SMU Data Science Boot Camp | Project 4 Write Up

Homicide in the Past Decade

I. Introduction

“Homicide is defined as the killing of one human being by another. Although homicides are generally thought of as criminal acts, such as murder or manslaughter, some homicides are considered lawful because they are “justified” for reasons such as self-defense. According to the Department of Justice, the rate of homicide in the United States has declined dramatically in recent years, falling to almost half in the past decade. The rate of homicide remains the greatest amongst young adults age 18-24, and males are over three times more likely to commit a homicide than females.” (Homicide Under the Law | Criminal Law Center | Justia)

II. The Datasets

For our project, we employed a total of three datasets. Our main one is Homicides from Kaggle. In this dataset, the Washington Post gathered info on over 52,000 criminal homicides from the past decade (2007-2017) in 50 of the biggest U.S. cities. They looked at where the incidents happened, whether arrests were made, and got basic info about the victims. Reporters got the data in all sorts of forms, even on paper, and spent months cleaning and organizing it. They double-checked the numbers and closure rates against FBI data to make sure everything was on point. Sometimes, police departments only give partial info about the cases. So, the reporters dug into public records like death certificates, court files, and medical examiner reports to fill in the gaps. This dataset goes into more detail than the yearly FBI data that covers homicides from police all over the country. The Post also made a map for each homicide, showing arrest rates in different parts of each city. They even shared their analysis with local police departments before putting it out for everyone to see. This dataset is 5.4MB, contains 12 columns, and has around 52.2k of data on victims. The homicide data from the Kaggle dataset contained demographic data, location data, time data, and most importantly, disposition data which shows whether the cases were open or closed and if an arrest was made. Apart from basic demographic visualizations, our intent is to create a predictive model to determine whether a specific homicide victim’s case would end in an arrest or not. Since the data is from 50 different US cities with different population sizes, we will us Census data from the timeframe of the data set to incorporate the number of homicides per city by population size. This combination should improve our ability to observe patterns based on city size as well as incorporate population as a variable in our predictive model.

1) Homicides: https://www.kaggle.com/datasets/joebeachcapital/homicides
2) SubEst2010: https://www.census.gov/data/datasets/time-series/demo/popest/intercensal-2000-2010-cities-and-towns.html
3) SubEst2020 https://www.census.gov/programs-surveys/popest/technical-documentation/research/evaluation-estimates/2020-evaluation-estimates/2010s-cities-and-towns-total.html

III. Data Cleaning

In the execution of our project, we used Jupyter Notebook along with libraries including pandas, pathlib, and numpy. Our data cleaning were divided into three key stages: rewriting or removing data, binning data, and prepping data for merger.

Under the 'Rewriting or Removing Data' phase, we carried out a series of pivotal adjustments. Notably, we rectified the state abbreviation for Wisconsin, transforming 'wI' to the standardized 'WI'. Additionally, we identified an entry for a city erroneously labeled as Tulsa, AL, which was subsequently corrected to Tulsa, OK. To streamline the dataset, we opted to omit the 'victim_last' and 'victim_first' columns, deeming them non-essential for demographic analysis. Furthermore, we addressed anomalies in the 'reported_date' field, rectifying instances with extraneous '1's. A significant transformation entailed converting disposition options to a binary outcome, specifically from "Closed without arrest" and "Open/No arrest" to "No Arrest", and from "Closed by arrest" to "Arrest Made", a task accomplished using the df.column.replace method. The 'reported_date' was further dissected into year, month, and weekday components, using datetime features.

In the 'Binning Data', we categorized victim ages into five brackets: '0-17', '18-29', '30-44', '45-64', and '65+'. Additionally, a function was used to derive a 'season' column from the 'month' column, providing a seasonal perspective on the data.

In the 'Merging Data' phase, we took steps to ensure seamless integration across datasets. This included generating a unique identifier for each city through the creation of a 'LOCATION' column derived from the combination of 'city' and 'state'. Subsequently, we established two distinct data paths: one tailored for Machine Learning and the other for Tableau visualization.

