Predictive Crime Analysis

Group 23

Vatsal Parsana
Parth Gojiya
Yash Shah
Revathy Sivasankaran
Shabari Girish
Hari Vintha

INTRODUCTION

A smart city is an urban development vision to integrate multiple information and communication technology solutions with the help of various data collection methods to ensure security and efficient management of a city's resources and assets – the city's assets includes, but are not limited to transportation systems, schools, libraries, law enforcements, power plants and other community services. The goal of building a smart city is to enhance quality of life, performance and interactivity of urban services, to reduce costs and resource consumption and to increase contact between citizens and government by using urban informatics as a tool.

PROBLEM AT HAND

Providing safety is crucial to improve the quality of life in the cities and has a significant impact on the promotion of economic growth. The world's cities are bursting at the seams, civic resources are under pressure and crime is harder than ever to police. Safe neighborhood is the one of the fundamental factor for developing city to become a smart city. After two years of steady increase in the rate of crime in America, this past year the crime rate decreased by 0.1% across the entire country. But there is an overall increase in violent crimes across the US and it has plagued the US for the past decade. Public security is a growing problem for cities worldwide and to remediate this issue our team has tried to come up with an approach to predict crime before it happens. To measure the effectiveness of our model's approach, we have chosen to work with the Los Angeles Crime Dataset obtained from LAPD.

The dataset contains information of the crime committed from the year 2010-present in the city of Los Angeles. To better understand the current and relevant crime trends, we have truncated the dataset to include the crimes committed for year 2017-present. The dataset consists of information such as the date on which the crime occurred, the date on which it was reported, crime code, crime code description (type of crime), age and sex of the victim, latitude and longitude and the location where the crime occurred. The total number of crimes occurred for the year 2017-present was 417,539. This number can be reduced with better methods to prevent and stop crime from happening in the first place.

GOAL OF THE PROJECT

The goal of our project is to empower the city officials and the residents of the city to create a smart city by making it a safer city. Our project aims at developing a predictive model by implementing advanced data analytics solutions to curb/reduce the crime rate (theft) in the city of Los Angeles. The predictive model will be based on the previous 2 years crime dataset which details the type of crime, location of the crime, date occurred, premise, area and address of the crime.

APPROACH

Our vision behind this problem statement is to reduce the crime rate in the city to make it safe and smart. In order to reduce the crime rate we need to not only respond to the crime but prevent the crime from occurring in first place and our idea is to predict if a particular crime will occur on a day based on the area location, time of the day, month, latitude and longitude, and the premise code and hence achieve low crime rate in the city. For our project we are only considering the crime of theft and the categories that comprises under theft. This is done with keeping in mind the business aspect of this project as our stakeholders comprise of security companies, retailers of all sizes, business chains and departmental

stores. Our goal is to help these retailers and the public living in those areas of the LA county where the crime of theft is highest by providing with insight and recommendations to prevent crime from occurring at their locations.

BUSINESS PROBLEM

The success of a business (e.g. retail) thrives when it has a constant flow of customers wanting to spend money for a product of their choice. But the same business could suffer if the crime rate (especially theft) is higher in that neighborhood resulting in loss of customers to the business. To remediate this problem, our team has tried to come up with a predictive model which gives the possibility of theft occurring on the shift (morning, evening, night) of the day based on historical crime dataset of the city of Los Angeles obtained from LAPD. The focus of our predictive model is only on a particular crime i.e. Theft. Crimes like theft can be particularly devastating to small businesses, who lose both customers and employees when crime and fear claim a neighborhood. According to National Retail Federation survey Americans lose nearly $50 billion dollars annually due to theft related activities. Of that total amount, approximately $100-150 million dollars are lost every year in Los Angeles from thefts such as shoplifting, petty and grand theft and stealing. Identifying this problem and minimizing the impact of theft on small and large businesses would pave a long way in making Los Angeles a smarter city. Our stakeholders for this project are security companies that are hired by businesses and business owners to whom we will pitch our recommendations.

