Investigating Crime and Communities

A. Vishnyakov, J. Ahmar, and R. Aridi

***Abstract*— Crime is a complex phenomenon whose outcome is of social, economic, and family conditions [1]. Our project's goal is to identify the most influential features and explore which features are the most correlated with regard to a community's crime rate. The principal idea is that this information can be used by communities with a disproportionately large crime issue to help curb the crime rates. Our dataset includes a wide range of demographic and socioeconomic features, spanning 146 columns. It represents communities across the United States in 2018 [2].**

I. Background And Introduction

DATA analysis has a rich history when it comes to crime prediction and prevention. Even as early as 1842, Adolphe Quetelet writes in his book [3] *A Treatise on Man and the Development of his Faculties*: "supposing men to be placed in similar circumstances, I call the greater or less probability of committing crime, the propensity to crime. My object is more especially to investigate the influence of season, climate, sex, and age, on this propensity."

Statistics began to boom around this time, in the mid-19th century, with great use particularly in astronomy. Quetelet, though, had decided to use statistics to predict crime instead, and his investigation of features of a person and their environment on crime is still the underlying method of crime analysis today, over 180 years later. Quetelet was very skeptical of overreliance on direct predictions, being cognizant of the fact that crime's inherent nature of being underreported would make conclusive results difficult to obtain, even to researchers well into the future where we are today.

Predicting and preventing crime, particularly violent crime, is of extreme importance to law enforcement, policymakers, politicians, and the residents affected by it. The use of data analysis techniques can be effective when it comes to predicting crime at scale among different counties and states. Unfortunately, using data available to the public has some challenges, particularly with the normalizing of data obtained from different jurisdictions, as each law enforcement agency reports crimes differently. To address this, the FBI created the National Incident-Based Reporting System (NIBRS) specifically to assist with standardizing data entries, and to ensure incidents are reported with high quality data across all jurisdictions. The NIBRS calls this standardization the Uniform Crime Reporting program [4] and it serves as a useful tool for future machine learning on American data, as records will be consistent and easily aggregated even if one record references a crime in Hawaii and the other in Washington.

Today, we have a lot more tools, and a lot more data at our disposal (higher quality data, at that) than initial ML efforts to predict crime. More modern methods can dive deeper and build a limited identity for each individual. Or they can consider significantly more factors when analyzing crimes. One of the largest examples of the former, using ML to predict crime on an individual level, was in Chicago using their Strategic Subject List. This list [5], in use from 2012-2019, looks at roughly 400,000 individuals and assigns each of them a score that defines how likely they are to commit violent crime, or be a victim of crime. Unfortunately, many important details are not public. There are differing reports on the usage of the list online – police seem to have used it quite often, even on daily patrols. But it appears that its overall objective to reduce violent crime by targeting individuals had failed. Our work does not look at individual persons for analysis, assigning each a risk score; instead, we are analyzing crime in a community overall as Quetelet did centuries ago, by analyzing what difference in a community's features lead to higher crime, with the underlying idea that a community struggling with crime can influence these features in the hopes of reducing crime.

II. Data Pre-processing

A. Dataset

The dataset we will be working with is a Kaggle dataset with a wide range of features including demographic and socioeconomic information. We want to compare these features with our target variable, violent crimes per 100,000 people. Our dataset contains 2215 entries across 146 columns. For our analysis, only demographic and socioeconomic features were considered; all other columns were filtered out, locally, before we ran our models or generating our preliminary visualizations. After preprocessing our data, we ended up with 1993 entries across 97 columns.

B. Treatment

Several features had excessive missing entries; these were features such as police information, community, and country codes. We suspect this is a result of the heterogeneity mentioned in the background that the UCR is looking to address. Police information accounted for 33 columns, each with a varying amount of missing data, and so we omitted all of them. Features that include geographical location and population sums were also omitted for our analysis, as we are looking at percentages of crime across communities to see what features of a community are conducive to violent crimes. After these problem features were dropped, we implemented a threshold function for missing data; this safety measure then removed any records that had had an invalid entry, to ensure the data is cleaned thoroughly.

III. Preliminary Analysis

For our preliminary analysis, we initially decided to use a heatmap and a correlation bar plot to get a quick idea of what features we should focus on during the modelling stage. By fetching the 20 most and 20 least correlated features (with respect to violent crime rate), we can effectively visualize the features which are most strongly correlated with violent crime, giving us a starting point on what to expect in our analysis. Because our dataset still contained 97 features, even after our treatment, we needed to reduce the number of features further; so, we looked at reducing dimensionality by cutting features with little to no correlation, or that were highly correlated with one another.

