On this page you will find links to two files:

'train_data.csv' contains data similar to what you would work with at BuildZoom. Each of the 100,156 rows represent a building permit and there are 7 variables - licensetype, businessname, legaldescription, description, type, subtype, and job_value - associated with each permit.

'train_data.csv':

'xtest_data.csv' contains 25,149 building permit observations with 6 variables - licensetype, businessname, legaldescription, description, subtype, and job_value. The variable type, which is present in the training set, is omitted from the test set.

'xtest_data.csv':

We would like you to:

Build a classifier that predicts whether a building permit's 'type' is 'ELECTRICAL' or not. Note that there are many different types of permits but we are only interested in 'ELECTRICAL'. You are free to use any algorithm(s). You are also free to use all or a subset of the information included.

Use your classifier to predict whether each building permit contained in 'xtest_data.csv' is 'ELECTRICAL' or not.

How to return your results:

Predictions should be uploaded in a file named 'ytest_pred.csv'. Rows in 'ytest_pred.csv' map directly to rows in 'xtest_data.csv', e.g. if the first building permit in 'xtest_data.csv' is predicted to be of type 'ELECTRICAL' then this prediction is written to the first row of 'ytest_pred.csv' as a 1; if the second building permit in 'xtest_data.csv' is not predicted to be 'ELECTRICAL' then map this result to the second row in 'ytest_pred.csv' as a

0. Please note that 'ytest_pred.csv' is a file that contains *only* zeros or ones, each contained on a new line, and that these rows map to rows in 'xtest_data.csv'.

Your code should be uploaded in a file named 'buildzoom_classification.py'. Please write in python and comment appropriately to justify your feature selection, model(s) choices, validation, etc. Your script should be executable from the command line using 'python buildzoom_classification.py'. Please read the data in using the following path structure: './data/train_data.csv' and './data/xtest_data.csv'. Similarly, your script should write its results to './data/y_pred.csv'.

Ideally your code and comments should tell a complete story but you are also welcome to write a short note that highlights your work on this project. For example, feel free to identify something you found particularly challenging and how you tackled it, or what avenues you might pursue given substantially more time.

Good luck and have fun!