Dallas Crime:

An Investigation into Analyzing and Predicting Dallas Crime for 2020

Alex Arnold
Samantha Lane
Rachel Podemski

Executive Summary

"Half the crime in the city [Boston] came from 3.6 percent of the city's blocks…And every place they looked, they saw the same thing: Crime in every city was concentrated in a tiny number of street segments. Weisburd refers to this as the Law of Crime Concentration. Crime is tied to very specific places and contexts."[1]

So what are these places and contexts that crime is tied to? In this report, we investigate crime in Dallas for the year 2020 by uncovering trends through visual analysis and creating a machine learning model to predict the status of crime incidents. We'll go on to explain:

- The Dallas crime data we explored.
- How we used feature engineering on this data to create our target and features for our machine learning model.
- Our prediction model of whether an incident will result in an arrest, clearance, or suspension.
- Criminal activity trends visualized by time and location.
- How our findings will support local law enforcement efforts.

The Problem

Crime and criminal activity are prevalent issues that exist in any city. Can we then organize our efforts and focus on certain parts of the city, like Gladwell posts in his book, to better counter crime and criminal activity?

Data

We used open-sourced Police Incident data from the city of Dallas for the entire year 2020. This dataset provided the date and time, location information, demographic information, and type of incident in addition to much more detail within each of these categories. We then cleaned the data by removing nulls, dropping columns, changing object types, and removing incidents due to human-error when entering in information to the dataset.

Visualization Analysis

We developed two dashboards based on time and location to further investigate Dallas Crime. The "Time" Dashboard looks deeper into when crimes occur, what type of crime occurs, and who is involved in each crime incident. Filtering by Crime Category and Month allowed us to determine that most crimes occur in the month of August and on Fridays. The "Location" Dashboard discovers where crime occurs by analyzing crime incidents by division and density.

[1] Malcolm Gladwell, Talking to Strangers: What We Should Know About the People We Don't Know

These maps are also filtered by Crime Category and Month. There is a high crime density in the Central and Southern Area of Dallas, but for the month of February it has the lowest density of overall crime for the city. When filtering by crime, Assault, Larceny/Theft, and Miscellaneous crimes have the greatest density across the city.

Machine Learning Model

Based on our visualization analysis, we were interested to see whether we could predict the status of a crime incident based on it's time and location data. We created a two-step machine learning model using the XG-Boost Classifier that allows us to predict the probability that a reported criminal incident will result in either an arrest, clearance, or suspension.

Conclusion

We accepted our alternative hypothesis and determined that South Dallas experiences the most crime out of each division in our dataset. We also discovered that April had the lowest percentage of crime incidents out of every month in 2020.

After investigating what type of crime occurs the most, we failed to reject our null hypothesis and determined that the most common crimes that occur throughout Dallas are miscellaneous crimes, larceny and theft, followed by assault crimes.

Recommendations

Based on our analysis, we recommend that the Dallas police force have more officers on duty in the month of August and allow officers vacation time in April. We also recommend providing more officers in patrolling in Central and South Dallas areas due to the increase of percentage of crime incidents in those areas.

Limitations

When dealing with live data we often have to deal with discrepancies in our data due to human error. We also were only able to use a subset of data for our visualizations due to Tableau limits on file size for documents being uploaded. Due to scope and time frame, we were only able to analyze data for the year 2020.

Future Work

To further investigate the vast amount of crime data it would be vital to look into trends overtime. It would also be interesting to investigate other types of machine learning models using different targets such as crime category or watch. This would allow us to see what else we could potentially predict and compare the different scores between each model.

Table of Contents

Dallas Crime: An Investigation into Analyzing and Predicting Dallas Crime for 2020

Introduction

"Half the crime in the city [Boston] came from 3.6 percent of the city's blocks…And every place they looked, they saw the same thing: Crime in every city was concentrated in a tiny number of street segments. Weisburd refers to this as the Law of Crime Concentration. Crime is tied to very specific places and contexts."[2]

We were inspired to investigate crime incidents in Dallas after reading Malcolm Gladwell's book *Talking to Strangers: What We Should Know About the People We Don't Know* which is quoted above. We wanted to determine what types of crimes occur, when these crimes occur, and if we could potentially predict the outcome of each crime. This report will explore what impacts the status of a crime and how we can predict this status. It will also investigate and visualize where crime occurs, what type of crime occurs, and when crime occurs.

