**Predicting Vehicle Crime In LA Neighborhoods**

Julian Shomali

Project URL: https://github.com/jshomali/CSE482Final

**ABSTRACT**

This project is about trying to predict when the next vehicle crime could happen at based on the location, the time, whether it is a weekday or weekend, and the number of crimes. I would like to achieve where and when the next vehicle crime could be at. I formulated the problem as a binary classification to begin with then I used k means clustering, then using the prediction algorithm to achieve my goal. The results I obtained shows that I am able to achieve this goal.

# INTRODUCTION

For introduction, you need to include the following information:

1. The prediction algorithm mentioned in the abstract is essentially using collaborative filtering and a recommender system. The recommender system helps filter out information to provide suggestions and in this case, using location, day of week, time of day, and number of crimes, it narrows down to what the choices are for each location based on day of week and time of day with number of crimes. Collaborative filtering is essentially the process of filtering information to overcome the overload problem. It is used to predict if a crime will most likely happen during the weekday or weekend, on location and time of day.
2. The goal of this project is to predict where and when the next vehicle crime could happen. In order to achieve this goal, I trained my data to determine if the predictor could assume if it is a weekday or weekend for when a crime could happen. I applied the k-nearest-neighbors technique to split up the data to determine clusters of accuracy in validation and testing. The test size was 30% and the training size was 70%.
3. The type of data I collected before preprocessing was the Date Occurred, Time Occurred, Type of Vehicle Crime, and Location (Neighborhood). The preprocessing data is Date, Location (Neighborhood), Time of Day, Day of Week (Weekend or Weekday), and Number of Crimes.
4. Some of the challenges I encountered when collecting was that there was a Date Occurred column and a Date Reported column so there was some discrepancy that not all crimes were reported so I had to work with my data. Another challenge was determining my preprocessing data and had to restart my data preprocessing multiple times to fit in what was necessary to achieve my goal.
5. The findings of my project were successful. I was able to achieve my goal of determining when and where the next vehicle crime rate could happen at a particular location, given a time frame and if it is a weekday or weekend. I approve of the hypothesis since it shows the possibility of a crime happening.

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 first dataset I got was from a website provided from the government of California for the city of Los Angeles. The second dataset I got was also from a website but from Los Angeles specifically. It essentially gave the same dataset so I did not merge them to avoid duplication of some sort. The attributes that I was targeting were when the crime occurred (Date and Time), I filtered out the crime data just to only be vehicle crime rather than all types of crimes, and I was also searching for the neighborhood and picked out 7 of the biggest neighborhoods in Los Angeles.

The data I collected was all from Los Angeles and it is spatial data (crime) for different neighborhoods in Los Angeles. The dataset I mostly focused on had many unnecessary features such as police code, gender, victim description, and what kind of weapon used. There was a lot of missing data in those unnecessary columns but nonetheless, it was not used. The dataset had data from the year 2010 to present and had over millions of rows of information.

The preprocessing data steps I took were to take the raw data and divide them by the number of vehicle crimes in a particular day and time. I set up my columns to be Date, Location, Time of Day, Day of Week, and Number of Crimes. The data I collected were all from this year (2019), based it on 4 time frames (0000-0600, 0600-1200, 1200-1800, and 1800-2359), then I determined if it was a weekday or weekend, then I finally counted the number of crimes that happened in a frame. I did find outliers and stated that in my python notebook file. The data ended up being a month of a dataset (1/1/19-1/31/19) because it was to have a certain amount of data within a given date frame. I applied a discretization method for time of day and day of week. These methods were to determine to have a time frame for time of day and also find out if it was a weekend or weekday. To predict if a vehicle crime could happen, I used the collaborative filtering and the recommender system to determine when a vehicle crime could happen.

For the collaborative filtering and the recommender system, I used Time of Day and Day of Week to determine if a vehicle crime was going to happen at a Location, by basing it off the number of crimes. The regression model I used is K-nearest neighbors and my y-value (target) is the Day of Week (Weekend or Weekday) and my x-value (predictor) is the rest of the dataset. Since the recommender system was transformed as a predictor system for the crime data, I used 7 neighborhoods and there were many types of vehicle crimes, but was narrowed down to how many vehicle crimes happened.

The attributes used in the table was:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Date | Location | Time of Day | Day of Week | Number of Crimes |

The final dataset size for preprocessing was 869 rows and 5 columns because it was filtered out from the number of crimes and the raw dataset extracted was only one month based. I used both scikit learn and python surprise for data analysis. The dataset size of the preprocessing data was 29KB while the raw dataset size was 79.3MB.