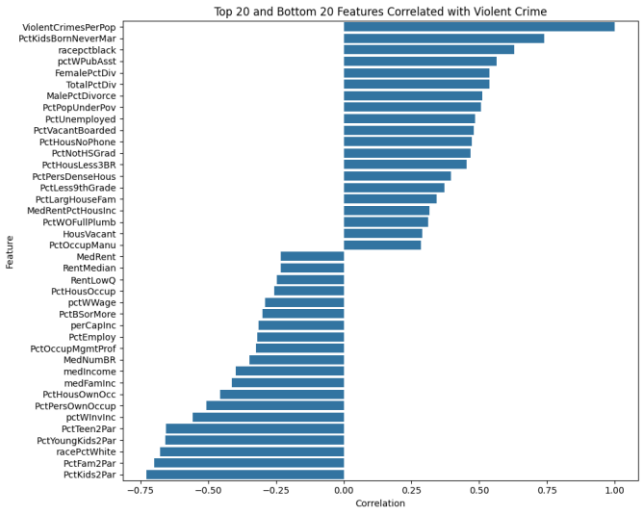

Fig. 1. Correlation bar plot, showcasing the features that had the strongest correlation to ViolentCrimesPerPop.

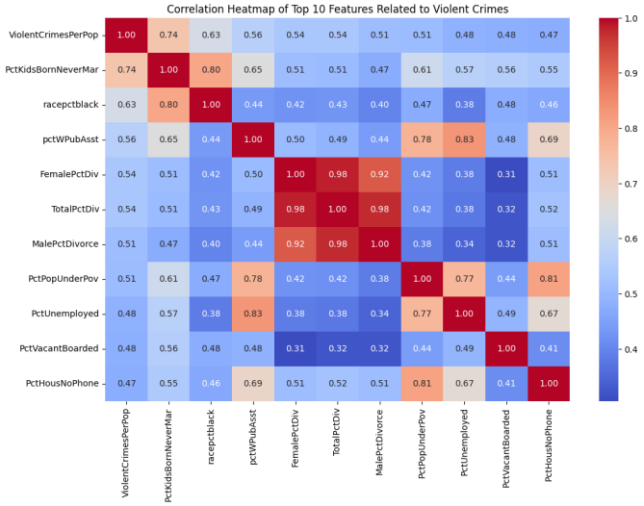

Fig. 2. Heatmap of the 10 features that were the most correlated to ViolentCrimesPerPop.

Figure 1 displays the features that were positively correlated and negatively correlated to ViolentCrimesPerPop. We then hand-selected 15 of the most significant features from this list, avoiding co-linear features, before proceeding to full modelling. This list of features ended up being: violent crimes per 100,000 people; percent kids born to parents who were never married; the percent of Black residents; the percent of White residents; percent of Hispanic residents; the percentage of people on public assistance; divorce rate percentage; percent of the population under the poverty line; unemployment rate; percent of the population that haven't graduated high school; median income; percent of houses that are vacant and boarded; percentage of immigrants who entered in the last 10 years; the percentage of people who only speak English; and the population density.

IV. Modeling and Analysis

A. Model Selection

For our study, we investigated both regression and classification models. Since the predictor, violent crimes per 100,000 population is continuous, it was converted to a categorically-based variable on whether the value is above or below the median, with above or equal to median being positive and below being negative. This was done for inputs to the categorical models and outputs of regressor models, so both create a classification. The models we investigated were regressors: Linear Regressor (LR_R), Random Forest Regressor (RF_R), multidimensional Support Vector Regressor (SVR) and Linear SVR (LinSVR). And our classifiers were: Logistic Regression (LR_C), Random Forest Classifier (RF_C), and multidimensional Support Vector Classifier (SVC), and Linear SVC (LinSVC). The results are visualized in figure 3. It was seen that performances were similar between classifiers and regressors. Classifiers were slightly better, but regressors provide more data in the form of a continuous value as opposed to a classification. Therefore, we chose to proceed with regressors, since they provide a purer prediction which can then be interpreted.

