best two models were the Extra Tree Classifier with an f1 score of 37.9% and Random Forest with a f1 score of 37.1%. The variable importance values for all four models can be seen in Figure 12A in the Appendix. The model performance increased from 8.6 % up to 37.9 for the Extra Trees and 34.9 up to 37.1 for the Random Forest. These two models will be tested with a balanced approach for class\_weight to improve the performance of the model and find the champion model.

**Figure 34: Extra Trees and Random Forest Best Model Parameters.**



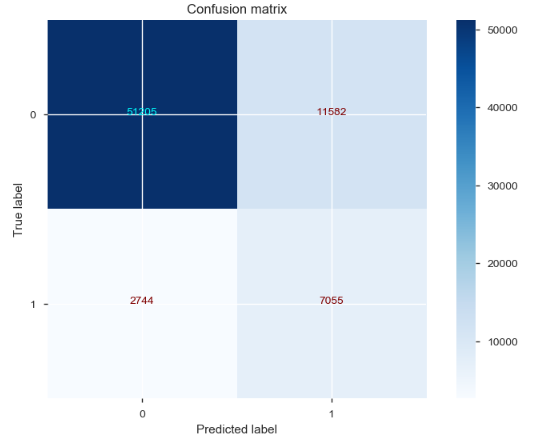
**Figure 35: Hyperparameter Tuning.**



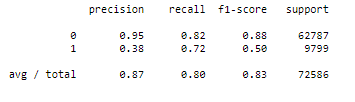
## **6.5 Champion Model**

The Extra Tree and Random Forest Model used the best parameters from the previous run and the balanced approach for the class\_weight. The model that had the best performance was the Random Forest model Extra Tree Classifier and the results from the Confusion Matrix and Classification Report can be seen in Figures 36 and 37 (A and B). The Random Forest had an accuracy rate of 80.3% and the f1 score was 46.9%. The Random Forest baseline model had an f1 score of 38.6% and optimized model increased f1 performance and did a better job in classifying the TP values for the data set. The model did not do well in classifying the False Positive (FP) values for the data set and is an area that needs to be improved on to increase the f1 scores. The Extra Tree Model had an f1 score of 57% and the accuracy rate was 83% and they both did a good job and were a better model than the base line model.

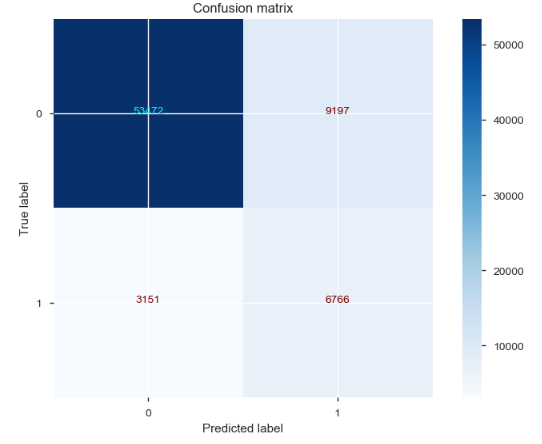
**Figure 36A: Confusion Matrix for RF Champion Model.**



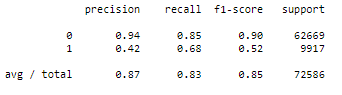
**Figure 36B: Classification Report for the RF Champion Model.**



**Figure 37A: Confusion Matrix for Extra Tree Champion Model.**



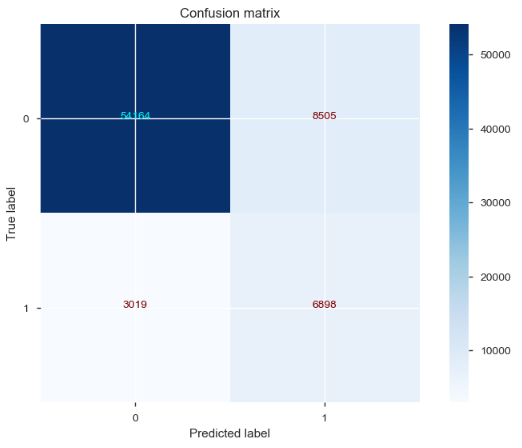
**Figure 37B: Classification Report for the Extra Tree Champion Model.**



The correlation matrix had the date-time variables with the highest values compared to the other input variables. The day, week and month variable attributes were created from the open\_dt time variable and tested with the previous Random Forest Model and the same parameters. The only change for this model was the addition of the week input variable and the model f1 score increased to 54.5% and the Accuracy Score was 84%. This model did a better job in selecting only 8,505 FP compared to 9,197 FP (Extra Trees) values and had 132 less False Negatives (FN) values than the previous model. The results can be seen in Figures 38 and 39.