# METHODOLOGY

This ended up being a predictive modeling problem and my training and test sets were based on my preprocessing data attributes. For my K-nearest neighbors training and testing sets, I used my y-value (target) attribute as the Day of Week (Weekend or Weekday) and my x-value (predictor) attributes were my Location, Time of Day, and Number of Crimes. I ended up doing cross-validation to select model hyperparameters to determine my accuracy with the number of neighbors. I decided the number of clusters is 100 for the data instances.

* project1.ipynb: I had one Jupyter notebook file and it was loading the preprocessing data, then using the K-nearest neighbors clusters method, determining the outliers, finding the area under ROC for the Day of Week (Weekend or Weekday), and then using the predictive modeling system to determine where the next vehicle crime could happen.

# EXPERIMENTAL EVALUATION

This section describes the experimental setup and results you obtain.

## Experimental Setup

This section should include:

1. I used MacOS Mojave and manipulating the data was run by Jupiter Lab (Notebook)
2. The baseline methods I used to compare my approach was using machine learning (sklearn) and randomness to determine my results.
3. The evaluation metrics I used to report my results were test accuracy, area under ROC, mean absolute error (MAE), and root mean squared error (RMSE).

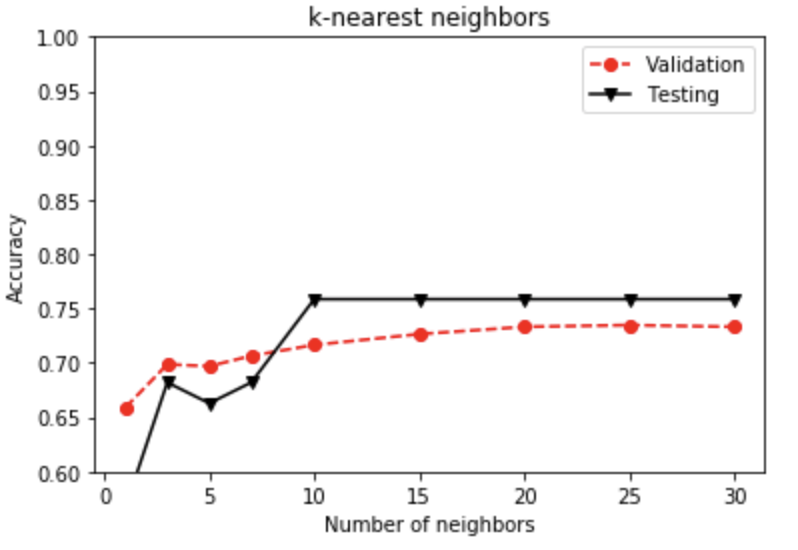
## Experimental Results

These are the tables and figures:

This table shows the first 5 rows for my preprocessed data:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Date | Location | Time of Day | Day of Week | Number of crimes |
| 1/1/19 | 77th Street | 0 | Weekday | 0 |
| 1/1/19 | 77th Street | 600 | Weekday | 2 |
| 1/1/19 | 77th Street | 1200 | Weekday | 2 |
| 1/1/19 | 77th Street | 1800 | Weekday | 6 |
| 1/1/19 | Devonshire | 0 | Weekday | 1 |

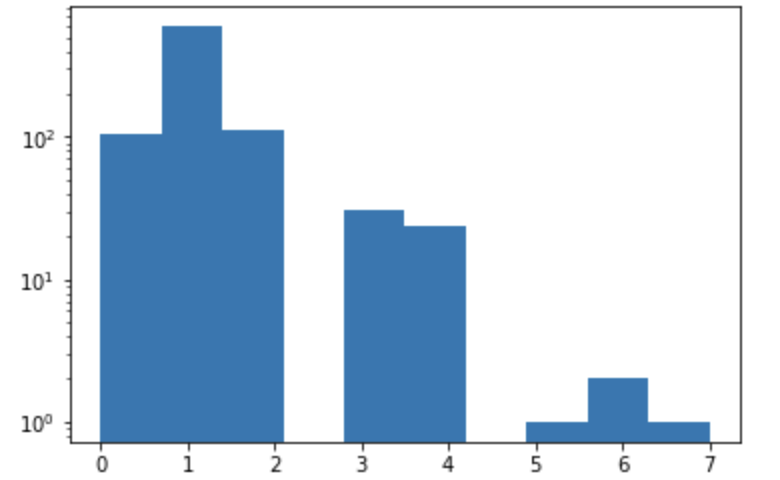
K-nearest-neighbors:



Best hyperparameter, k = 25

Test accuracy = 0.75862

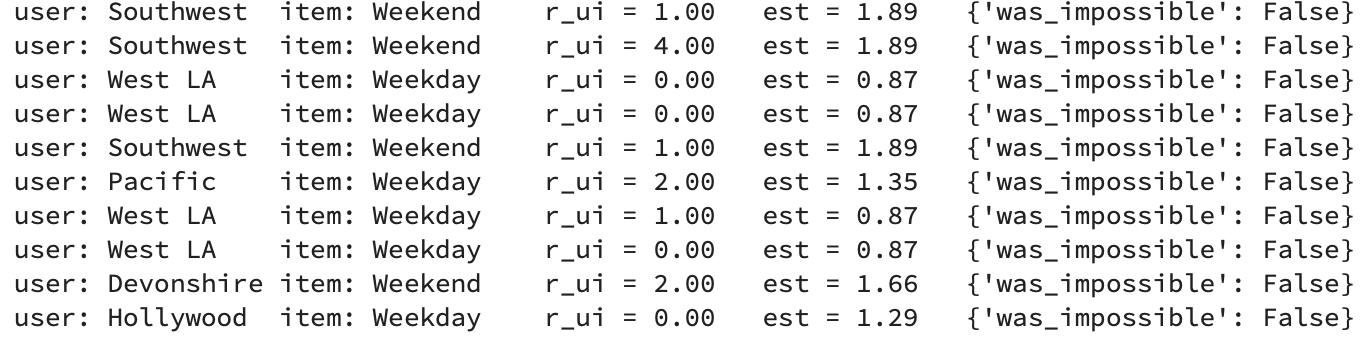
Outliers:



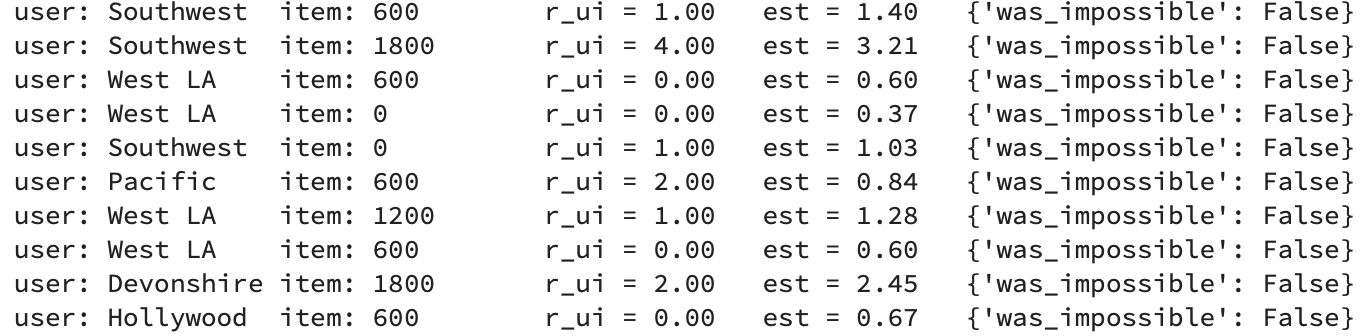
RMSE: 1.5634

MAE: 1.2055

Predictive results using day of week (first 10 rows):



Predictive results using time of day (first 10 rows):



Finally, state whether the project was successful. If not successful, explain the reasons that could affect the poor results you have obtained. What steps you could have taken to alleviate the problem.

This project was successful because the results I got was conclusive enough to determine where and when the next vehicle crime could happen. For example, using both of the predictive results, I can determine that the next vehicle crime could happen in the Southwest neighborhood, between the time of 0600 and 1200, during the weekend and also how many crimes could potentially happen.

# CONCLUSIONS

All in all, the findings of the project shows that the predictive system worked. I ended up finding out where the next vehicle crime could happen and how many could happen in a given time frame and if it is a weekday or weekend. For future work, I could use a larger dataset to have more accurate testing. I could also improve this project by adding other predictive attributes for my target such as having the number of each type of vehicle crime, rather than the total amount of vehicle crime.

# REFERENCES (at least 3 references)

1. <https://catalog.data.gov/dataset?q=los+angeles&sort=views_recent+desc&res_format=JSON&tags=crime&as_sfid=AAAAAAVfCFwsNKz5VFCofHCu0_C7kpNmlv5xJOdKAUtVjGKdp2ogofxq6I3R76sD9XLjgTDNAf452ZgjJeXrHClRLwqoFLtNyEgL67VjSJ1siBUyJTETAICq98dATbUjkdJpOV0%3D&as_fid=a0d7c0f968d54561be77904384117bf06221cd2f&ext_location=&ext_bbox=&ext_prev_extent=-183.515625%2C-55.37911044801048%2C-17.578125%2C59.17592824927138>
2. <https://data.lacity.org/A-Safe-City/Crime-Data-from-2010-to-Present/y8tr-7khq>
3. Professor Tan, *CSE 482 Lectures, Exercises 6 and 8*

**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)?