CSE587 DATA INTENSIVE COMPUTING-PROJECT PHASE 1

COST AND APPROVAL RATE PREDICTION FOR BUILDING PERMITS IN SAN FRANCISCO

AKSHEY RAM MURALI
50442206

VAISAALI MURALI KRISHNA
50468239

PROBLEM STATEMENT:

To Clean, Explore and Analyse the dataset to Predict the cost and approval rate in San Francisco, given the permit type and location of the construction.

BACKGROUND:

Each city or county has its own office related to buildings, that can do multiple functions like issuing permits, inspecting buildings to enforce safety measures, modifying rules to accommodate needs of the growing population etc. Getting permit approval is a cumbersome and crucial step in the process of construction. Hence, we need to have a system that will help us get an estimate of the cost and approval rate in an area. The permit application fee is costly and hence this model will help by guiding us in making better decisions.

DATASET:

https://www.kaggle.com/datasets/aparnashastry/building-permit-applications-data/download?datasetVersionNumber=1

EXPLORATORY DATA ANALYSIS:

1. BASIC NATURE OF THE DATASET:

The above dataset is a multivariate dataset which has 43 features and 1,98,900 rows. The dataset originally has features of data types - object, float and int. Some of the features are Permit Type, Permit Number, Block, Lot, Street Number, Current Status, Estimated cost, Zip code etc. The Record ID is the value that uniquely identifies every record submission in the dataset. The bigger the ID value, the more recent the entry is. Similarly, Permit Number is the feature that uniquely identifies a permit request. The permit number is either in the regex form “20*” or “M8*”. Since they cannot be converted to a common type and there is no requirement to do so as of now, we are not disturbing the column to make any alterations. We can observe that the location is of the form (latitude, longitude). The features Current Status and Estimated Cost would play a main part in the prediction process.

2. STATISTICS:

The statistics associated with the entire dataset is given in a form of a table below. From the below table we can infer the min, max, std , count etc for each feature. This

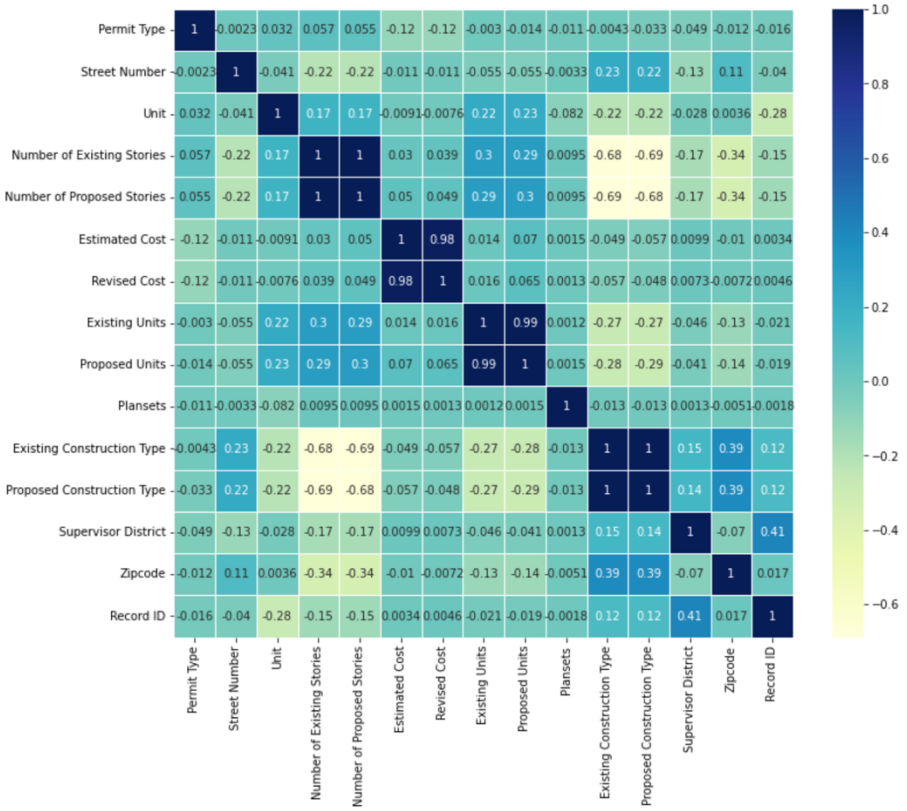
	Permit Type	Street Number	Unit	Number of Existing Stories	Number of Proposed Stories	Estimated Cost	Revised Cost	Existing Units	Proposed Units	Plansets	Existing Construction Type	Proposed Construction Type	Supervisor District	Zipcode	Record ID
count	198900.000000	198900.000000	29479.000000	156116.000000	156032.000000	1.608340e+05	1.928340e+05	147362.000000	147989.000000	161591.000000	155534.000000	155738.000000	197183.000000	197184.000000	1.989000e+05
mean	7.522323	1121.728944	78.517182	5.705773	5.745043	1.689554e+05	1.328562e+05	15.666164	16.510950	1.274650	4.072878	4.089529	5.538403	94115.500558	1.162048e+12
std	1.457451	1135.768948	326.981324	8.613455	8.613284	3.630386e+06	3.584903e+06	74.476321	75.220444	22.407345	1.585756	1.578766	2.887041	9.270131	4.918215e+11
min	1.000000	0.000000	0.000000	0.000000	0.000000	1.000000e+00	0.000000e+00	0.000000	0.000000	0.000000	1.000000	1.000000	1.000000	94102.000000	1.293532e+10
25%	8.000000	235.000000	0.000000	2.000000	2.000000	3.300000e+03	1.000000e+00	1.000000	1.000000	0.000000	3.000000	3.000000	3.000000	94109.000000	1.308567e+12
50%	8.000000	710.000000	0.000000	3.000000	3.000000	1.100000e+04	7.000000e+03	1.000000	2.000000	2.000000	5.000000	5.000000	6.000000	94114.000000	1.371840e+12
75%	8.000000	1700.000000	1.000000	4.000000	4.000000	3.500000e+04	2.870750e+04	4.000000	4.000000	2.000000	5.000000	5.000000	8.000000	94122.000000	1.435000e+12
max	8.000000	8400.000000	6004.000000	78.000000	78.000000	5.379586e+08	7.805000e+08	1907.000000	1911.000000	9000.000000	5.000000	5.000000	11.000000	94158.000000	1.498342e+12