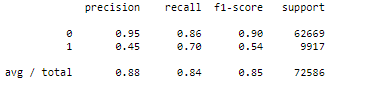
**Figure 38: Confusion Matrix for Random Forest with the addition**

**of the week input variable.**



**Figure 39: Classification Report for Random Forest Model with the**

**Addition of the week input variable.**



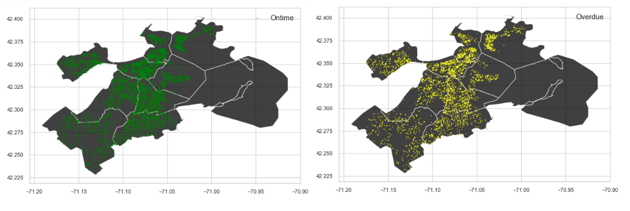
The SMOTE Model over-sampling approach was used on the same eight attributes (this includes week) and run to optimize model performance. The model results can be seen in Figure13A in the Appendix and did not perform as well as the other models. The f1 score was 17% and accuracy score was only 70.2%. The model did a good job in selection the TN and only selected 3,151 FN. The model did not do well for the FP and misclassified a total of 9,197 cases.

# **Geo Pandas**

The Geo Pandas was the graphing program initiated to determine if location may be a factor between ontime and overdue cases for the Boston Area. The base map selected for the background is the Fire District Map and was easier to visualize the data compared to the Street Level Map. The entire data set was used to plot for the department and the DISB attribute. The other variables did not show much variation between the different districts and were not used in the analysis. The ontime variable is a circle that is green in color and the overdue variable is a triangle with a yellow or orange color.

The Inspectional Services had 9,658 overtime cases out of a total of 20,581 cases and was almost half of the number of cases that did not get finished ontime. The GEO Panda was plotted the results can be seen in Figure 40. The overall spread of the data points for the ontime and overdue cases shows not specific hotspots for either group of data points. The empty Fire District on the upper right side is called Fire District 1\_3 and did not contain any samples from this study.

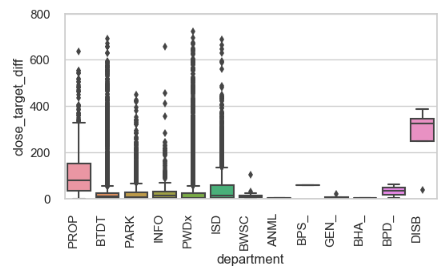
**Figure 40: Geo Pandas for Department filtered for ISD Cases only (Inspectional Services).**



The one interesting Geo Map for this study was the department DISB attribute because this had the largest mean number of days when compared to all other departments and this can be seen in Figure 41. The DISB attribute only had a total of 17 cases and nine were ontime compared to eight that were overdue for the three-year study. The first five variables starting

**Figure 41: Box Plot for all overtime Department Variables for the mean number**

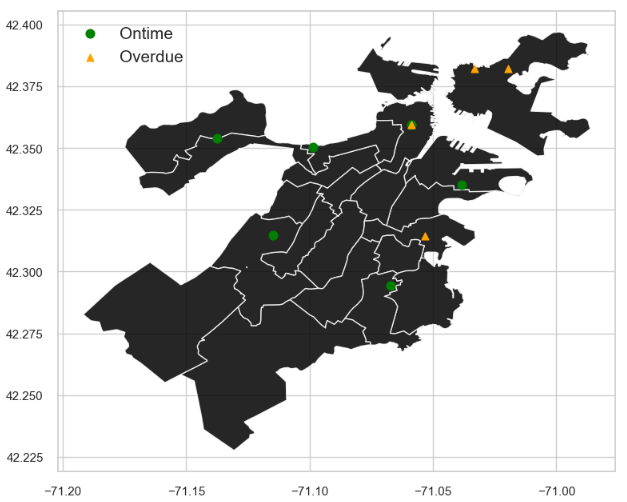
**of days calculated between the target date and the closing date for 311 calls.**



