YOUR PROJECT TITLE

Megha Erappa

Project URL: <https://github.com/meghaerappa/CSE_482_Final_Project>

**ABSTRACT**

The goal of this project is to create a framework to predict the likelihood that a crime takes place in large cities. Diverse types of data will be used to predict the pattern such location, type of crime and arrest reported and auxiliary data about the location. The prediction model will be able to help law enforcements to predict the kinds of crime and its frequency in different cities**.**

# INTRODUCTION

For introduction, you need to include the following information:

1. Crime prediction is very important application of big data for law enforcement especially in large cities. For example, law enforcements can predict when the next time will occur, at which particular location and the time as well. Prediction models such as scikit- learn: machine learning and collaborative filtering and recommender system. The prediction tool analyzes the type of crime that took place within a location and outputs if an arrest was made for that crime.
2. The goal for this project is to predict the type of crime that will occur within a specific location and see if an arrest will be made or not. Decision Trees, K Neighbors Classifier, Logistic regression and Collaborative Filtering and Recommender System. The prediction results are definitely better than the historical data because it helps to predict the next crime within a location whereas with historical data, a prediction can not really be made.
3. I used historical crime data for Chicago to create the prediction model.
4. The challenged encountered when creating the prediction model was preprocessing large amount of data. The historical data collected had to be reduced in size significantly so that it could be preprocessed. And many different training algorithms were used to create the prediction model.
5. Yes I was able to achieve the goal as my accuracy rate for the prediction model turned out to be 0.74.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

*CSE881-2015*, Month 1–2, 2004, City, State, Country.

Copyright 2004 ACM 1-58113-000-0/00/0004…$5.00.

# DATA

The data sets used to generate the model is from Chicago, Illinois. The reason for choosing these cities is because they’re metropolitan cities with huge population count. And crime rates in these cities are normally high compared to other cities. Additionally the quantity of data found within these cities are massive compared to other states. The source used for the raw data (<https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2>). This data was exported and saved as a csv file. I collected the complete data set after which is reduced the columns to location, type of crime, year, date, time, arrest, longitude and latitude. And then preprocessed them to two different csv files, such as, arrest, Type of Crime and count and the other is Location, arrest and count.

For the preprocessing steps, I grouped data using excel queries and created two different csv files from it. I applied classification and regression to the datasets. Different features were used to calculate it. I also used the recommender system problem (explained as above) to create my prediction model. Multiple neighborhoods and crimes occurred across Chicago was used. The historical crime data for each of the states were collected through the websites/API provided by each of the state. Location, type of crime and the date are taken into consideration when preprocessing. In terms of auxiliary data, the location in terms of where the occurred such as residential, business or entertainment district will be considered to draw up the prediction model.

Two final tables were used once containing three attributes: location, arrest and count (how many were arrest for committing the crime). The other table contained, the type of crime, arrest and the count (number of individuals that committed that crime within the location. The raw data file size is 124.37Mb and the preprocessed file are 0.03Mb and 0.01Mb.

Note that this is an important section to demonstrate to the instructor the amount of work you did for the project. If all you did was to simply download a dataset that has already been nicely prepared (e.g., Kaggle data) and there is little evidence you had performed much preprocessing, then you will not receive much credit for the project.

# METHODOLOGY

I used decision tree classifier, K Nearest Neighbors Classifier to perform my classification. And Python Surprise toolkit to do the recommendation. I also used logistic regression.

I used both the predictive modeling and recommendation problem, so to create the training and test sets, I used logistic regression to get the model. The distance measured used for clustering is 100.

* Sklearn model selection was used to test and train the data. The best hyperparameter and maxdepth recorded was 1 and the test accuracy was 0.763.
* Sklearn linear model was also used for logistic regression where the best hyperparameter and C equaled to 0.1 and the test accuracy was 0.736 as well.
* Decision trees was used to for my predictor attribute, in my case the arrest attribute.

