PROJECT CAP - Crime Analysis & Prediction

Author: Frank R. Serafine, December 2021

INTRODUCTION

Given the statement, "*We live in a time of high crime*," what is your gut reaction? Whether you agree or disagree, it is unlikely to be a scientifically informed snap judgment. We commonly intuit our sentiment one way or another on statements like this all the time, but who is truly responsible for the accuracy of such a statement? Law enforcement agencies have the highest stake in such a statement due to their responsibility to public safety, but free societies do not rely solely on such an arm of local, state, or national government.

Herein lies the business use case of having an informed answer to this question: The fourth estate (the media) has a vested interest in cross-checking any person, agency, or paradigm and informing the populace without bias. Journalists have a long history of employing the highest-tech methods available to do their work, and machine learning algorithms should be used as one tool toward this end.

My project aims to showcase machine learning modeling of time series data as a lens to contextualize crime. The best attempts by others at using machine learning models to specifically predict crime are no more accurate than a coin flip at their best, so I began this project assuming that my own models would likely be wrong. I believe that the measure of their wrongness, however, is a valid and accurate heuristic for knowing the truth value of such a statement, and that is the crux of my goal – to classify periods of criminal activity.

DATA SOURCE

The United States Federal Bureau of Investigation (FBI) employs the National Incident-Based Reporting System (NIBRS), a central database for retaining and analyzing crime, and encourages law enforcement agencies to report their incidents for this purpose. Data from this system is made publicly available via the agency's Crime Data Explorer website in both CSV and visual form by state and year. For my project, I used a ten-year span (2010-2019) of incident data from a large city in Texas that consistently reported their data to the NIBRS – **Fort Worth**.

PREPROCESSING

The amount of information was vast (43 CSV files of data per year), but I began by creating a SQLite database, importing the first year's data into tables, and assessing which spreadsheets were absolutely critical for the following years' imports. This ended up being 20 code lookup spreadsheets and 9 dynamic data spreadsheets.

Using strategic SQL JOINs, I made a table from a custom view unifying the most relevant information from the scattered imported tables. I further cleaned the data for consistency and Jupyter notebook/Pandas DataFrame compatibility (see 02. DBMods-FtWorth.sql). At this point, I imported the table into a Pandas DataFrame in a Jupyter notebook (03. CrimePred.ipynb) to begin analysis and modeling. For a full recounting of the extensive preprocessing and beginning EDA, please reference 01. EDA.md.

DECOMPOSITION

Although my data contained mostly non-numeric categorical information, I transformed all of the non-numeric categorical features into numeric values using one-hot-encoding and aggregated them into daily totals, using the dates as an index. In order to view the daily totals as meaningful plots across time, I fed the data into the Seasonal and Trend decomposition using Loess (STL) algorithm to split it into three views: Trend, Seasonal, and Residual.

The **trend view** showed consistent yearly patterns of rise and fall, decreasing slightly overall from 2010 through 2018, with small weekly peaks and troughs throughout, and a sharp decline in late 2018 followed by an unusual rise in crime levels following. The **seasonal view** provided a look at the levels of variation in the weekly patterns, showing that crime levels rarely remained constant for more than a month, with 2010-2012 having higher volatility than other years. The **residuals** that didn't fall into the other views showed an even occurrence and intensity in their deviations from the bulk of the data, but showed relative volatility in 2010 and 2011, as well as late 2013 and early 2014.

Together with a partial autocorrelation plot identifying time lags 1-7 as important, this analysis helped me decide on the algorithms I would use for prediction – a **SARIMAX** model would capture the robust consistent weekly patterns, a **Prophet** model would respond well to the multiple seasonalities observed, and a **Vector Autoregression** model would leverage the close correlation of incidents and offenses to describe the relationship between simple crime and multiple-offense crime incidents.

MODELING

I divided the data into 8 years of offenses to train the models and 2 years to test their predictive capacities. Optimizing for the relative best combinations of AIC, Ljung-Box, and Jarque-Bera scores, I tried many **SARIMAX** settings. The final model showed the autoregressive component's lag 1 and the moving average component's lags 1, 7, and 14 all to be statistically significant; a sign of a good fit for seasonalities, if a bit inflexible overall. I further used R^2, $Mean^2$ Error, and Mean Absolute Error calculations to find that the model could explain 76% of the variance in the data and that it generalized well without overfitting.

The **Vector Autoregression** settings were comparatively easy due to the simplicity of the model's parameters, although its predictions were inflexible and clung to the mean. The VAR's summary was more useful and showed that complex multi-offense criminal

incidents occurred more frequently in the late-week, whereas simple crime incidents reliably increased early and late each week.

Prophet's final settings were informed by the previous models rather than tuning. I factored in weekly and monthly seasonalities and had the model respond quickly to trend changes by using a high fourier order value. Using its built-in holiday functionality, I had it also take into account unique US holiday seasonalities. I had the model perform Bayesian MCMC sampling to get reliable prediction intervals. Overall, it arrived at an accuracy close to the SARIMAX model on the R^2, $Mean^2$ Error, and Mean Absolute Error calculations.

I averaged the predictions of the models into a final set to reach an agreement point between them. These **ensemble predictions** tame Prophet's highly variable weekly trends and SARIMAX's relatively inflexible weekly trends while incorporating the monthly and yearly trends from Prophet and using VAR's points to lessen the tendency of outlier data points to distract the model.

CONCLUSION / NEXT STEPS

To properly classify the test set's data as normal or abnormal, I compared the number of actual crime data points that were more than two standard deviations away from the mean of the ensemble averages to the number of those that weren't. **68.77% of the the data was abnormally high** and **4.66% of the data was abnormally low**, leaving 26.57% as normal levels. This allowed me to conclude that **2018 and 2019 were together years of abnormally low crime** relative to the eight years that preceded them.

To do this classification at scale, a formal ETL pipeline could be built to take the FBI NIBRS data each year, organize it, retrain the models with it (splitting the data before the period in question), and use the ensemble predictions to contextualize crime for the time period. Journalistic organizations would do well to use this methodology to ground their societal assessments about crime in informed research.

The biggest takeaway from this project is that crime levels are far too variable to reliably predict even with top technologies. Machine learning's best role in applied crime analysis is therefore as **a lens through which to view periods of time**, at least until algorithms can better understand the criminal psyche at scale and incorporate the numerous factors that influence criminal behavior. My models took no external factors into consideration, making them more naive than I would recommend for thorough research. The models can be improved with reliable external factors, but can still be used in their flawed capacity to classify periods of time, as they are still more reliable than simple intuition.

SUPPLEMENTALS

All files used in Project CAP can be found on my GitHub. I have put together several interactive dashboards from visualizations used in my exploratory data analysis, published through Tableau Public – three of which deep-dive into the data and one of which shows the final ensemble predictions compared to the actual crime levels. Feedback is greatly appreciated and suggestions on improvements will be taken kindly.