I. Problem statement:

Title: Predicting the time required for obtaining a building permit in NYC

Problem: Obtaining a construction permit can be a major hurdle for new developments in New York City. The process is highly uncertain and can take from several weeks to a year depending on numerous factors, including the zone of construction, project type, total construction cost, and many other factors. We propose to use data from New York City open data on building permit approvals to explore these factors and build a machine learning model for predicting the likelihood of approval of building permits as well as the estimated time required for the approval.

Who might care? Construction companies, developers, building contractors, architects, real estate investors, city administration (department of buildings), and other stakeholders can use such a model to predict the likelihood of the approval of the building permit and to estimate the time required for the approval. They can then incorporate the time estimate in their project management processes, investment return models, as well as total cost estimates. For the city administration, having an accurate estimate of the approval process of the applications in the pipeline can help to allocate resources more efficiently.

II. Description of the dataset

The data on New York City's building jobs (as reported by New York City's Department of Buildings) were downloaded from NYC open data in CSV format and was 1.2 Gb in size. The below steps were taken to clean up the data and prepare it for the analysis.

1. Taking a specific subset from the larger data
    - The data were filtered to include only a specific subset that had NB (new buildings) code in the 'Job Type' column. This was achieved by using pandas dataframe's filtering methods.
    - The data were filtered to include only a subset of columns that are relevant to the analysis. Only 21 columns were selected out of the existing 96 columns.
    - The filtered and reduced data was saved as a separate CSV file under the name "new_buildings.csv" to be used as the primary data for the analysis. These steps were saved in a separate jupyter notebook file named "samping.ipynb".
2. Cleaning the data
    - The reduced (filtered) data was loaded to a new jupyter notebook named "preliminary_analysis.ipynb".
    - The columns that had dates were parsed as datetime columns using the required argument in the pandas read_csv call.

- All column names were converted to title case for consistency.
- Cost/Fee data were converted into numeric form by removing dollar signs and converting to integers.
- A dependent variable column (Duration) was created by measuring the duration of the approval process as the difference between the "Approved" and "Fully Paid" columns.
- The data was further filtered to select only values that make sense - by removing values that are below 0 (meaningless), as well as zero values.
- "Boroughs" were transformed into a category to increase the speed of execution.
- Several categorical variables included more than 100 categories with sample sizes ranging from several thousand to as low as 5 or 10 in each category. We have aggregated these smaller categories into “OTHER” category to reduce the noise and increase the informative value of the dataset.

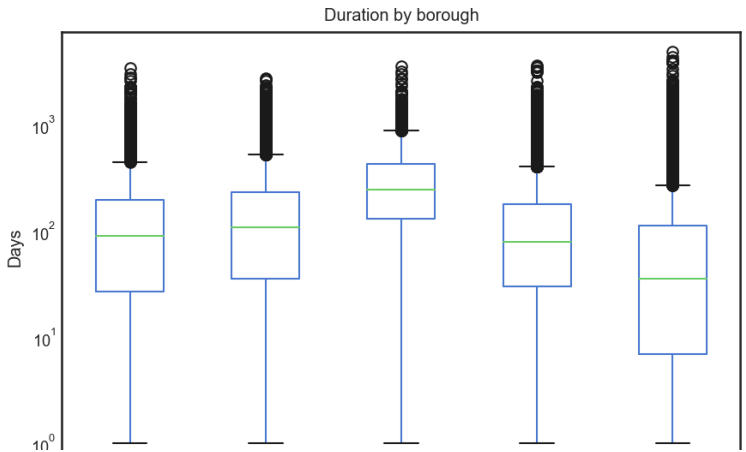
3. Dealing with missing values
   - The original data had 2 columns with YES/NO data where only YES responses were recorded as Y, and the NO values were left blank. We have converted the blank cells to NaNs and then converted NaNs to explicit N values, indicating NO.
4. Outliers
   - The dependent variable has outliers, constituting about 1% of the data. For the purposes of the capstone project, we have defined outliers as values that were 3 standard deviations above the mean. We have not removed them from the dataset for now as they might be informative for the statistical analysis section.