DATA ANALYSIS

Data Visualization

After the data cleaning operation, the data was visualized using multiple scatter plots, bar charts and pie charts. This step was performed to better understand the data and remove any extraneous information from the dataset that doesn't concern our problem. For the exploratory analysis and keeping our scope of the project in mind we are only going to consider theft and since there are many different forms of theft mentioned in the dataset, we have combined the ones which closely match our problem statement. The figure on the left shows the pie chart distribution of the consolidated theft crime vs the other top crimes. The distribution of theft is approximately 56% (66911) out of all the other crimes which deduces that people of Los Angeles are more prone to getting mugged, robbed or burglarized than being assaulted or murdered. The percentage of theft being above 50 % out of all the crimes is one of the main reason why we are only targeting theft and not any other crimes. The figure on the right shows a base map of Los Angeles which is superimposed with all the thefts (green color) occurred in the past 2 years

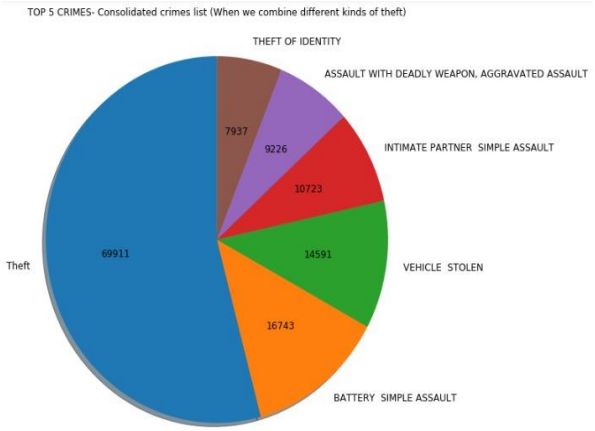


Figure 2: Pie chart distribution of theft and other crimes in LA

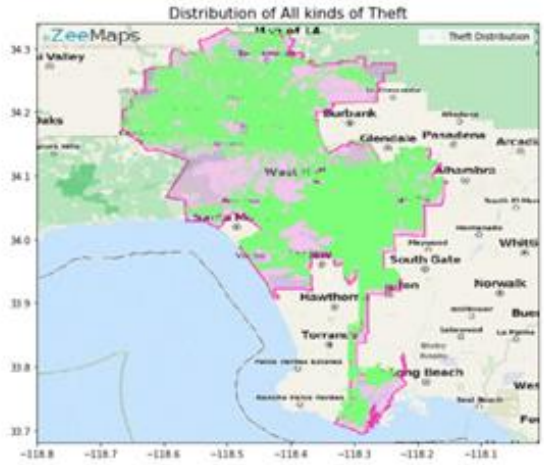


Figure 1:Spatial distribution of thefts in the LA county

DISCUSSION AND INSIGHT

Cleaning the raw data and performing exploratory analysis on the dataset provided us with a clear perspective of which area of crime we want to focus more on for our predictive model. From the business statistics and the figures above, we were able to deduce that theft is a big problem for businesses (small and big) and the citizens of Los Angeles. In our prediction model we have combined crimes that implicitly and explicitly fall under the general crime of theft and have used the aggregate data to formulate our predictions.

Classification problem

Since our predictive model requires us to predict the shift of a day on which theft might occur at a particular location, at a particular time of the week or the time of the year, our inherent problem deals with classification. For our classification problem, the primary algorithm that we have used is gradient boosting. Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees built sequentially on the whole dataset to create a strong learner.

For the prediction problem we chose our input variables as area id (which area in Los Angeles), crime code (different thefts), premise code and day of the year (100^{th} or 150^{th} day) and for our output variable, we are predicting on what shift of the day (Morning, Evening or Night) is the possibility of theft occurring in that area. Using gradient boosting on our predictive model we found the accuracy of theft occurring on a particular shift of a day to be in the range of 56 – 61%. What this means is that when a person or an entity using this prediction model inputs the variables, he/she will get the shift of a day as an output with a surety of approx. 56% that theft would occur on that shift at that particular day.

Model relation to the business problem

As we discussed above in the business problem section, businesses which include small and high-end retailer's loss due to theft for the past 2 years in the city of Los Angeles have been in the range of $100-150 million dollars. We believe that this amount can be greatly reduced by implementing our team's proposed idea.

Since the availability of the monetary data is limited to a general amount of money lost by businesses during the span of 2 years ($100-150 million), for the cost benefit analysis, we have come up with the approximated cost of implementing our solution and the benefits that the businesses would reap from our solution. Taking a small example of shoplifting from the theft categories, we will calculate the approximate loss faced by businesses in the city of Los Angeles. From the National Retail Security Survey, the average cost of 1 shoplifting incident was found to approximately $800. From our Los Angeles crime dataset, the total shoplifting incidents for the years 2017 and 2018 was 11,846.

