Forecasting Crime Rates on the Chicago Dataset

Ruchira R Vadiraj
B.Tech ,Computer Science and Engineering
PES University, Bangalore, Karnataka
ruchvad@gmail.com

Nikhil K R
B.Tech, Computer Science and Engineering
PES University, Bangalore, Karnataka
nikhilkramesh1999@gmail.com

Roshan Daivajna
B.Tech,Computer Science and Engineering
PES University, Bangalore, Karnataka
roshan.daivajna@gmail.com

***Abstract*—Burglary is one of the most common crimes. The variation in type of crime and the place where the crime is committed is important to the law enforcement to analyze and understand to impose measures and steps to curb or stop the perpetrators as quickly as possible. Here, we take a look at how burglary impacts the society and try to forecast and predict how this crime fluctuates through the year.**

Keywords—burglary, crime, forecasting, ARIMA

I. INTRODUCTION

Prediction markets have been proposed for a variety of public policy purposes, but no one has considered their application in perhaps the most obvious policy area: crime. This paper proposes and examines the use of prediction markets to forecast crime rates and the impact on crime from changes to crime policy, such as resource allocation, policing strategies, sentencing, postconviction treatment, and so on. Most of the approaches underestimate the market trends and it's relation with crime and vice versa. No police department or other law enforcement agency has deployed any of the emergent forecasting models being developed by academic criminologists. But on the other hand, the investments are rather scarce in communities tarnished by criminal activity due to low cash flow and underdevelopment due to insufficient funding. One way to change face in this regard is by formulating and analyzing models and helping the police by providing sufficient evidence of the working of such models which in turn is stimulates growth and drastically help the community to fast track itself from the underpinned position it is in with illegal drug use and high school dropout rates to a more balanced side which will then stimulate market growth by bettering it's cashflow.

II. IMPORTANCE

A. Betterment of Society

"What makes a good society is sound economy. Without it all the rest falls apart."- Llewellyn Rockwell Jr. Economy brings in companies and companies brings in employment which in turn stimulates growth and development in education, policing, public spaces and so on.

The necessity for reducing crime rates is instrumental for the upliftment of society[1]. When a child is born, he or she normally has parents and extended family members who seek to raise them with the necessary values and morals needed in order to develop a good character for the future. The family's behavior should be a prime and positive example of what the child should be striving to emulate. But, due to the fact that we do not live in an ideal world, this is not always possible. Influence of the present generation infiltrates the minds of the youth which deeply depreciates the chance for any improvement of the community in various aspects, mainly education and urban development.

B. Use of Forecasting

The practice of forecasting is as old as crime regulation and policing, since all policy makers, from the governor to local police chiefs, must make decisions about how to allocate scarce resources. In the absence of specific forecasting tools, the most likely method of predicting crime is human-based pattern recognition and the gut instincts of decision makers. In other words, police chiefs have a feeling about where crimes will occur or the governor has a prior about crime patterns or the impact of tools like the death penalty or parole, and these are used to make policy decisions.

This kind of decision making may be successful in some instances given experienced decision makers and some ex post political accountability checks, but undoubtedly these techniques will be crude, subject to political biases, and possibly systematically skewed by decision making heuristics and errors by well-meaning decision makers.

The lack of systematic forecasting may also have to do with something we might call "politics" or "public choice". There are many possible explanations that fit under this rubric: it may be politically difficult to justify (to colleagues, constituents, or subordinates) an investment in forecasting over, say, another cop on the beat; police unions or police management may resist changes that are socially efficient but have high private costs for them; certain political constituencies might not want crime predictions, especially for certain areas of a city or state, to be publicly available in such a conspicuous form; decision makers too might not want information on crime in particular neighbourhoods to be forecast, since this might scare off developers or new residents, and might discourage current residents by revealing that current policies are failing despite (or because of) city efforts; the whole concept of betting on crime might be normatively troubling to some citizens; and so on.