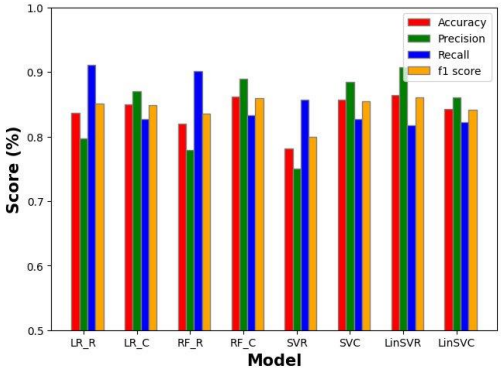

Fig. 3. Performance of Regressors compared to Classifiers.

B. Model Examination

The performance of the regressor models is visualized in figure 4, as well as the classification metrics for the classified

prediction in figure 5.

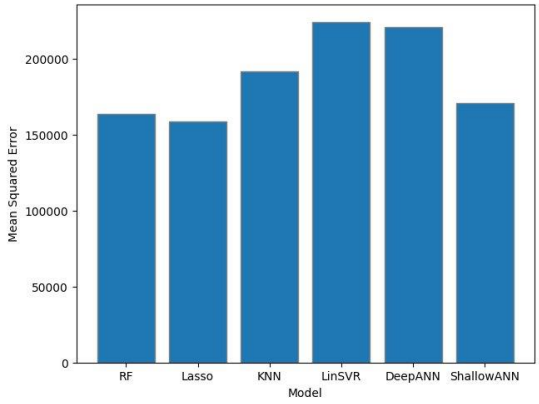

Fig. 4. MSE of the regressor models' predictions.

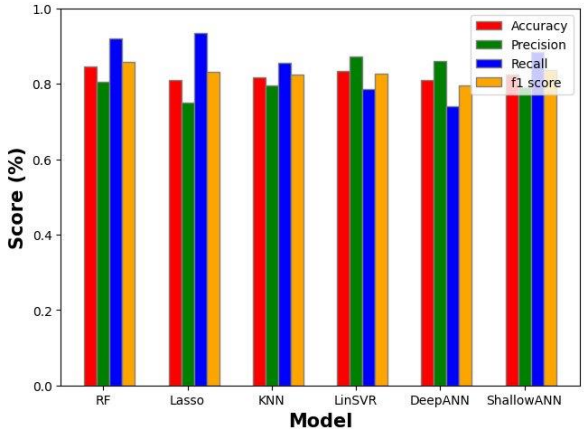

Fig. 5. Performance of regressor models for classifying.

It is seen that performance is comparable between the models for both MSE and classification, without significantly better or worse models. Quartile classification was attempted, but did not give adequate results. This is possibly due to the classification models not being able to capture the ordinal context of each quartile classification (i.e. that the 1st quartile is closer to the 2nd quartile, than the 1st to the 3rd). As for the regressors, the performance of quartile classification was poor, perhaps because ViolentCrimePerPop's quartiles ended up being significantly right-skewed. The histogram with quartiles marked is seen in figure 6.

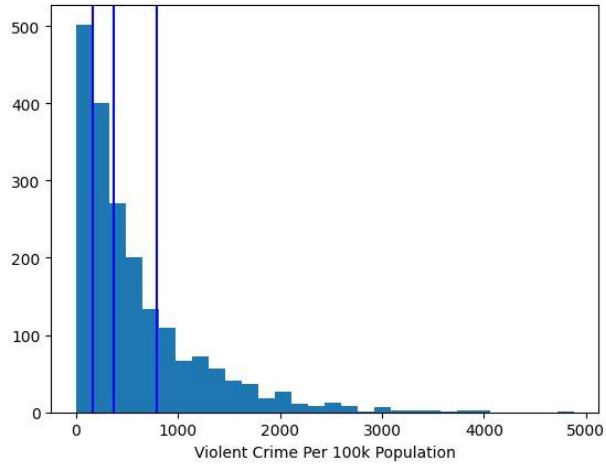

Fig. 6. A histogram on violent crime per 100,000 people, with quartiles marked.

C. Feature Importance

Feature importance analysis was conducted on the interrogatable models, the resulting feature importance plots are seen in figures 7, 8, 9, 10.

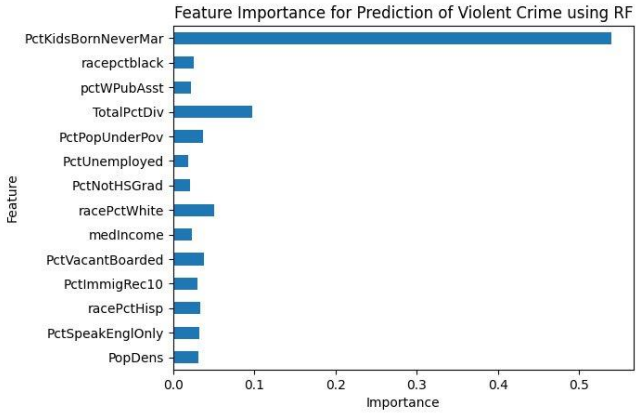

Fig. 7. Feature importance for the Random Forest model.

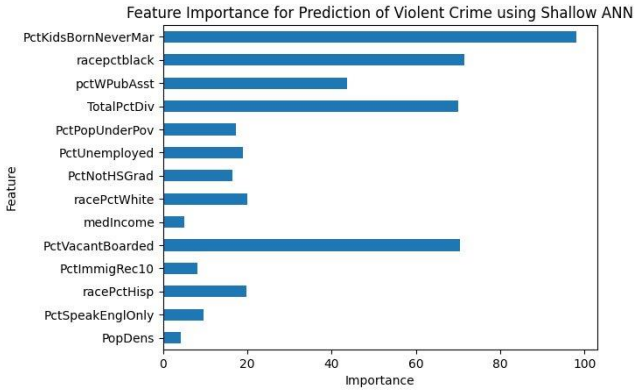

Fig. 8. Feature weights for Shallow ANN model.

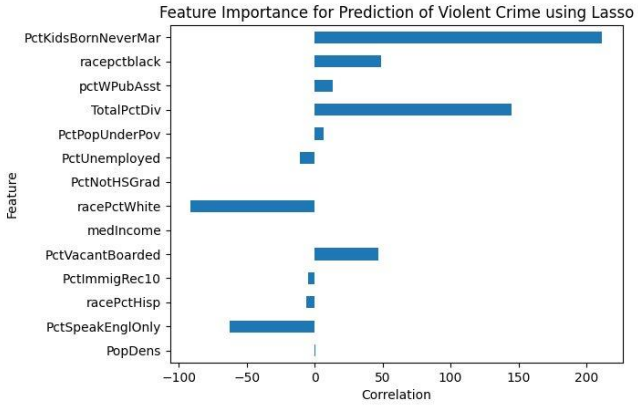

Fig. 9. Feature correlation for Lasso model.

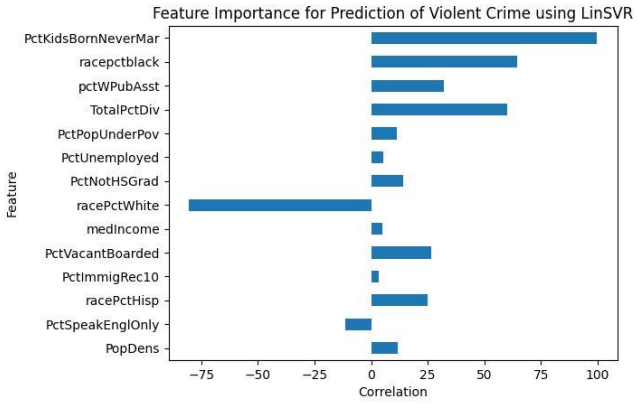

Fig. 10. Feature correlation for Linear SVR model.

It is seen that the random forest model relies heavily on the percentage of kids born to parents who were never married, with around a 55% reliance. For the shallow Artificial Neural Network (ANN), the input weights of the shallow ANN were extracted, since it didn't have a hidden layer. This can be used to see which features lead to a higher crime prediction. It was seen that the percentage of kids born to parents who were never married is the strongest, with the percentage of Black residents, the percentage on public assistance, the percentage of divorced people, and the percentage of vacant, boarded homes also being significant for predicting high crime rates. Both lasso and the linear SVM models are similar in the way they predict, by using coefficients on each feature. As mentioned earlier, lasso was chosen over linear regression due to the reduced reliance on a single feature, as shown by its reliance being distributed across multiple features. Additionally, we saw that it performed its own feature selection and excluded the percentage of people who did not graduate high school, and also excluded median income, which further shows why lasso was selected over linear regression. For both lasso and linear SVR, we see feature outputs that are positively correlated to violent crime that were similar to what ANN reported. But the difference with the former two models is that they could also produce a set of features with a negative correlation, which correlated to a lower crime rate. Those features were the percentage of White residents, and the percentage of the population that only speak English.