Approximate loss to businesses due to shoplifting (2017 & 2018) is

$$Total\ incidents * Avg\ cost\ per\ incident = 11{,}846 * 800 = \$9{,}476{,}800$$

Since the model has not been implemented yet, there are no practical values to verify how much would the reduction in shoplifting be. Hence, we chose a conservative number of 20% reduction in shoplifting crime due to the implementation of our model and the following calculations are done.

Shoplifting incidents for the year 2019 would be

$$(1 - 0.2) * 6500 = 5{,}200$$

Approximate loss to businesses due to shoplifting for year 2019

$$Total\ incidents * Avg\ cost\ per\ incident = 5{,}200 * 800 = \$4{,}160{,}000$$

Money saved from the implementation of the model

$$9{,}476{,}800 - 4{,}160{,}000 = \$5{,}316{,}800$$

This means that on an average, businesses in the city of Los Angeles could save approximately 5.4 million dollars from shoplifting crime for year 2019 with the implementation of the predictive model. This example can also be related to other crimes in the theft category thereby increasing the potential to save money for the businesses. Not only that, saving money for the businesses would mean an increase in the revenues of the security companies that are hired to protect these businesses. Depending on the crime that has the highest possibility of occurring at a particular location, a business at that location should be well prepared in advance to prevent it from occurring. The business can do this by alerting the security company that it hires to patrol its grounds and stores to employ more personnel on that day, be more attentive to the CCTV cameras and increase the number of employees for that shift.

CHALLENGES

One of the biggest challenge governments and communities face to achieve smart city goals is dealing with enormous amount of data available. While historical data is generally helpful in prediction modeling problems, in some cases the data needs to update continuously to assess the current trends of the problem that is to be identified. Taking the example of crime prediction, the more historical data you

have, the harder it becomes to create an accurate model. This can be attested to the fact that as time progresses, criminals become much more smarter and try to find innovative ways to skirt the system. Using historical data in this case to predict current trends would lead to inaccurate results as the data is no more relevant to the current trends/patterns of crimes being committed. So, there is a constant need for adaptation and update of the model based on the current trends. Another challenge one might face is the genuinity of the data itself. Is the data correct or has there been any wrong inputs in the dataset? This will affect the predictions of the model and will give incorrect results.

RECOMMENDATIONS

With the world getting smarter, crime detection should also be made intelligent so that it can analyze wide data pattern and lead to better decision making and optimized actions. There are lots of areas where chances of improvement are very high. Community leaders could develop proactive strategies in reducing crime and re-deploy law enforcement resources in places and during times of greatest crime risk. This could be achieved to some extent with the implementation of our model. Enterprises and businesses can help in providing video footages of the crime scenes, which can be analyzed using real time video analysis to proactively prevent the crimes in vulnerable areas. Data from people's survey report or reviews can be processed to determine the public expectations. Police patrolling could be increased to high risked areas where crimes are frequent on a particular day to prevent the crime from occurring beforehand. This model can be converted to a flexible system wherein users can contribute and provide data related to crimes. Instead of a centralized system, this predictive model could be developed into a UI interface like IOS/Android application that can be used by general public apart from the targeted businesses.

CONCLUSION

Mining of data, cleaning it and then performing exploratory analysis on it provides us with an array of solutions to tackle a given problem. With the access to large amounts of data and with increasingly smarter statistical analysis, it has enabled us to foresee and percept various types of criminal acts at a particular time and location, before they even occur. Intelligent Traffic Operations and Prediction is just one of many solutions for Smarter Cities that continue this tradition, providing real solutions that can facilitate sustainable growth and offer a robust foundation for building a smarter city. By making cities more instrumented, integrated and intelligent, we can help government and city officials to meet and exceed their citizens expectations through innovation. With the research currently being done on data analytics for prediction modelling**,** we believe that data mining for crime prediction has a promising future for increasing the effectiveness and efficiency of criminal and intelligence analysis.