III. EXPLORED FRONTIERS

A. Spatio-temporal prediction

The paper's objective is to predict crime levels for different types of crimes and a given year based on previous years criminal activity and other social aspects. It states that augmenting the crime data with other social aspects such as education, and economic conditions give few insights for

predicting the crime and improve the quality of prediction. Also, social aspects of other similar communities give out critical information for predicting the crime pattern for next year.[2].

A critical component of the proposed method by the above is to fuse various types of social and historical information in a network which helps to find relationships between different communities within the city. The extracted relationships are incorporated in our prediction model.

Although market trends aren't enhanced in this paper, the spatial and temporal importance that is must be incorporated for predicting market forecasts must be reinstated here.

Fig 1. Schematic representation of the effects and interlaps between communities [2]

One of the promising research direction is fusing social network data that can provide socio-behaviour "signals" for crime prediction. This is one of the shortcomings in the paper, which fails to incorporate the "actual" data which is mainly on social media.

B. Data Mining – Various Models

Several techniques have been proposed in the recent years for solving a problem of extracting knowledge from explosive data adopting different algorithms[3]. One of such applications is that of finding knowledge of criminal behaviour from its historical data by studying the frequency of occurring incidents. P.Thongtae [4] studied and gave a comprehensive survey of effective methods on data mining for crime data analysis.

The paper concludes with Random Forest Classifier giving the most balanced results with respect to accuracy, precision, recall and F1 score out of three models (NAÏVE BAYES CLASSIFICATION, RANDOM FOREST CLASSIFICATION, DECISION TREES) for prediction of 'Per Capita Violent Crimes' feature. While Linear Regression gave the lowest values in these performance measures, the data could not fit well to the straight line considered using target and remaining features. only with units.

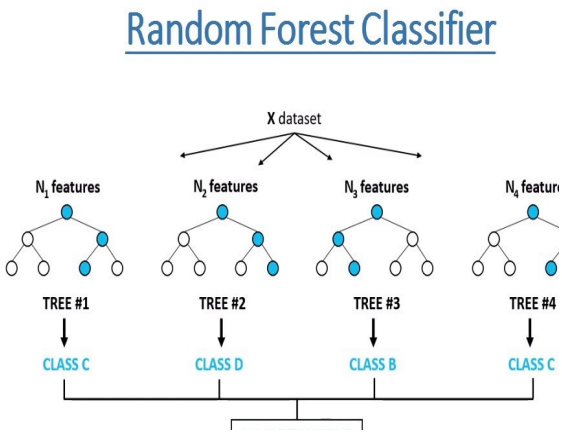

Fig 2. Random Forest Classifier

The paper concludes that reduction of overfitting using cross validation improves performance by enough training and testing samples that seemed to help in this analysis by giving correct and consistent performance measures. These predicted features will be useful for the Police Department to utilize their resources efficiently and take appropriate actions to reduce criminal activities in the society.

However, the paper doesn't consider the variation in types of crimes and their differences, only the overall crime activity is explored.

C. Recurrent Neural Network

This paper follows systematic approach for using deep learning neural networks to analyze crime data. By proposing this new framework, we aim to be able to leverage the deep learning neural network for prediction of crime rates and possible hotspots and for proactive policing and prevention measures. It proposes a generic framework for the day to day crime handling and analysis for the police personnel. [5]

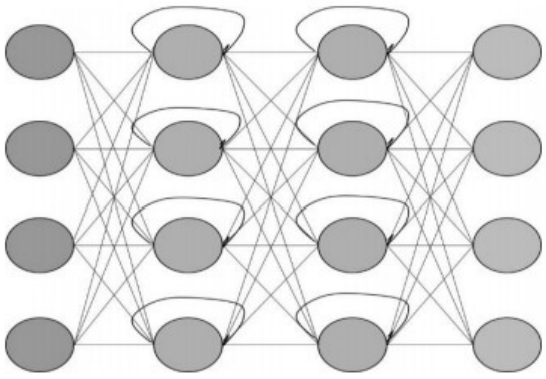

Fig 3.Deep Recurrent Neural Network [5]

The model proposes an interesting way of data acquisition, where documenting evidence and other related information that may not be available to the general public,

is used in their analysis. This might not be feasible and therefore this approach is not enforced.

