

# ZS - Machine Learning Challenge

By - Kumar Ritu Raj Singh

# The problem

## Statement

Given details of a construction/demolition site like its permit type, location, description, application status, work type etc, we need to identify if the task is related to one of the following building type:

- Single Family / Duplex
- Commercial
- Institutional
- Multifamily
- Industrial

## Metric


It is multiclass classification problem and is evaluated on the **weighted average F1 score**.

# Section A

Approach, preprocessing,  
observations, model used and  
results.

---

# The features we have

- Permit Type - Categorical - 0% missing
  - Address - Key value pairs - 0.06% missing
  - Description - Textual - 0.04% missing
  - Action Type - Categorical - 2.25% missing
  - Work Type - Categorical - 0% missing
  - Applicant Name - Textual - 0.4% missing
  - Application Date - Date/Time - 20.45% missing
  - Issue Date - Date/Time - 31.42% missing
  - Final Date - Date/Time - 53.33% missing
  - Expiration Date - Date/Time - 31.27% missing
- 

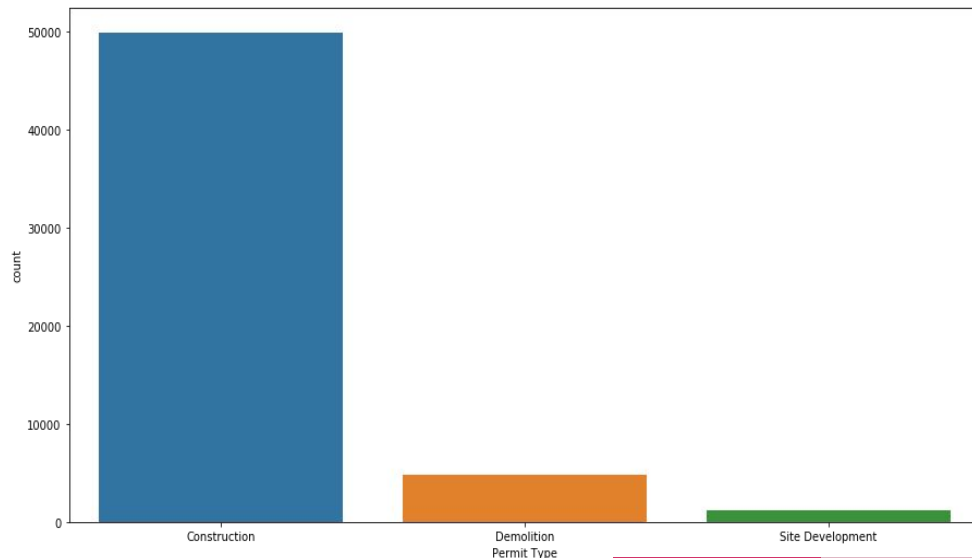
# The features we have

- Status - Categorical - 4.66% missing
- Contractor - Categorical - 81.84% missing
- Master Use Permit - Categorical - 86.57% missing
- Latitude - Numeric - 0.07% missing
- Longitude - Numeric - 0.07% missing



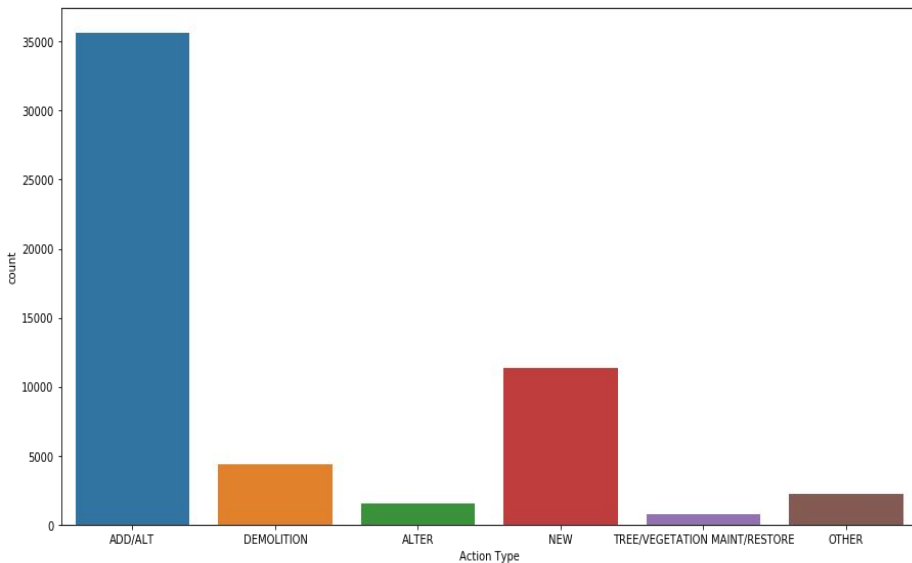
# Processing - Permit Type

- Categorical - 3
- Types:
  - Construction
  - Demolition
  - Site Development
- One-hot encoded this column
- 'Construction' is the most common permit type.



