Springboard Capstone Project 1:

Predicting Building Permit Issue Time

By David Tse

Executive Summary:

Building permits are oftentimes a long and less predictable item in a construction project schedule. A data-driven prediction of the issue time would benefit real estate developers, homeowners and building permit officials by ultimately reducing financial risk. To provide some guidance on building permit issue times for these permit applications, I used classification algorithms to classify permit issue times as short, medium or long time duration. I found that location, time and work/permit type are some of the most important features in predicting building permit issue times. Decision trees or random forests may be used to predict building permit issue time durations (medium or long duration) for non-trivial work items based on a 70% AUC metric; non-trivial items may include new buildings and major alterations that will change the use, egress, or occupancy of the building. For minor work (e.g. electrical work and demolition) involving single building departments, a heuristic may be used where issue times of less than a month may be expected.

1. Problem Statement

What is the function that best estimates whether a building permit will be more than several months late? Building permits are often times the longest item in a construction project schedule; e.g. 2-3 years for a newly constructed low-rise building. Cities review applications on a first-come, first-serve basis and must check relevant plans to ensure they meet local building codes. A building permit is an official approval issued by the local governmental agency that allows you or your contractor to proceed with a construction or remodeling project on your property. Those who would like to construct, enlarge, alter, repair, move, remove, improve, convert, or demolish a building or other structure typically apply for a building permit. In this problem, the issue time is the dependent variable, which is the difference between the issuance date and filing date in days. The issue time may also be grouped into a binary classes or multiple classes as needed. Some important use cases for this problem may include:

- **Real estate developers** may use the prediction to better optimize their project portfolio to consider reduced project timelines, which should improve their bottom line.
- **Homeowners**, especially those inexperienced in the building permitting process, will obtain more accurate time estimates for permitting and reduce reliance on anecdotal advice from building contractors/building department.
- **Building departments** could better triage requests to potentially reduce response turnaround time.

2. Data Wrangling

a. Cleaning Steps

The data was downloaded as of May 9th, 2018 from NYC Open Data. The .csv data contains a list of permits issued for a particular day and associated data. Prior weekly and monthly reports are archived at DOB and are not available on NYC Open Data. The raw building permit data contains 60 unique columns and nearly 3.37 million rows of data. A separate .csv file provides additional detailed descriptions about the data.

The data was then imported into a Pandas DataFrame for ease of data manipulations. Feature names were adjusted to be short yet meaningful, free of spaces via replacement using underscores and converted to lowercase. Twenty features out of the original 60 were kept for use in the EDA and machine learning stages to keep only relevant features in the scope of the study. Feature relevance was based on visual inspection of the records and partially due to the proportion of null/empty records. The raw data contained building permits issued dating back to the 1980s, however I decided to only use data from the past five years because it is likely to be more relevant from a prediction standpoint; processes have likely changed to reduce building permit issue times. However, I may use permits issued up to 10 years ago if additional data is required to correct class imbalance issues at the machine learning stage.

https://data.cityofnewyork.us/Housing-Development/DOB-Permit-Issuance/ipu4-2q9a

b. Missing Values

Missing values were primarily due to whether the feature is required or optional for a building permit. Even if the data is required, sometimes null values were still present as a small proportion of all of the records. The standard practice is to mark missing values as NaN. Values with a NaN value are ignored from operations like sum, count, etc.

After marking these records as NaN, I decided that it was acceptable to remove records where the proportion of nulls to total records is extremely small, e.g. the nulls in the latitude feature were <1% of the total number of records. For the remaining categorical features, dummy variables of 0 or 1 will be generated in the machine learning stage to place NaN values into a separate column.

c. Outliers

Outliers may be due to measurement/input error, data corruption, or perhaps actually represent legitimate behavior. As a simple way to determine outliers, a record is considered an outlier if it is 1.5 multiple of the interquartile range. The dependent variable, issue time (days), is heavily right skewed based on its histogram. From the histogram, I observed negative values for the issue time, which I subsequently removed from the analysis because of suspected input error.

On the other hand, outliers are present for permits requiring multiple years to be issued, and I suspect these represent legitimate records. For example, new buildings, especially high-rise and skyscrapers, typically require a much longer time to obtain a building permit compared to an average project. From the data, I notice that in the last five years some of these projects needed up to four years to obtain a building permit. One approach is to simply trim the bottom 1-5% of data, but there may be value in understanding the factors that lead to multi-year building permit issue times. However, it would be a good idea to trim permits with less than a month to issue. Permits with issue times less than a month are primarily from "quick-fix" projects, e.g. minor electrical work, which are generally more predictable and likely don't provide much value to the client if included in this study.

3. Exploratory Data Analysis

a. Initial Trends and Questions Explored

I initially explored the categorical data for any potential outliers and trends by creating and answering questions, including:

- **What are the most frequent permit types submitted and are there outliers?** From the subplots below, electrical work is the most popular permit type. Permit types for DM (Demolition and Removal) and FO (Foundations) may also be considered outliers. They are considered outliers because their share of the data is <= 5% of the category with the max count.
- **Are most permits issued?** Yes, 3287778 out of 3349782 permits are issued, or roughly 98%. Permit statuses that are in-process and re-issued may be considered outliers.
- **Are most submittals first-time or renewals?** Initial submittals are more than two times than higher than renewals. Submittals are comprised of 73% initial submittals.
- **How do the boroughs rank in building permit frequency?** 1) Manhattan, 2) Brooklyn, 3) Queens, 4) Bronx, 5) Staten Island. Based on this ranking, it'd be interesting to further explore geographic hotspots for building permits and issue times.

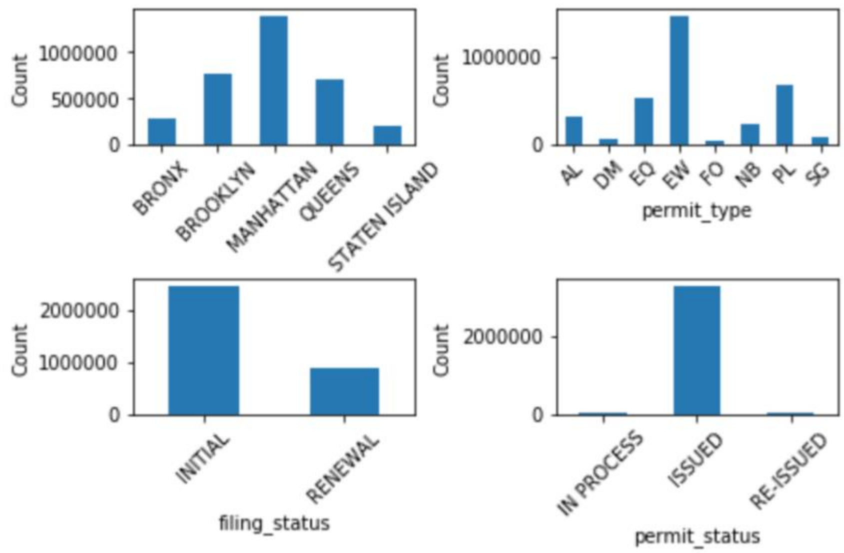

Figure 1. Categorical Features Bar Graphs

Furthermore, I explored numerical features to answer the questions below. I've also plotted an overlapping time series to inspect any anomalies in the frequency of building permits over the years and identify trends.

- **Are most filed permits issued as well?** The issuance and filing are practically following the same trend showing that almost all of the building permits that are filed are also issued.
- **How has building permits issuance frequency changed over time?** From visual inspection I notice that since the 2008 U.S. recession, building permits have been relatively chaotic in their issuance and filing, but generally on a slow monotonically increasing trend. This trend makes sense because the building industry is a notoriously cyclical industry that seems to have a strong correlation with the U.S. economy. I may explore how the recession impacted permit issuance time; I hypothesize that building permit issue times are lengthier due to the economic slowdown.

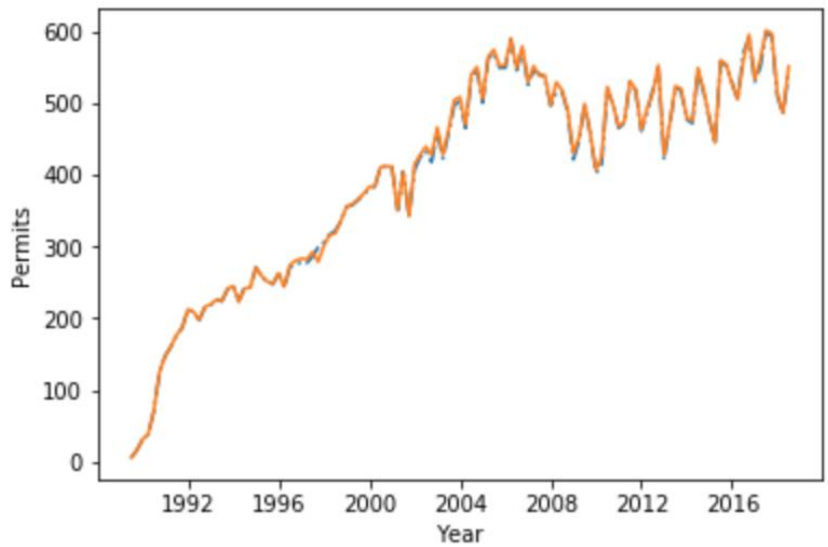

Figure 2. Issued and Filed Permits Time Series

When I initially explored the issue time feature, I found that a majority of the permits were issued on the same-day, and that the data was highly skewed to the right. I decided to omit same-day issue times because I do not believe there is much value in knowing if they could get a permit the same-day compared to permits that take months if not years. I then broke up the issue time data into four discrete time ranges to plot as histograms: 1) one week, 2) one week to one month, 3) one to six months, and 4) over six months. Based on the visualization, I asked and answered the following questions:

- **Is there enough data to continue EDA after filtering the data?** Yes, there are thousands of records still available even after filtering
- **Is the data still skewed to the right and why?** Yes, and this is likely because buildings that require longer issuance times generally require a more careful inspection of plan sets. I hypothesize that the data follows an exponential distribution which is the time taken between two events occuring (filing time to issuing time).