Natural Language processing is used to convert texts to regional languages which might help in multi-lingual countries where policing is enforced regional rather than state/federal.

The model proposed uses classification of independent crime related variables, forgoing the dependent ones. Furthermore, clustering/matching using Deep Neural Networks is instated, using MO and helps in predicting future offenders if they match a given MO. This mainly helps in finding criminals who haven't been caught yet or who've escaped, and does very little in predicting the right areas that must be explored for reducing criminal activity in the long term.[6]

IV. PROBLEM STATEMENT

The problem statement formulated for this report is the finding the differences in criminal activity during a normal year, a pandemic year and a year hit with recession. The results can help the police force employ strategic responses beforehand to curb criminal activity. The importance of economic activity with spatial and temporal differences and the different models that can be explored is what is learnt so far by the two research papers analysed.

In this report, for performing predictive analysis, the Communities and Crime dataset from UCI repository [6] has been used which consists of crime data in Chicago, a city having highest crime rate in the United States of America. It includes features affecting crime rate like population, race, sex, immigrants etc. Many features involving the community like, such as the percent of the population considered urban, and the median family income, and involving law enforcement, such as per capita number of police officers, and percent of officers assigned to drug units are included so that algorithms that select or learn weights for attributes could be tested [6]. The attribute or feature to be predicted is 'Per Capita Violent Crimes' which was pre-calculated in the data using population and the sum of crime variables considered violent crimes in the United States: murder, rape, robbery, and assault.

The reason for doing predictive analysis is that existing tools may simply be insufficient to provide meaningful results. Technical forecasting, using futuristic models, computer technology, and mapping tools, is a modern phenomenon, and these methods are proven, occasionally difficult to interpret, and occasionally expensive to set up and operate, especially for communities with tight budgets.

The differences with other predictive models proposed seem to revolve around a particular year or a range of 3-4 years. The variedness in crime activity and the range of differences that emerge in the growing decades hasn't been explicitly stated. Exploiting and tapping into this resource may give a wider view on how policing is enforced. The Black Lives Matter movement voiced it's response on policing measures taken which are reflected to be racially biased. This revolution made us take a step back and analyze what wrong and what can be done to fix it.

A. Challenges

The proposed model that is going to be enforced is forecasting. Forecasting usually has a strong stance when used with stock market prediction, materials requirements planning (MRP), warehouse capacity, supply and demand variations and so on. Explicit usage of forecasting for crime analysis is infrequent and is argued to be potent if applied seamlessly.

The challenges involved in predicting crime rates or the impact of different crime policies are very similar to those in other forecasting domains. Classic examples include predicting: sales of a product, changes in interest rates, the likelihood of a terrorist attack, or the outcome of political elections. In each of these cases, the inputs needed to generate a reliable forecast may be skewed by a variety of factors, some of which might be benign but disruptive, and some of which may be selfish and opportunistic. These include the fact that relevant information may be quite dispersed; true experts may be difficult to discern from over-confident charlatans; a wide variety of alternative models may exist, but no obvious "best" model may prevail; political or social concerns may lead some to misrepresent their information; decision making heuristics and barriers to information flow to key decision makers may inhibit analysis; and there may be few incentives for uncovering new information and developing or identifying improved models. Given the similarities in the forecasting problems across these varied areas, it is not surprising that a recent forecasting innovation—prediction markets—have been used in each of these cases

A new forecasting tool—crime forecasting prediction markets—that may be more affordable, accessible, and accurate for the relevant decision makers. Two models can be explored : the first is a simple crime rate prediction model that is similar to those being used currently to predict everything from the weather to political elections to flu outbreaks; the second are various contingent markets or prediction market event studies that can be used to inform policy decisions on subjects like sentencing, the death penalty, the use of surveillance technology, and so on.