Research Questions

- What area of Dallas experiences the most crime?
- What time of day does most crime incidents occur?
- What status do most crimes end in?
- What type of crime is most popular in the Dallas area?

Hypotheses

We hypothesize that South Dallas experiences the most crime out of any other area in Dallas and that murder is one of the most common crimes commited in the city of Dallas.

Data Gathering

We sourced our data from the Dallas Open Data website. We chose to analyze the Police Incident Dataset as it is rich with information regarding each incident that is called into 911 for police attention. "An 'incident' is defined for NIBRS reporting purposes as one or more offenses committed by the same offender, or group of offenders acting in concert, and at the same time and place."[3] This dataset provided the date and time, location information, demographic information, and type of incident in addition to much more detail within each of these categories.

This dataset has its own API that we used. The API only allows for 1000 incidents to be reviewed at one time, so we first gathered all of the incidents recorded for 2020 by looping

[2] Malcolm Gladwell, Talking to Strangers: What We Should Know About the People We Don't Know

[3] https://ncc.nebraska.gov/sites/ncc.nebraska.gov/files/pdf/manuals/nibrs/nibrs_man.pdf

through each month in the API. We then saved all of those incidents for 2020 to a CSV to explore and clean for our data analysis and modeling.

Data Cleaning

Pulling all of the Police incidents recorded for 2020 provided us with a very large dataset, so we started by cleaning the data. We first cleaned the data to drop and remove all data pertaining to any location or information outside the city of Dallas. For example, our dataset provided incidents throughout all of Dallas county, which included additional cities. We removed those cities to help narrow our scope.

We cleaned the columns in which we were interested in separating the data within them. For example, the API provided a geocoded column of the latitude and longitude for each incident and we wanted to separate those into their own columns for further analysis.

We changed all string columns to date times that provided us with date, day, year, and time information. We also cleaned the columns with discrepancies in reporting. For example, some districts reported had D in front of the district number while some only reported the number. We added the D in front of all districts to standardize. We also dropped the status columns that were not significant such as “L”, “Returned for Correction,” and “Open” in addition to dropping all nulls provided for this column.

Data Manipulation for Tableau

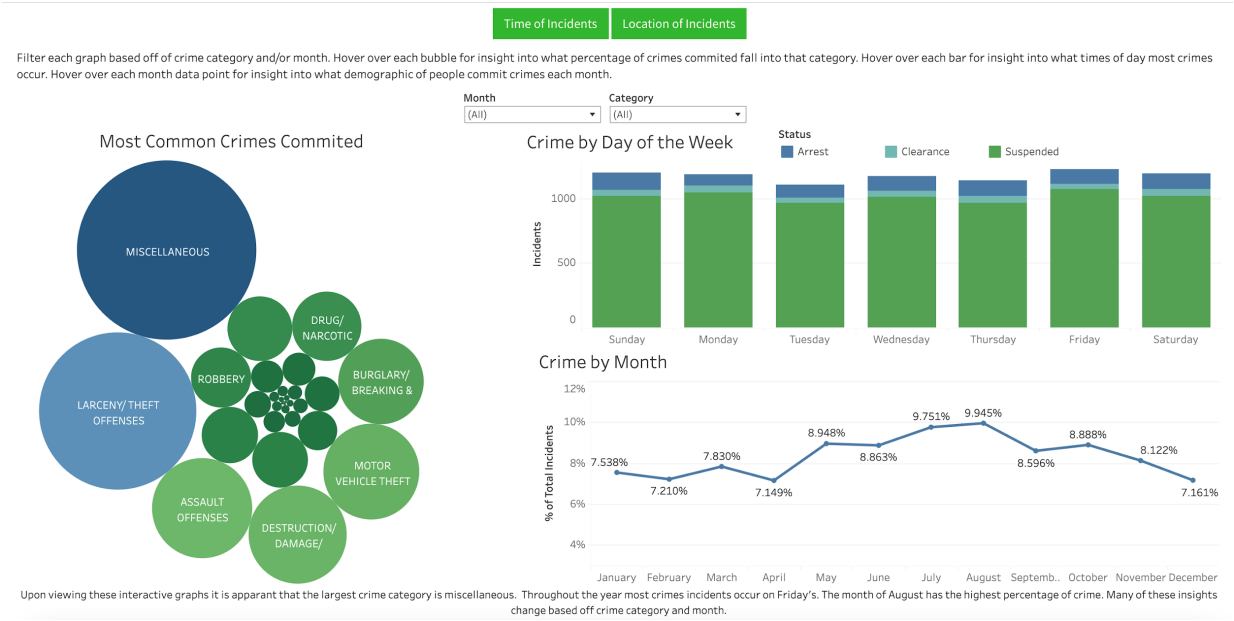
Based on all of the cleaning techniques we performed, we utilized a subset of this data for our Tableau visualizations. After the aforementioned cleaning, our data set was still larger than allowed for Tableau; therefore, we took a random subset of 0.6 of the dataset to be used with Tableau.