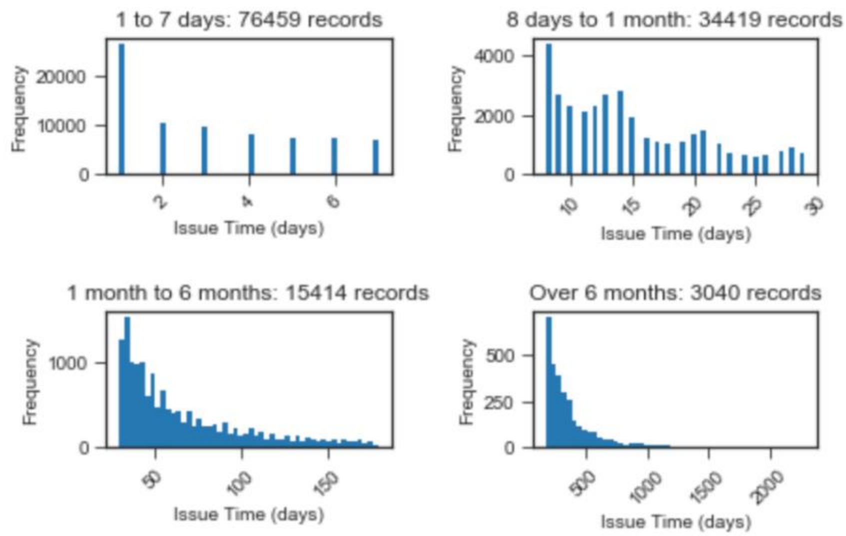

Figure 3. Issue Time Divided into Four Time Ranges

b. Statistical Analyses

I started with exploring the summary statistics and plotting boxplots, bar graphs, and empirical CDFs to better understand the data. In the figures below, permit type in particular was of particular interest since building permit applicants typically use this feature as a way to intuitively approximate issue times. The box plots and CDFs show a high skewness to the right and several high leverage points, which would make it difficult to satisfy regression assumptions; classification-based ML algorithms should be used instead. Furthermore, a two and three class feature was feature engineered based on the continuous dependent variable, issue time. Another key finding was that the wait time is on average four months and that the distribution of issue times is highly variable with a standard deviation to mean, or coefficient of variation, greater than 1.0.

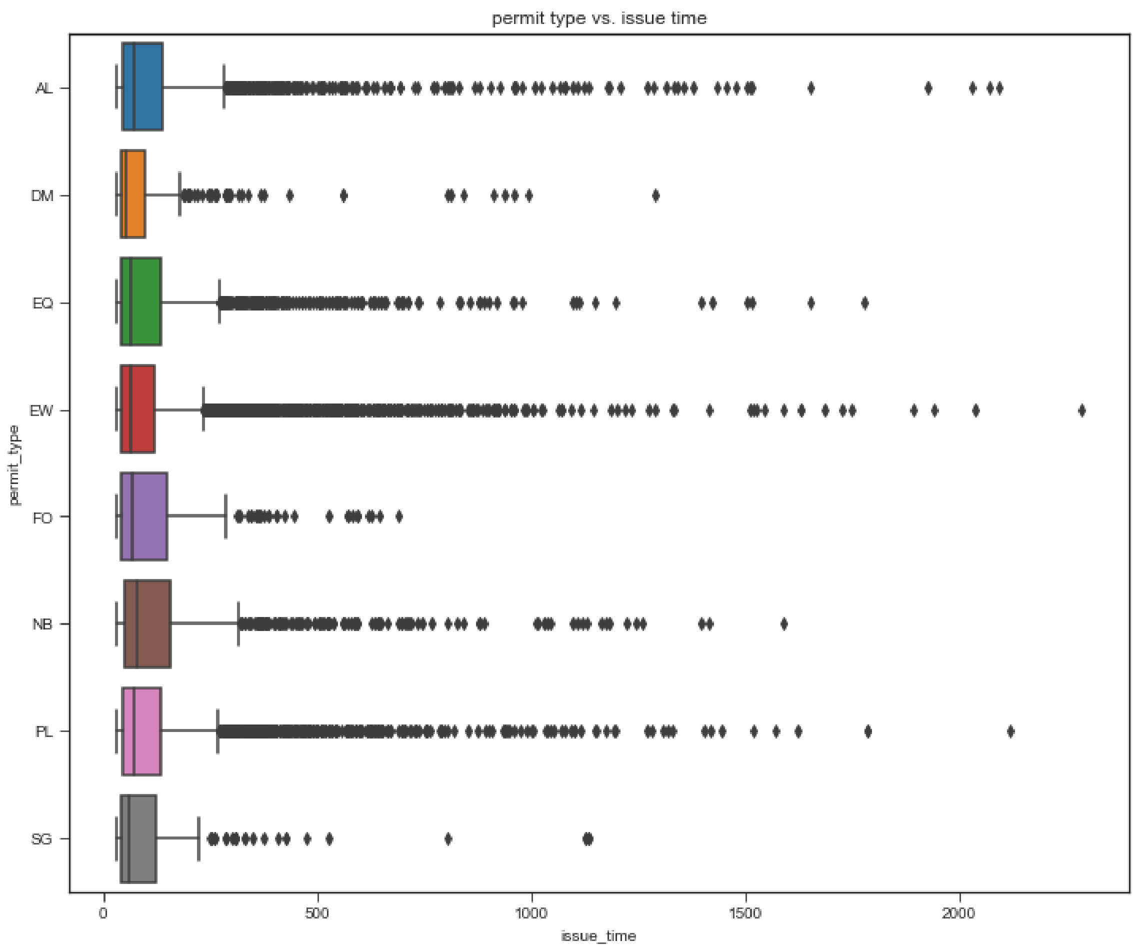

Figure 4. Permit Type Box and Whisker Plots

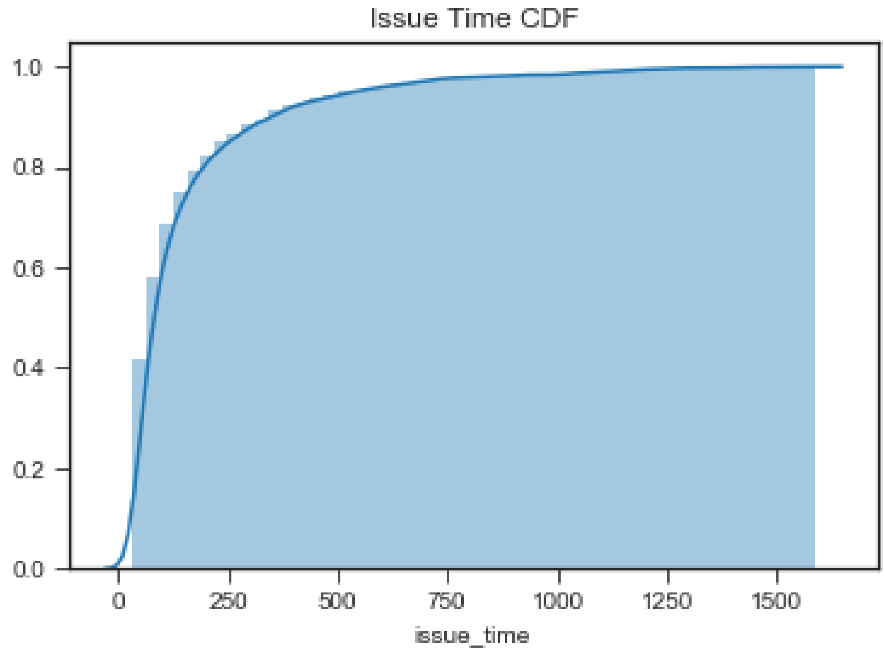

Figure 5. “New Building” Permit Type Empirical CDF

Next, I looked for strong correlations between pairs of independent variables or between an independent and a dependent variable for the numerical features. Plotting a correlation coefficient heat map, we can see that the most negative correlation between council district and latitude. The most positive correlation is between the census tract and the longitude. I found that location-based features, e.g. Council district and longitude, seem to have a statistically significant correlation with the issue time. I used a Spearman rank correlation coefficient instead of the Pearson rank correlation coefficient. A Pearson’s correlation works well if the variables are roughly normal and outliers are not present. It is a measure of the linear relationship between variables. Spearman’s rank correlation, however, is a better alternative that mitigates the effect of outliers and skewed distributions.

