Would Your Murder Be Solved?

Laura Godleski
Lindsay MacDonald
Natalie VanDyke

Overview

WHO IS OUR TEAM?

Lindsay MacDonald

Natalie VanDyke

Laura Godleski

WHAT IS THE TOPIC?

Influence of demographics on the homicide solve rate in California.

WHY THIS TOPIC?

Data Sources

U.S. Homicide Reports, 1980-2014

https://www.kaggle.com/jyzaguirre/us-homicide-reports

US Census Demographic Data

https://www.kaggle.com/muonneutrino/us-census-demographic-data/data?select=acs2015_county_data.csv

Questions to Answer

- Is there a correlation between the personal attributes of the victim and the solve rate?
- Is there a correlation between county demographics where the murder occurred and the solve rate?
- Can we develop a machine learning model that predicts whether or not a crime would be solved given hypothetical sets of circumstances?

Technologies Used

Data Exploration: Data Cleaning

Kaggle Dataset #1 (US Homicide Reports, 1980-2014)

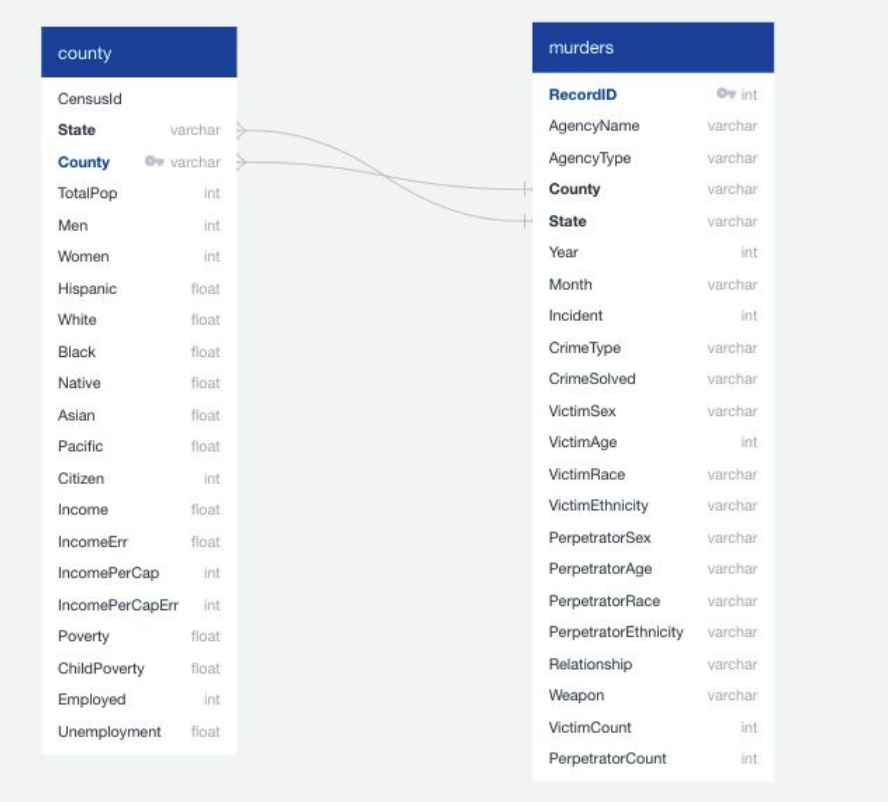
- Filtered ‘State’ column for California
- Renamed ‘City’ column as ‘County’ in order to merge with with the second dataset
- Dropped irrelevant columns
- Checked for null values (0)

Kaggle Dataset #2 (US Census Demographic Data)

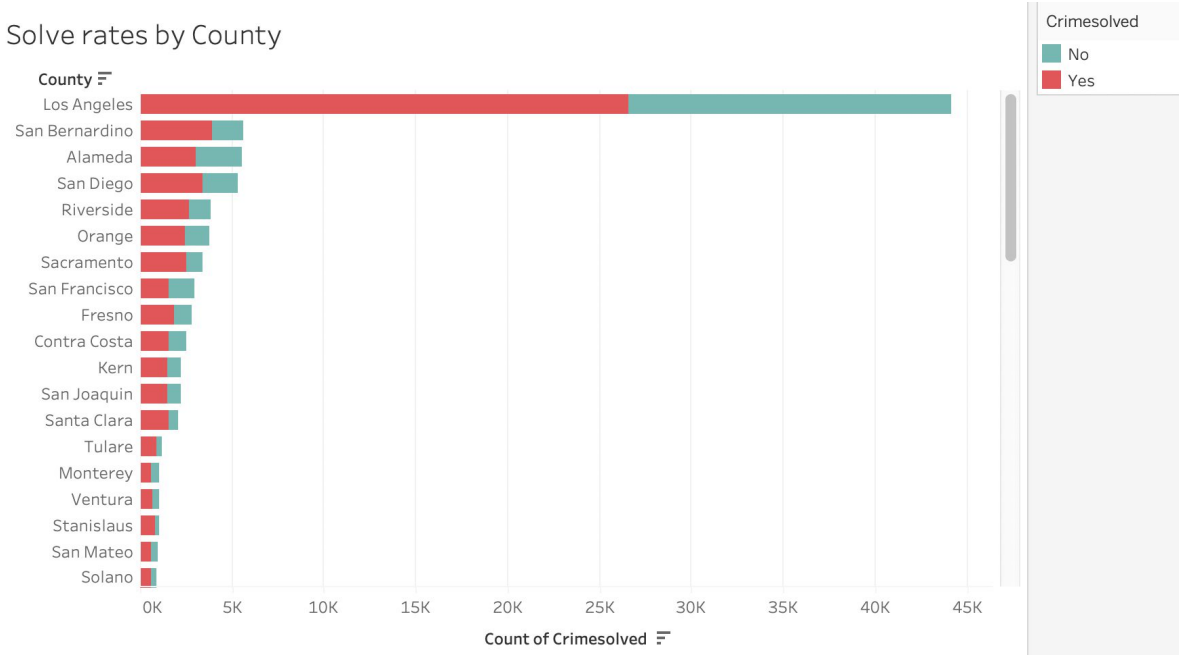
- Filtered ‘State’ column for California
- Dropped irrelevant columns
- Checked for null values (0)

Datasets were imported into Postgres and joined; merged dataset was then uploaded to AWS S3 bucket

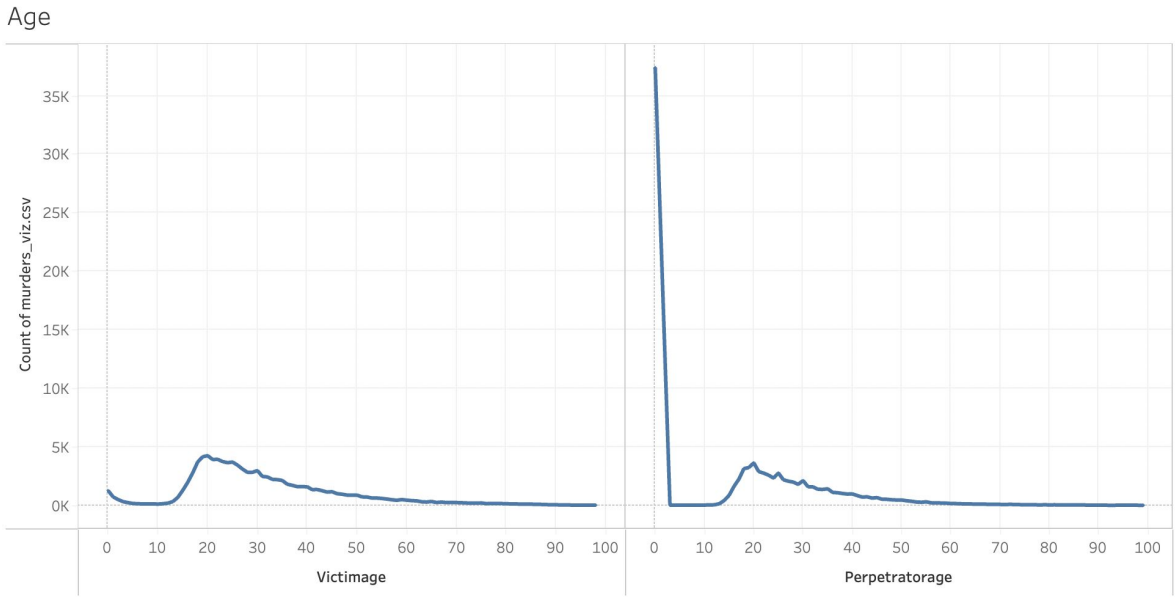
Database Structure



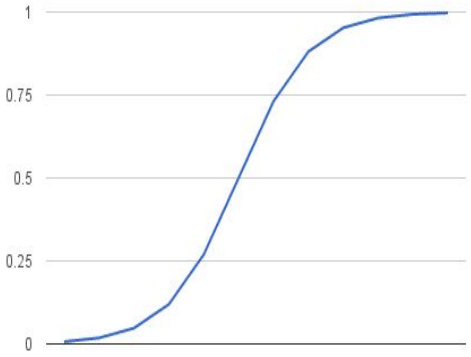
Data Exploration: Initial Findings

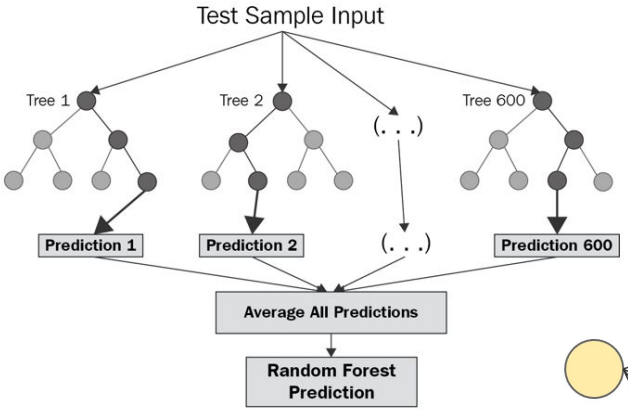


Data Exploration: Initial Findings



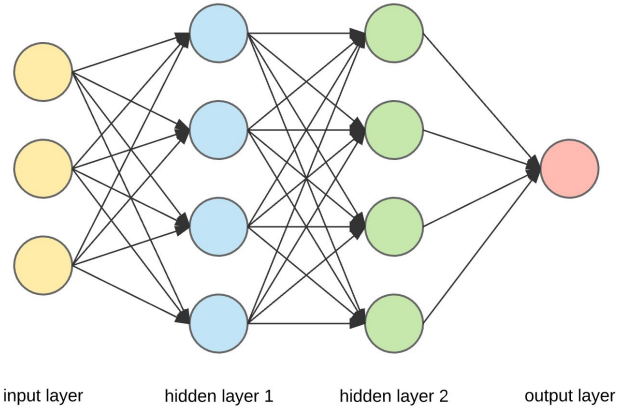
Analysis: Machine Learning





Model Selection:

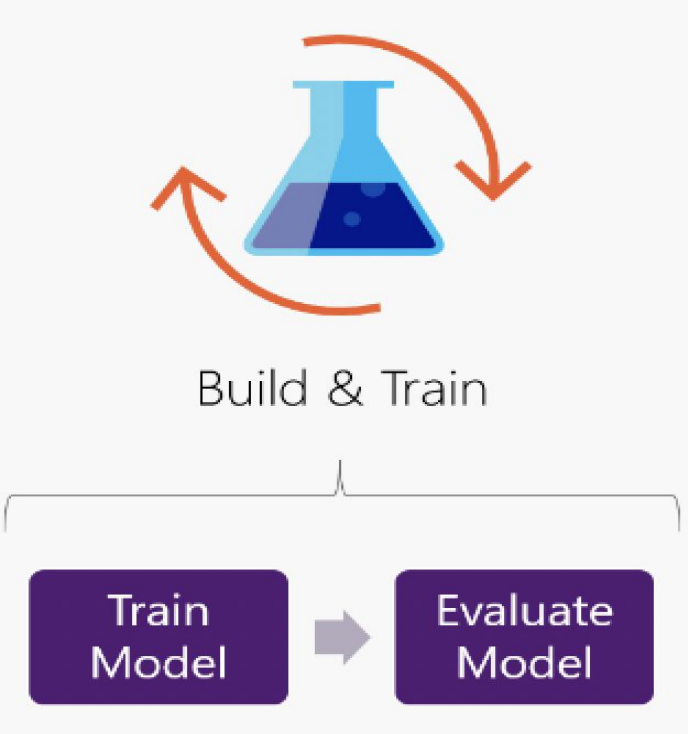
Pros and Cons



Analysis: Machine Learning

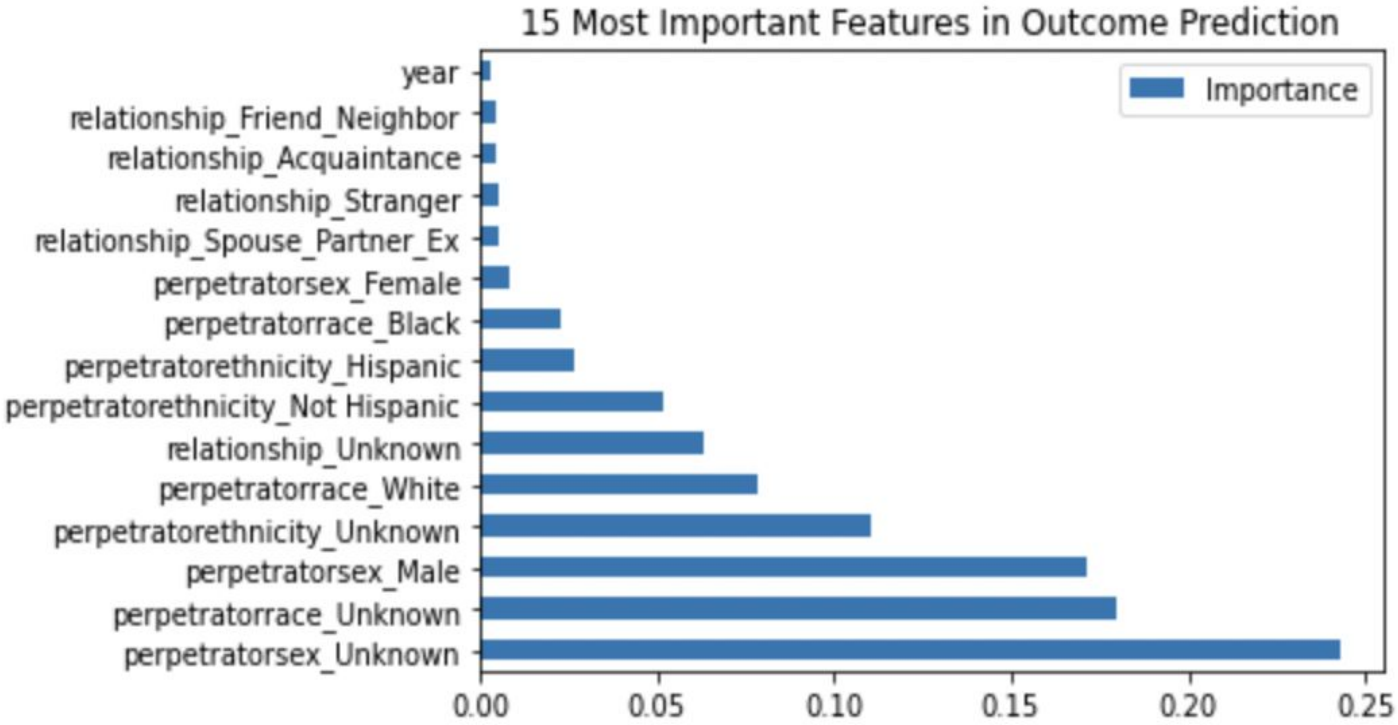
- Import data from AWS
- Preprocessing steps
- Feature Extraction
- Train_Test_Split() & StandardScaler()

Analysis: Machine Learning

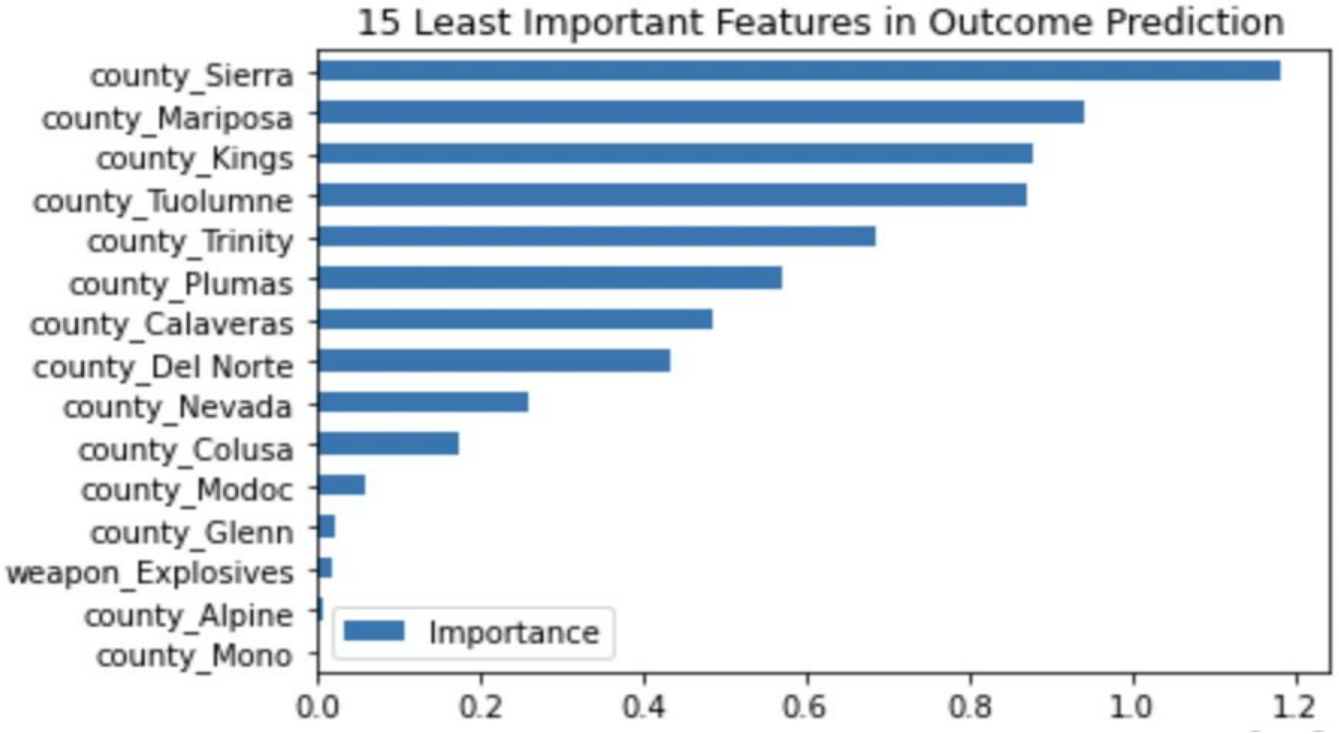


- Accuracy Scores
- Confusion Matrix Results
- Feature Ranking
- Rinse & Repeat

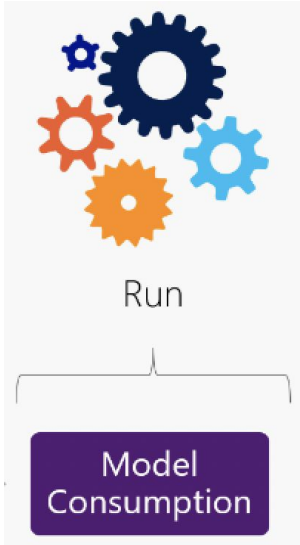
Analysis: Machine Learning



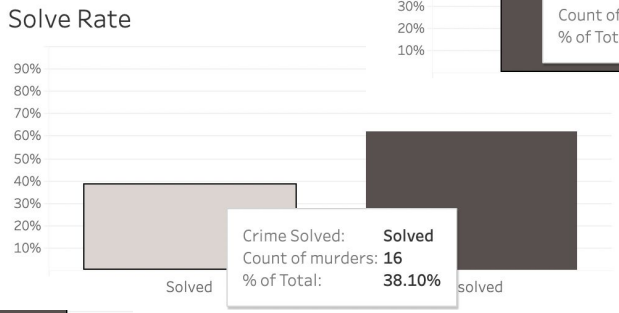
Analysis: Machine Learning



Analysis: Machine Learning



Predictions for hypothetical scenarios



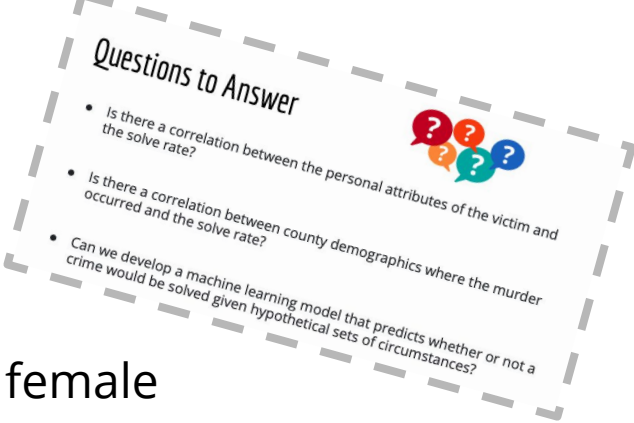

```
# Run the model on new data
new_reg_pred = regressor.predict(new_scaled_data)
new_reg_pred
```

```
array([0.99414333, 0.29538321, 0.99975    ])
```

Analysis: Dashboard

Link to Interactive Dashboard

Key Findings



- Crimes are solved more frequently when the victim is female
- The younger the victim, the more likely it is that the crime will be solved
- The crime is more likely to be solved when the victim is white
- Counties with higher rates of citizenship have a higher crime solve rate
- Counties that are predominantly white have a higher crime solve rate

Unexpected Results

- As regional income increased, solve rates decreased
- When normalized by per capita, homicide rates were much higher in Alpine County and much lower in San Diego than when compared to the raw numbers
- While the majority of perpetrators were male, the majority of victims were female

Recommendation for Future Analysis

- Interactivity for machine-learning portion of the analysis
- Tableau stories and filter table based on perpetrator demographics in addition to victim demographics
- Census data expanded to include the entire time frame present in the homicide data

Things We Would Have Done Differently

- More data exploration in the beginning
- Coordinated dashboard and machine learning earlier on