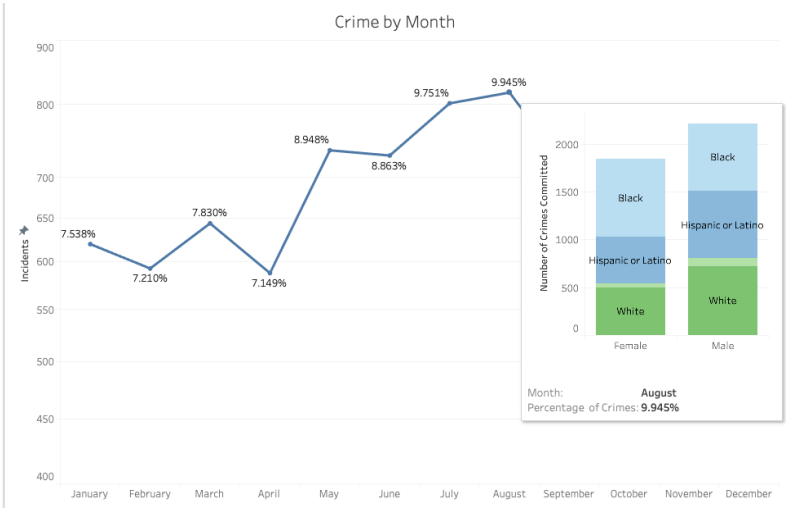
Visualization Analysis

We developed two dashboards based on time of incident and location of incident to further investigate Dallas Crime.

Time Dashboard

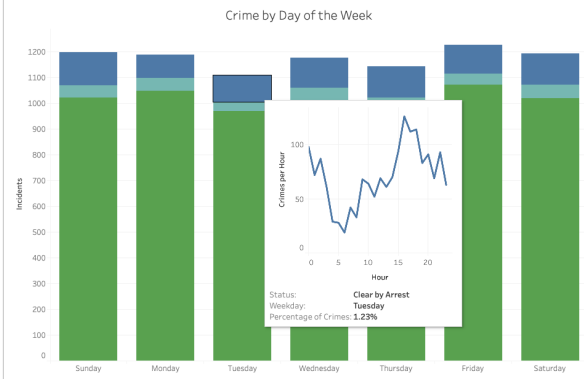
The “Time” Dashboard looks deeper into when crimes occur, what type of crime occurs, and who is involved in each crime incident. You can filter each graph based on the type of crime and the month the crime occured.