- longitude: test-statistic = 0.023   p-value = 0.0026
- council_district: test-statistic = 0.031   p-value = 0.0000
- census_tract: test-statistic = 0.020   p-value = 0.0083

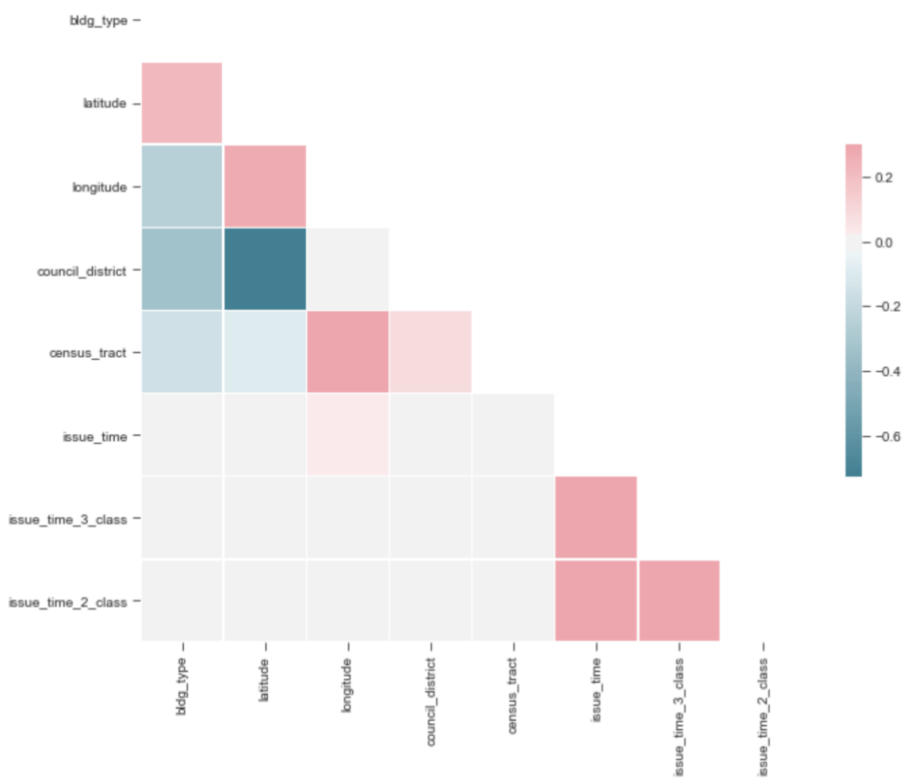

Figure 6. Correlation Matrix

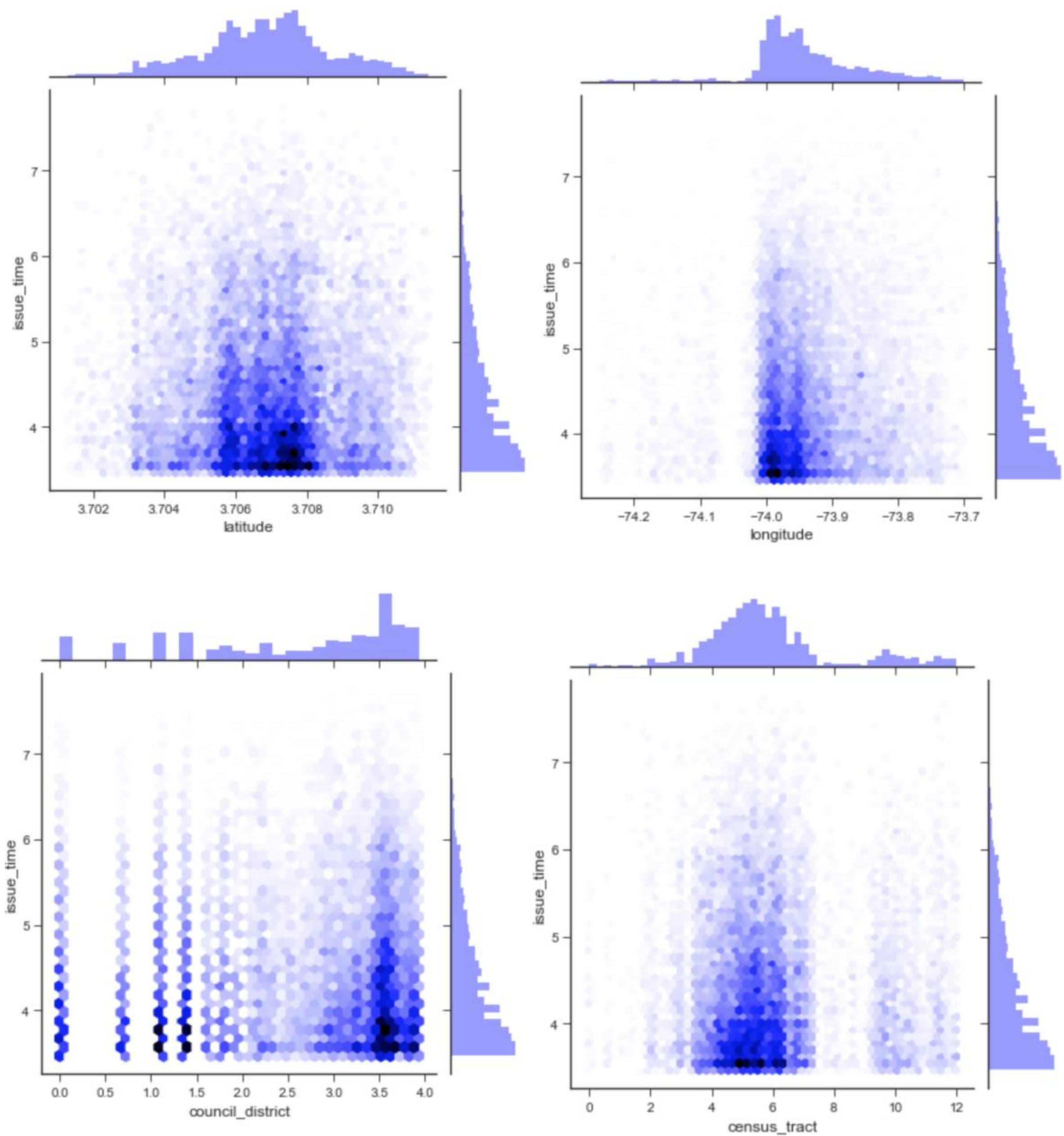

Figure 7. Hex-binned Scatter Plots

I also performed hypothesis testing to test a claim that concerning average wait time for building permits. Quoting a [building permit expediting services]() company, the company claimed that "complex projects," i.e. new building construction, could take 6 months or more to obtain an issued permit. Using a left-tailed Z-Test and bootstrapping methods, I found that building permits for "complex projects" actually require less than 6 months on average based on a sample size of 1109 and p-value of 0.000.

To gain insight into the neighborhood hotspots for building permits, I used a heat map visualization based on the Google Maps API. These neighborhoods seemed to coincide with regions of high economic activity based on my own domain knowledge. Economic data could be a great addition to this project and may help with the prediction of building permit issue times. The heat map below shows that the following neighborhoods within each borough have active building permit applications.

- Brooklyn: Park Slope, Bushwick, Brighton Beach, Williamsburg
- Manhattan: Majority of neighborhoods
- Queens: Flushing, Jackson Heights, Elmhurst, Astoria, College Point
- Bronx: N/A
- Staten Island: N/A