III. Initial findings from exploratory analysis

Duration of building permit issuance by borough

- On average it takes the longest in Manhattan and shortest in Staten Island to obtain a building permit.
- In Manhattan, it takes on average 11 months (333 days) to obtain a building permit, which is almost twice as long as in the rest of the 4 Boroughs.
- However, there is significant variance in all boroughs in the duration.
- The median value is higher in Manhattan and lowest in Staten Island.

Figure 1: Duration of building permit approvals by Borough



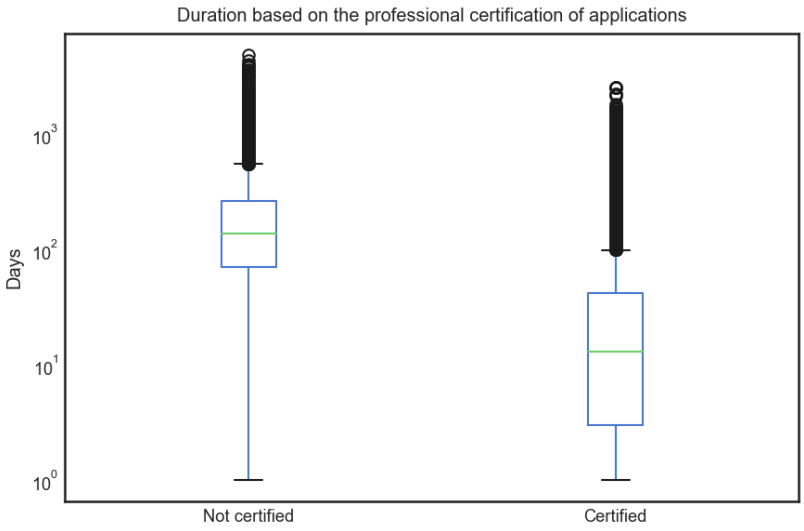
The type of buildings approved

- The majority (>70%) of building permits were issued for 1,2 or 3 family buildings. However, this ratio varies significantly across boroughs (figures A and B).
- Manhattan is the only borough where the majority of the building permits where issued for the "Others" category, meaning anything other than 1,2 or 3 family houses (figure B).
- The duration for obtaining a building permit is longer for "Others" category buildings than "1, 2 or 3 family" houses, by on average 79 days (figures C and D).

Professional certification of the application materials for building permits

- From data, we can observe that professionally certified applications are approved significantly faster than those that are not certified.
- This makes sense because professionally certified application materials are less prone to errors and reduce the back and forth between the issuing agency and the applicant.

Figure 2: Building permit applications submitted by certified engineers have a shorter approval time