from PROP going to ISD shows a lot of outliers for these variables. The PWDx and BTDT were the attributes that contained 70.7% of all overdue calls for the data set. The DISB variable for the Department was Geo Pandas shows a good separation between the ontime and overdue calls (Figure 42). When the missing times were removed from the data set only six ontime and four overdue variables were left and seven still have not been finished at this time. A total of three overdue cases were only found in Wards 1 and 13 and number 3 was shared with one ontime case. Wards 1 is East Boston and Ward 13 is North Dorchester. The Wards that only contained ontime cases are 5, 6, 17, 19 and one on the line for 21 and 22. This is another area that needs to be looked into because this variable has the greatest number of mean days compared to the other variables and there is clear separation between three of the four overdue cases for the data set.

**Figure 42: Geo Pandas Wards for Department filtered for**

**DISB only (Disabilities – General).**



## **Conclusion**

The is the completed report for Milestone 1B and contains the model development, time analysis and Geo Pandas mapping for select attributes of the Boston 311 calls. The Final Report will contain the condensed version of the Milestone Report 1A and 1B, the discussion of the significant findings from the reports and the recommendations to reduce overdue completion time rates.

# **References**

Albon, C. (2018). Machine Learning with Python Cookbook, Practical solutions

preprocessing to deep learning. O’Reilly Media Inc in Boston, First Addition.

Analyze Boston. (ND). Welcome to Boston Maps, Open Data, Geospatial Datasets. Retrieved

from: <http://bostonopendata-boston.opendata.arcgis.com/>

Analyze Boston. (ND). 311 Service Requests. Retrieved

from: <https://data.boston.gov/dataset/311-service-requests/resource/2968e2c0-d479-49ba-a884-4ef523ada3c0>

Chicago Data Portal. (ND). 311 Service Requests. Retrieved from:

<https://data.cityofchicago.org/Service-Requests/311-Service-Requests/v6vf-nfxy>

DataSF. (ND). 311 Cases. Retrieved from:

<https://data.sfgov.org/City-Infrastructure/311-Cases/vw6y-z8j6/data>

Hastie, T., Tibshirani, R. and Friedman, J. (2017). The Elements of Statistical Learning: Data

Mining, Inference and Prediction. Springer Series in Statistics, 2nd Edition.

Hendrickson, J. (2019). Property Taxes Need Accountability and Transparency.

Retrieved from: <https://caffeinatedthoughts.com/2019/04/property-taxes-need-accountability-and-transparency/>

OPENGOV. (ND). Analyze Boston. 311 Service Requests. Retrieved from:

<https://data.boston.gov/dataset/311-service-requests>

NYC Data. (ND). NYC OpenData. Open Data for all New Yorkers Retrieved from:

<https://opendata.cityofnewyork.us/>

Patch.com. (ND). 2019 Residential Property Tax Rates for the 351 MA Communities

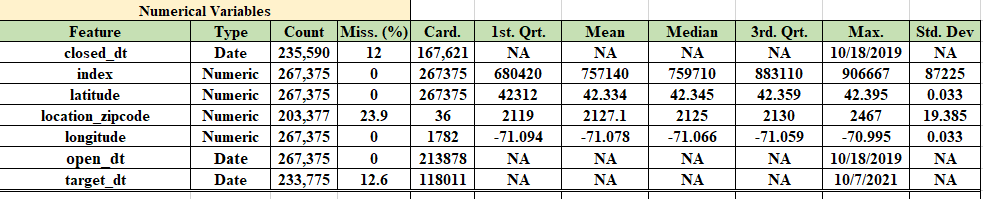
Retrieved From: <https://patch.com/massachusetts/boston/ma-residential-property-tax-rates-each-community>

Pichhi, A. (2019). Which homeowners around the U.S. pay the highest property taxes?