data gives us an insight into what kind and ranges of data we are going to be dealing with in our process. We can accordingly plan to apply appropriate data types and cleaning operations on the features. These table values also give us an idea about what values we can expect to be present within each of these features, so that in case there is any evident error, we can easily identify and correct them.

3. CORRELATION:

The pairwise correlation of all the features in the dataset is given below. Each square in the matrix shows the correlation between the corresponding features from each axis. Correlation ranges from -1 to +1. Values closer to zero means that there is no linear trend between the two variables.
The closer the value is to 1, the more positively correlated they are and stronger their relationship is. Revised cost and estimated cost are more positively correlated. We can also note that the proposed and existing construction types exhibit a high correlation.

This data will help us when we find missing values in any one of the correlated features. In this case, we can find what kind of relation the two features display and accordingly fill missing values. Also, if we find a particular trend among the two features being followed across a reasonable number of records, we can fix errors in records that show differences from the pattern.



4. CHECKING FOR NULL VALUES:

One important step in EDA is to check the missing/null values in the dataset. Since this dataset being a real-world dataset and frequently getting updated, there are many null values when compared to other datasets.

The features containing null values are dropped or filled with an estimate in the data cleaning phase based on the feature's significance. In our Dataset, only 12 out of 43 features have 0 missing values. The features: TIDF Compliance and Site Permit both the features are entirely empty. Dropping these won't affect our model since it has the least impact on what is to be predicted.

Refer the pynotebook for the exact numbers of the missing values corresponding to the features.

5. CHECK AND DROP DUPLICATES:

Duplicates are nothing but the repeated data in our dataset. Duplicates are an extreme case of nonrandom sampling, and they bias your fitted model. Including them will essentially lead to the model overfitting this subset of points. In our dataset, we found around 17405 duplicate records. hence approx 8% dropped from the dataset.