Instead of decision makers relying on certain individuals, a certain theory, or certain information, they can rely on a forecasting-based aggregation of available information.

Crime predictions occur, as they must, within every public safety agency at every level of government. Every agency from the local sheriff's office to the FBI must make forecasts about how much crime and how much of each particular crime is likely to occur in the future. These forecasts help in determining the amount of crime fighting resources needed and how they should be allocated across the jurisdiction. Our review of current practices reveals that

there is an abundance of tools and methods for forecasting, but these are rarely if ever used.

Substantial scholarship exists within the field of criminology discussing methods of forecasting future trends in criminal activity. Existing prediction mechanisms, however, tend to focus on extrapolating from data collected on past crimes, rather than attempting to aggregate information from crime professionals on estimates of future developments. To be sure, the past can be used to help predict the future, but the existing methods for doing this are not commonly used, let alone systematic or theoretically sound.

There are some relatively new forecasting models related to crime trends. These models use multiple-regression analysis to determine the statistical significance of different factors in determining the crime rate. A pioneer in this field is James Alan Fox who, in his book Forecasting Crime Data, attempts to predict future crime trends using a range of variables, including violent crime rate, property crime rate, size of police force, police force expenditure, unemployment rate, consumer price index, and estimates of the resident population by race and by age. The models are complex and the results ambiguous. In their book Is Crime Predictable?, Carolyn Block and Sheryl Knight attempt to predict future trends in specific types of crime based on data gathered from past criminal activity taking place in the Chicago-land area. The predictive accuracy of their model varied widely depending on the type of crime in question.

Here, we take a look on how burglary fluctuates through a span of almost 20 years and try and predict the rate at which it can vary by using complex forecasting models. ARIMA is used as the model to build and drive the data forward on full analysis of the data in hand.

V. Aggregating and Refining

We first try to find the right parameters that are supposed to be given in to the model with the data provided by the Chicago Police Department. We first, split the input timestamp in the original dataset to Year, Month and Day and discard the time. Further, we refine the dataset to only contain Burglary crimes.

The data is cleaned through by simply ignoring rows which contain insufficient data as many of the parameters at hand are categorical, we choose this pathway, which might lead to further anomalies that are discussed later.

No standardization or normalization is required as the data at hand has to be transformed and worked upon without the use of any input numerical data provided in the dataset.

The data is then encompassed with only the Year, Month and the aggregate sum of all the burglary crimes committed in that particular month.

VI. Exploratory Data Analysis

With the data in hand, we now move on to exploring the varying trends that the data can tell us.

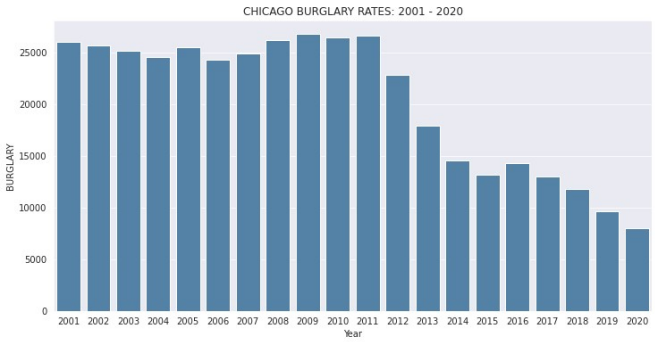

Fig 4. Burglary rates throughout the years

As seen in Fig.4, we can make notes about the trends seen. The most eye-catching part of the given graph is the highest rate of burglary seen in the year 2009. The repercussions of the 2008 recession which led to staggering rates of unemployment and low rates of cash flow, might have an influence in what is seen. The decent improvement from 2006 to 2009 might be due to the fact that savings were usually kept in safes at home and people generally avoided the banks due to the high risk it entailed at that time. As and when the economy started to pick up after 2009 and people realized the importance of protection by using entry cameras, front door traps we see the gradual decline in the rates of the crime committed and we can only make the assumption that the other types of crimes were enhanced by this.