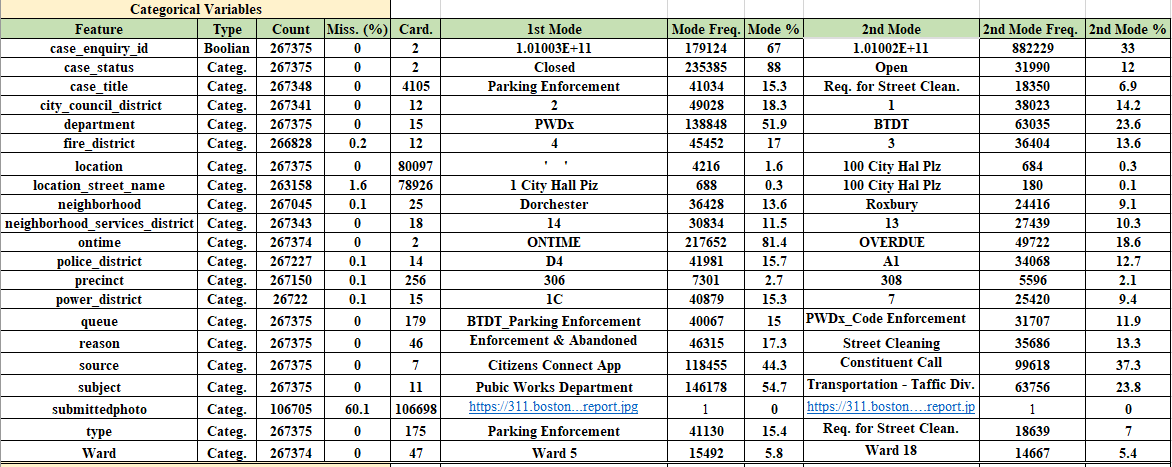
Retrieved from: <https://www.cbsnews.com/news/property-tax-which-homeowners-around-the-u-s-pay-the-highest/>

# **Appendix**

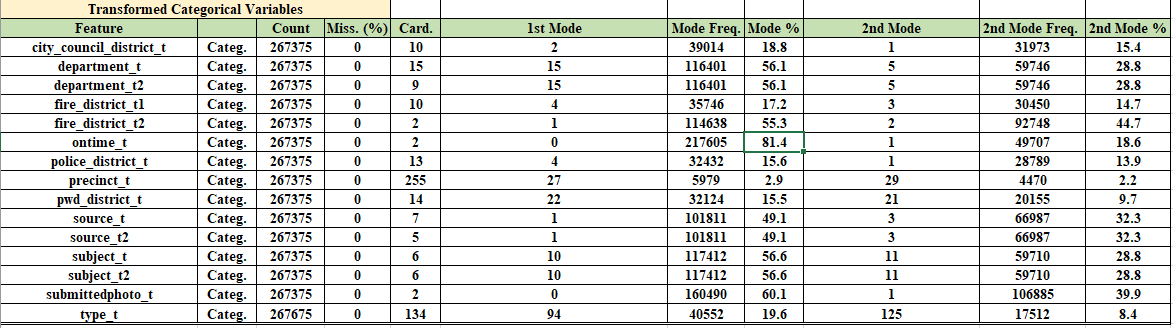
**Figure: 1A Data Quality Report for Numerical Attributes.**



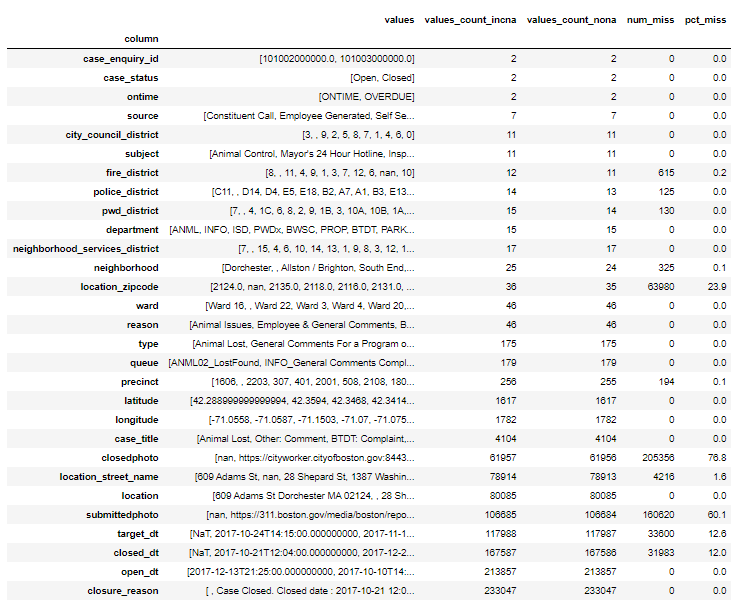
**Figure 2A: Data Quality Report for Categorical Attributes.**