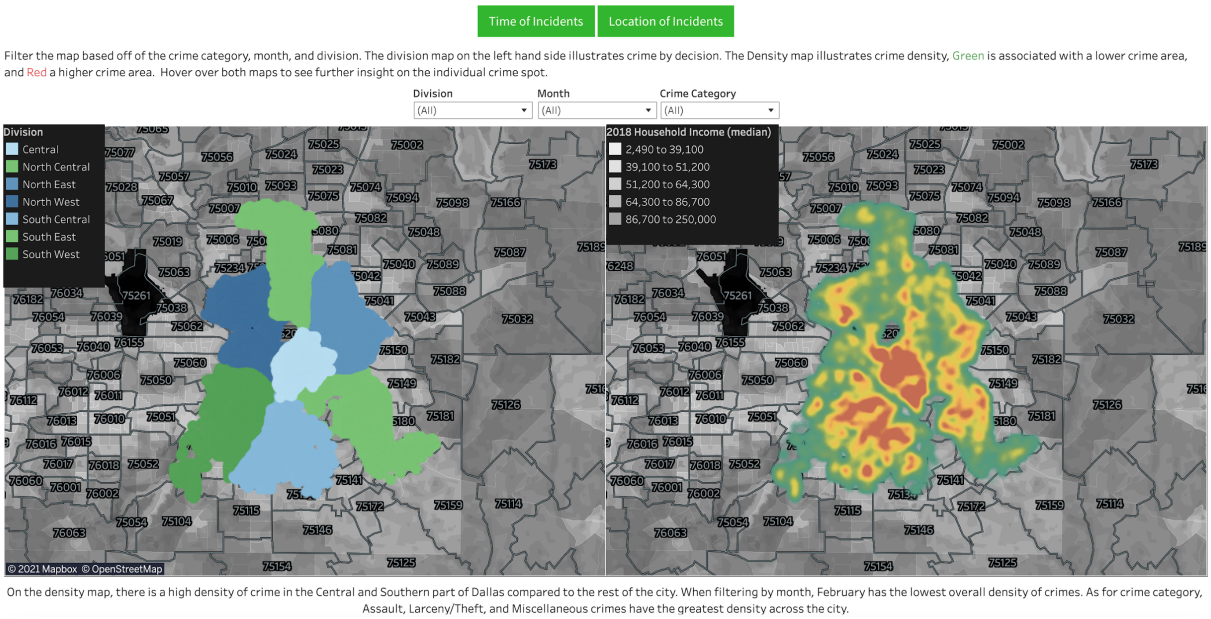
In the image to the left, it is apparent that most crimes occur in the month of August. We can also see in the tooltip graph that more of the crimes that occured in August were committed by men.

In the image to the right we can see that most crimes are committed on Friday with a smaller tooltip graph that represents the times of day that most crimes occur. The smaller graph shows that on Friday's most crimes occur at midnight and 5:00 pm depending on the status of the crime itself.



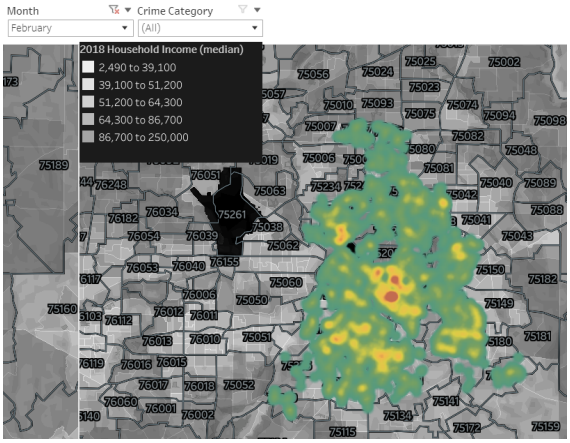
Location Dashboard

The Location Dashboard provides two maps of the Dallas area. The map on the left in the below image shows the areas of Division used by Dallas police. Interestingly enough, each division also has a police station. One can hover over this map and see the data information regarding each particular crime incident. The density heat map on the right in the image below shows the concentration of crime incidents as they pertain to the city of Dallas as well as the divisions.