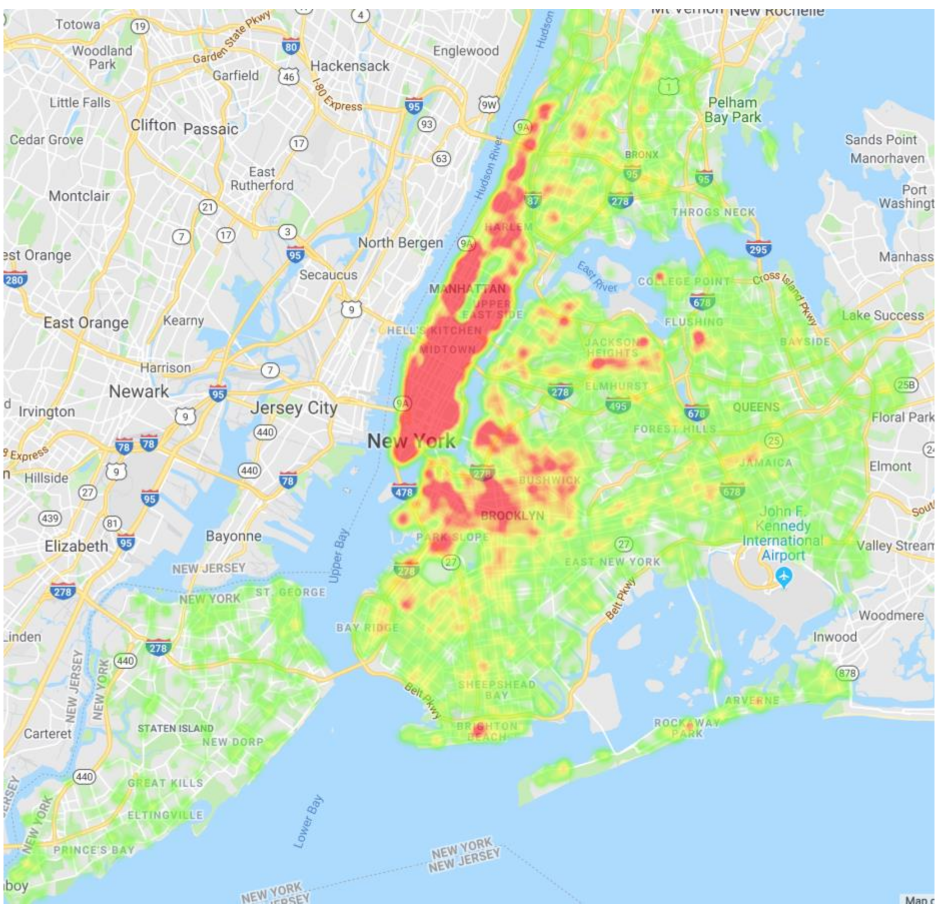

Figure 8. Building Permit Heatmap

4. Framing the Machine Learning Problem

Supervised classification algorithms were used to classify the building permit issue times into several time ranges. A classification approach was used because regression-based approaches are sensitive to outliers, which are present in the majority of independent variables: extensive cleaning and transformations would be necessary to achieve a reasonable and reliable result.

The number of classes could also largely affect model performance due to class imbalance. A binary class and three-class outcome were first considered for this project, but a binary class was preferable starting point. The class time ranges were chosen to reduce class imbalance issues:

- Binary outcome:
  - class 0: 1 month to 3 months
  - class 1: > 3 months
- Multi-class outcome:
  - class 0: 1 month to 3 months
  - class 1: 3 months to 6 months
  - class 2: >6 months

Some classification algorithms to be explored in this project include:

- Logistic Regression (LR)
- K-Nearest Neighbors (KNN)
- Classification and Regression Trees (CART)
- Support Vector Machine (SVM)
- Random Forest (RF)
- Naive Bayes (NB)
- Gradient Boosted Trees (XGBoost)

a. Additional Data Transformations

The filing date and the job start date were feature engineered to capture information about the month and day of week. This may be important because construction is typically more active during warmer months when the weather allows for labor crews to work effectively. Features were then converted to appropriate data types to reduce computation time and allow for calculations to be performed on these features.

Nulls are not acceptable inputs in some machine learning algorithms. Dummy variables were created for the categorical features, even the null values. Missing/null values for the continuous variables (latitude and longitude) were imputed with the most frequent values since only 23 out of 17589 records were missing.

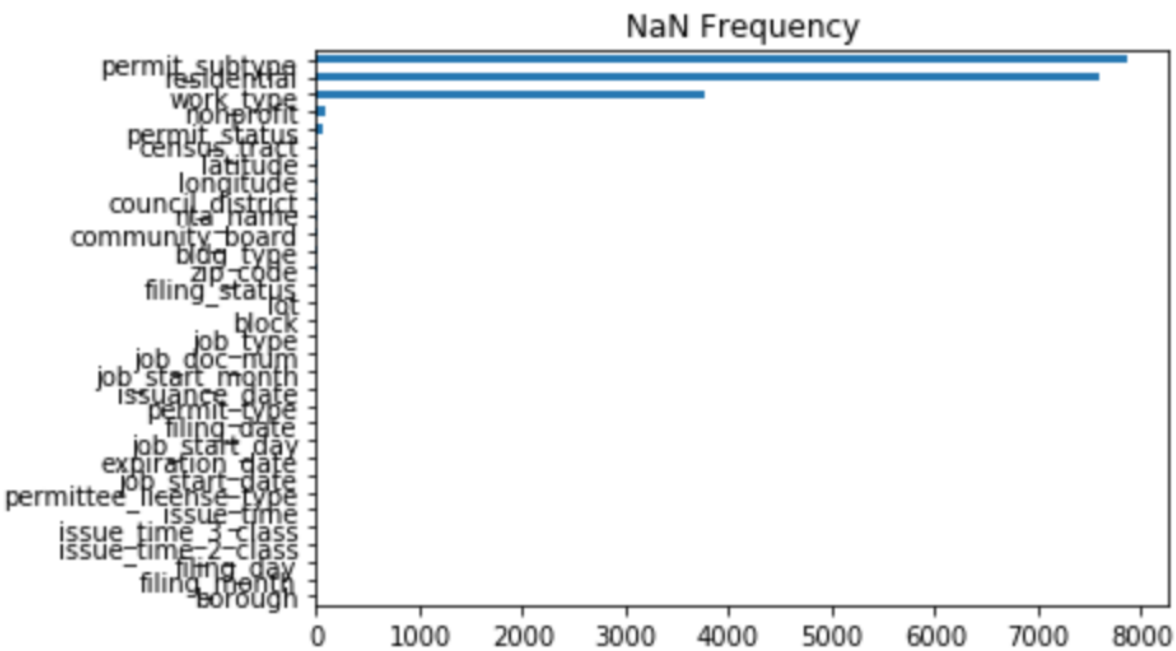

Figure 9. Null Frequencies for Each Feature

Features were then scaled using the standard scale, which results in values between -1 and 1 to ensure features are not artificially adjust the importance of a particular feature.

a. Compare Machine Learning Models

Before performing algorithm optimizations, a quick comparison of multiple algorithms using an accuracy metric was completed to understand the range in algorithm performance. The analysis showed that Random Forest (67%), CART (68%) and SVMs (66%) have the highest median accuracy using 10-fold cross-validation. A dummy classifier was used as a benchmark for comparison, which resulted in a median of 53% accuracy.

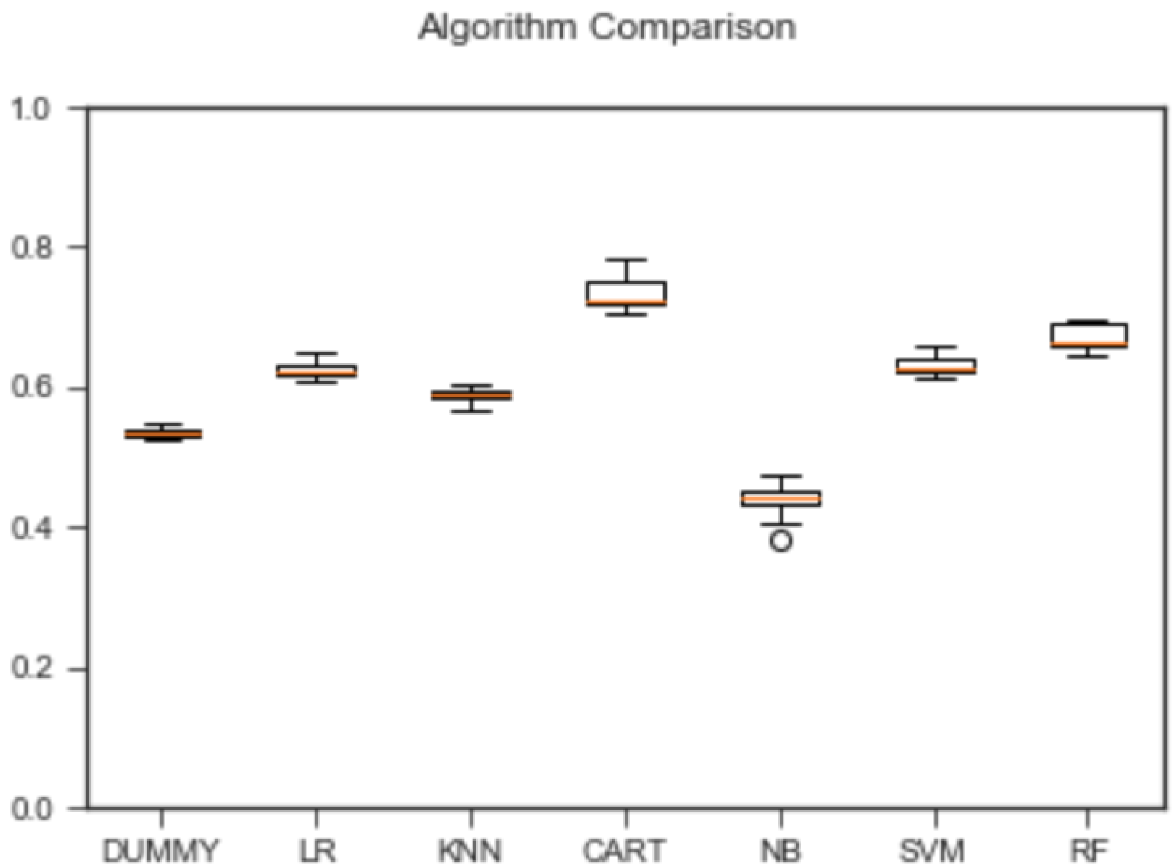

Figure 10. Machine Learning Algorithm Comparison Using the Accuracy Metric

b. Feature Importance

The building permit dataset is abundant in the number of features to explore with the original dataset contained 60 features. Honing in on the most important features would be useful in three ways:

- Reduces Overfitting: Less redundant data means less opportunity to make decisions based on noise.
- Improves Accuracy: Less misleading data means modeling accuracy improves.
- Reduces Training Time: Less data means that algorithms train faster.