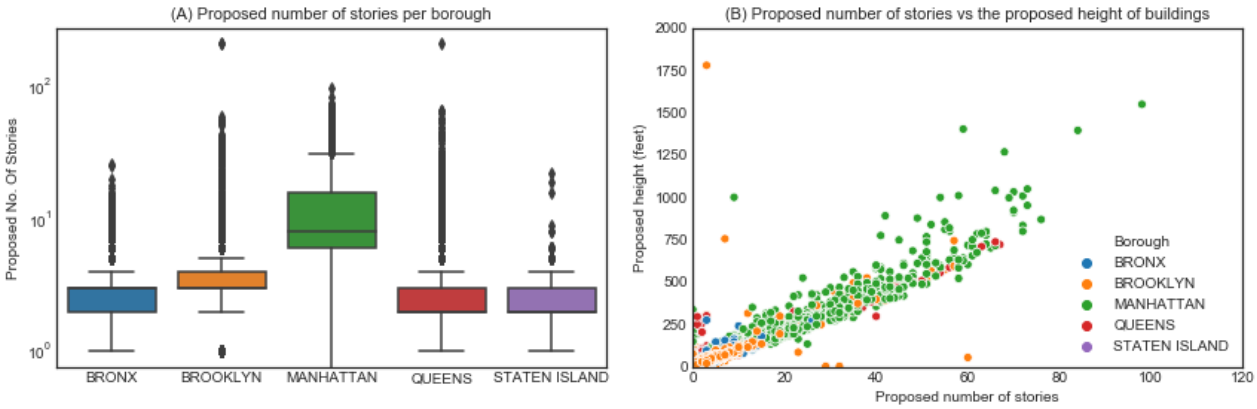


Proposed number of stories and height of buildings

- The highest buildings are proposed in Manhattan. The shortest are in Staten Island and the Bronx.
- From figure B we can observe that the relationship between the proposed height and the number of stories is quite strong and linear. However, from the same figure, we can see

that there are quite a few anomalies, such as buildings as high as 1750 feet but with less than 5 floors. Or buildings with more than 60 stories but with height less than 10 feet. This scatter plot helps us to evaluate the quality of data and possibly filter out values that do not make sense.

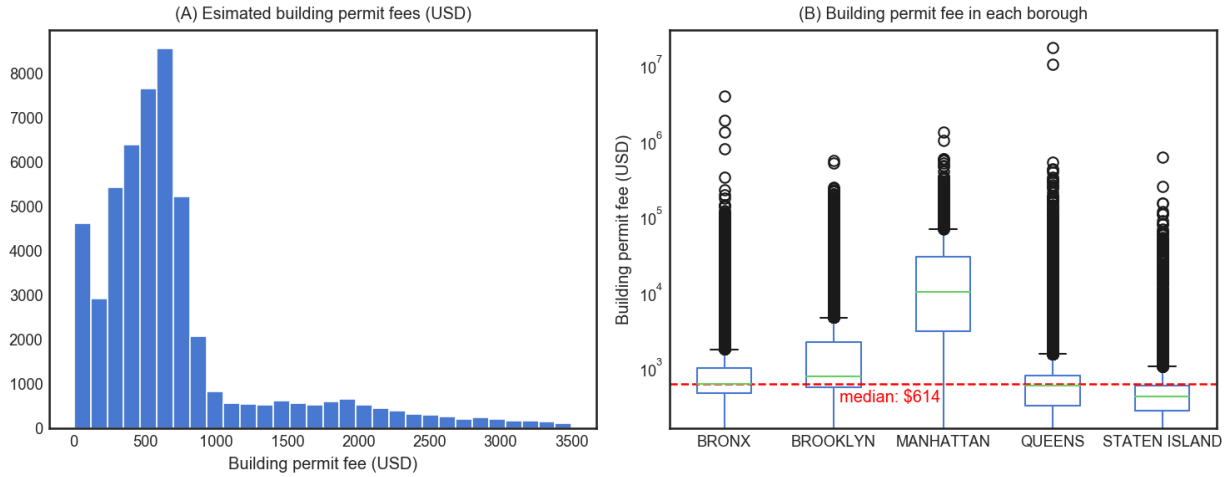
Figure 3: Number of stories and heights of proposed buildings per borough



Estimated fees for obtaining a building permit

- The highest fees are in Manhattan and the lowest fees are in Staten Island.
- Building permit fees range from $0 to more than several million dollars. However, most of the building permit applicants pay between $0 and $3,500 with a median fee of $614 (figure 5A).

