Will this Case be Solved?

Laura Borton

Project Design:

The Murder Accountability Project is dedicated to educating Americans on the importance of accurately accounting for unsolved homicides within the United States. Their website (http://www.murderdata.org) gives the public access to the Supplementary Homicide Report, a database maintained by the FBI. Under the Freedom of Information Act, the Murder Accountability Project has obtained data on more than 27,000 homicides that were not reported to the Justice Department.

Because the Project's website reports that approximately 30% of homicides are unsolved within a year, predicting the outcome of a case seemed interesting. The police could use this information to make a decision on how to approach a homicide investigation. If the case is predicted to be unsolved, law enforcement needs to choose the number of resources to dedicate to it. Does it dedicate few detectives and little time? Or does it do the opposite to increase the probability of it being solved?

Tools and Data:

The homicide database contains over 750,000 murders from 1976 to the present. Some of the features were irrelevant, unexplained or redundant, such as "Case File #", "Number of Records", and "State", and the "ethnicity" information was sparse. These features were removed. In addition, any information on the offender and his/her relationship to the victim also had to be withdrawn as unsolved cases would not have this data. This left the dataset with details that would be obvious to the police coming upon a homicide crime scene (or within a very short amount of time). These include the victim's age, race, and gender; the Metropolitan Statistical Area (MSA) and type of law enforcement; the number of people killed and weapon; and the month and year.

Python was used for loading, preprocessing, and modeling the data, The Project's website had a lot of interactive exploratory data tools for analysis. This decreased the amount of Exploratory Data Analysis that needed to be done, however the remainder of the exploratory analysis was also done in Python.

Tableau was used to make two charts for its ease of use. Unsolved Murder Charts in Tableau

Metrics:

Before starting modeling, metrics were established. Ideally, all solved cases would be predicted correctly, as would all unsolved cases. The unsolved cases that are predicted correctly means that although a killer is free, the police can reopen the case later. However, an unsolved case that is considered solved not only means that a criminal is free, it also means that the case will never be re-examined. Thus, Recall, which measures the true positive prediction against all positive predictions was chosen as the primary metric. Accuracy was not considered as the data set was imbalanced (70/30), and the accuracy score would be skewed by the large number of solved crimes. The Area Under the Curve (AUC) was chosen as a secondary metric to compare the best overall models in terms of precision and recall.

Modeling, Results, and Interpretation:

Numerous models were used to optimize the recall score. These include: knn, logistic regression, LinearSVC, SVC, random forest, and gradient boosting. An optimized random forest gave the highest recall score. Gradient boosting produced the highest AUC.

The non-default hyper-parameters used in the random forest were model max_depth: 100 and n_estimators: 50. This means the highest recall was the model with 50 random trees each 100 branches deep. The ultimate recall score was 0.43. Although this is quite low, it is better than the baseline recall of 0. The model gave an AUC score of 0.675. This is not a great score, but it is in line with the other models, and it is better than the baseline of 0.5 The gradient boosting model had the highest AUC of 0.72.

The features that were most important were year and age of the victim.

Recall gives the percentage of correctly predicted unsolved cases out of all the unsolved case predictions. As an example, with 100 cases, about 30 would be unsolved. With the model, only 13 (30 x 0.43) of these would be classified correctly. However, with no model and just guessing everything will be solved gives no correctly classified unsolved cases. These are both bad situations -- killers are left free. With the random forest model, 13 killers would be classified as free, but law enforcement would be aware of the situation and could reopen the cases. The other 17 unsolved cases mean killers are free, and law enforcement would think the case is closed. Hence, unsolved cases predicted as solved are the worst type of predictions. The baseline model would not predict any unsolved cases correctly. It would leave all 30 killers free, but law enforcement would believe the case is closed.

A Comparison:

This data set is on Kaggle. One user created a similar project a year ago. Murder Comparison . The approach is similar, but there are differences with models, metrics, and some features . One interesting difference is that year was not an important feature, as it was in this project. Although the years were grouped by decades, this should not have made a large difference. It is possible the difference is the data set. The 70's had the lowest unsolved rates of all the decades by quite a significant percentage, and 2016 had the second highest unsolved rate since 1976. The user only had data from 1980 - 2014, which did not have the extremely low unsolved rate early and high rates in recent years.