The most important features seem to be heavily influenced by location (e.g. borough and zip codes), time related features (e.g. job start month and filing day) and the type of work listed on the permit (e.g. job type and work type). After several iterations, 17 features were kept and when the data is transformed with dummy variables (0 or 1 for categorical features), I ended up with 354 features.

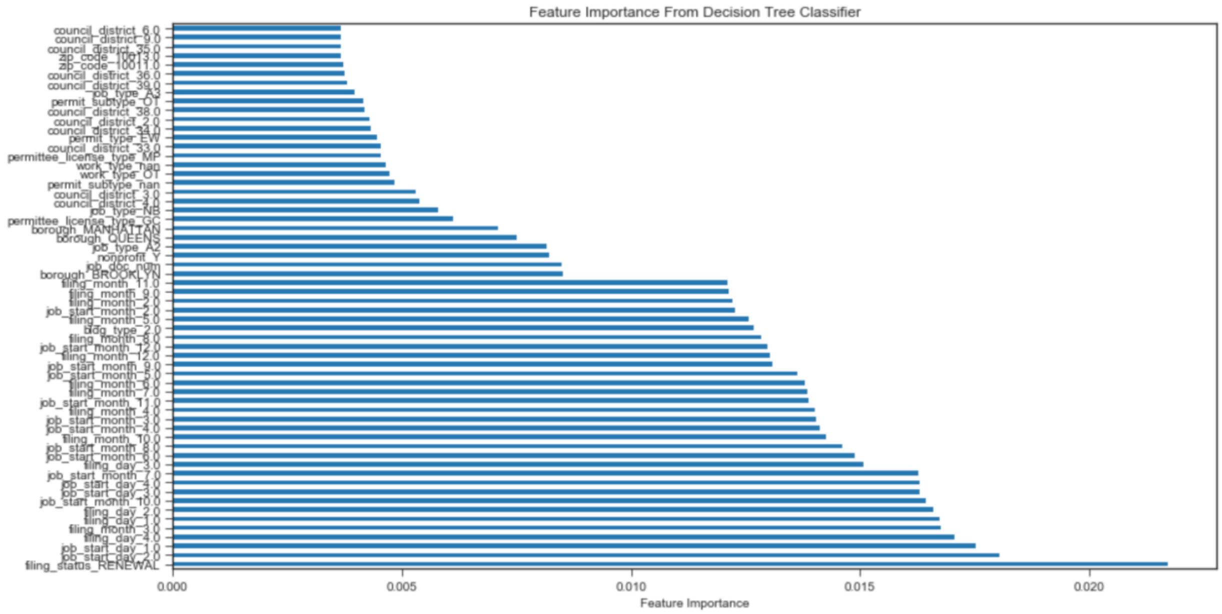

Figure 11. Top 25 Important Features

c. Class Imbalance

The classes were purposefully chosen to reduce the effects of class imbalance. Reduction of class imbalance avoids the "accuracy paradox" where a high accuracy may be achieved with a dataset suffering from class imbalance; a close inspection of the predictions show hidden issues such as low precision and recall in the minority class.

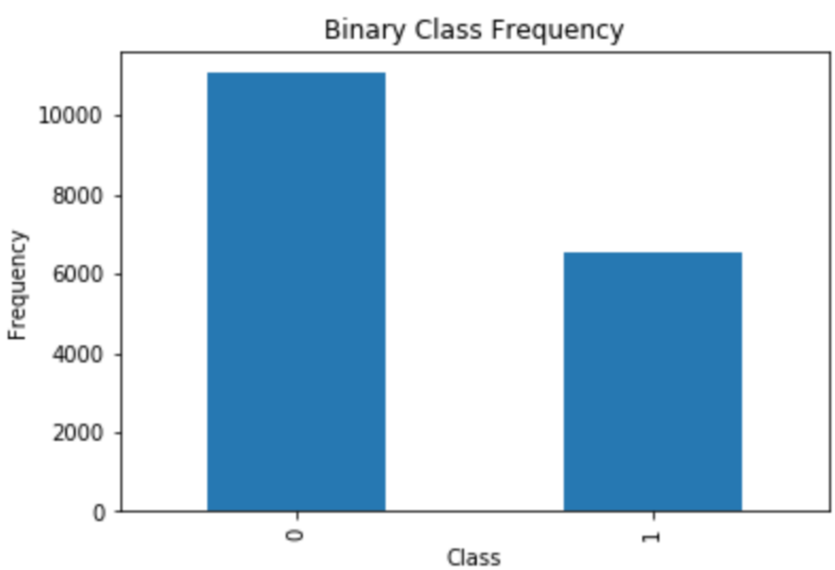

Figure 12. Binary Class Output: Comparison of Record Frequency

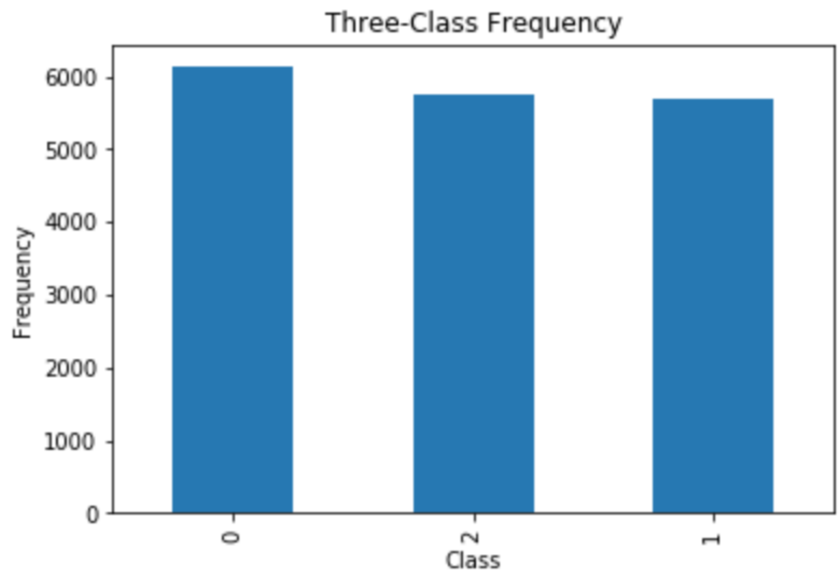

Figure 13. Three-Class Output: Comparison of Record Frequency

d. Model Enhancements

Based on the initial non-optimized runs of several kinds of algorithms earlier, I focused on improving tree-based classifiers as they seem to have the best initial accuracy metrics. Five-fold cross-validation was used for the model enhancement studies.

Logistic regression would be a dubious choice of algorithm since there is a chance that records may not be independent since applicants may re-submit applications at a later date. Outliers are also present in the data and would require extensive pre-processing to remove in the time span of this project. Lastly, although a binary output was feature engineered, the problem could be framed as a multi-class problem with more than two classes, so it would be a good idea to leave that option available.

KNN and Naive Bayes did not have an acceptable improvement in accuracy compared to the dummy classifier so these were not considered further.

Instead of just relying on an accuracy metric, some other classification metrics that were explored include:

- **Classification Accuracy**: Number of correct predictions made as a ratio of all predictions made
    - Note: Suitable when there are an equal number of observations in each class. Predictions and prediction errors are equally important
- **Logarithmic Loss**: Evaluates the predictions of probabilities of membership to a given class. The scalar probability between 0 and 1 can be seen as a measure of confidence for a prediction by an algorithm. Predictions that are correct or incorrect are rewarded or punished proportionally to the confidence of the prediction.
    - Note: Smaller logloss is better with 0 representing a perfect logloss.
- **Area Under ROC Curve**: Evaluates a model's ability to discriminate between positive and negative classes
    - Note: Used for binary classification problems. 1.0 = Perfect, 0.5 = Random
- **Confusion Matrix**: Presents the accuracy of a model with two or more classes
- **Classification Report**: Displays the precision, recall, F1-score and support for each class

Starting with a Dummy Classifier, we may use this classifier as a benchmark for other supervised classification algorithms.

```
accuracy : 0.534 (0.011)
Classification Report:
             precision    recall  f1-score   support

          0       0.62      0.62      0.62      2730
          1       0.39      0.38      0.39      1668

avg / total       0.53      0.53      0.53      4398
```

Figure 14. Dummy Classifier: Accuracy and Classification Report

The Receiver Operator Characteristic (ROC) curve shows an almost 45 degree line, which is to be expected when a classifier is as good as flipping a coin.