References

- https://data.lacity.org/A-Safe-City/Crime-Data-from-2010-to-Present/y8tr-7khq
- https://github.com/ddhaval04/Smart-Cityhttps://www.shopify.ca/retail/119922883-retailers-lose-45-billion-annually-to-theft-heres-how-signage-can-offset-that
- https://losspreventionmedia.com/insider/shoplifting-organized-retail-crime/shedding-light-on-retail-theft-statistics/
- https://www.kaggle.com/aceconhielo/london-crime-data-analysis
- https://towardsdatascience.com/analyzing-crime-with-python-8b28252559ce
- https://stackoverflow.com/questions/44474570/sklearn-label-encoding-multiple-columns-pandas-dataframe/46290943
- https://machinelearningmastery.com/implement-random-forest-scratch-python/
- https://en.wikipedia.org/wiki/Crime_in_Los_Angeles
- https://github.com/Vidyashree22/Chicago-Crime-Data-Analysis

Appendix

The graph below shows the predicted shift of the day for the test data.

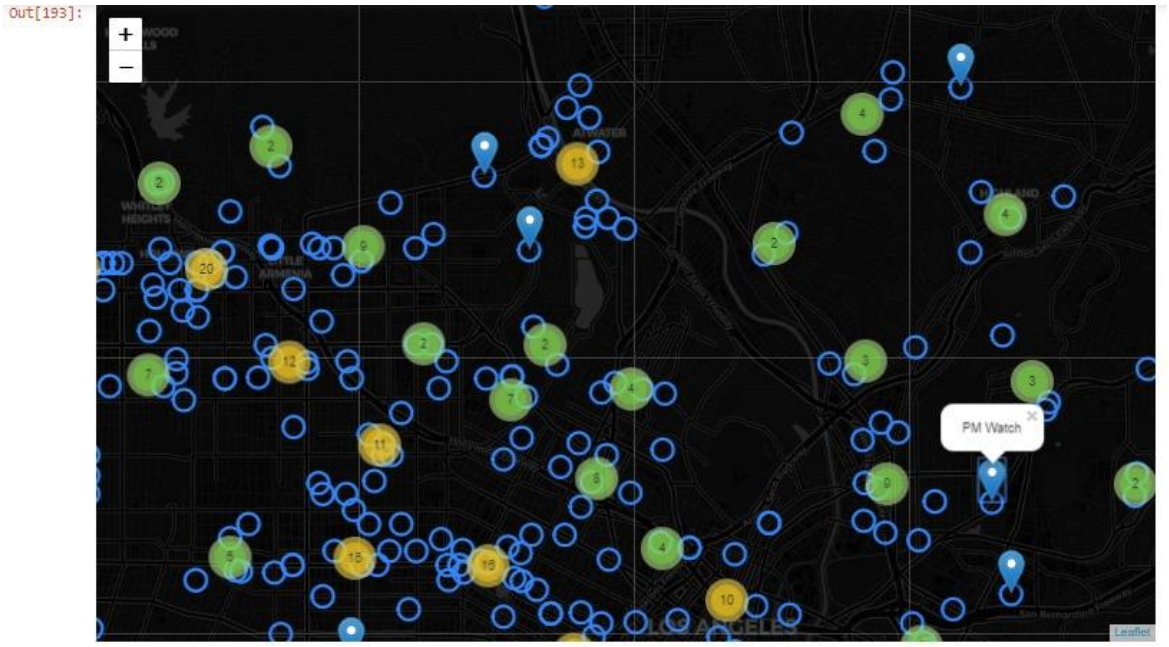


Figure 3: Predicted shift of the day for test data

The following graph below shows distribution of crime according to time distribution.