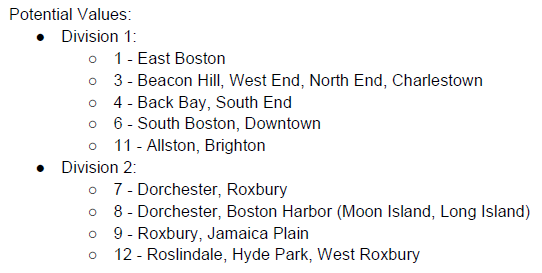
**Figure 3A: Data Quality Report for Transformed Categorical Attributes.**



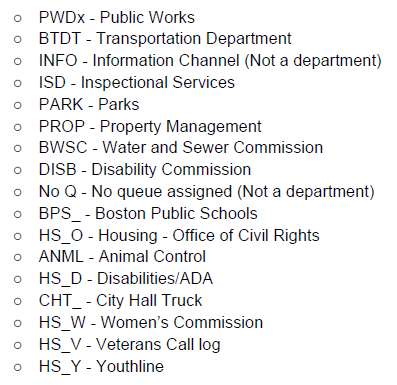
**Figure 4A: Unique Counts and Missing Value Report.**



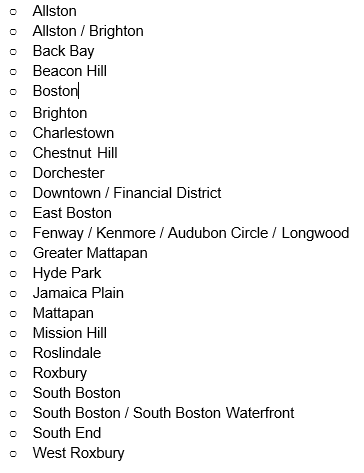
**Figure 5A: Description of the two Boston Divisions.**



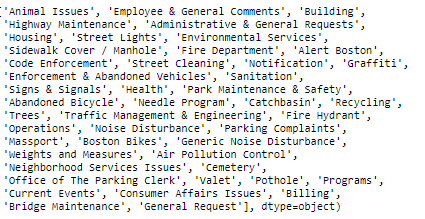
**Figure 6A: Description of the Department Names.**



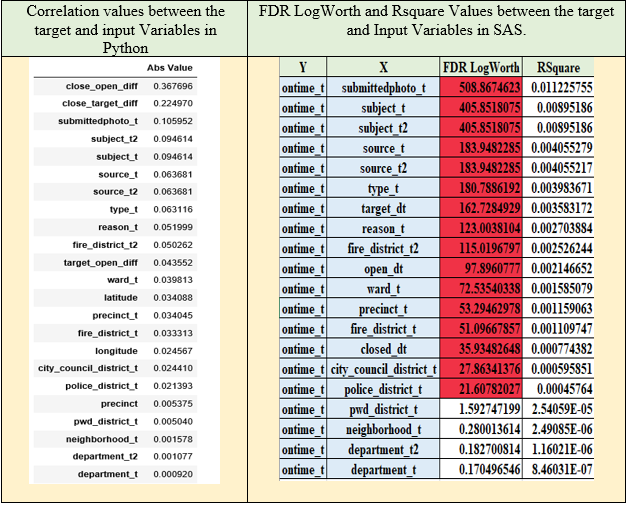
**Figure 7A: Description of the Boston Neighborhoods.**