6. CHECK FOR MISMATCH BETWEEN TWO CORRELATED FEATURES:

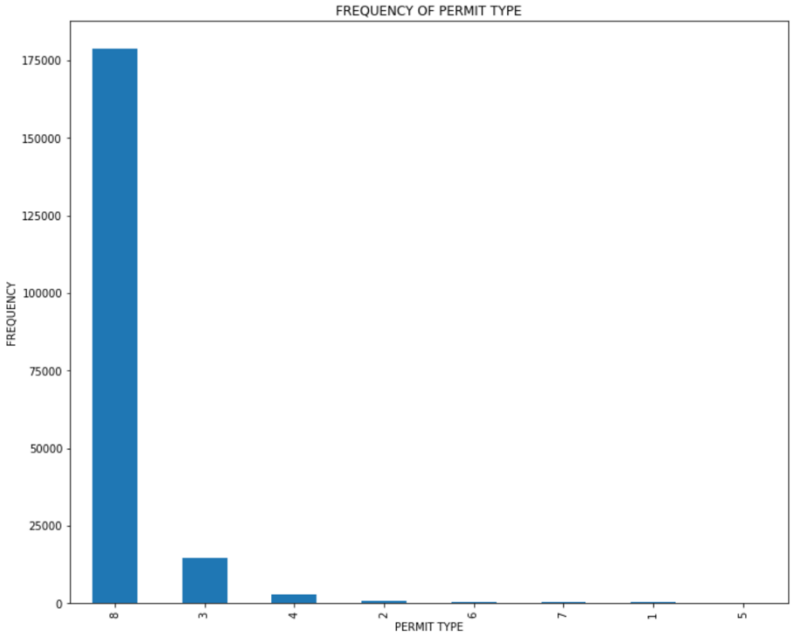
We have two features namely "Permit Type" and "Permit Type Definition" which should be equivalent to each other in meaning. Hence to confirm if the definition is unique to each permit type, we have used Group By Clause to display the combinations. Since each Permit Type was associated with only one Permit Type Definition, we have confirmed that they both are the same, only the form is different. We check this because we need to understand if there is any repetition of columns, one of which we can ignore during the process making it easy to interpret. This helps in cleaning the dataset w.r.t eliminating redundancy in features.

7. ANALYSING "PERMIT TYPE FEATURE":

The frequency/ count distribution for the feature "Permit Type" is shown below. From the graph we can infer that the Permit type 8 has the highest instances when compared to all other features, followed by the type 3, type 4, type 2, type 6, type 7, type 1 and type 5. Type 8 is "otc alterations permit" (from the Permit Type Definition Feature). The permit type 5 is "grade or quarry" is the least of all the permit types.

This gives us an understanding of the distribution of Permit Type which is one of our fixed input datas(Permit type and Location), in our training set. This understanding will help us process data to different extents for Permit Type that has a large number of records to learn from and for Permit type with less number of records. The

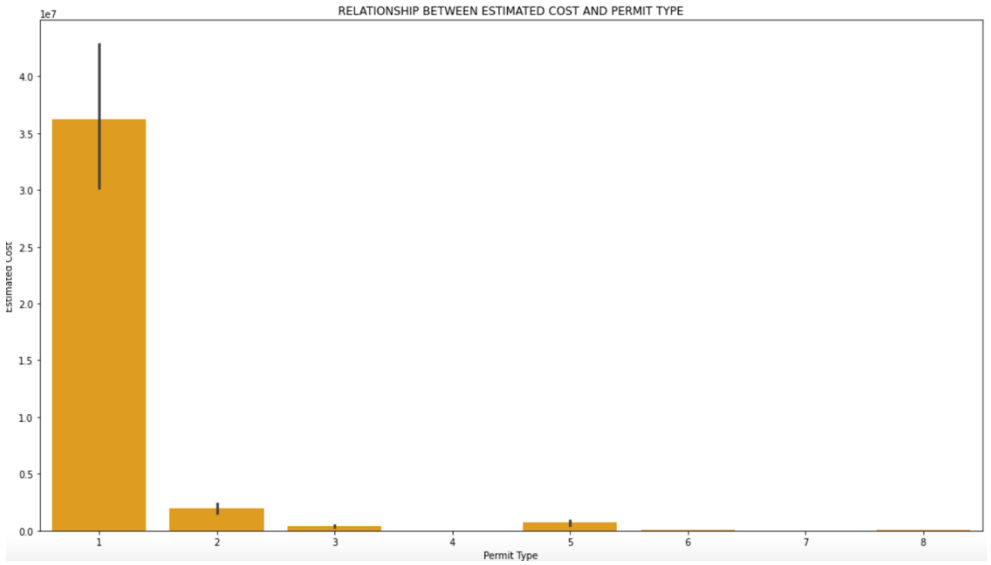
accuracy also can be expected to vary from one Permit type to the other since the frequency difference is huge.