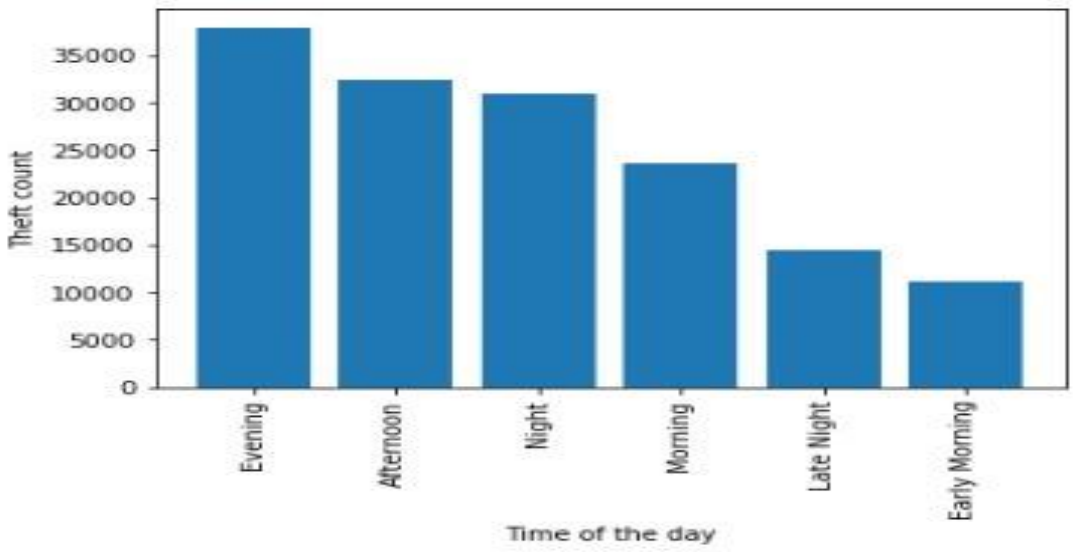
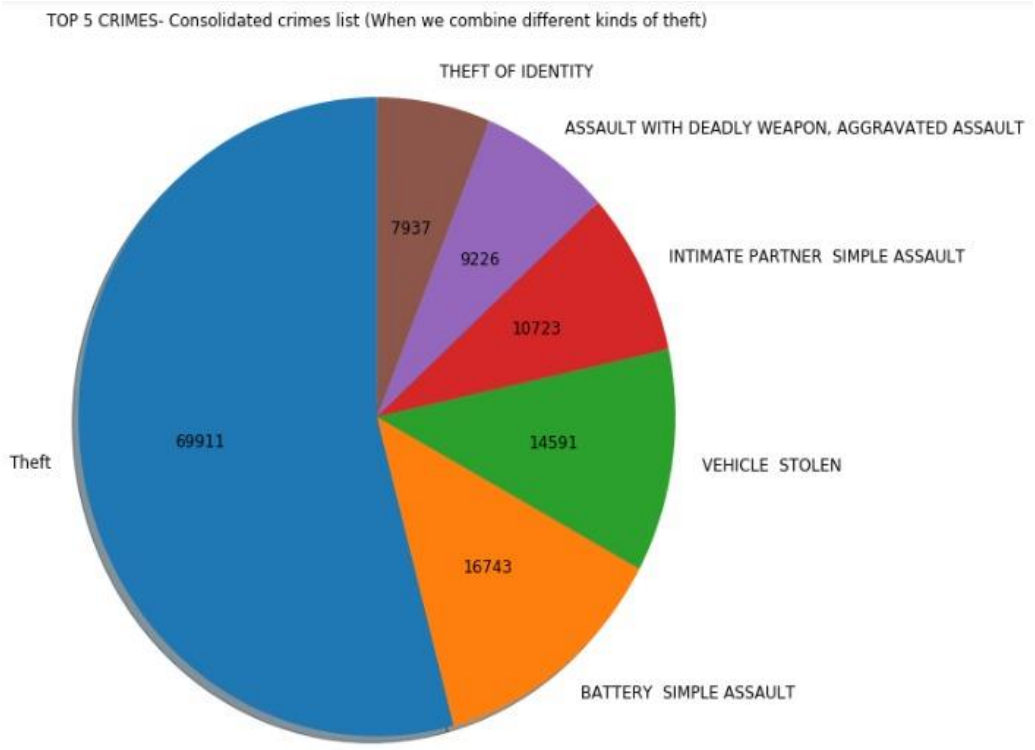
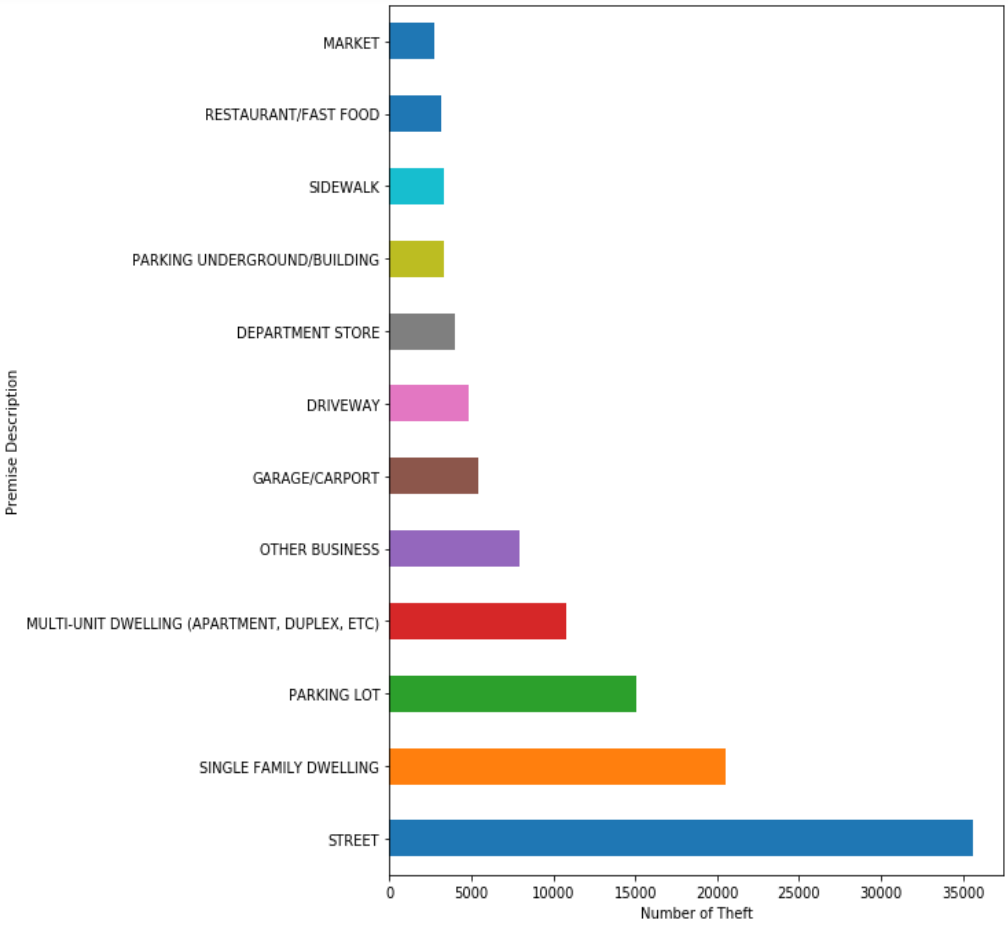


