Capstone 1 Proposal

- Predicting Property Maintenance Fines (Blight)

- **Problem Statement**

The City of Detroit has been continuously suffering from Blight. Blight Violation Notices (BVN), or Blight Tickets, are issued by the city to individuals when property owners have violated City of Detroit ordinances that govern how property owners must maintain the exterior of their property - that is, issued when properties remain in a deteriorated condition. Every year the city of Detroit issues millions of dollars in fines to residents and every year, many of these fines remain unpaid. Enforcing unpaid blight fines is a costly and tedious process. So the city wants to know: 'how can we increase blight ticket compliance?'

The first step in answering this question is understanding when and why a resident might fail to comply with a blight ticket. This is where predictive modeling comes into play. Blight Prediction Model (BPM) will help the city to take proactive action by predicting who is likely to fail to pay off the fine on time by focusing on those people and, furthermore, to develop improved system that can raise blight ticket compliance.

- **Data:**

All data for this project has been provided through the Detroit Open Data Portal. The data contains all the information about who, when, why, to whom each ticket was issued and others such as address, hearing date, and violation code. Data size is of about 388K rows and 40 columns from 2004 to 2018 (present) excluding some atypical date such as 3016 or 1938.

- **Approach**:

The target variable is compliance that will be added, which is True if the ticket was paid early, on time, or within one month of the hearing date, False if the ticket was paid after the hearing date or not at all, and Null if the violator was found not responsible. Because our goal is to predict compliance as correctly as possible, the evaluation metric for this project is the Area Under the ROC curve (AUC). Also, since we do not have separate train and test data. To validate BPM, I will split data into twofold: tickets issued from 2004 to 2014 and the rest (approximately 3:1 ratio).

At first glance, the data contains many missing values, which will be one of challenges to deal with. Not only does the data include numerical columns, it also text columns which will probably have to be converted to the categorical and it will involves decision of how unique it is to preserve or abandon the feature. Next, I will study correlation of each column with target, compliance and collinearity among features. Lastly, many different models will be explored in order to deliver the best result.

Deliverables:

1. Codes (notebook) for:
    a. Data Examination
    b. Data Cleaning / Feature Engineering
    c. Feature Selection
    d. Model Selection
2. Capstone project analysis report
3. Slide deck on the project