Figure 4: Estimated building permit fees

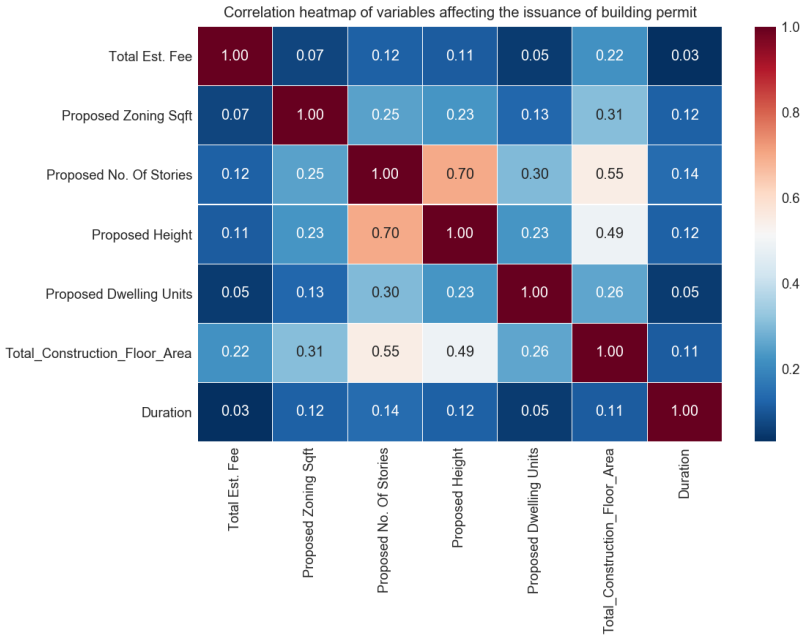


Heatmap of variables and their correlations with each other

- We have selected 21 variables that are related to the issuance of building permits, of which only 7 are quantitative.
- The dependent variable (Duration) does not seem to have a strong correlation with any of the independent variables.

- However, not surprisingly some of the area/floor/height variables are quite strongly correlated amongst each other.

Figure 5: Heatmap of correlations of continuous variables

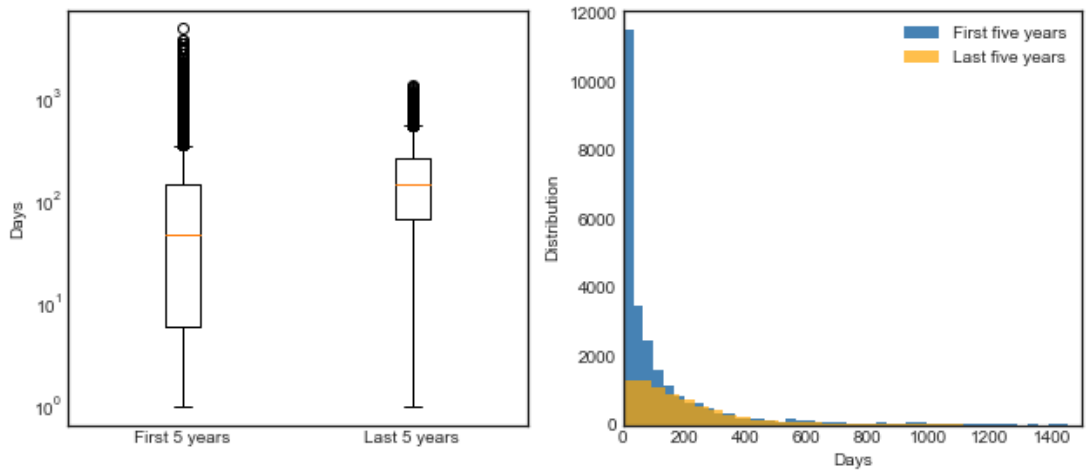


IV. Statistical analysis and findings

- The distribution of the first five years of data is significantly different from that of the last 5 years of data. Therefore, we decided to only use the latter in the predictive model.
- One quantitative variable and most of the categorical variables have significant explanatory potential for addressing the research question.
    - quantitative variables: Each additional floor added to the project is associated with some 9 days of additional time in obtaining the building permit.
    - Some of the most significant categorical variables are:
        - Whether or not the application was professionally certified by an engineer before the submission. If yes, the building permit issuance time is reduced by almost a third (91 days).
        - The building permit approval time is on average 67 days shorter for 1, 2 or 3 family houses.
        - Buildings in Manhattan have on average 68.5 days longer approval time than buildings in the other boroughs.