Figure 4: Distribution of crime according to the shifts



This graph shows the top 5 crimes in Los Angeles. Here the crimes that come under the category of theft are combined together and shown above

The map below is the representation of Los Angeles which is superimposed by the occurrences of theft for the past 2 years. Just to clarify the green color represents the spots where theft occurred.



The following graph above shows the distribution of theft according to the premise where it had happened.

The following table shows the crime code description for the particular crime code.

Description of theft crime codes	
Crime Code	**Crime Description**
310	BURGLARY
320	BURGLARY, ATTEMPTED
440	THEFT PLAIN - PETTY ($950 & UNDER)
442	SHOPLIFTING - PETTY THEFT ($950 & UNDER)
420	THEFT FROM MOTOR VEHICLE - PETTY ($950 & UNDER)
341	THEFT-GRAND ($950.01 & OVER)EXCPT,GUNS,FOWL,LIVESTK,PROD0036
320	BURGLARY, ATTEMPTED
330	BURGLARY FROM VEHICLE
331	THEFT FROM MOTOR VEHICLE - GRAND ($400 AND OVER)
343	SHOPLIFTING-GRAND THEFT ($950.01 & OVER)
480	BIKE - STOLEN
410	BURGLARY FROM VEHICLE, ATTEMPTED
421	THEFT FROM MOTOR VEHICLE - ATTEMPT
441	THEFT PLAIN - ATTEMPT
443	SHOPLIFTING - ATTEMPT
471	TILL TAP - PETTY ($950 & UNDER)
474	THEFT, COIN MACHINE - PETTY ($950 & UNDER)
473	THEFT, COIN MACHINE - GRAND ($950.01 & OVER)
485	BIKE - ATTEMPTED STOLEN
472	TILL TAP - ATTEMPT
470	TILL TAP - GRAND THEFT ($950.01 & OVER)
452	PICKPOCKET, ATTEMPT