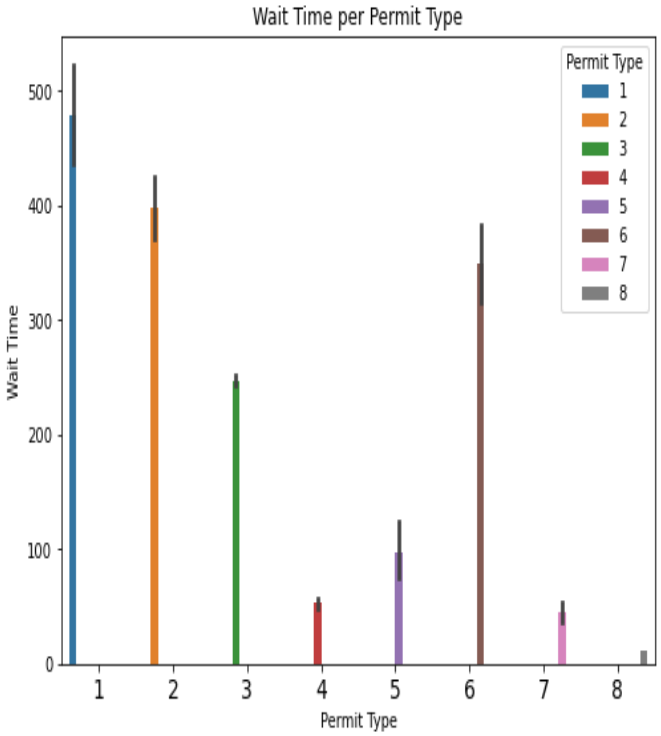
Dorcas Munoko
Big Data and Forecasting

Introduction

The aim of the project was to explore issues that lead to an increase in processing times of building permits in San Francisco. The data contained 198,900 observations with dates ranging from the years 2013-2018. Top features include permit type, filed date, issued date, number of proposed stories, revised cost, proposed units, existing construction type, proposed construction type, supervisor district, zip code, and plansets. The first step I took in data pre-processing was to find out the missing values in the dataset. Since the dataset contained a lot of missing values, deleting them would lead to loss of massive critical data. I created a threshold of columns containing at least 80% missing data to be deleted, dropped columns that were highly colinear, and those that had no effect on our outcome variable.

Description of Features

Permit type 1 had the highest number of wait time (approximately 480 days) as compared to permit type 8 which took approximately 10 days to be issued. Permit type 8 had the highest number of records (with otc alteration as the most popular) followed by permit type 3. The rest (2,4,5,6,7) had very few records. More permits had already been issued (80,000) as compared to revoked and incomplete permits. Construction Type 5 had the highest number of permits. I created a correlation matrix, and it indicates that existing construction and revised costs are positively correlated with the outcome variable while permit type, number of proposed stories, and proposed units are negatively correlated with outcome variable.



Models, Results, and Conclusion

I built different models to explore issues that lead to an increase in processing time of building permits. The features that I used include permit type, filed date, issued date, number of proposed stories, revised cost, proposed units, existing construction type, proposed construction type, supervisor district, zip code, and plansets. The regression tree model had the highest accuracy score (81%), Multinomial Naïve Bayes' (62%), LDA (56%), Gaussian Naïve Bayes' (53%), while Ridge regression had the lowest accuracy score. According to Ridge regression results, the best estimators of a longer wait time when alpha equals 2 are existing and proposed construction types. Finally, existing construction type appears to be the best predictor of delay in issuance of building permits in San Francisco.