Tigmanshu Chaudhary
tic48@pitt.edu/tighuchaudhary@gmail.com
www.linkedin.com/in/tigmanshu20

I have tried to explore the relationship between various factors that go behind a given permit in DOB permit Issuance dataset and the corresponding fees that are calculated as part of corresponding jobs which are related to that permit. To work on this relationship I have merged the given dataset with another dataset provided by NYC open data DOB Job Application Filings dataset. This additional dataset contains all the job applications submitted to DOB and contains columns which have information about jobs like whether electric work was involved, if fuel storage work was involved , whether the property is city owned , total estimated fees for a job which I have made as my target variable.My work involved identifying columns in the two dataset which might affect the target variable (Total estimated fees).

I have omitted basic data exploration/visualisation of dataset columns as they were already provided by NYC open data website and can be found here . I used this tool to better understand the data forsee some of the issues I might later face in working with the data.

Below are the tasks I did before I went on to model the data:(For more details please refer to the individual sections of the attached Jupyter Notebook)

- Sanitize the data: combine certain columns (eg Special District Column).
- Fill the null value in certain columns(eg latitude,census) by appropriate measure(mean, median)
- Simplify certain columns so they may be better suited to modelling.
- Encoding columns having ordinal values using appropriate integers.
- Impute certain columns with data based on my understanding what those columns represent(eg Self Certification)
- Remove non essential columns which I deemed were not necessary to model which is supposed to predict the fees for a job.(eg Street Name,Permittee's First Name etc)
- Drop rows which contain null values in essential columns which I couldn't impute values for. (Left with 3 Million rows down from 3.5 Million)
- Permit Issuance Dataset contain multiple rows for single JOB#, each row indicating a different work type that would be involved as part one permit(eg

construction, demolition etc), so If were to merge this dataset with another dataset based on JOB# numbers I would need a single row corresponding to a JOB#. To achieve this I aggregated the rows based on JOB# and sumed the values under the column WorkType. This resulted in a dataset which has a single row for a particular JOB# and whose Work type column contains integer reflecting how my different work types are part of this job. A value of 5 would indicate that this job would further require 5 different type of work types.(Aggregation resulted in a dataset with 1.4 million rows)

- Encoding the category variables using one Hot Encoding
- I performed similar steps of sanitizing and cleaning on the other dataset (DOB Job Application) and performed aggregation quite similar to the one performed on Permit Issuance.
- Merged the two datasets using the 'inner' join on JOB# column , resulting dataset has 1.1 million rows and 43 columns including the extra columns which resulted from one hot encoding.
- Split the dataset into train , test .

Below is the list of Final columns that would be used to fit the model for predicting the Total est Fees corresponding to a particular job along with the Target variable.

City Owned	Self_Cert	Job Type_A1	Permit Type_FO
Plumbing	Bldg Type	Job Type_A2	Permit Type_NB
Mechanical	Residential	Job Type_A3	Permit Type_PL
Boiler	Work Type	Job Type_DM	Other
Fuel Burning	Non-Profit	Job Type_NB	Professional Cert
Fuel Storage	LATITUDE	Permit Status_IN PROCESS	Total Est. Fee
Standpipe	LONGITUDE	Permit Status_ISSUED	BOROUGH_MANHATTAN
Sprinkler	COUNCIL_DISTRICT	Permit Status_RE-ISSUED	BOROUGH_QUEENS
Fire Alarm	CENSUS_TRACT	Permit Status_REVOKED	BOROUGH_STATEN ISLAND
Equipment	Special District	Filing Status_INITIAL	Permit Type_AL
Fire Suppression	BOROUGH_BRONX	Filing Status_RENEWAL	Permit Type_DM
Curb Cut	BOROUGH_BROOKLYN	Permit Type_EW	Permit Type_EQ

Machine Learning Models Used:

- Random Forest Regressor - The model which achieved the lowest absolute mean square error(approx 1150) when used with 500 estimators but took significant amount of time(~ 60 mins)
- Neural Networks MLPRegressor- The model which performed fastest(~30 mins) and gave an absolute mean error of approx 1200.
- Ridge Model Regression- Had the lowest accuracy among the three (approx 1600 absolute mean error) but a decent enough run time of 35-40 mins.

Summarisation :

Through this assignment I tried to relate the data from permit dataset with the fees involved for work types types that would be done as part a particular permit. On the face of it mean absolute error of order of 1000$ in the prediction might not seem much but one can appreciate it when one considers the cost involved in some of permits for bigger work permits goes in the magnitude of more that 50,000 $ frequently!