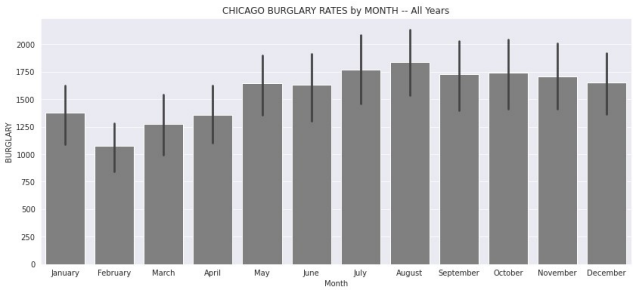

Fig 5. Burglary rates by month

The above figure, Figure.5, gives a broader perspective on when the crime might be committed, meaning, when do the burglaries plan and loot. We see that the highest rates are seen in the month of August, which can only be assumed is caused by unguarded houses during the summer break. The month of February is the safest the houses are from break-ins. The standard pause in any variations is seen from the month of September to December.

These two figures, help us gain useful insights and try and analyze the right model to be used to forecast the given data.

VII. ARIMA

The data now is check for stationarity and is passed through the usual checks with the standard, simple methods of forecasting used to analyse stock data. The mean, naïve and the seasonal naïve methods were first used to see how the data fits through. We see high rates of MAPE, MPE and ACF1 in these simple methods which makes us incline to jump over to using complex models which don't just operate on mean and historical data.

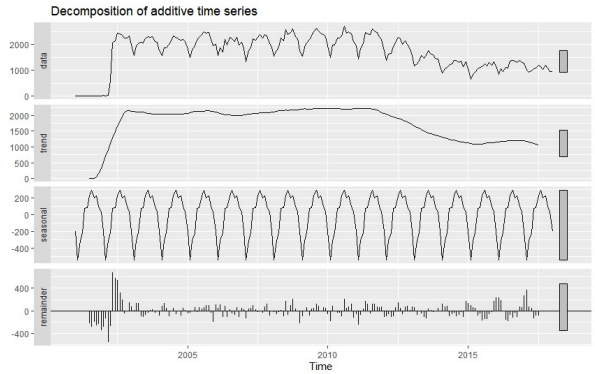

Fig 6. Decomposition of the time series data

On further analysis, we see in Fig. 6, the additive time series shows us that there is little to no trend seen and the seasonality component remains almost the same throughout the time frame. The remainder/residuals help us see the randomness the data has.
We know that the MA(moving average) models usually don't handle trend and seasonality well so we see high error rates using these models as well.

The Holt method is redundant as there is no trend component seen in the additive decomposition done. The Holt-Winters' method is skipped as it only applies to time series data with both trend and seasonality.

We therefore, explore the fields of complex models. The ARIMA model with steps of h as 24, (as the time series here is across years) gives us a good estimate and acceptable error rates. The MAPE and the ACF1 which are mainly used to understand the fitness of the data to the model, have very low values which indicates that the data is in fact par with the model.