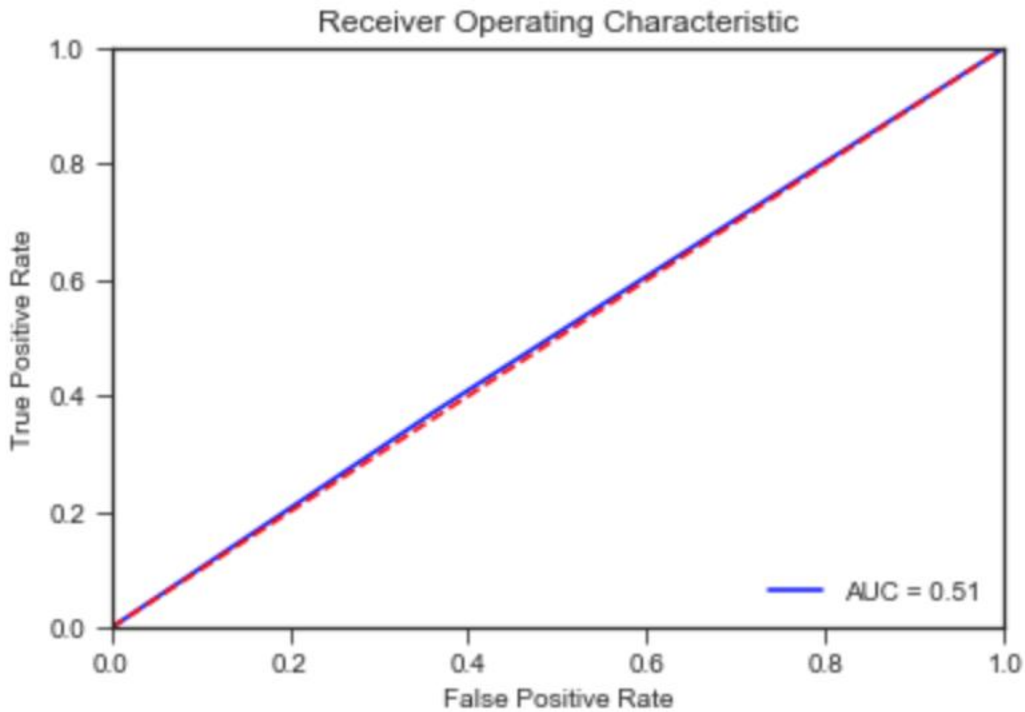

Figure 15. Dummy Classifier: ROC Curve

Decision Trees are a non-parametric supervised learning method used for classification and regression. Their key advantage is that they can learn non-linear relationships, and are fairly robust to outliers. The issue is that unconstrained, individual trees are prone to overfitting because they can keep branching until a full tree is developed.

The number of estimators was tuned, and the best model had a max depth of 75 and the minimum of 4 samples per leaf node to achieve the results below:

```
accuracy : 0.733 (0.024)
Classification Report:
             precision    recall  f1-score   support

          0       0.76      0.83      0.79      2730
          1       0.67      0.58      0.62      1668

avg / total       0.73      0.73      0.73      4398
```

Figure 9. Decision Tree: Accuracy and Classification Report

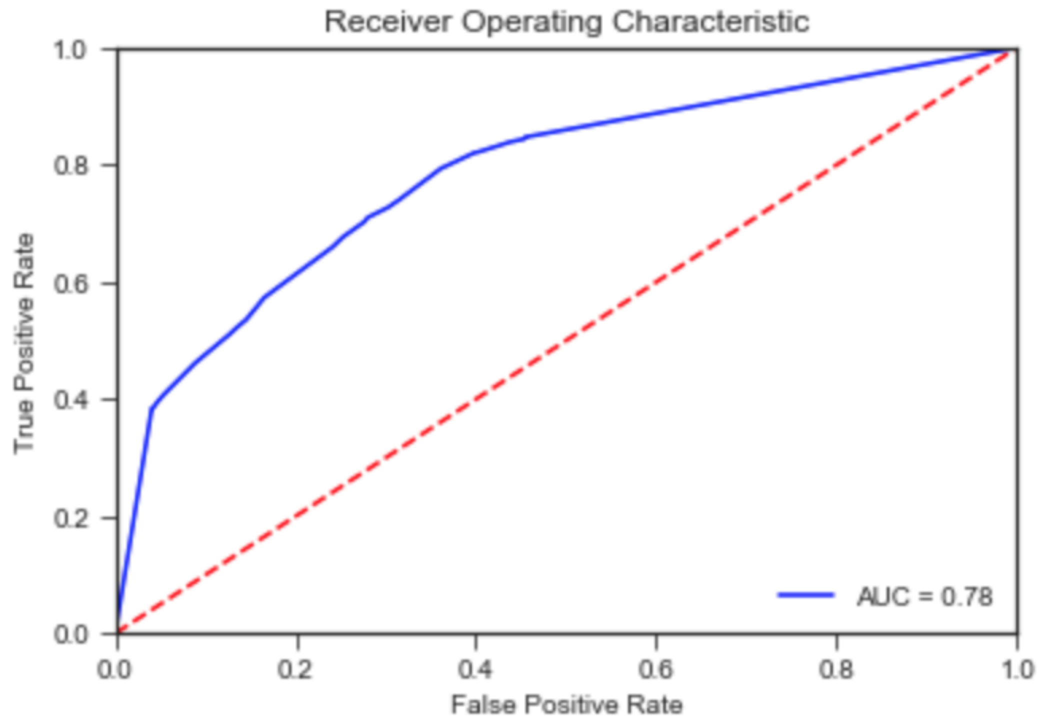

Figure 16. Decision Tree Classifier: ROC Curve

Random forest classifiers are an ensemble of decision tree classifiers. They have the key advantages of computational efficiency and an ability to handle high dimensions well. Because randomness is introduced in the selection of records and features, the random forest reduces overfitting, which decision trees typically suffer from.

The number of estimators was tuned, and the best model had 150 estimators, a minimum of 1 sample per leaf node and a max tree depth of 150 to achieve the results below:

```
accuracy : 0.692 (0.012)
Classification Report:
             precision    recall  f1-score   support

          0       0.70      0.92      0.79      2730
          1       0.73      0.35      0.47      1668

avg / total       0.71      0.70      0.67      4398
```

Figure 17. Random Forest: Accuracy and Classification Report

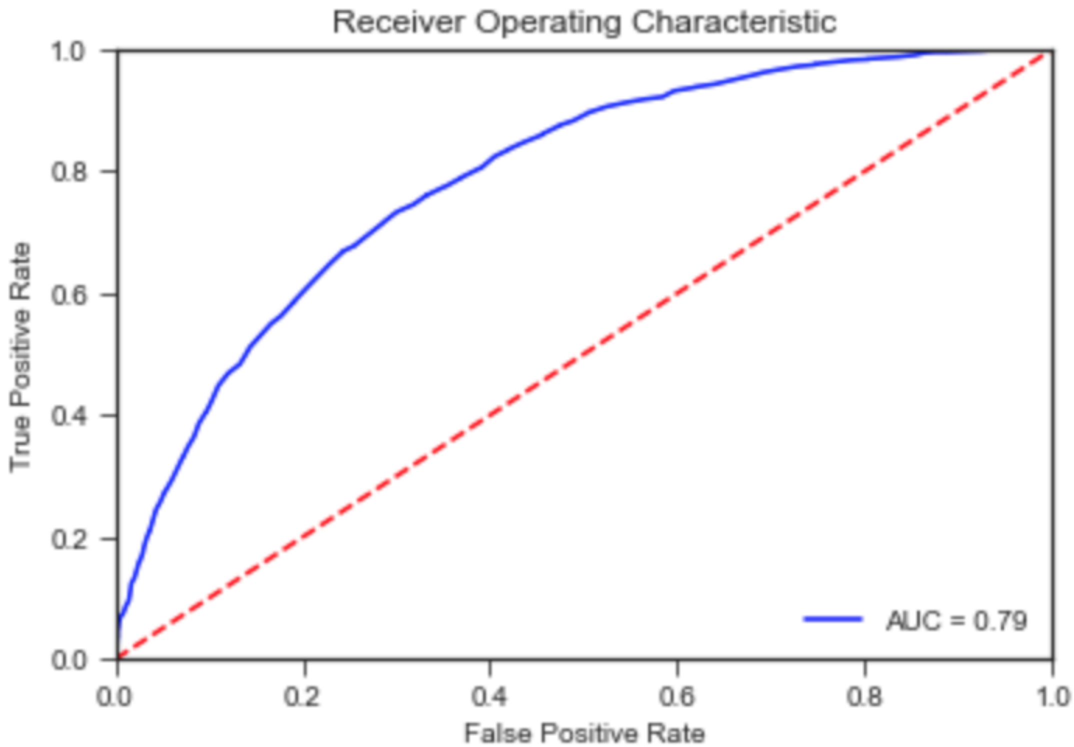

Figure 18. Random Forest Classifier: ROC Curve

Gradient Boosted Trees are sequentially grown trees that aim to improve predictions from previously grown trees. Because the growth of a particular tree takes into account the other trees that have already been grown, smaller trees are typically sufficient. Using smaller trees can aid in interpretability as well since a decision stump may be used; a decision stump uses a single decision rule to split the data. Random Forest on the other hand typically does not use decision stumps, but uses rather complex trees.

