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 seen a huge surge in usage over the past decade and has been seen to be used in a wide variety of cases;However, we have seen a lack of development in one of the areas with most potential: Crime. Most of the approaches underestimate the market trends and it's relation with crime and vice versa. We have repeatedly seen a lacklustre use of emergent forecasting models by law enforcement agencies. 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 stimulates growth and drastically helps the community 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 investment and investment 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]. All youth look and learn from the environment they are raised in, a supporting family's actions or the actions of people in his environment should be a prime and positive example of what they should be striving to emulate; However, this means the 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. By reducing crime rates we create a safe and secure environment from which the society of the future will see a dramatic improvement in these aspects.

This improvement of safety also attracts the populous which in turn attracts investment in the area, this acts as a positive feedback loop which helps development in all fields and so, with the same logic, it follows that if crime is left unchecked in the area, the consequences can be disastrous

B. Use of Forecasting

The huge benefit of forecasting can be truly understood while looking at the scarcity of resources available to a developing area and the large competition to those resources that comes with it. Different regions and sectors must be given ample staff and equipment based on the amount of crime that is expected of that region. Giving too many resources to a particular area will case a scarcity in resources to other areas. This problem is exacerbated when it comes to Law Enforcement as over-equipping officers without proper training procedure will increase the usage of violent strategies.

Without the use of well developed forecasting techniques this resource allocation problem is left for humans to decide. This produces very inconsistent results, stern experts may be very successful in decreasing crime rate based on the areas he allocates resources to ,where as someone with not much expertise will fail to do so; However , it is obvious these methods are crude and will be subject to biases

The severe under-development of forecasting techniques may be attributed to lack of understanding of the underlying technology used and the un-intuitive nature of the problem statement. Prediction that goes against officers intuition may be brought into scepticism. There is also the possibility of a public outrage when it is published that important policy decisions are being recommended by computers. These forecasts might also scare away new potential residents from a relatively high-crime area causing lack of further progress.

Though these concerns are valid, with sufficient education and demonstration it is possible to create real progress in this field, and the benefits for doing so are tremendous. Unbiased accurate predictions of crime rate not only improve the efficiency of resources used to contain problematic areas, it also allows provides for a much safer environment which leaves more budget to be spent on research, training, and other areas to further the improvement of Society.

III. EXPLORED FRONTIERS

A. Spatio-temporal prediction

The paper's preliminary objective to predict the level of crime for each of the various types and a given year based on criminal activity in the past and other social aspects. "augmenting criminal data with other social aspects give few insights for predicting the crime and improve the quality of prediction".It also states that the socio-economic aspects of similar communities can be analysed to extract critical

information for predicting the crime pattern in the upcoming years[2].

A huge part of the proposed method above is to merge the different categories of historical information in a network, this helps find existing relationships between different communities within the city. The extracted relationships are used 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

Numerous techniques have been proposed to solve the problem of extraction of information from high velocity, voluminous data using many different algorithms[3]. One of these approaches is that of "finding knowledge of criminal behaviour from its historical data by studying the frequency of occurring incidents."- P. Thongtae [4]

The paper provides a well documented survey of different approaches on data mining for crime data analysis and their respective effectiveness. The paper checks for 3 different models: NAÏVE BAYES CLASSIFICATION, RANDOM FOREST CLASSIFICATION, DECISION TREES and tests each for their precision, accuracy, and F1 score.These models and performance measures are selected with the intention of "finding the illegal activities of professional identity fraudsters based on knowledge discovered from their own histories". It also points out different problems in applied data mining in crime control and criminal suppression. The paper finds that a Random Forest Classifier model providing the most balanced outcome for prediction of the "Per Capita Violent Crimes" feature.Whereas Linear Regression seemed to produce the lowest values in the tested performance measures, the data could not fit well to the straight line considered using target and remaining features. only with units.

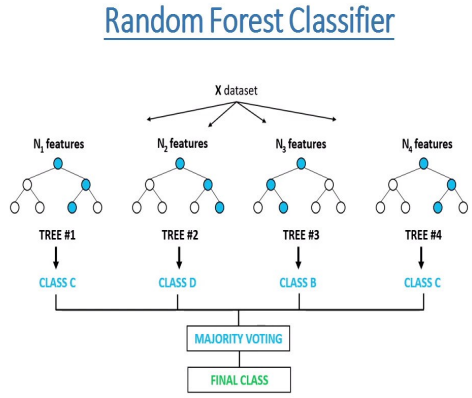

Fig 2. Random Forest Classifier

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 helpful for Law Enforcement agencies by allowing them to allocate their resources efficiently and reduce criminal activity in the area by taking appropriate actions

However, this paper acknowledges but does not consider the variation in different types of crimes and their respective severity, only the overall criminal activity is explored. ie. the number of criminal offences in a particular area.

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]

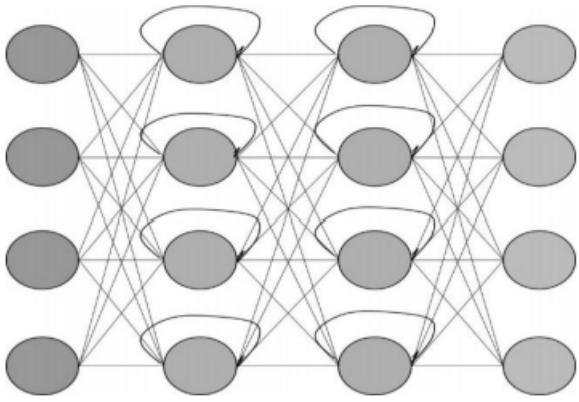

Fig 3.Deep Recurrent Neural Network [5]

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.

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 that consistently has one of the highest crime rate in the USA. It includes features that are considered to have an affect on crime rate such as population, sex, race 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 feature that is being predicted is "Per Capita Violent Crimes" which in statistics is calculated using population and the number of crime variables; the pre-calculated data considered violent crimes in the United States: murder, rape, theft, and abuse.

The explanation for predictive analysis is that there might simply be insufficient current instruments to provide meaningful results. Technical forecasting is a new concept, using futuristic models, digital technology, and spatial analysis, and these methods are proven, often difficult to understand, and sometimes costly to set up and run, particularly for tight-budget communities.

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 main challenge in generating a solid, reliable forecast is that the input data can be slanted by an assortment of factors and elements. These factors can be quite troublesome, some of which can be deeply coupled and hard to get rid of. One problem is what most forecasting models face: important, relevant data might be highly scattered, some of which might get classified as outliers. Then, there is the concern of the social and political impact of working on such a problem, which can lead to a slight degree of distortion in actual results. Lastly, there is no "best" model in this scenario. Various models can be used and must be properly tested and validated before finalizing. dynamic heuristics, boundaries to data stream and insufficient motivations become a challenge in revealing new data and creating or recognizing improved models. These challenges are not uncommon in other areas where forecasting can be applied, such as prediction of sales and elections. Given these similarities, the modern approach of forecasting prediction markets can be utilized in every one of these cases.

A new forecasting tool—crime forecasting prediction markets—that might be more reasonable, open, and precise for the applicable chiefs Two models can be explored : the first is a straightforward crime percentage expectation model that is like those being utilized at present to foresee everything from the climate to political races to influenza episodes; the second are different unexpected business sectors or forecast market occasion considers that can be utilized to advise strategy choices on subjects like condemning capital punishment, the utilization of surveillance innovation, etc.

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 need to, within every public safety agency at every level of presidency. Every agency from the local sheriff's office to the FBI must make forecasts about what proportion of crime and the way much of every particular crime is probably going to occur within the future. These forecasts help in determining the quantity of crime fighting resources needed and the way they must be allocated across the jurisdiction. Our review of current practices reveals that there's an abundance of tools and methods for forecasting, but these are rarely if ever used.

Within the field of criminology, there is extensive literature addressing methods of predicting future patterns in criminal activity. However, current prediction models appear to rely on extrapolating from data gathered on past crimes rather than seeking to aggregate knowledge on estimates of potential events from crime professionals. To be sure, the past can be used to help forecast the future, but there is no widespread use of the current methods to do this.

In relation to crime patterns, there are several relatively recent forecasting models. To assess the statistical importance of various variables in evaluating the crime rate, these models use multiple-regression analysis. A leader in this area is James Alan Fox, who tries to "predict future crime patterns using a number of variables" in his book "Forecasting Crime Statistics", including the "rate of violent

crime, property crime rate, police force size, police force budget, unemployment rate, and consumer price index". 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." Depending on the type of crime that was committed, the predictive performance of their model differed significantly.

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.

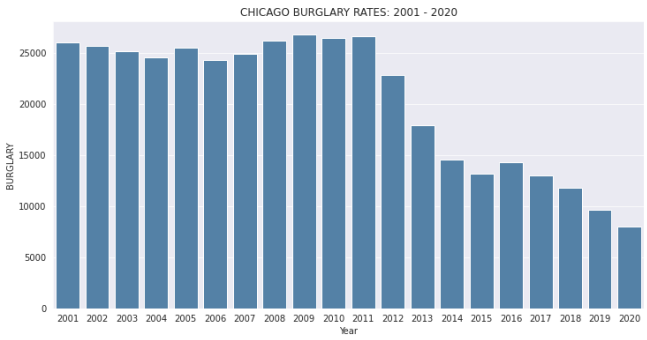

Fig 4. Burglary rates throughout the years

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.

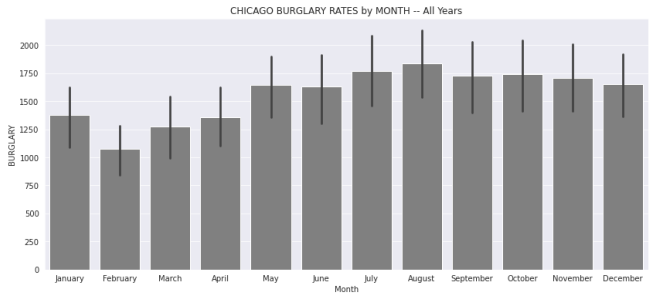

Fig 5. Burglary rates by month

The above figure, Figure.5, gives a broader perspective on when crime is likely to be committed. 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.

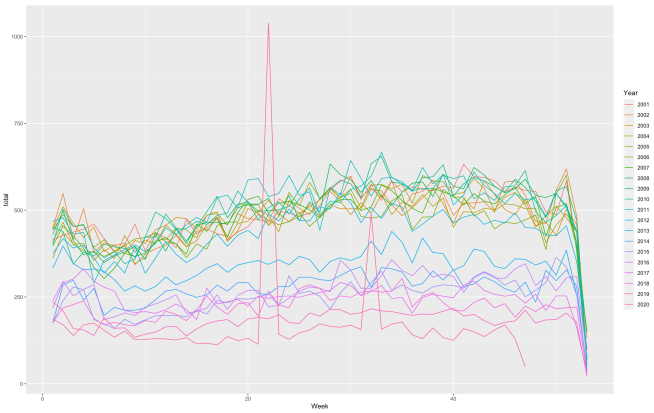

Fig 5-1. Burglary rates throughout the years

In the above figure we can see that burglary reports have seen a noticeable decrease after 2014, this may be due to higher security standards, improvement of social welfare fund , or changes in reporting. We do, however, see a sharp increase in the number of cases in the 23rd week of 2020. This is likely attributed to the BLM protests that were happening around USA during that period of time and the huge number of reports that came along with it. Even including this extremity, 2020 has seen record low levels of burglary reports so far, this is likely due to movement restrictions following the pandemic which led to a lot of people staying at home and thus reducing the opportunity of a burglary taking place.

These three 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.

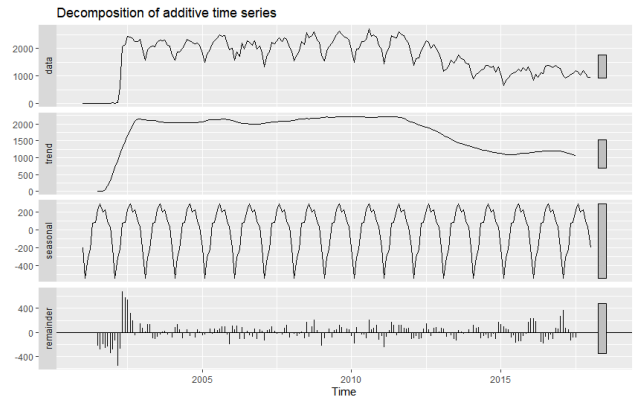

Fig 6. Decomposition of the time series data

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.

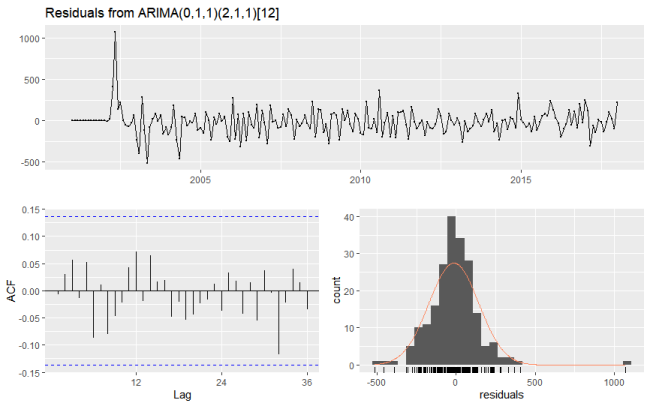

Fig 7. ACF and Residual plots

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.

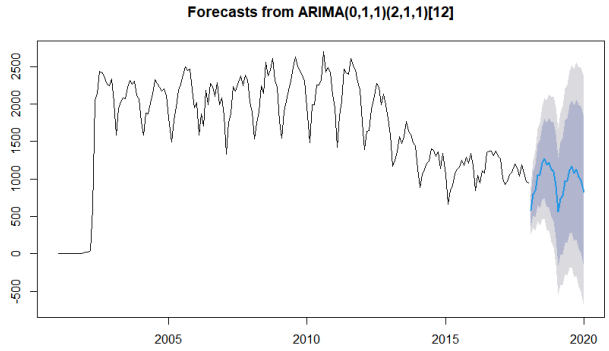

Fig 8. Forecasts from ARIMA

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