        - Buildings in Staten Island have on average 74.5 days shorter approval time than buildings in the other boroughs.
- There are strong correlations between several explanatory variables, including the building height, the total number of stories and the total construction floor area.
    - We have kept the total number of stories and have removed the other 2 explanatory variables that have strong correlations amongst each other and to keep the model unbiased
- What are the most appropriate tests to use to analyze these relationships?
    - Testing the distribution of the target variable for the first 5 years vs the latest 5 years to determine whether or not we can employ the entire dataset or not.
        - We have determined that the distributions are not normal by examining them graphically using histograms and boxplots (Figure 1). This has ruled out a t-test.
        - We have used the Mann-Whitney U test instead. The results were statistically significant and showed that the early years' distribution is different from the later years' distribution.
        - Hence, we have selected only the latest 5 years for conducting further statistical analysis.
    - Testing the significance of the association between the explanatory variables and the target variable.
        - We have used ordinary least squares (OLS) regressions to determine the correlations between the explanatory variables and the target variable, as well as their statistical significance.
        - For categorical explanatory variables, we have used dummies to conduct the OLS.

Figure 6: Boxplot and histogram of the first and last 5 years of data show visible differences in the distribution

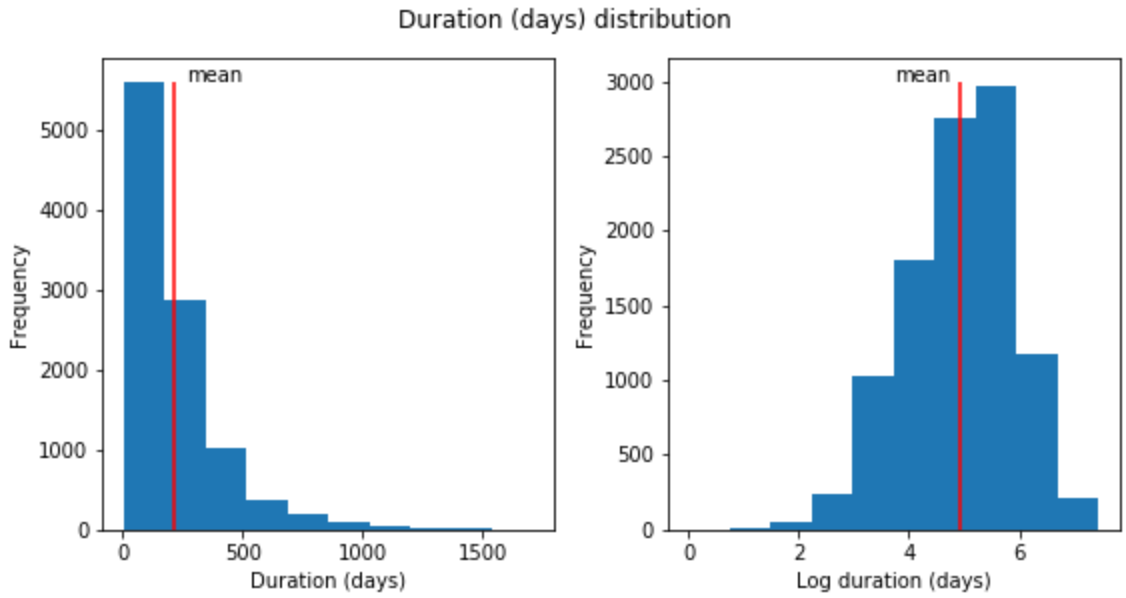


V. In-depth analysis with Machine Learning

Pre-processing and evaluating the data

- Data has both continuous and categorical variables. In order to conduct a regression analysis to predict the duration of the building permit, we convert the categorical variables into numeric dummies.
- Then we combine the relevant columns of the data into a single dataframe.
- We can see that the duration is not normally distributed, it is skewed to the right. The log presentation of the duration shows a more normal distribution. We use the log duration for the regression analysis.
- We apply sklearn's StandardScaler to standardize the numeric columns to improve the regression performance
- We apply SimpleImputer to fill in missing values

Figure 7: The histogram of the original duration in days (left) is skewed to the right. The log duration (right) shows a more normal distribution.