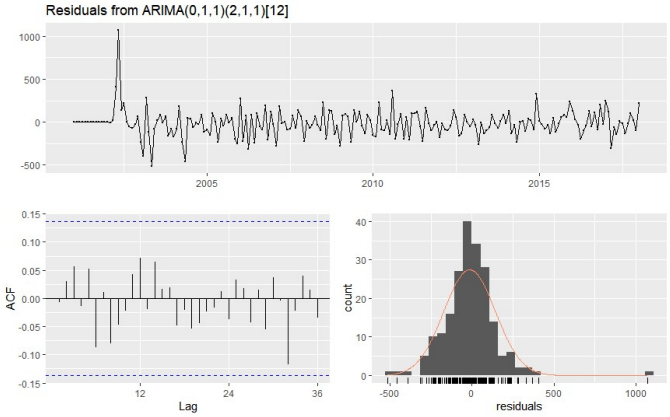

Fig 7. ACF and Residual plots

The above figure indicates that the residuals are normally distributed with a constant variance, apart from the one spike at the start. The values of p, d, q are set to 0, 1, 1 in order of trend AR, trend difference and the trend MA. The values of P, D, Q are set to 2,1,1 which correspond to the seasonal parameters.

The MAPE, ACF1 values are relatively low which indicates that the model is a good fit.

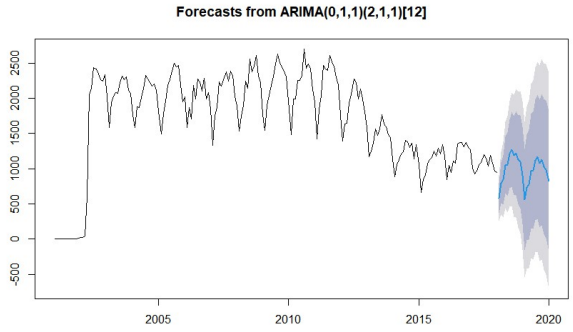

Fig 8. Forecasts from ARIMA

The forecasts suggest us that the predicted values are accurate. The above figure verifies the same. Therefore, we can strongly concur that ARIMA is the best model to analyze the burglary time series data.

VIII. CONCLUSION

From the above data collected and analyzed we can safely conclude that the best fit model is ARIMA. The data collected in section VI and VII help us identify the necessary steps that the law enforcement can take to help aid the community better and keep the society safe. The data analysis comes with caveats however, as the ignored rows due to insufficient values on further investigations acquiring those values can change how the model behaves. On frequent updation of the Chicago Crimes dataset, the

changes seen or observed may vary from what is concluded in this report. The model here has taken in account the randomness that may exists or may come to existance but drastic changes seen in the incoming data might influence how the model behaves, making us repeat the process all over again from the start. Future extensions may include GIS data to find the locations and impacts of the crime and the different patterns observed throughout the years over different months. Incorporating clustering and forecasting might be the next steps in increasing the effective usage of the given data.

IX. Acknowledgment

We thank Prof. Bharathi R in guiding us complete this report. The insights and understandings seen in this report is due the hard work of all the contributors mentioned at the beginning of the report. Data Acquisition and Data Preprocessing was handled by Mr. Nikhil KR. Exploratory data analysis and data model selection was done by Mr. Roshan Daivajna and Mr. Ruchira R Vadiraj. The final model selection with trial and error was done by all the contributors.

References

[1] Mariana MITRA, 2015. "**The Necessity And Importance Of Preventing Crime In Contemporary Society**," Management Intercultural, Romanian Foundation for Business Intelligence, Editorial Department, issue 33, pages 217-223, June.

[2] Dash, Saroj & Safro, Ilya & Srinivasamurthy, Ravisutha. (2018). Spatio-temporal prediction of crimes using network analytic approach. 10.1109/BigData.2018.8622041.

[3] Prajakta R. Yerpude, Vaishnavi V, Gudur,International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.7, No.4, July 2017 K. Elissa, “Title of paper if known,” unpublished.

[4] P.Thongtae and S.Srisuk, “An analysis of data mining applications in crime domain”, IEEE 8th International Conference on Computer and IT Workshops, 2008

[5] Lydia J Gnanasigamani, Seetha Hari, INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 8, ISSUE 11, NOVEMBER 2019

[6] UCI Repository Communities and Crime dataset, Retrieved from http://archive.ics.uci.edu/ml/datasets/communities+and+crime