For the Machine Learning part, we removed rows with missing demographic information. Unfortunately, this led to the exclusion of entire datasets from three cities, namely Dallas, Phoenix, and Kansas City. As a result, the dataset count was reduced from 52,179 to 47,478, comprising high-quality, usable data. As a whole, we lost about 10% of the dataset. Fortunately, Tableau did not drop these cities but focused on being usable for total counts.

Moving to the population data from 'SubEst2010.csv' and 'SubEst2020.csv', a rigorous cleaning process was applied. This involved identifying and extracting the requisite cities and years. To facilitate seamless merging with the homicide data, we standardized state names from their full form to the 2-letter abbreviation. Additionally, we rectified city columns by removing extraneous designators. The creation of a 'LOCATION' unique id column from the combination of 'NAME' and 'STNAME' further streamlined the dataset. Subsequently, we refined the dataset to include only the necessary tables and merged the 2010 and 2020 datasets. By utilizing the saved 'LOCATIONS' from the homicide data, we ensured that only the 50 (47) pertinent cities were retained. A melt function was then applied to generate a row for each year per city.

Finally, we performed a merge of the homicide and population datasets, incorporating population data into the homicide data, a crucial step in our analysis. This facilitated the creation of a dataset illustrating homicides per 100,000 by city and year.

Our data cleaning efforts resulted in three distinct, meticulously cleaned datasets: 'ml_clean_homicide_data.csv', 'tbl_no_drop_homicide_data.csv', and 'tbl_homicide_count_per_100000.csv'. These refined datasets serve as the foundation for our subsequent machine-learning analyses and visualizations.

IV. Tableau Dashboard

Tableau dashboard page 1 explains about the distributions of the homicides across US cities. The dashboard has filters based on Cities & a range of counts of homicides.

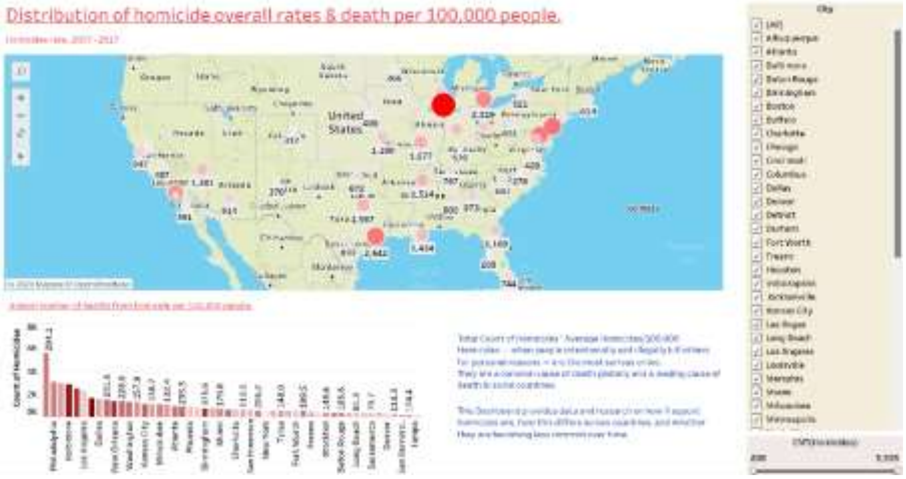
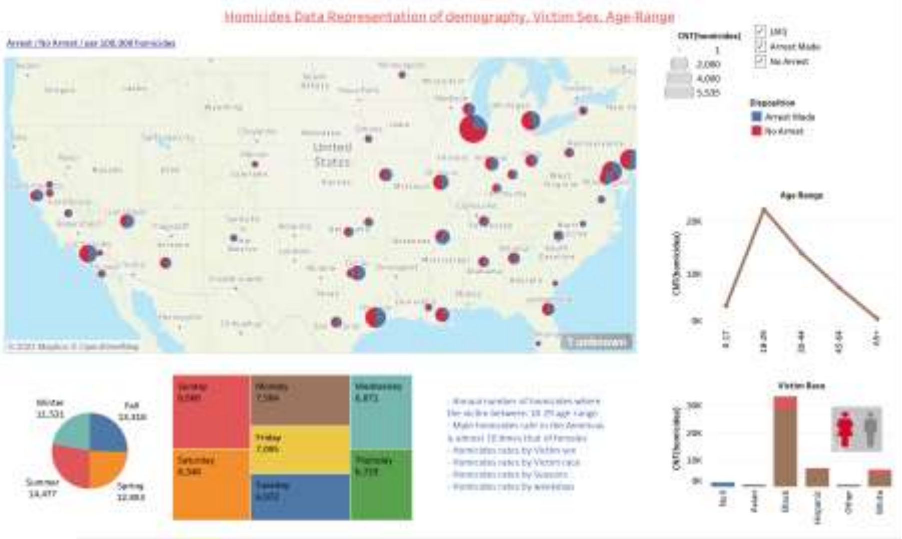


Tableau dashboard page 2 explains demography wise distributions of the homicides across US cities. The dashboard has filters based on “Arrest” “no arrest” & a range of counts of homicides



V. The Website

The creation of the homicide website is motivated by the aim of providing a valuable online resource for individuals seeking insights into homicide data. This website utilizes a dataset spanning from 2012 to 2017, with key features and technologies outlined below. The site's structure follows a standard navigation bar, focusing on two main elements:

1. Arrest Prediction Page: This section utilizes data from 2012 to 2017 to offer predictive insights concerning arrests in homicide cases.

2. Dataset Overview Page: We've dedicated a page to present a comprehensive overview of our dataset, integrating the Ydata library. This dataset overview was initially developed within Jupyter Notebooks and later exported as an HTML file.

Additionally, a designated 'Resources' page consolidates various materials that contributed to the website's development. Python and Flask were utilized to construct the website's framework. The user interface centers around a simple form that guides users in inputting data, with clear formatting instructions provided at the top.

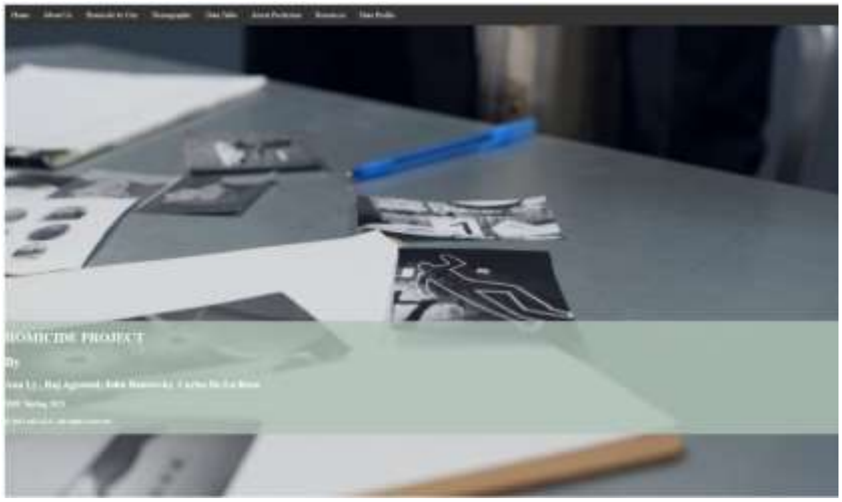
Data visualization is presented on two distinct web pages, both created using Tableau. These Tableau pages, originally designed on Tableau Public, have been seamlessly integrated into the website through the Tableau API. This ensures quicker response times compared to embedding HTML code.

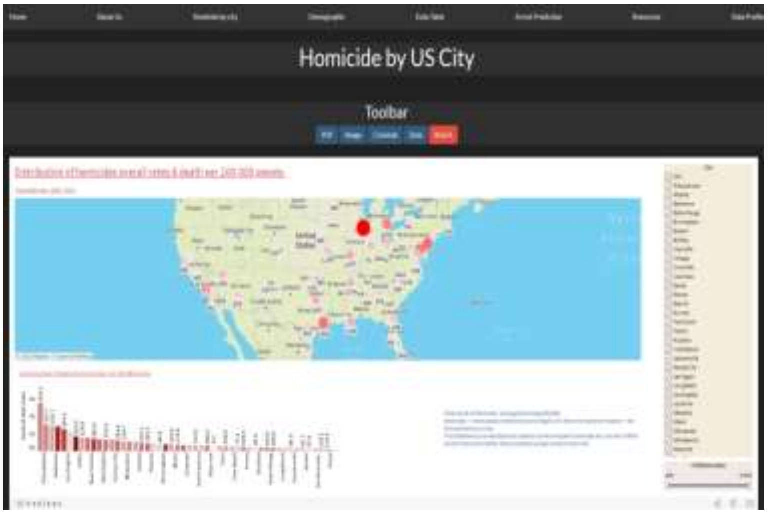
Deployment of the website was achieved through PythonAnywhere, mirroring the same website structure established during local development. Although the deployment process encountered minimal technical hurdles, a notable issue emerged due to the absence of a required library for the Light GBM (LGBM) machine-learning model on the PythonAnywhere platform. Ultimately, with the assistance of our professor, we successfully deployed the machine learning model on our website.

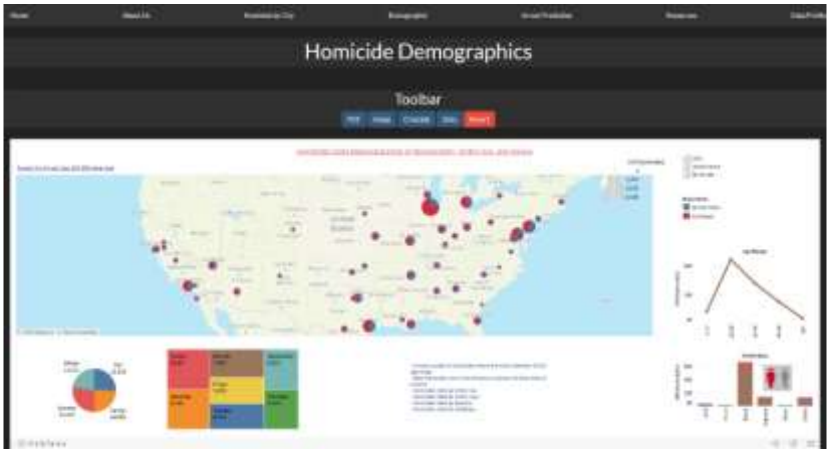
The Flask-based website comprises ten pages in total. The main page, designed to immediately capture attention, features a dynamic background—a looping video clip. This choice aims to create a visually engaging backdrop, emphasizing the sensitivity of the homicide topic.

The website's color scheme was purposefully selected from Bootswatch's 'Darkly' theme. This decision was made to maintain a tone of seriousness, respecting the gravity of the subject matter—representation of homicide data. It underscores our commitment to ensuring that the importance of this data, which represents victims of homicide, is never overshadowed by colors that do not align with the gravity of the topic.

Our Website: http://datadan4310.pythonanywhere.com/







IV. Machine Learning

Our primary goal in this project is to conduct a comprehensive analysis of our dataset, followed by meticulous data cleaning, merging, and refinement processes. This refined dataset will serve as the cornerstone for creating an informative visualization dashboard on Tableau, effectively narrating the intricate story within the numerical data.

We've also developed an interactive website to offer our audience vivid graphical representations of the homicide dataset. Additionally, we've embarked on building a classification machine-learning model. This model is designed to predict the likelihood of a suspect being apprehended based on victim-related data. Our evaluation involved various machine learning models like K-Nearest Neighbors (KNN), Random Forest (RF), AdaBoost (Ada), Extra Trees (ET), Gradient Boosting (GB), XGBoost (XGB), and LightGBM (LGBM) using Jupyter Notebook.

After thorough testing, we've singled out LightGBM as our primary predictive model. We then conducted an in-depth study to see if fine-tuning its hyperparameters and applying cross-validation could improve its predictive accuracy. Despite our best efforts, the resulting accuracy remained consistent with the original model. This was further confirmed through cross-validation using K-Fold methodology.

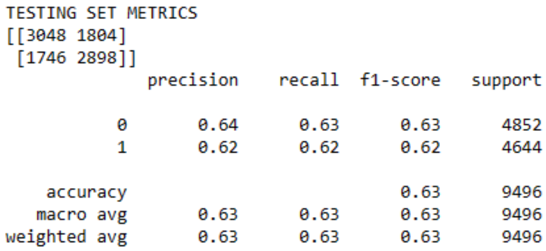
It's important to acknowledge certain limitations in our dataset. It covers a decade, from 2007 to 2017, offering a snapshot of that time frame. Additionally, it's limited to 50 major cities in the United States. Notably, New York City contributes only two years of data compared to the ten years provided by other states. We believe a more comprehensive, uniformly covered dataset across a broader range of U.S. cities would lead to a more nuanced and accurate predictive system.

In summary, our focus has been on meticulous data preparation, model selection, and thorough evaluation. We recognize the potential for improvement with a more expansive and representative dataset. This has the potential to significantly enhance the accuracy and depth of our predictive system.

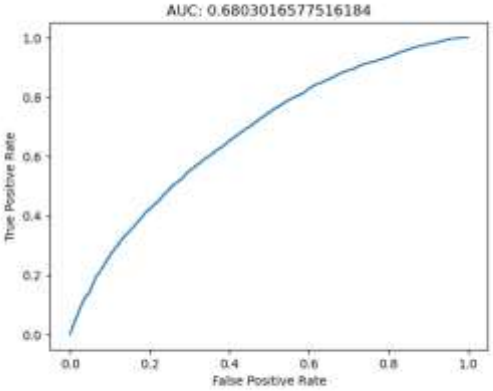
The following slides illustrate the stages of our machine learning process, systematically progressing from target selection through model training, culminating in the critical phases of hyperparameter tuning and cross-validation."

Machine Learning (cont.)

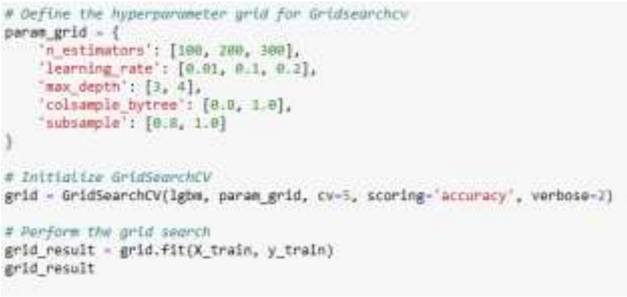
LGBM Machine Learning



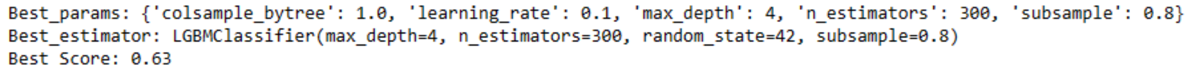
```
TESTING SET METRICS
[[3048 1804]
 [1746 2898]]
              precision    recall  f1-score   support

           0       0.64      0.63      0.63      4852
           1       0.62      0.62      0.62      4644

    accuracy                           0.63      9496
   macro avg       0.63      0.63      0.63      9496
weighted avg       0.63      0.63      0.63      9496
```



Hyper Parameter Tuning on LBGM using GridSearchCV

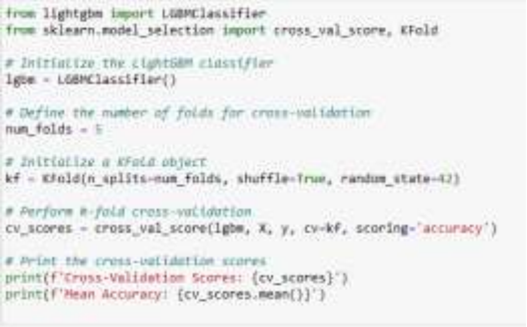


```
# Define the hyperparameter grid for Gridsearchcv
param_grid = {
    'n_estimators': [100, 200, 300],
    'learning_rate': [0.01, 0.1, 0.2],
    'max_depth': [3, 4],
    'colsample_bytree': [0.8, 1.0],
    'subsample': [0.8, 1.0]
}

# Initialize GridSearchCV
grid = GridSearchCV(lgbm, param_grid, cv=5, scoring='accuracy', verbose=2)

# Perform the grid search
grid_result = grid.fit(X_train, y_train)
grid_result
```

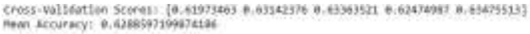


```
Best_params: {'colsample_bytree': 1.0, 'learning_rate': 0.1, 'max_depth': 4, 'n_estimators': 300, 'subsample': 0.8}
Best_estimator: LGBMClassifier(max_depth=4, n_estimators=300, random_state=42, subsample=0.8)
Best Score: 0.63
```

Cross Validation Using KFold



```
from lightgbm import LGBMClassifier
from sklearn.model_selection import cross_val_score, KFold

# Initialize the LightGBM classifier
lgbm = LGBMClassifier()

# Define the number of folds for cross-validation
num_folds = 5

# Initialize a KFold object
kf = KFold(n_splits=num_folds, shuffle=True, random_state=42)

# Perform k-fold cross-validation
cv_scores = cross_val_score(lgbm, X, y, cv=kf, scoring='accuracy')

# Print the cross-validation scores
print(f'Cross-Validation Scores: {cv_scores}')
print(f'Mean Accuracy: {cv_scores.mean()}')
```



```
Cross-Validation Scores: [0.61973463 0.63542376 0.63363521 0.62474987 0.63475513]
Mean Accuracy: 0.6288597199874186
```

VI. Summary

In summary, this dataset does not provide enough information to give us an insight into how minority, and racial profiles and the roles that they played in homicide cases and the outcome of the case whether the suspect is arrested or not. It only covers 50 major cities in 50 states. The data only covers homicide information from 2007-2017. Here are some of the results that we derived from this dataset.

- The outcome of both dispositions is very close to each other, 51% of no arrests made and 49% of arrests made based on all cases.
- 85% of victims are male, and the other 15% are female.
- The victim race is breaking down as follows:
  - Black: 70%
  - Hispanic: 14%
  - White: 13%
  - Asian/Other: 3%
- The age range of the victims is breaking down as follows:
  - 45% in the 18-29 age group
  - 28% in the 30-44 age group
  - 15% in the 45-64 age group
  - 8% in the 0-17 age group
  - 3% in the 65+
- Summer months (June, July, August) are the leading months with most homicides
- Sunday, Saturday, and Monday are the most reported weekday where homicides occurred.
- Chicago, Philadelphia, and Houston are the 3 leading cities with most homicides reported.
- California, Illinois, and Texas are the 3 leading states with most homicides reported.

References:

Homicide Under The Law: Homicide Under the Law | Criminal Law Center | Justia
Homicides: https://www.kaggle.com/datasets/joebeachcapital/homicides
SubEst2010: https://www.census.gov/data/datasets/time-series/demo/popest/intercensal-2000-2010-cities-and-towns.html
SubEst2020 https://www.census.gov/programs-surveys/popest/technical-documentation/research/evaluation-estimates/2020-evaluation-estimates/2010s-cities-and-towns-total.html

Project Team Student Members:

- Carlos Delarosa
- John Banowsky
- Raj Agrawal
- Ann Ly

- **Faculty**
- Alex Booth - Instructure
- Sherhone Grant - TA
- Sean Fleming - SSM