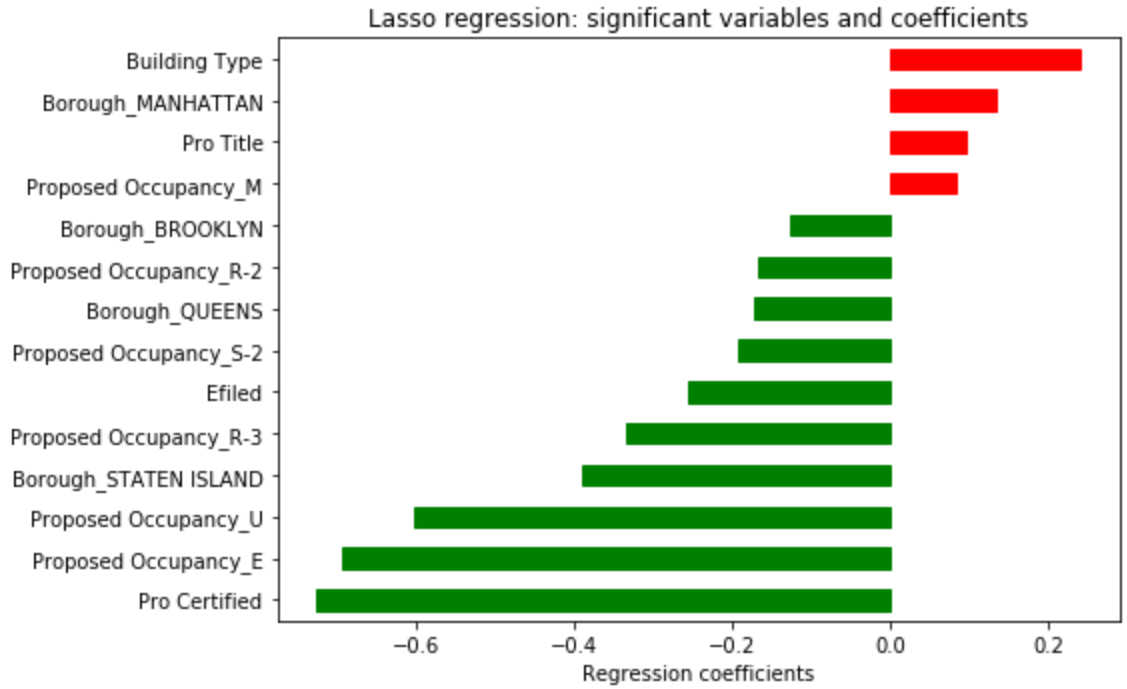
Conducting Machine Learning analysis

- We apply linear and ensemble models to predict the duration of the building permit issuance.

Linear models

- We apply Ordinary Least Squares (OLS) regression from StatsModels, Linear, Lasso, Ridge regressions from the Scikit Learn library.
- Lasso regression shows that the most important variable that reduces the duration of the building permit issuance is the Professionally Certified application materials. The most important variable that increases the duration of the issuance is the building type. In this case, the type of the building indicates buildings that are NOT 1,2 or 3 family houses.

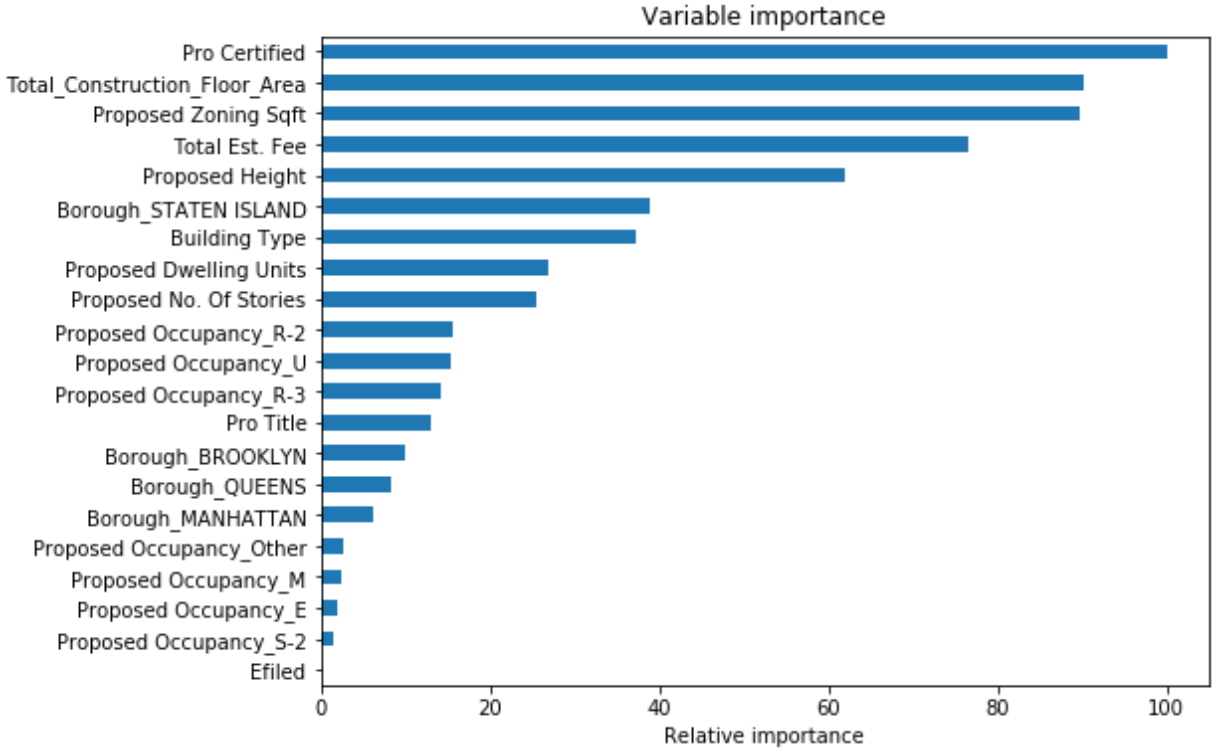
Figure 8: Lasso regression shows the most important variables. Variables in red increase the time for issuance of building permit, and the green colored variables reduce the time.