**Figure 8A: List of all 46 Reasons assigned to the 311 Calls.**

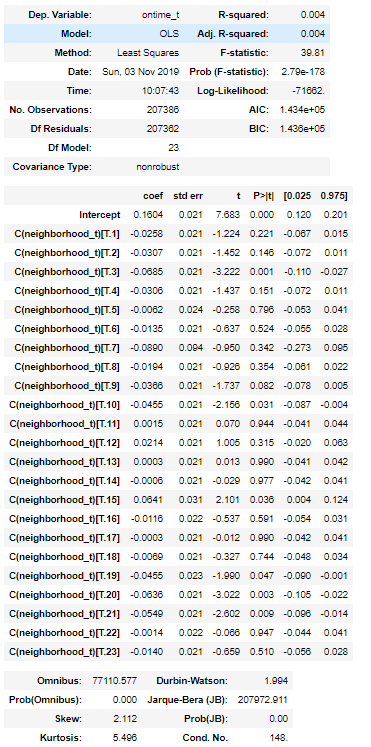


**Figure 9A: Correlation values, Logworth and Rsquare Values.**



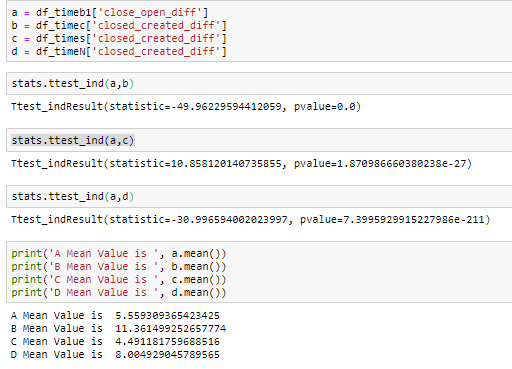
**Figure 10A: OLS Regression Model for**

**neighborhood\_t.**

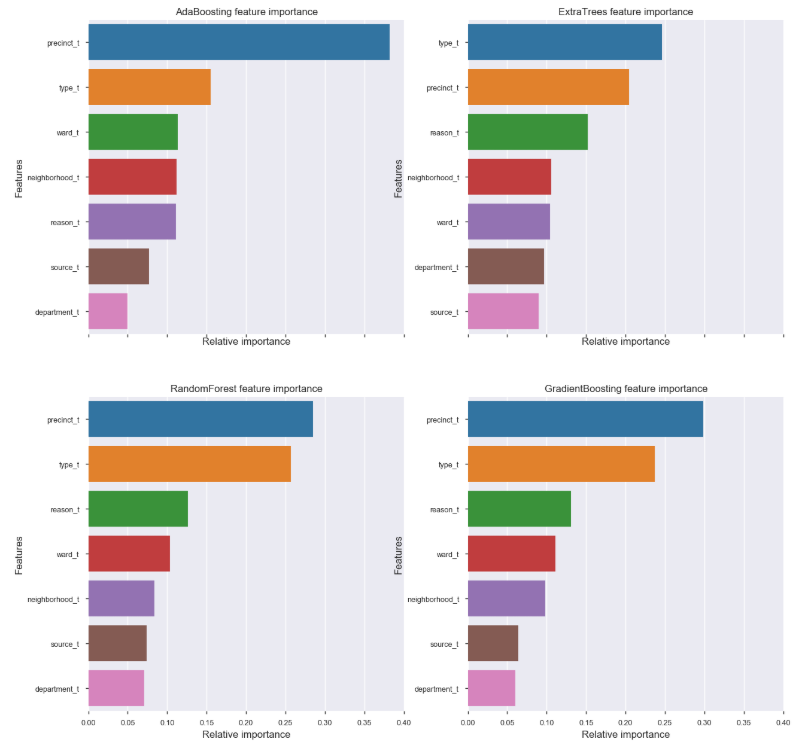


**Figure 11A: Basic Test Results for Completion Time (a=Boston, b =Chicago,**

**c=San Francisco and d = New York City).**



**Figure 12A: Feature Importance for Ada boost, Extra Trees, RF and Gradient Boosting.**



**Figure 13A: Smote Model Results with the same parameters as the**

**Random Forest Model.**