# EXPERIMENTAL EVALUATION

This section describes the experimental setup and results you obtain.

## Experimental Setup

This section should include:

1. I am using Jypyter lab to code for the project on a Windows machine.
2. I compared it to the logistics online with the accuracy acquired after creating the prediction model.
3. Accuracy was used to report the results.

## Experimental Results

These are the experimental results:

Prediction based on how many arrests were made based on the crime:

[('DECEPTIVE PRACTICE', True, 8.0, None),

('CRIMINAL DAMAGE', True, 7.0, None),

('OTHER OFFENSE', True, 21.0, None),

('CRIM SEXUAL ASSAULT', True, 2.0, None),

('WEAPONS VIOLATION', True, 17.0, None)]

Prediction based on Location and how many arrests were made:

[('CURRENCY EXCHANGE', True, 4.0, None),

('JAIL / LOCK-UP FACILITY', True, 3.0, None),

('APARTMENT', True, 33.0, None), ('STREET', True, 44.0, None),

('ALLEY', True, 6.0, None)]

And the final processed for predicted result is below (top 10 results):

user: TAVERN/LIQUOR STORE item: 0 r\_ui = 3.00 est = 16.38 {'was\_impossible': True, 'reason': 'User and item are unkown.'}

user: GROCERY FOOD STORE item: 1 r\_ui = 6.00 est = 0.00 {'was\_impossible': False}

user: CHA HALLWAY/STAIRWELL/ELEVATOR item: 0 r\_ui = 1.00 est = 16.38 {'was\_impossible': True, 'reason': 'User and item are unkown.'}

user: CONVENIENCE STORE item: 1 r\_ui = 1.00 est = 16.38 {'was\_impossible': True, 'reason': 'User and item are unkown.'}

user: RESIDENCE PORCH/HALLWAY item: 0 r\_ui = 14.00 est = 16.38 {'was\_impossible': True, 'reason': 'User and item are unkown.'}

user: AIRCRAFT item: 0 r\_ui = 1.00 est = 16.38 {'was\_impossible': True, 'reason': 'User and item are unkown.'}

user: MEDICAL/DENTAL OFFICE item: 0 r\_ui = 2.00 est = 16.38 {'was\_impossible': True, 'reason': 'User and item are unkown.'}

user: COMMERCIAL / BUSINESS OFFICE item: 1 r\_ui = 1.00 est = 0.00 {'was\_impossible': False}

user: RESIDENTIAL YARD (FRONT/BACK) item: 0 r\_ui = 13.00 est = 5.44 {'was\_impossible': False}

user: SIDEWALK item: 0 r\_ui = 34.00 est = 18.54 {'was\_impossible': False}

The project was successful. I used different regression clustering methods and algorithms to test and train the prediction model.

# The potential challenges was that the data provided was either incorrect or missing. I believe, this may have altered the model since the information is not complete and mis-reported .

# CONCLUSIONS

In conclusion, the project was successful however the prediction model could have been more trained. Overall the model is now able to at least predict the type of crime that will occur within a location and if the offender will be arrested or not.

# REFERENCES (at least 3 references)

1. <https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2>
2. <http://www.cse.msu.edu/~ptan/CSE482/>

Grading criteria

Note that the project accounts for 10% of your final grade. The project will be graded based on the following criteria:

1. Presentation - structure/organization and clarity of writing (including tables and figures).
2. Technical - Correctness and thoroughness of the analysis performed. What are the challenges faced and how well did you address them? How do you evaluate the performance of the method you'd applied to the data? How much detailed discussion you provide to explain the results you'd obtained (e.g., discussion about why the method works or didn't work on the data)?
3. Difficulty level - How large is the dataset used? How much effort you had to spend to collect, integrate, preprocess, and analyze the data? Are you implementing the project on a cluster or a single machine? What tools did you use (do you have to implement them or are you simply using existing libraries)?