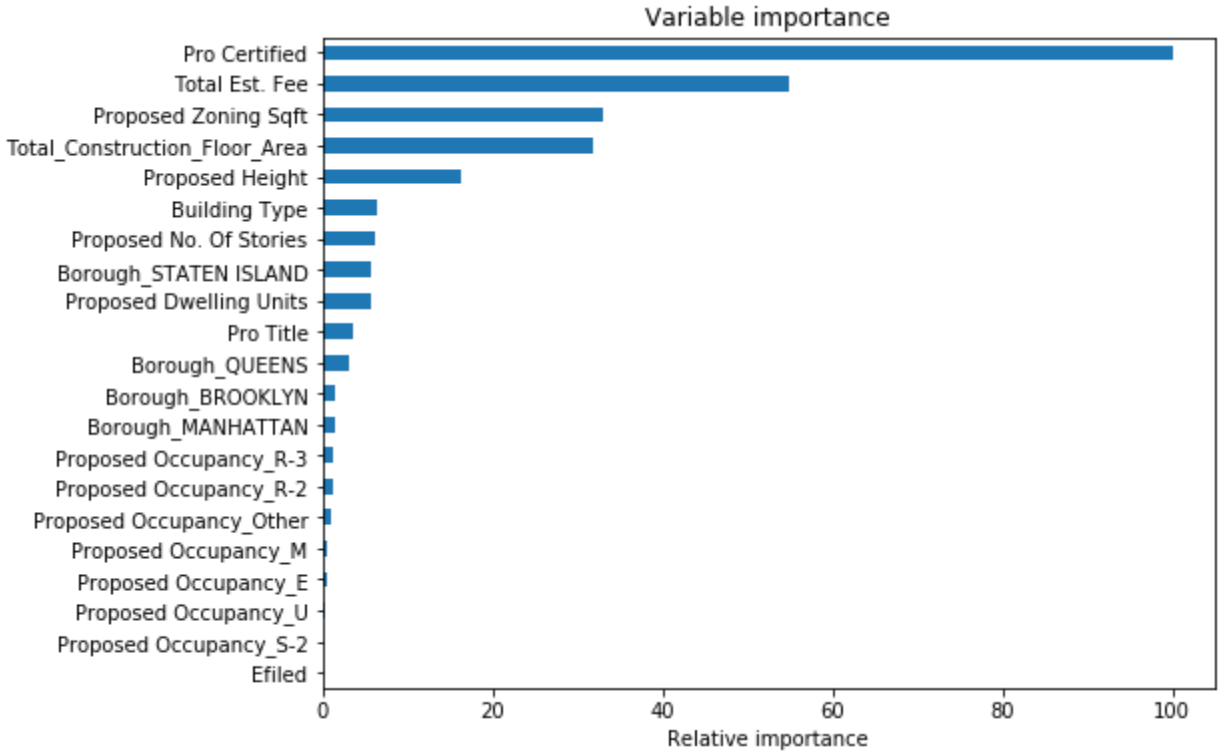


Ensemble models

- We apply Gradient Boosting and ExtraTrees models. The ExtraTrees model shows that the three most important variables in predicting the duration of the building permit issuance are:
    - Whether or not the application materials are professionally certified
    - The total construction floor area
    - The proposed zoning square footage

Figure 9: Relative importance of independent variables in explaining the duration of the building permit. Top: ExtraTrees model; Bottom: GradientBoost model





Evaluation

Overall, the ensemble models perform better in predicting the duration of the building permit. The R square for the ensemble methods is about 50%, vs around 37% for the linear models. This means that the ensemble models explain about 50% of the variation in the duration of the building permit. The mean squared errors of ensemble methods are also better (about 20% lower) compared to the linear models (Figure 9).

Figure 10: Ensemble models have higher explanatory power (R squared) and lower error rate (meas squared errors) compared to the linear models.

VI. Conclusion

Predictors

It seems that the best predictor for the duration of the building permit is the professional certification of the materials submitted for approval. Both the linear models and the ensemble models show the “Professionally certified” variable to be the strongest predictive variable. This makes sense, because if the construction project plans are professionally certified, this means that the building permit issuing authority will have less comments or adjustments and will be able to approve the plan faster.

Other important predictors include the building type, which in our case indicates whether or not the building is considered a 1, 2 or 3 family house or something else (“Other” category). We can see from figure 8 that the buildings that are in the “Other” category take longer to approve. Other variables of interest include the proposed occupancy of the building (type R2, R3, E, etc.), which indicates the nature of the occupancy. The location of the building also affects the duration of the approval. We can see from the Figure 8 that buildings in Manhattan are approved slower than buildings in Staten Island.

Models

The ensemble models perform much better, both in terms of the error of prediction (MSE) as well as the R squared measure (the percentage of variance in the target variable that is explained by the model). However, the linear models provide better interpretability. For example, by looking at the lasso regression coefficients (figure 8), we can conclude that buildings that are of the “Other” type tend to take 1.3 days longer to approve (the results are in the log form. So e^0.24 = 1.27 days). Similarly, the buildings permit applications that are professionally certified take on average 0.5 days less to approve.