The number of estimators was tuned, and the best model had 200 estimators to achieve the results below:

```
accuracy : 0.643 (0.017)
Classification Report:
             precision    recall  f1-score   support

          0       0.64      0.96      0.77      2730
          1       0.62      0.12      0.20      1668

avg / total       0.63      0.64      0.55      4398
```

Figure 19. Gradient Boosted Trees: Accuracy and Classification Report

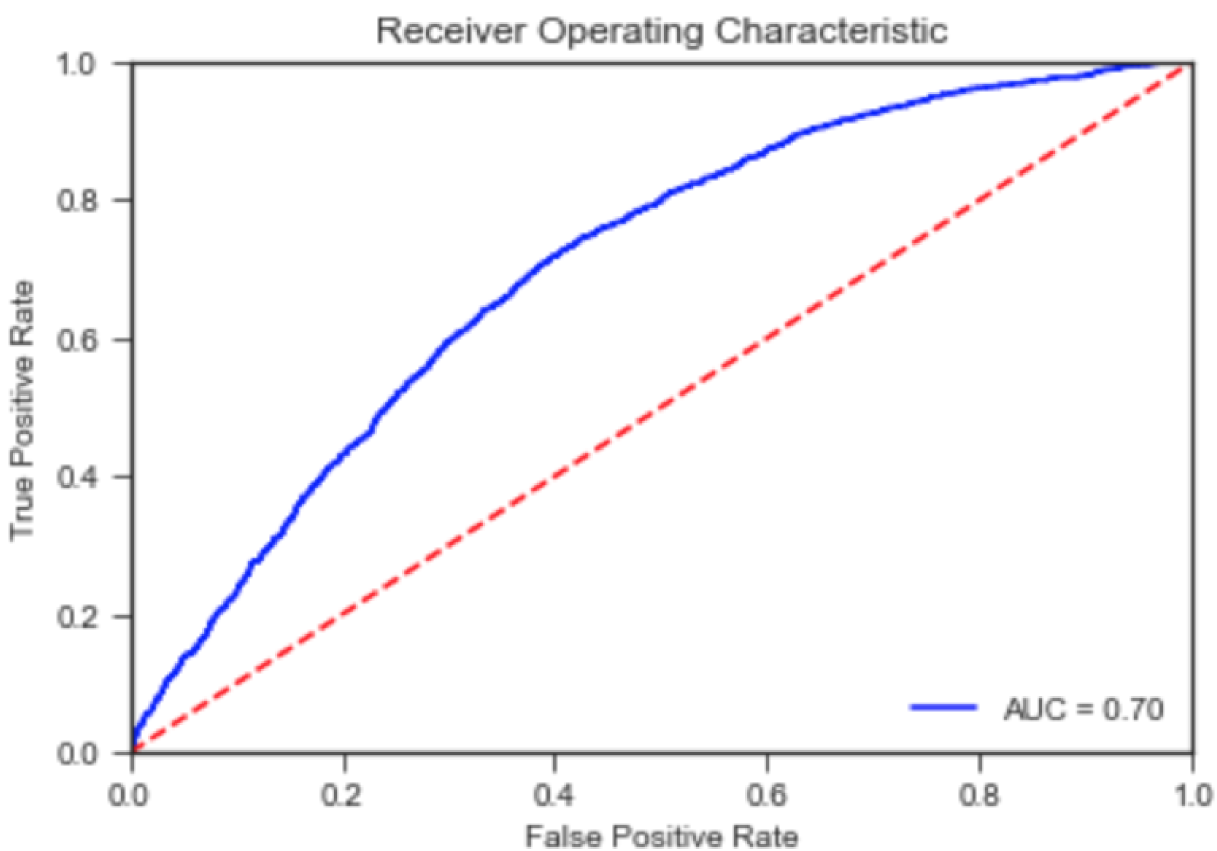

Figure 20. Gradient Boosted Trees: ROC Curve

I've decided to use the F1 score and AUC of the ROC curve as the classification metrics to consider. The F1 score ensures that each class has a balanced precision and recall. The AUC is insensitive to data sets with unbalanced class proportions and end users may want to adjust the decision thresholds based on their business needs. From the analysis, the Random Forest algorithm achieved the highest AUC of 0.79, but the decision trees had the highest F1-score of 0.73.

The business problem requires a mixture of interpretability and AUC since the end user may be a real estate company, homeowner or the NYC building department. For example, homeowners would probably just want a quick and highly accurate answer for the few properties they own. However, a real estate company may need interpretability slightly more than high AUC so they could better allocate resources at a more granular level for multiple properties.

Two models may be considered. Decision trees offer both interpretability and an acceptable AUC and would be best suited for real estate developers. However, random forests have the highest AUC in the analysis and tend to overcome overfitting at the cost of lower interpretability, which would be more suited for homeowners.

e. Limitations

- The building permit data did not include the cost of the proposed work. It may be the case that proposed work with large economic value to the city may get pushed ahead of the queue.
- Building permit expeditors typically reduce turnaround time for permit applications, which may introduce bias into the dataset.

f. Future Work

- Investigate framing the problem into a multi-class problem.
- Try incorporating other related datasets that may boost the signal, such as population and household income datasets to better understand why location is a strong predictor.

g. Major Findings and Client Recommendations

Real Estate Developers/ Homeowners:

1. Location, time and work/permit type are some of the most important features in building issue times.
2. Use decision trees (real estate developer use case) or random forests (homeowner use case) to predict building permit issue time durations (medium or long duration) for non-

trivial work items, e.g. new buildings and major alterations that will change the use, egress, or occupancy of the building.

3. For minor work, such as electrical work and demolition, involving single building departments expect issue times of less than a month (short duration).

NYC Building Department:

Minimal information is provided about issue times on the NYC building department website. One solution is to provide a web service that allows applicants to take an online survey, which provides a building permit issue time range based on what was checked on the survey. The issue time range could be based on statistics from the past five years to ensure it is robust.

h. Acknowledgements

I would like to thank Springboard, especially my mentor Scott Clendaniel for the advice and support throughout the capstone project.

i. Appendix

Data Dictionary:

Column Name	Column Description	Additional Notes (where applicable, include the range of possible values, units of measure, how to interpret null/zero values, whether there are specific relationships between columns, and information on column source)	Data Type (Plain Text, Number, Date)	Required	Cardinality	Used for Dependent Variable	Reason for Removing Column	Transformation
BOROUGH	The name of the NYC borough where the proposed work will take place.	Expected values = Manhattan, Bronx, Brooklyn, Queens, Staten Island.	Text	Yes	5	No		Category
Bin #	Building Identification Number assigned by Department of City Planning.		Number	Yes	381121	No	High cardinality, specific to application	
House #	The house number for the building where the proposed work will take place.	In the address, "280 Broadway", 280 is the house number. This number may contain a dash (-), as in "8-15 27 Avenue".	Text		32555	No	Use lattitude and longitude instead	
Street Name	The street name for the building where the proposed work will take place.	In the address, "1050 Park Place", Park Place is the street name.	Text	Yes	26368	No	Use lattitude and longitude instead	
Job #	The DOB Job Application Number assigned when the applicant begins the application.	This is the unique identifier for the application submitted to the Department. It may contain several work types, and more work types may be added as the application review and the work continues. It is a 9-digit number where the first digit indicates the borough where the building is located. • 1 = Manhattan• 2 = Bronx• 3 = Brooklyn• 4 = Queens• 5 = Staten Island	Number	Yes	1481882	No	High cardinality, specific to application	
Job doc. #	A sequential number assigned to each of the documents that make up a job application.	Every job application should have a 01 document. Every additional document receives a number that increases by 1 (ex: 02, 03, 04)	Text	Yes	12	No		Category
Job Type	2-digit code to indicate the overall job type for the application.Note: You should also look at the Permit Type field to find out what this specific permit has been issued for. An NB job, for example, can have several different Work/Permit Types such as PL, SP, EQ, etc. And each Work/Permit Type will be issued a seperate permit.	Expected values are:• A1 = Alteration Type I, A major alteration that will change the use, egress, or occupancy of the building.• A2 = Alteration Type II, An application with multiple types of work that do not affect the use, egress, or occupancy of the building.• A3 = Alteration Type III, One type of minor work that doesn't affect the use, egress, or occupancy of the building.• NB = New Building, An application to build a new structure. "NB" cannot be selected if any existing building elements are to remain—for example a part ofan old foundation, a portion of a façade that will be incorporated into the construction, etc.• DM = Demolition, An application to fully or partially demolish an existing building.• SG = Sign, An application to install or remove an outdoor sign.	Text	Yes	6	No		Category
Self_Cert	Indicates whether or not the application was submitted as Professionally Certified. A Professional Engineer (PE) or Registered Architect (RA) can certify compliance with applicable laws and codes on applications filed by him/her as applicant.	Expected values are: • Y = Yes. • N = No.• blank = No information for this record.	Text	No	5	No	Optional Field, high number of missing values	
Block	Tax Block assigned to the location of the proposed work by the Department of Finance.		Text	Yes	23789	No		Category
Lot	Tax Lot assigned to the location of the proposed work by the Department of Finance.		Text	Yes	3075	No		Category
Community Board	Community Board Number for the building's address.		Text	Yes	220	No		Category
Zip Code	ZIP Code for the building's address.		Text	No	229	No		Category
Bldg Type	Legal occupancy classification.	Expected values are:• 1-2-3 Family• OtherDOF classification has 36 types - this field just simplifies it for DOB purposes. Anything above 3 family we consider multiple dwelling and is regulated more heavily. Other agencies might have different cutoffs.	Text	No	2	No		Category
Residential		Expected values are: • Y = Yes. • blank = No information for this record.	Text	No	1	No		Category
Special District 1			Text	No	95	No	Optional Field, high number of missing values	
Special District 2			Text	No	8	No	Optional Field, high number of missing values	
Work Type	The specific type of work covered by the permit.	This is a two character code to indicate the type of work. You can find the plain language definition of the work type in our acronym glossary at http://www1.nyc.gov/site/buildings/about/acronym-glossary.page. If the Work Type is blank, check the Permit Type.	Text	No	13	No		Category
Permit Status	The current status of the permit application.	Expected values are IN PROCESS, ISSUED, RE-ISSUED, and REVOKED.	Text	Yes	4	No		Category
Filing Status	Indicates if this is the first time the permit is being	Expected values are INITIAL or RENEWAL.	Text	Yes	2	No		Category
Permit Type	The specific type of work covered by the permit.	This is a two character code to indicate the type of work. You can find the plain language definition of the work type in our acronym glossary at http://www1.nyc.gov/site/buildings/about/acronym-glossary.page.	Text	Yes	8	No		Category
Permit Sequence #	A sequential number assigned to each issuance of t	Every initial permit should have a 01 sequence number. Every additional renewal receives a number that increases by 1 (ex: 02, 03, 04)	Text	Yes	29	No		Category
Permit Subtype	A more specific designation for the type of work in	This is a two character code to indicate the type of work. You can find the plain language definition of the work type in our acronym glossary at http://www1.nyc.gov/site/buildings/about/acronym-glossary.page.	Text	No	15	No		Category
Oil Gas	If the permit is for work on fuel burning equipment	Expected values are OIL, GAS, or blank.	Text	No	2	No	Optional Field, high number of missing values	
Site Fill	This indicates the source of any fill dirt that will be	Expected Values are:• USE UNDER 300 CU.YD = Less than 300 cubic yards (on-site or off-site) is being used. • NOT APPLICABLE = Not applicable to proposed work.• ON-SITE = 300 or more cubic yards of fill is being used from on-site.• OFF-SITE = 300 or more cubic yards of fill will be transplanted from an off-site location.• blank	Text	No	5	No	Optional Field, high number of missing values	
Filing Date	The date the permit application was filed with DOB.		Date	Yes	8141	Yes		DateTime
Issuance Date	The date the permit was issued.		Date	Yes	8138	Yes		DateTime
Expiration Date	The date that the permit expires.		Date	Yes	10819	No		DateTime
Job Start Date	The date that the initial permit was issued for this Permit Type.		Date	Yes	9235	No		DateTime
Permittee's First Name	First name of the person that the permit was issued to.		Text	Yes	37922	No	Irrelevant	
Permittee's Last Name	Last name of the person that the permit was issued to.		Text	Yes	90212	No	Irrelevant	
Permittee's Business Name	Business name of the person that the permit was issued to.		Text	Yes	361384	No	Irrelevant	
Permittee's Phone #	Phone number of the person that the permit was issued to.		Text	Yes	171437	No	Irrelevant	
Permittee's License Type	Professional license type of the person that the per	Expected values are:DM = FS = Fire Suppression ContractorGC = General ContractorHI = MP = Master PlumberNW = OB = Oil Burner InstallerOW = PE = Professional EngineerRA = Registered ArchitectSI = Sign Hanger	Text	Yes	12	No	Irrelevant	
Permittee's License #	Professional license number of the person that the permit was issued to.		Text	No	61893	No	Irrelevant	
Act as Superintendent	Indicates if the permittee acts as the Construction S	Expected values are Y, blank.	Text	No	2	No	Optional Field, high number of missing values	
Permittee's Other Title	Another license number for the person that the permit was issued to, if any.		Text	No	3156	No	Irrelevant	
HIC License	NYC-Registered Home Improvement Contractors (HIC) need a license for all alteration work in 1-, 2-, 3- , 4-family homes or in		Text	No	6291	No	High cardinality, specific to application	
Site Safety Mgr's First Name	The Site Safety Manager's first name.	A certified Site Safety Manager is required on new construction or demolition sites 15 stories and above or 100,000 square feet or greater.	Text	No	638	No	Irrelevant	
Site Safety Mgr's Last Name	The Site Safety Manager's last name.		Text	No	1548	No	Irrelevant	
Site Safety Mgr Business Name	The Site Safety Manager's business name.		Text	No	1563	No	Irrelevant	
Superintendent First & Last Name	The Construction Superintendent's first and last na	A registered Construction Superintendent is required at new buildings and buildings under demolition nine stories and below.	Text	No	157062	No	Irrelevant	
Superintendent Business Name	The Construction Superintendent's business name.		Text	No	309495	No	Irrelevant	

Owner's Business Type	Indicates the type of entity that owns the building where the work will be performed.		Text	No	14	No	Irrelevant	
Non-Profit	Indicates if the building is owned by a non-profit.	Expected values are Y, N, blank.	Text	No	4	No		Category
Owner's Business Name	Business name for the owner of the building where the work will be performed.		Text	No	457996	No	Irrelevant	
Owner's First Name	First name of the owner of the building where the work will be performed.		Text	Yes	90606	No	Irrelevant	
Owner's Last Name	Last name of the owner of the building where the work will be performed.		Text	Yes	166843	No	Irrelevant	
Owner's House #	House number for the address of the owner of the building where the work will be performed.		Text	Yes	39578	No	Irrelevant	
Owner's House Street Name	Street for the address of the owner of the building where the work will be performed.		Text	Yes	133015	No	Irrelevant	
Owner's House City	City for the address of the owner of the building where the work will be performed.		Text	Yes	12514	No	Irrelevant	
Owner's House State	State for the address of the owner of the building where the work will be performed.		Text	No	57	No	Irrelevant	
Owner's House Zip Code	ZIP Code for the address of the owner of the building where the work will be performed.		Text	Yes	12622	No	Irrelevant	
Owner's Phone #	Phone number of the owner of the building where the work will be performed.		Text	Yes	391474	No	Irrelevant	
DOBRunDate	Date when query is run and pushed to Open Data. Could be used to differentiate report dates.		Date	Yes	160	No	Irrelevant	
PERMIT_SI_NO			Text	Yes	3360294	No	High cardinality, specific to application	
LATITUDE	Latitude for the building where the proposed work will take place.		Text	Yes	220576	No		Float
LONGITUDE	Longitude for the building where the proposed work will take place.		Text	Yes	230068	No		Float
COUNCIL_DISTRICT	Council District for the building where the proposed work will take place.		Text	Yes	51	No		Category
CENSUS_TRACT	Census Tract for the building where the proposed work will take place.		Text	Yes	1326	No		Category
NTA_NAME	Neighborhood Tabulation Area for the building whe	Neighborhood Tabulation Areas (NTAs) were created to project populations at a small area level, from 2000 to 2030 for PlaNYC, the long-term sustainability plan for New York City. For more info, go to https://www1.nyc.gov/site/planning/data-maps/open-data/dwn-nynta.page.	Text	Yes	194	No		Category

Summary Statistics of continuous variables:

- Issue Time:

```
count    17589.000000
mean       124.519529
std        167.362093
min         32.000000
25%         43.000000
50%         66.000000
75%        129.000000
max       2284.000000
Name: issue_time, dtype: float64
```

- Latitude:

```
count    17566.000000
mean        40.726829
std          0.072279
min         40.499227
25%         40.682072
50%         40.728893
75%         40.768984
max         40.911082
Name: latitude, dtype: float64
```

- Longitude:

```
count    17566.000000
mean       -73.937831
std          0.076697
min        -74.252245
25%        -73.985089
50%        -73.952962
75%        -73.901479
max        -73.701445
Name: longitude, dtype: float64
```

Independent Variables vs. Dependent Variable Scatter Plots: