Catching a serial killer with Data Mining

Sindhu Kishore
skishor2@ur.rochester.edu

Yudhisteer Chintaram
ychintar@ur.rochester.edu

***Abstract*— In this research we propose to create both an analytic and predictive model. Firstly, we intend to provide an analysis that can help law enforcement to narrow their search on possible offenders. Secondly, we aim to develop a predictive model based on the nature of the crimes and other relevant features, that will help the authorities spot similar recurrent crimes in the future and the characteristics or traits of the person behind the crimes.**

Keywords—Serial Killer, Frequent Pattern, Random Forest, Decision Tree.

I. INTRODUCTION

According to the FBI, a serial murder is defined as "unlawful killing of two or more victims by the same offender(s), in separate events". Though it is a rare event, estimated to comprise less than one percent of all murders committed in a year, serial murder cases present numerous challenges to law enforcement. These cases involve multiple victims, offenders' behavior, series may span from days, weeks to months or even years. Thomas K. Hargrove, a retired investigative journalist and founder of nonprofit Murder Accountability Project said that serial murders go unsolved due to linkage blindness, i.e., the inability of law enforcement to communicate or share information in a way that connects similar unsolved crimes. Although a lot of literature on serial killings exists, the application of data science or any kind of data analysis, is minimal.

II. RESEARCH QUESTIONS

A. Research Question

In order to build the analytic and predictive models described above, we will need to answer the research questions below:

- How to identify if a serial killer exists and his pattern of killing?
- How to predict a serial killer's profile?
- What led to the decline of serial killers?
- Are mass murderers the new facet of serial killers?

III. DATASET

A. Choosing Our Dataset

Public dataset regarding Serial Killers is limited for analysis. We came across Murder Accountability Project (MAP) by Thomas K Hargrove where the data maintained by FBI was made publicly available. We also opted a private resource for our data, the Radford/FGCU Serial Killer Database.

The Murder Accountability Project is a nonprofit group based in Alexandria, Virginia, and operated by a Board of Directors that includes veteran homicide investigators, investigative journalists and homicide scholars. Using the Freedom of Information Act, MAP obtained data on more than 28,200 homicides reported from 1976-2020 with 32 variables per homicide consisting of information about Victim age, sex, location, Weapon etc.

Radford/FGCU serial killer database: A cross-university collaboration of Radford University with Florida Gulf Coast University built this non-governmental private database of serial killers in the world. The database contains 5724 subjects, comprising of serial killers, mass murderers and more. There are over 14773 victims profile, 500 documents and about 191 variables per subject including the background information, victim preference, victim treatment and information about the crimes committed.

B. PREPROCESSING

Both Radford/FGCU and MAP database required lot of preprocessing. We restricted the homicide timelines to 1976-2020.

Handling Missing Data

Radford data had five of the columns with greater than 5% and less than 10% missing values. Four of the variables have greater than 10% missing values. Target variables such as Education level, sexual orientation and marital status are greater than 30% missing values. The latter were discarded as doing any sort of data imputation would distort the data distribution

MAP Dataset had around 2% missing values for Victim Sex and Victim Race, hence these records were discarded.

Frequent Category Imputation

Frequent Category Imputation can be used when data are missing at random. They are easy to implement, and it is a fast way to obtain a complete dataset. The disadvantage of this technique is that it can lead to an over-representation of the most frequent label if there is a big number of null values. Hence, this technique was used for the categorical features with less than 5% of missing data.

Adding a missing Indicator

Similar to the Frequent Category Imputation, the Missing Indicator is easy to implement and it also helps in capturing the importance of "missingness". The disadvantages of this method is that it increases the feature space. Since we are using 27 features and all of them has missing data then we end up with 54 features where 27 of them are the Missing Indicator feature. Another problem is can lead to similar or highly correlated missing indicators. Note that we are using the mode imputation to fill the missing values plus adding another missing indicator feature vector for each variables.

Missing Values in Target

In Case of Radford for the target variables Age, Gender and Age Group we have less than 5% missing data. The whole point of supervised learning is to have the real label both to train and then to evaluate the model. Hence, we dropped the rows where the target labels had missing values. Multivariate Imputation by Chained Equation (MICE) method can also be used where the initial samples are randomly drawn from the observed data.

For MAP dataset we used data imputation by Mode to fill in the missing age of victims. Further we created a new variable Age Group of victims to determine the killer's target age group and another variable Active Range to determine the timeline of killer's activity.

C. CATEGORICAL ENCODING

One Hot Encoding

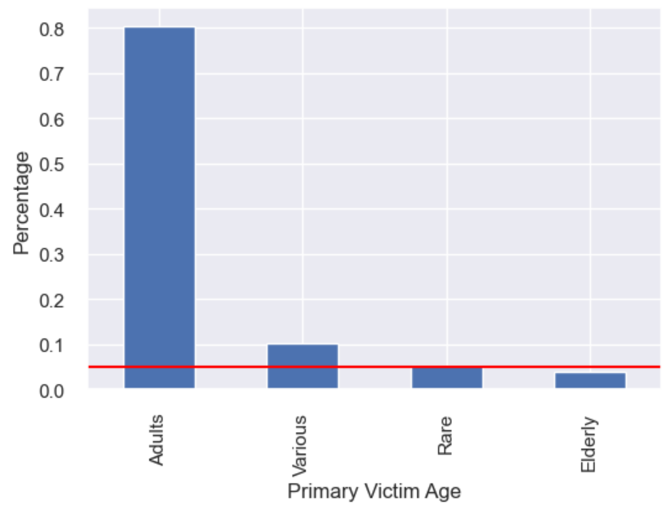
One Hot encoding consists in encoding each categorical variable with different boolean variables which take values 0 or 1, indicating if a category is present in an observation. We have the option to encode into K or K-1 binary variables. The advantages of this model is that it makes no assumption about the distribution or categories of the categorical variable and it is suitable for linear model. However, it is also expanding the feature space and many dummy variables may be identical, introducing redundant information. This method is best for neural networks or when building tree-based algorithms. Both methods of encoding into K and K-1 are used and their performance metric is described below.

Integer Encoding

Integer encoding consists in replacing the categories by digits from 1 to n (or 0 to n-1), where n is the number of distinct categories of the variable. Contrary to One Hot Encoding, this technique does not expand the feature space. However, it is not suitable for linear models and it does not handle new categories in the test set automatically.

Rare Label Encoding

Rare labels are those that appear only in a tiny proportion of the observations in a dataset. We use this if we have cardinality, a small number of categories or one predominant category. Note that grouping categories into rare for variables that show low cardinality may or may not improve model performance. However, the opposite is true for variables with high cardinality. In summary, what we are doing is creating another category call "Rare" and grouping all these infrequent labels into that new category. Note that Rare labels should be identified in the training set only.

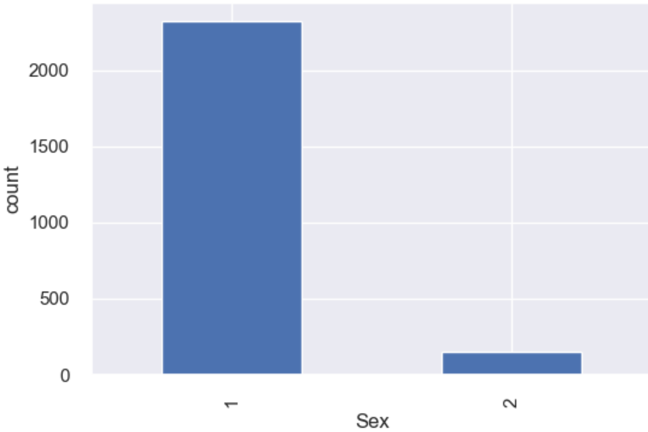


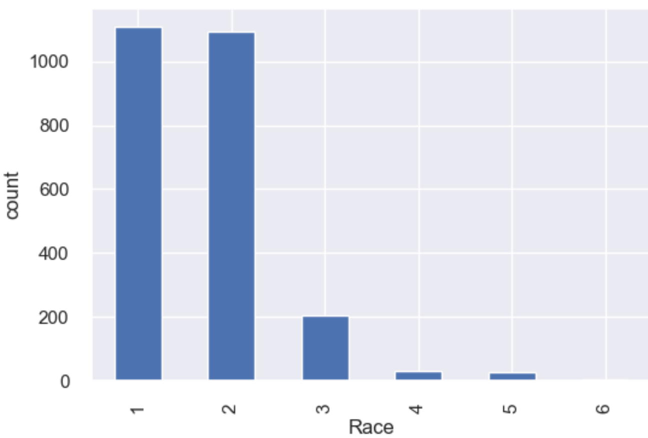
Grouping Variables

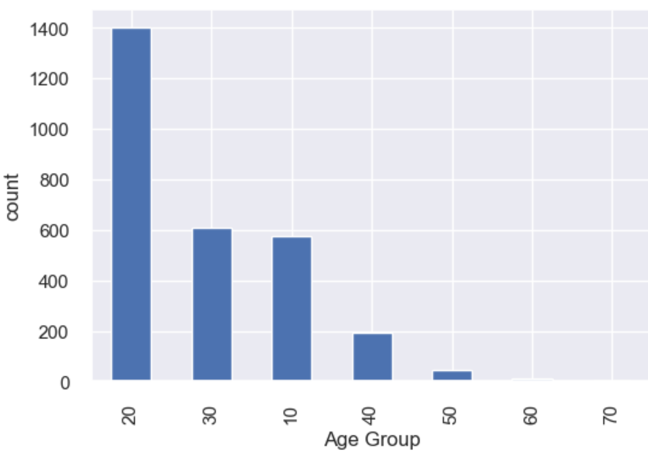
Similar to creating a Rare category as described above, we can also group variables with infrequent observations in an already existing category. The variable "Age Group" was used for such technique as we found that we have many low serial killers where their age group on their firls kill was above 40 years old. The variable "Race" was also grouped as the predominant catgories were White, Black and Hispanic. Categories such as Asian, Natives and Other consisted of less than 5% of the data. Grouping variables to reduce cardinality showed an improvement in performance metric but not by much.

D. IMBALANCED DATA

For our 3 target variables,Sex, Race and Age Group, we have an unbalanced amount of data. Hence, this may be a problem when doing prediction *as the algorithm will not have enough data to learn from the minority class. It may also result in biasness from* the model, for example, if we have more "Male" serial killers, the model will be bias in predicting "Male" most of the time. The graphs below shows the imbalance in our target variables. Grouping was also done in the infrequent categories, as explained above, to reduce cardinality.





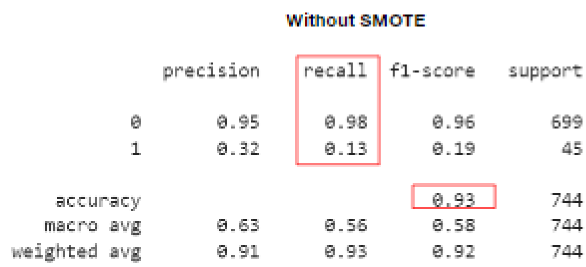


Sample Extraction for Radford

One way to counter the imbalance problem is to use a Sample Extraction technique. The method extracts samples at random from the minority class. We can use Random Oversampling technique which will extract observations at random from the minority class until we reach a balancing ratio of 1. This is a Naive technique as no assumptions for the data is made. Since we will be duplicating samples from the minority class, there may be a likelihood of overfitting.To counter this duplication of data problem, we can use Random Oversampling with Smoothing. i.e. we create new examples based on the minority class but we add some noise to it. The noise will be informed by the class distribution. In that way, we avoid the data duplication problem

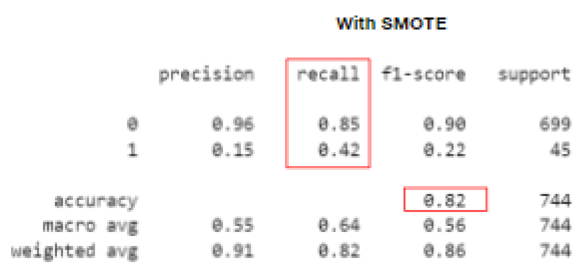
Sample Generation

Another method of oversampling is to take existing samples as template from the minority class to create new samples. The new samples will be different from the existing ones. The technique used is Synthetic Minority Oversampling Technique (SMOTE). The SMOTE will take a random data point from the minority class and its k-nearest neighbors. Then it will create a new data point from this chosen data point and one of the random k-data points. A new information point is added to the minority class.Undersampling also can be used where data from the majority class is removed till we reach a desired balancing ratio with the minority class. However, with undersampling we will be losing information. Hence, such method is good only when we have a large dataset.Below are the recall and accuracy values for predicting the Gender with and without SMOTE. We can observe that without SMOTE we have a low recall of 13% for the minority class but with SMOTE the recall increases to 42%. However, we decreases in accuracy from 93% to 82%.



Without SMOTE

	precision	recall	f1-score	support
0	0.95	0.98	0.96	699
1	0.32	0.13	0.19	45
accuracy			0.93	744
macro avg	0.63	0.56	0.58	744
weighted avg	0.91	0.93	0.92	744



With SMOTE

	precision	recall	f1-score	support
0	0.96	0.85	0.90	699
1	0.15	0.42	0.22	45
accuracy			0.82	744
macro avg	0.55	0.64	0.56	744
weighted avg	0.91	0.82	0.86	744

IV. PREDICTIONS

A. *Cross Validation*

In order to check for overfitting and to assess our model performance, we used cross-validation. Since our target variables are categorical, we used Stratified Cross-validation, which is a technique of cross-validation only for classification purposes. With stratification, we ensure that each fold has a similar proportion of each class. With a hold-out method coupled with the Pipeline function from Sklearn, we have no overlap with the test set, hence avoiding any sort of data leakage. Below is the process of evaluating our model using cross-validation:

1. Divide the data into train-test sets.
2. The train set is used for cross-validation:
    - Separate k-fold of data
    - Over or Under sample k-1 folds
    - Train the model on the re-sampled k-1 folds
    - Evaluate the model on the left out kth fold with the original data distribution
1. Evaluate the final model on the test set.

B. Metrics

One metric that we will use is the Accuracy on our test set. The accuracy is the percentage of fraction of correct predictions.

Accuracy = (Total numbe rof correct predictions)/Total number of predictions

Accuracy for binary classification:
Accuracy = (True Positive + True Negative)/Total number of predictions

However, for imbalanced dataset, accuracy may not be the optimal metric as it does not distinguish between the numbers of correctly classified examples of different classes. The minority class will have very little impact on the accuracy as compared to that of the majority class. In summary, accuracy will not let us know how much of the minority class were actually classified correctly.Balanced accuracy would be a more appropriate metric as it takes both classes, majority and minority, into consideration:

Balanced Accuracy = (Recall0 + Recall1)/2

Another metrics for imbalanced datasets would be:

- True Positive Rate (Recall):

Recall = True Positive/(True Positive + False Negative)

- Positive Predictive Value (Precision):

Precision = True Positive/(True Positive + False Positive)

- F-measure (F1 score):

F1 score = 2 * (Precision * Recall)/(Precision + Recall)

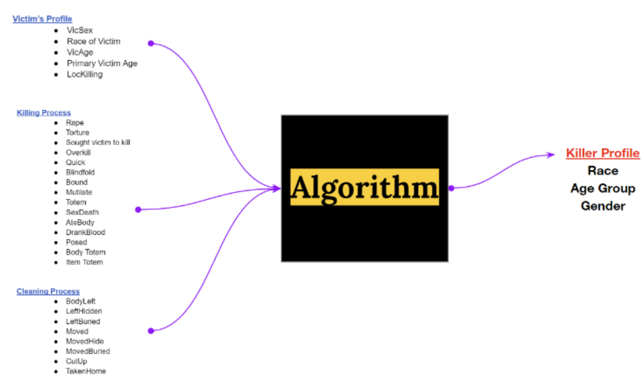
Both precision and recall vary between 0 and 1. Our goal is to maximize these both metrics. Note that these two metrics depend on a probability threshold which is a probability value above which we consider the sample belongs to the minority class.

Another metric which we will consider is the Receiving Operating Characteristic (ROC) Area Under Curve score (AUC). The ROC curve plots the Recall or True Positive Rate and False Positive Rate at different classification thresholds. It evaluates how well a classifier can separate positive and negative examples.

AUC provides an aggregate measure of performance across all possible classification threshold. Higher AUC means the model is better at predicting both classes. For an AUC of 1, we have a perfect model which classifies correctly all the observations. For an AUC of 0.5, we have a model which predicts randomly. That is it returns the same TP and FP for evry single threshold of probability. For every batch of prediction, half of them will be classified incorrectly.
Note that ROC plots the TPR and FPR for one class only at various thresholds. Hence, it is designed for binary classification. But since the target variable "Race" and "Age Group" are multiclass, we need to extend this metric for multi-class classification problems. We need to train our classifier in a 1 vs Rest strategy. i.e. we train one model per class and we output one ROC curve per class for one model at a time.

C. MODELS

Previous papers only used the victim's age, gender and race to predict the Killer's profile.Our approach is to use the victim's profile but on top of that use information on how the victims were killed and how the killer managed to cover their traces.Studies showed that there have been patterns in the killing process and the killer's profile. For example, a killer who hunt at night is most likely to be unmarried.



Other targets such as Marital Status, Education Level and Sexual Orientation could have been predicted for the killer but the percentage of missing data is too high for data imputation.

Initially different combination of imputation techniques that were described above were done along with different combination of encoding techniques to study which one yields the best accuracy. This strategy was used to predict Gender and Race. Tree based classifiers, ensemble method and Neural Network were considered for prediction. We are doing this step so as to choose which data imputation technique and which categorical encoding technique to use in the final models. The results are as shown below:

Predicting Race							
Data Processing		**Model**					
Encoding	Imputation	Random Forest	Logistic	AdaBoost	KNN	NN	Voting
OHE: K	Frequent Category	0.685483871	0.63172	0.541667	0.697581	**0.778**	0.727151
OHE: K-1	Frequent Category	0.678763441	0.63172	0.541667	0.709677	**0.7728**	0.737903
Rare + OHE: K	Frequent Category	0.677419355	0.629032	0.489247	0.712366	**0.746**	0.736559
Rare + OHE: K-1	Frequent Category	0.680107527	0.63172	0.528226	0.737903	**0.7446**	0.733871
OHE: K-1	Missing Indicator	0.693548387	0.637097	0.564516	0.727151	**0.7419**	0.729839
Rare + Group Race(4)	Frequent Category	0.692204301	0.63172	**0.75**	0.712366	0.746	0.727151
OHE: K + Group Race(4)	Frequent Category	0.686827957	0.633065	**0.736559**	0.716398	0.7366	0.731183
Rare + Group Race(3)	Frequent Category	0.720430108	0.626344	**0.744624**	0.721774	0.7433	0.737903

Test Accuracy

From the results below, we observe that we do not have much discrepancy in accuracy either for k or k-1 one hot encoding. Grouping variables and Rare Label encoding do yield better accuracy in general. With Missing Indicator imputation, we increase our space from 27 features to 54 and we risk overfitting for too little increase in accuracy. We also observe that Ensemble methods such as AdaBoost provides better accuracy. Hence, k-1 One Hot Encoding with Rare Label and grouping of categories will be used for Categorical Encoding. Frequent Category Imputation is selected for further testing on other models.

Predicting Gender							
Data Processing		**Model**					
Encoding	Imputation	Random Forest	Logistic	AdaBoost	KNN	NN	Voting
OHE: K	Frequent Category	0.939516129	0.942204	0.939516	**0.944892**	0.9227	0.942204
OHE: K-1	Frequent Category	0.942204301	0.942204	0.943548	0.944892	0.9089	**0.946237**
Rare + OHE: K	Frequent Category	0.940860215	0.94086	**0.944892**	0.938172	0.9395	0.942204
Rare + OHE: K-1	Frequent Category	0.942204301	0.94086	**0.946237**	0.94086	0.9395	0.94086
OHE: K-1	Missing Indicator	0.942204301	0.943548	0.943548	0.943548	0.9395	**0.944892**

Test Accuracy

D. Results

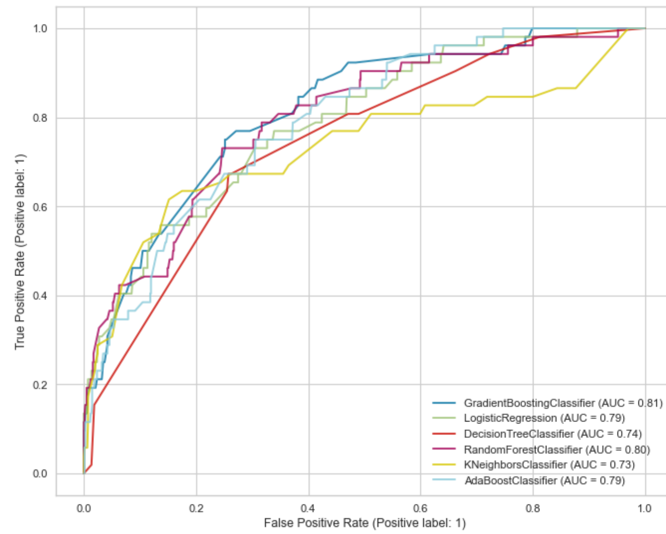
A couple of tree based classifiers model were used such as Decision Tree Classifer and Random Forest. Ensemble methods such as AdaBoost and Gradient Boosting also proved to have a good performance metric. Simpler models such as KNN and logistic regression models were also utilized.

d) Predicting Gender:

To predict gender, we observe that not one algorithm was good at all the metrics. Gradient Boosting model has the highest AUC score but Decision Tree has the highest balanced accuracy, precision for the majority class and recall for the minority class.

Predicting Gender									
Model	**Accuracy**	**Balanced Accuracy**	**Precision**		**Recall**		**F1 Score**		**ROC AUC**
Logistic Regression	0.93	0.613902589	0.95	0.42	0.98	0.25	0.97	0.31	0.787466
Decision Tree	0.74	0.68968747	0.97	0.14	0.74	0.63	0.84	0.23	0.743716
Gradient Boosting	0.89	0.687612634	0.96	0.26	0.91	0.46	0.94	0.33	0.813051
Random Forest	0.95	0.575690031	0.95	0.8	1	0.15	0.97	0.26	0.796642
KNN	0.91	0.628568719	0.96	0.28	0.95	0.31	0.95	0.29	0.733081
AdaBoost	0.85	0.661351133	0.96	0.19	0.88	0.44	0.92	0.27	0.790856

From the ROC curve, we observe Gradient Boosting to be the better classifier with an AUC of 0.81. Hence, we choose this model to predict Gender of serial killers.

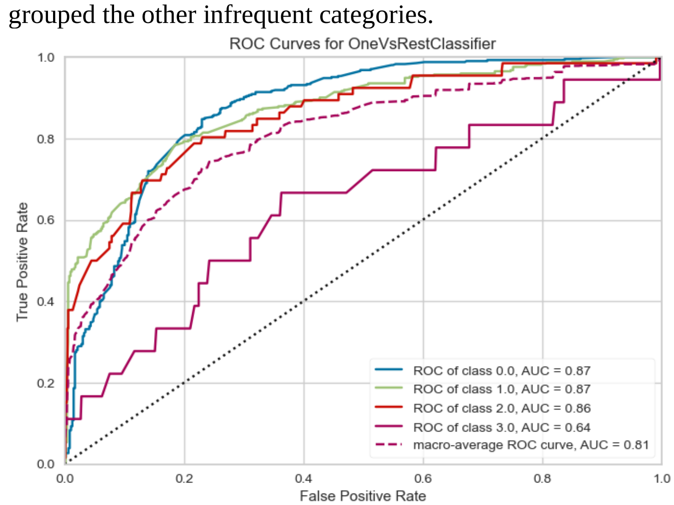


d) Predicting Race

To predict Race we have one model which has all four best metric. Gradient has the best ROC AUC score and the best balanced accuracy.

Predicting Race				
Model	**Accuracy**	**Balanced Accuracy**	**One-vs-Rest ROC AUC scores (Weighted)**	**One-vs-Rest ROC AUC scores (Macro)**
Random Forest	0.69	0.511685072	0.840517	0.771952
KNN	0.49	0.494566414	0.783614	0.738222
Gradient Boosting	0.74	0.540769509	0.86537	0.809433
Decision Tree	0.65	0.534914313	0.848278	0.773999
Logistic Regression	0.43	0.35991013	0.686675	0.642656
AdaBoost	0.74	0.539981881	0.830734	0.7797

As seen from the graph, Gradient Boosting is predicting class 0, 1 and 3 with a score greater than 85%. However, it is only predicting class 3 with 64% score since this is the class with the most infrequent labels and also the class in which we grouped the other infrequent categories.

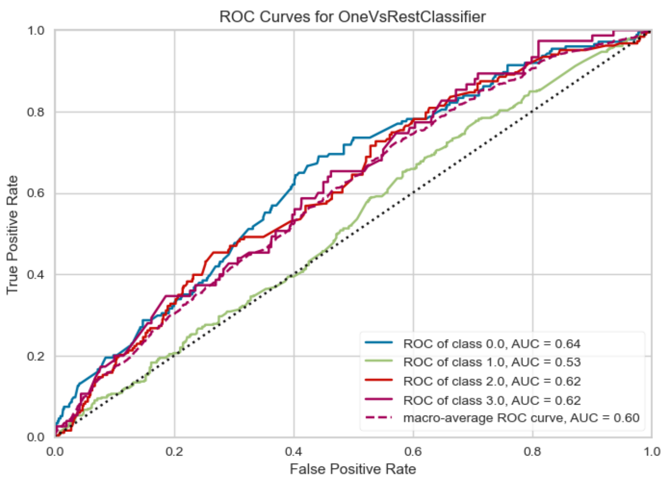


d) Predicting Age Group

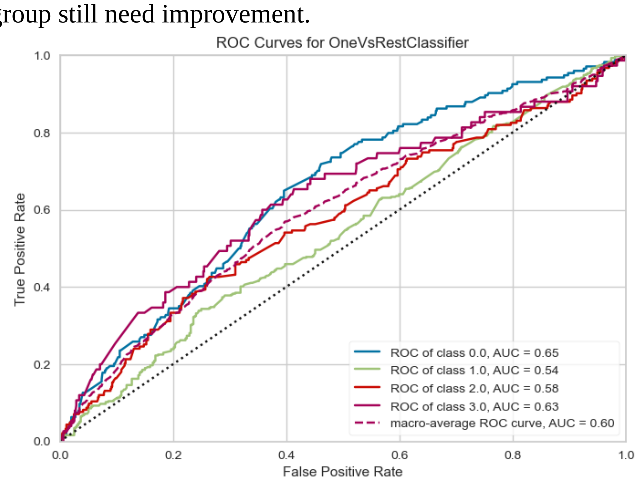
Predicting Age Group was the most difficult variable to predict. The Gradient Boosting provided a Balance Accuracy score of only 33%. However, it is the Random Forest which has the highest Macro ROC AUC score of 60%.

Predicting Age Group				
Model	Accuracy	Balanced Accuracy	One-vs-Rest ROC AUC scores (Weighted)	One-vs-Rest ROC AUC scores (Macro)
Gradient Boosting	0.41	0.333470013	0.577928	0.599364
Logistic Regression	0.28	0.310393315	0.567747	0.580533
Decision Tree	0.26	0.318275222	0.577649	0.586732
Random Forest	0.31	0.318322904	0.57614	0.600234
KNN	0.27	0.314076551	0.570018	0.587074
AdaBoost	0.34	0.302064067	0.559145	0.580329

Below is the ROC curve for the Random Forest. We observe that it does not do a good job at predicting class 1 with a score of 53%.



Below is the ROC curve for the Gradient Boosting. It does classify class 1 better than Random Forest but only by 1%. Overall, we observe that predicting Gender and Race can be done with quite a good accuracy. However, predicting the age group still need improvement.

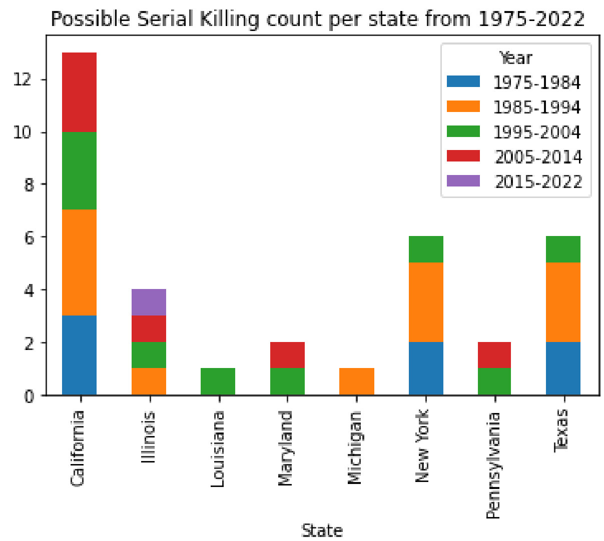


d) *Frequent Pattern*

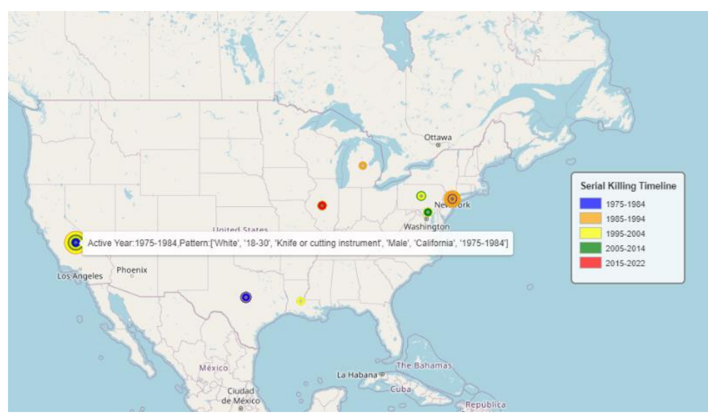
To Determine the frequent pattern from MAP dataset, we extracted field of interest like State, Killer Active Range,Age Group,Victim Sex,Victim Race and Weapon and converted it to a list for each victim entry. Further sequential pattern algorithm Apriori, Fpgrowth and FPmax was applied on this dataset. FPMax gave us the promising results. For minimum support of 0.002 , we obtained 35 itemset containing all the targeted variables.

Support	State	Itemsets	Year
0.00722089	California	['White', '1985-1994', '18-30', 'Handgun - pistol, revolver, etc', 'Male', 'California']	1985-1994
0.00667472	California	['1995-2004', 'White', '18-30', 'Handgun - pistol, revolver, etc', 'Male', 'California']	1995-2004
0.005020037	New York	['1985-1994', '18-30', 'Handgun - pistol, revolver, etc', 'Male', 'New York', 'Black']	1985-1994
0.004496261	California	['1985-1994', '18-30', 'Handgun - pistol, revolver, etc', 'Male', 'California', 'Black']	1985-1994
0.004492528	California	['White', '18-30', 'Handgun - pistol, revolver, etc', '2005-2014', 'Male', 'California']	2005-2014
0.003820702	California	['White', '1985-1994', 'Handgun - pistol, revolver, etc', '31-50', 'Male', 'California']	1985-1994
0.003484789	California	['White', '18-30', 'Handgun - pistol, revolver, etc', 'Male', 'California', '1975-1984']	1975-1984
0.003468615	New York	['White', '1985-1994', '18-30', 'Handgun - pistol, revolver, etc', 'Male', 'New York']	1985-1994
0.003362865	California	['1995-2004', '18-30', 'Handgun - pistol, revolver, etc', 'Male', 'California', 'Black']	1995-2004
0.003127725	Texas	['White', '1985-1994', '18-30', 'Handgun - pistol, revolver, etc', 'Male', 'Texas']	1985-1994
0.002933642	California	['1995-2004', 'White', 'Handgun - pistol, revolver, etc', '31-50', 'Male', 'California']	1995-2004
0.002923689	Texas	['White', '18-30', 'Handgun - pistol, revolver, etc', 'Male', '1975-1984', 'Texas']	1975-1984
0.002677353	California	['18-30', 'Handgun - pistol, revolver, etc', '2005-2014', 'Male', 'California', 'Black']	2005-2014
0.002596485	New York	['1995-2004', '18-30', 'Handgun - pistol, revolver, etc', 'Male', 'New York', 'Black']	1995-2004
0.002544232	New York	['18-30', 'Handgun - pistol, revolver, etc', 'Male', '1975-1984', 'New York', 'Black']	1975-1984
0.002493223	Texas	['1985-1994', '18-30', 'Handgun - pistol, revolver, etc', 'Male', 'Texas', 'Black']	1985-1994
0.002345172	Texas	['White', '1985-1994', 'Handgun - pistol, revolver, etc', '31-50', 'Male', 'Texas']	1985-1994
0.002342684	Maryland	['1995-2004', '18-30', 'Maryland', 'Handgun - pistol, revolver, etc', 'Male', 'Black']	1995-2004
0.002302872	Illinois	['Illinois', '1985-1994', '18-30', 'Handgun - pistol, revolver, etc', 'Male', 'Black']	1985-1994
0.002285454	Pennsylvania	['Pennsylvania', '2005-2014', '18-30', 'Handgun - pistol, revolver, etc', 'Male', 'Black']	2005-2014
0.002230713	Pennsylvania	['1995-2004', 'Pennsylvania', '18-30', 'Handgun - pistol, revolver, etc', 'Male', 'Black']	1995-2004

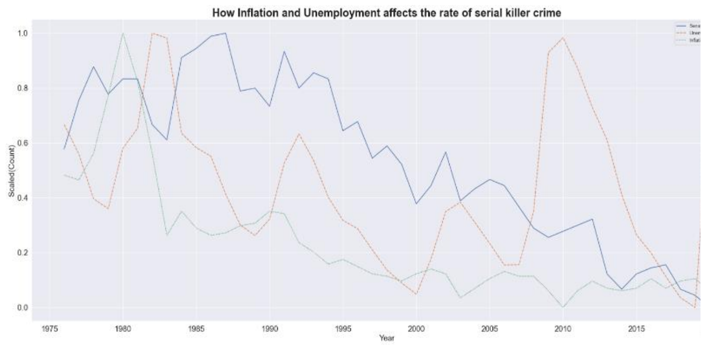
Possible serial killing from 1975-2022



clusters indicating serial killing from 1975-2020 based on MAP dataset with minimum support count as 0.002

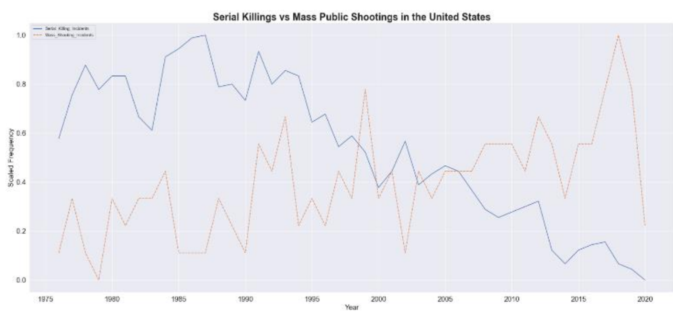


What led to the decline of serial killers?



Overall, the number of incidents by serial killers had been on the decline.Granger causality test was used to check if Inflation or Unemployment can be used to predict the number of incidents. Note that a predictive relationship does not imply a causal relationship.P-value is not significant for number of incidents and Unemployment. But p-value is significant for number of incidents and Inflation rate at lag 2 and 5.We can also observe from the graph that both inflation and the number of incidents has a similar decreasing trend.

Are mass murderers the new facet of serial killers?



The number of serial killer and mass shooting incidents were filtered from 1976 to 2020.As the number of serial killer incidents has been on the decline, the number of mass public shootings has been increasing.Part of the reason for the rise of mass shootings is due to the ease of availability of guns and lack of gun control in the U.S.We can confidently conclude that we now have less serial killing incidents than mass public shooting incidents. The current laws, culture and environment is giving birth to more mass shooting murderers.

FUTURE WORK

If we pursue this study further we plan on predicting more features of the killer like IQ, family background etc. In our current approach we predict the age group of the killer this can be further improved by predicting the right age.

CONCLUSION

To Predict if a serial killer exist we used frequent pattern mining algorithm with minimum support as metric to obtain the killing pattern unlike other studies where only clustering was used to predict if a serial killer exists.To make a serial killer profile we looked at different Machine Learning models in order to predict the Gender, Race and Age Group. Our approach was different to previous studies where they considered only the victim's profile for prediction. For our model, we also considered the way the crime was committed and how the killer interacted with the victim's body after killing. Previous studies showed that we could deduce much more about the killer based on the type of crime committed. Hence, the reason for us choosing this approach.

Classifier ensemble method proved to be the most reliable model for predicting this categorical variables. We can successfully predict the Gender with 95% accuracy and the Race with 74% accuracy. Clasifying the Age Group is still below moderate and better feature selection technique can be used to have a better accuracy.

Building a killer's profile based on Age, Gender and Age Group will help funneling down the number of possible suspects and can help authorities have a better idea of the culprit.

REFERENCES

[1] Yaksic E. et al., (2020). Statistical study of time intervals between murders for serial killers. Journal of Criminal Justice.

[2] Martinez F. & Garza M. Using Machine Learning to Analyze Patterns in Serial Killers. (2020). University of Texas at Rio Grande Valley.

[3] Kim, Suhong, et al. "Crime Analysis Through Machine Learning." 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON). IEEE, 2018.

[4] Mariani S. (2020). Profiling Serial Killers Using Multiple Supervised Machine Learning Approaches. Utrecht University.

[5] Thomas K. Hargrove. (2015). The Murder Accountability Project.

[6] Radford/FGCU Database