8. RELATIONSHIP BETWEEN ESTIMATED COST AND PERMIT TYPE:

The below graph exhibits a relationship between Permit Type and estimated cost. From the graph we can infer that in the permit type 1 the takes the highest estimated cost when compared to all other permit types and hence can be considered as the costliest one. Type 4 and type 7 have the least estimated cost when compared to the rest. Type 2 is the second highest type in terms of the estimated cost. The estimated cost is scaled down to a smaller value for easy interpretation.

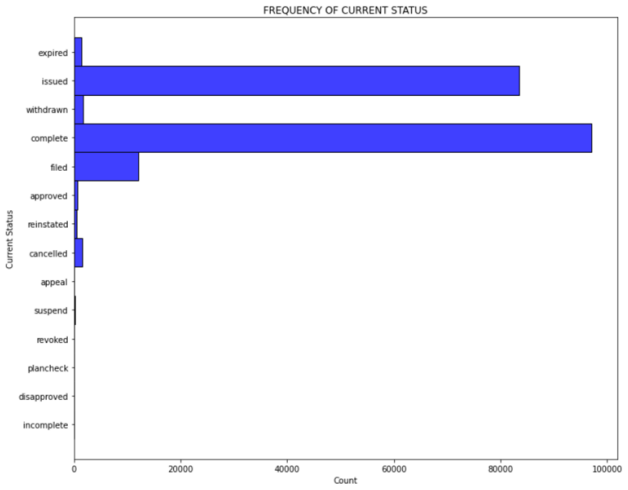
This relation will help us get an overall idea about the estimates, one of our primary results(estimate and approval rate) of our goal. We can evaluate our results with these values as a basic evaluation metric. i.e., If ever we come across a value that does not follow this trend, we can correct it.



9. ANALYSING "CURRENT STATUS":

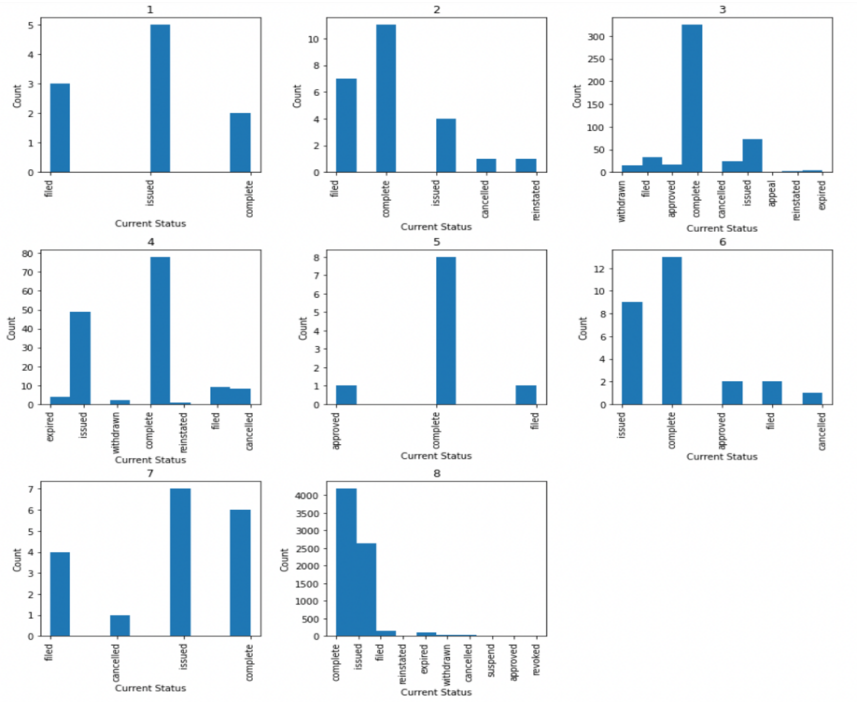
The frequency/ count distribution for the feature "Current Status" is shown below. From the graph we can infer that most of the applications in San Francisco are in status "completed". Around 90,000 applications are in the completed status. Around 85,000 applications are issued. Remaining statuses in the feature "current status" contribute to 1% of the total applications and those with status "incomplete" are less than 100 in number.