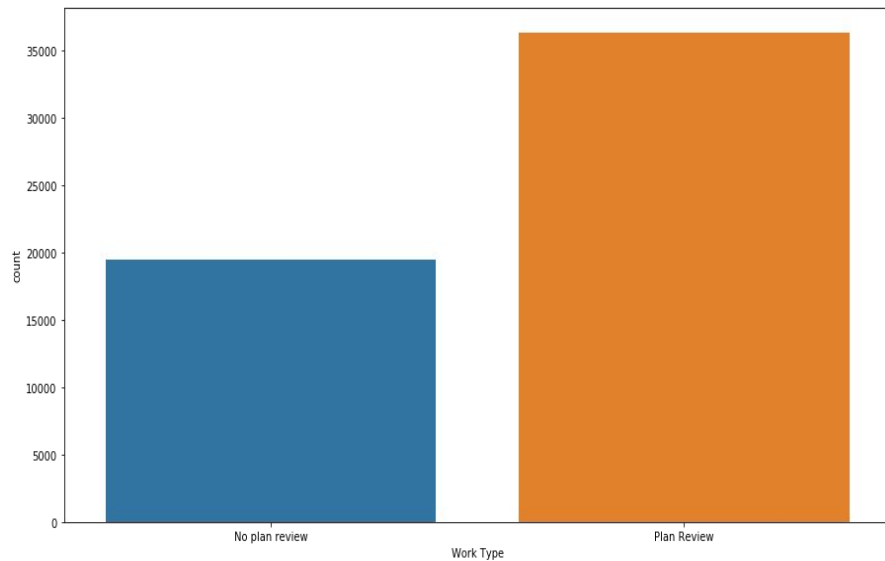
# Processing - Action Type

- Categorical - 16
- Picked Top-5 most popular action types and marked remaining as 'Others'.
- Types :
  - ADD/ALT
  - DEMOLITION
  - ALTER
  - NEW
  - TREE/VEGETATION MAINT/RESTORE
  - OTHER
- One-hot encoded the resultant column.



# Processing - Work Type

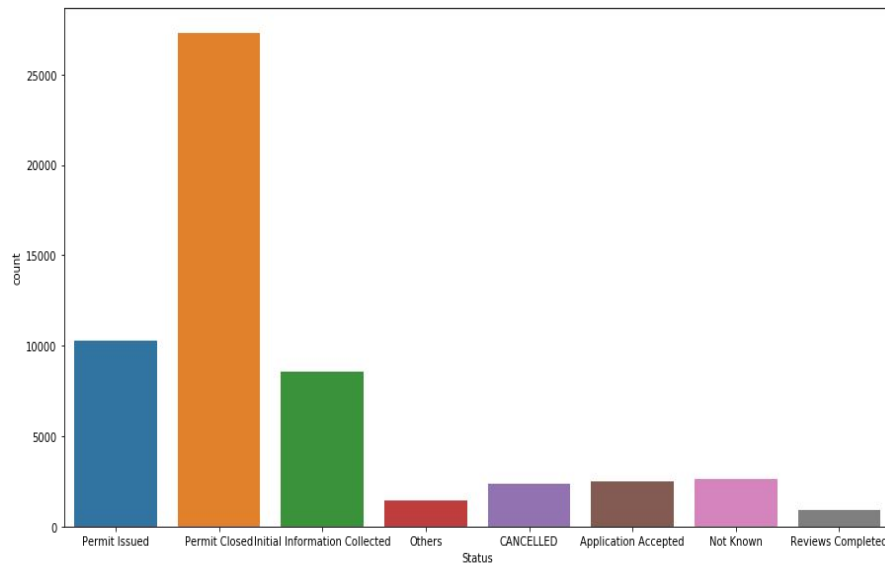
- Categorical - 2
- Types:
  - Plan Review
  - No Plan Review
- Marked 'Plan Review' as 1 and 'No plan review' as 0.





# Processing - Status

- Categorical - 12
- Replaced missing values with 'Not Known'.
- Picked Top-7 most most popular 'Status' and marked remaining as 'Others'.
- One-hot encoded the resultant column.



# Processing - Longitude and Latitude

- Location of the site given in terms of latitude and longitude.
- Mapped latitude and longitude to its respective position in 3D coordinate system.
- Mappings :
  - $X = \cos(\text{latitude}) * \cos(\text{longitude})$
  - $Y = \cos(\text{latitude}) * \sin(\text{longitude})$
  - $Z = \sin(\text{latitude})$
- Filled in missing values with -1



# Processing - Address

- Addresses are in the form : '9434 DELRIDGE WAY SW'
- Separated the initial number in address with the remaining address.
- Converted address number into a categorical column by keeping the address numbers with frequency  $> 100$  and marking remaining as 'Others'.
- Converted remaining address into a categorical column by keeping the remaining address with frequency  $> 300$  and marking remaining as 'Others'.
- One-hot encoded both of these resultant columns.



# Processing - The Date Columns

- Calculated number of days between:
  - Application date and Issue date
  - Issue date and Final date
  - Issue date and Expiration date
- Replaced missing values with the mean of columns.
- Standardized each of these three columns



# Processing - Descriptions

- Converted all words to lower-case.
- Removed punctuation marks and other symbols.
- Removed stopwords.
- Stemmed the words.
- Calculated TF-IDF values for to 500 relevant words in the description.



# Other observations

- Column of 'Applicant's name' was dropped because names were almost unique for each data point.
- Column 'Master Use Permit' was dropped because it had very high missing values, and the remaining values were also unique.
- In the end total 569 feature columns.
- The X, Y and Z coordinates calculated by mapping latitude and longitude into 3D space, were the most relevant features (tf-idf values of words from 'Description' column was not taken into consideration).



# Results

- Decision Tree Classifier gave an weighted F1 score of 0.8234
- Multi-Layer Perceptron gave an weighted F1 score of 0.8595
- Random Forest Classifier gave an weighted F1 score of 0.8688
- **Gradient Boosting Classifier** gave an weighted F1 score of **0.87765**
- eXtreme Gradient Boosting Classifier gave an weighted F1 score of 0.8758

All these scores are averaged over 2 cross validation runs.



# Section B

Code Walk-Through

---



# Thank You