V. Future Works

There are areas for improvement in our investigation. Some limitations of this project include the dataset being from Kaggle as opposed to a credible source such as the government. There was also a lack of consistent police information. Finally, the features still have correlation between each other, even if it isn't clear in the analyses. Societal factors influence each other and cannot be assumed to be orthogonal. For future works on this study, we could perform a similar analysis in other continents such as Europe, Asia, or Africa, as well as expand upon which familial factors influence crime. It was seen that a community with a larger number of single parents had higher crime rates, but it's unknown if there is a difference between single motherhood as opposed to single fatherhood; this could be explored.

VI. Conclusion

In conclusion, our analysis demonstrates that family factors play the most significant role in contributing to violent crime. Between the various models we have used, we come to a similar result each time, in that communities with a higher number of families without two married parents have more violent crime. During our preliminary analysis we have shown that there is a strong correlation between the percentage of kids born to parents who were never married to one another, and certain racial features. This may explain why race related features are consistently found to be an important feature in our analysis – the two features are correlated. Ultimately, we began our project with the intention to seek out the feature that is most strongly correlated with crime, so that we could suggest a starting point to addressing crime, from its root. Our findings suggest that a lack of a strong familial structure at home is the most correlated feature with higher crime.

APPENDIX

A. Vishnyakov

All of the code and execution for Modelling and Analysis, as well as part of the preliminary analysis and feature importance extraction and selection. Worked on all report and presentation sections relating to modelling, results, and conclusions, and helped with other sections.

J. Ahmar

Code for preliminary analysis. Visualization and analysis for project proposal. Slides and notes for brainstorming video. Data pre-processing and visualization for report draft. Project objective and description of applied problem on group presentation. Data pre-processing and conclusion for final report.

R. Aridi

Managed the entire project's structure, including general ideas and brainstorming, and writing a structured outline for the deliverables. Built the structure for report draft, the final report, and our Microsoft Loop page. Edited both video deliverables. Wrote a portion of every part of the report, and dealt with proofreading to ensure the quality was up to standard.

Microsoft Loop page
https://loop.cloud.microsoft/join/eyJnIjoiT0lEOjJlNmExNzEyLTI0MjUtNGVmOS05NGU3LTJiODRjYmE3ZTMwZEBhZDkzYTY0ZC1hZDBkLTRlY2QtYjJmZC1lNTNjZTE1OTY1YmUiLCJpIjoiM0M2ZkxuV0x3VU9JbVpETDRHaGFsQSIsInciOiJUMFJUVUh4MWQyOWpZUzV6YUdGeVpYQnZhVzUwTG1OdmJYeGlJWEV6TlhoT05qbDRXRlZIVGtoMWFFaGxiVUp3U1dSTFdsY3diRzlxZVRWUGF5MURWV3BKUVhsa01IcE1UVTR6U0ZSWlRVTlVTVlpaYjJ0bVV6VTJkSFY4TURGTldVZFBTbGRaUTFSTlUwTTBNMGRWTjBaSVdrcEdVRTh6V1V4Q04wVkxVZz09In0%3D

GitHub: https://github.com/RibalAridi/DS3000

REFERENCES

[1] CS&CPC, "CS&CPC Statement on the Root Causes of Crime," 1996. [Online]. Available: https://preventingcrime.ca/wp-content/uploads/2014/08/Causes_of_Crime.pdf. Accessed: Dec. 6, 2024.

[2] M. Bryant, "Crimes in US Communities Dataset," 2018. [Online]. Available: https://www.kaggle.com/datasets/michaelbryantds/crimedata/data. Accessed: Dec. 6, 2024.

[3] A. Quetelet, *A Treatise on Man and the Development of His Faculties*, Chapter III, p. 82, 1842. [Online]. Available: https://openlibrary.org/works/OL1547156W/A_treatise_on_man_and_the_development_of_his_faculties?edition=key%3A/books/OL23350005M. Accessed: Dec. 6, 2024.

[4] Federal Bureau of Investigation (FBI), "National Incident-Based Reporting System (NIBRS)," 2012. [Online]. Available: https://ucr.fbi.gov/nibrs/2012/resources/variables-affecting-crime. Accessed: Dec. 6, 2024.

[5] City of Chicago, "Strategic Subject List - Historical," 2020. [Online]. Available: https://data.cityofchicago.org/Public-Safety/Strategic-Subject-List-Historical/4aki-r3np. Accessed: Dec. 6, 2024.