Now that we have an overall idea about the distribution of approval status in our dataset, From this, we can go forward in analysing why the distribution is in such a way. If there is a status with negligible frequency, we can check if it is an invalid feature value that needs to be removed or if it's a synonym to another status so that it can be combined. We need to categorise them into final status and intermediate status so that we can use final status for our prediction.



10. ANALYSING "CURRENT STATUS" W.R.T PERMIT TYPES:

Shown below is a graphical representation of relation between Status and Permit Type. Within each permit type, the approval status behaviour varies. We can see that some status is absent in some of the permit types.
This gives us the question that some statuses are specific to a permit type. The rate of issuance and cancellation varies with each permit type.
This graph helps us gather the correlation info among these two features for prediction, which is our primary goal.



DATA CLEANING:

1. DROPPING UNWANTED COLUMNS :

We find columns having more than 50% of entries that are missing. After confirming that the columns are insignificant to our end goal, we drop those unwanted columns. In our case, we have dropped 9 out of 43 columns leaving us with 34 columns to work with.

2. DROPPING DUPLICATE RECORDS :

We found 17405 duplicate records that have the same permit Number and differ slightly by the record ID and one/two attribute values, which gives us an idea that there have been multiple entries instead of one. In this case, we will keep only the recent record and drop the remaining entries.

3. SPLITTING COLUMN INTO TWO :

The attribute Location contains both latitude and longitude, the values of which will be a primary requirement for our use case. Hence, we split the Object value by “, “ and fill two new separate columns accordingly.

4. REMOVING UNWANTED CHARACTERS FROM LAT AND LON COLUMNS:

Cleaning the value of the two columns that have ‘(‘ and “)’, we strip the characters from these columns.

5. DROPPING INSIGNIFICANT ROWS :

Rows that do not contain either location or zip code cannot be filled or predicted with any other attribute from the dataset. Hence they are insignificant as they are the primary attributes we will be using for our project. We do this because we do not want to keep any data that won't be of any use to us and might cause unwanted confusion to our algorithm.

6. FILLING MISSING COLUMNS :

If we have missing values in location columns, we make use of zip code to determine the lat and lon value using geolocator function and fill them. We will later be using the same method to do conversions between zip code and locations depending on the input data type and to find distance between any two data points in data models by finding radius.

7. COMBINING COLUMNS TO ONE COLUMN WITH MORE MEANINGFUL DATA:

We have two columns namely "Estimated Cost" and "Revised Cost". As the name suggests, revised cost is the final rate. For our processing, we need the most accurate expected cost for entries. Hence, Revised Cost is given preference over Estimated Cost. In case of missing values in Revised Cost, we fill the combined column "Cost" with Estimated Cost.

8. CONVERSION OF DATA TYPE OF COLUMN :

The attributes 'Filed Date' and 'Issued Date' need to be converted to timestamp for performing numerical operations on the columns. We can later combine the two to a single column called "Issuance Duration" if required.

9. ENCODING COLUMN :

We encode Current Status to numerical type for performing easy analysis and processing of the column's data. Each status is defined by a unique natural number ranging from 1 to the total number of statuses.

10. SCALING OF COLUMN VALUES :

We scale down the Expected Cost by 1000 since most of the values are in 1000s and hence, it'll be easy to interpret the data.

11. SETTING PROPER PRECISION :

We have changed the precision of location columns to 10 digits from decimal point. This is because when we use these values to find an estimate of cost/approval rate, we

won’t need those many digits to the right of the decimal point. Hence, it is better to remove which is of negligible value.

REFERENCES:

https://numpy.org/doc/

https://pandas.pydata.org/docs/

https://matplotlib.org/stable/index.html

https://seaborn.pydata.org/

https://www.itl.nist.gov/div898/handbook/eda/section1/eda11.html

https://mathshistory.st-andrews.ac.uk/Biographies/Tukey/,

C. O’Neill and R. Schutt. Doing Data Science., O’Reilly. 2013.

CONTRIBUTION:

AKSHEY RAM MURALI - 50%
VAISAALI MURALI KRISHNAN - 50%