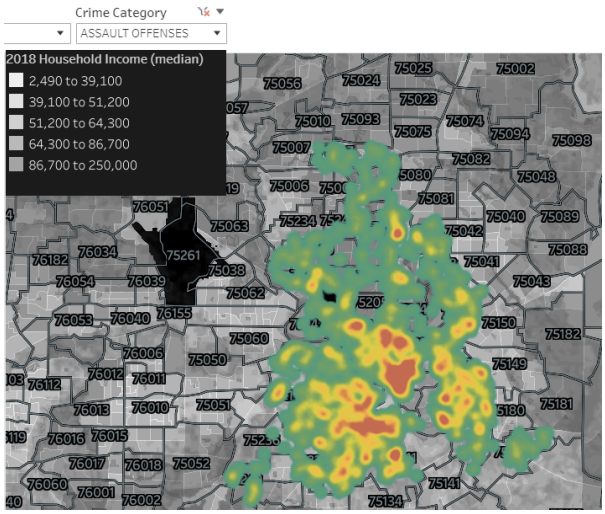
You can further filter the data by crime category and month. There is also an underlying area of household income to visualize if there is a relationship between income and crime.

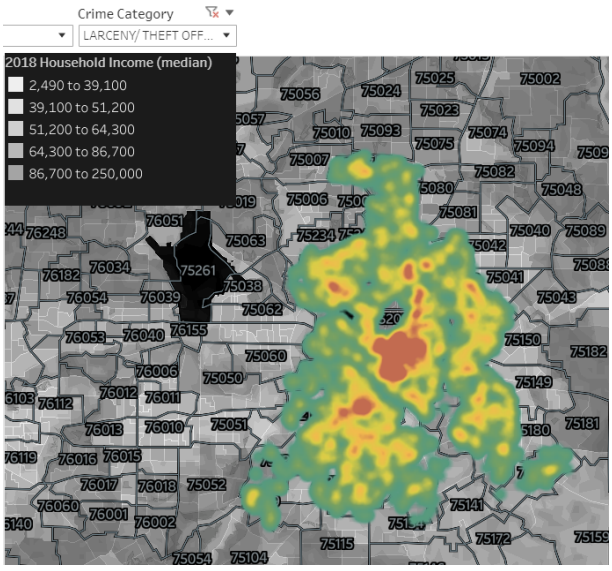
From the density map we can see that there is a high crime density in the Central and Southern Area of Dallas. Zooming in further to the high density crime areas, they are located in areas that are at the lower end of the spectrum for household income.

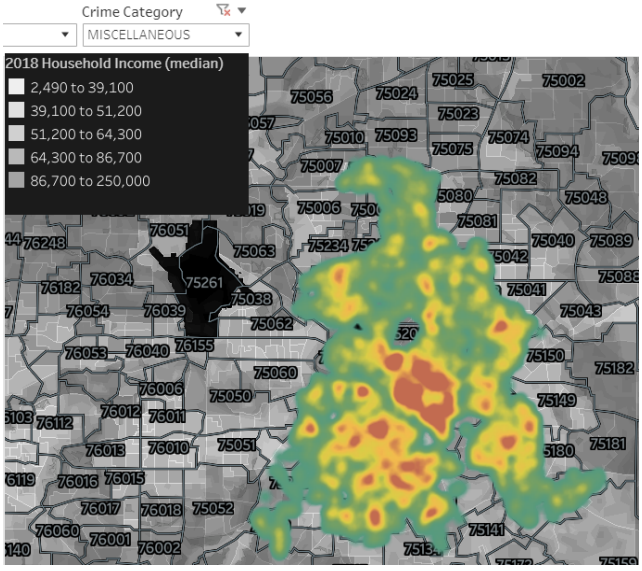


In the image to the left, filtering by month, the month of February can be seen to have the lowest density of overall crime for the city. There is still a large density of crime in the Central part of the city.

In the following images to the right and below, we filter the Crime Category by Assault, Larceny/Theft, and Miscellaneous. These crime categories compared to the rest have the greatest density across the city.







Machine Learning

When exploring the data through our visualization dashboards, we found trends in when, where, and what crimes occurred throughout the city of Dallas. We also found it interesting how the status of each crime differed based on these trends in when, where, and what crimes occurred. Therefore, we wanted to see if we could predict the result of these crimes by predicting whether the status of these crime incidents resulted in an arrest, a clearance, or a suspension.

Feature Engineering for Machine Learning Modeling

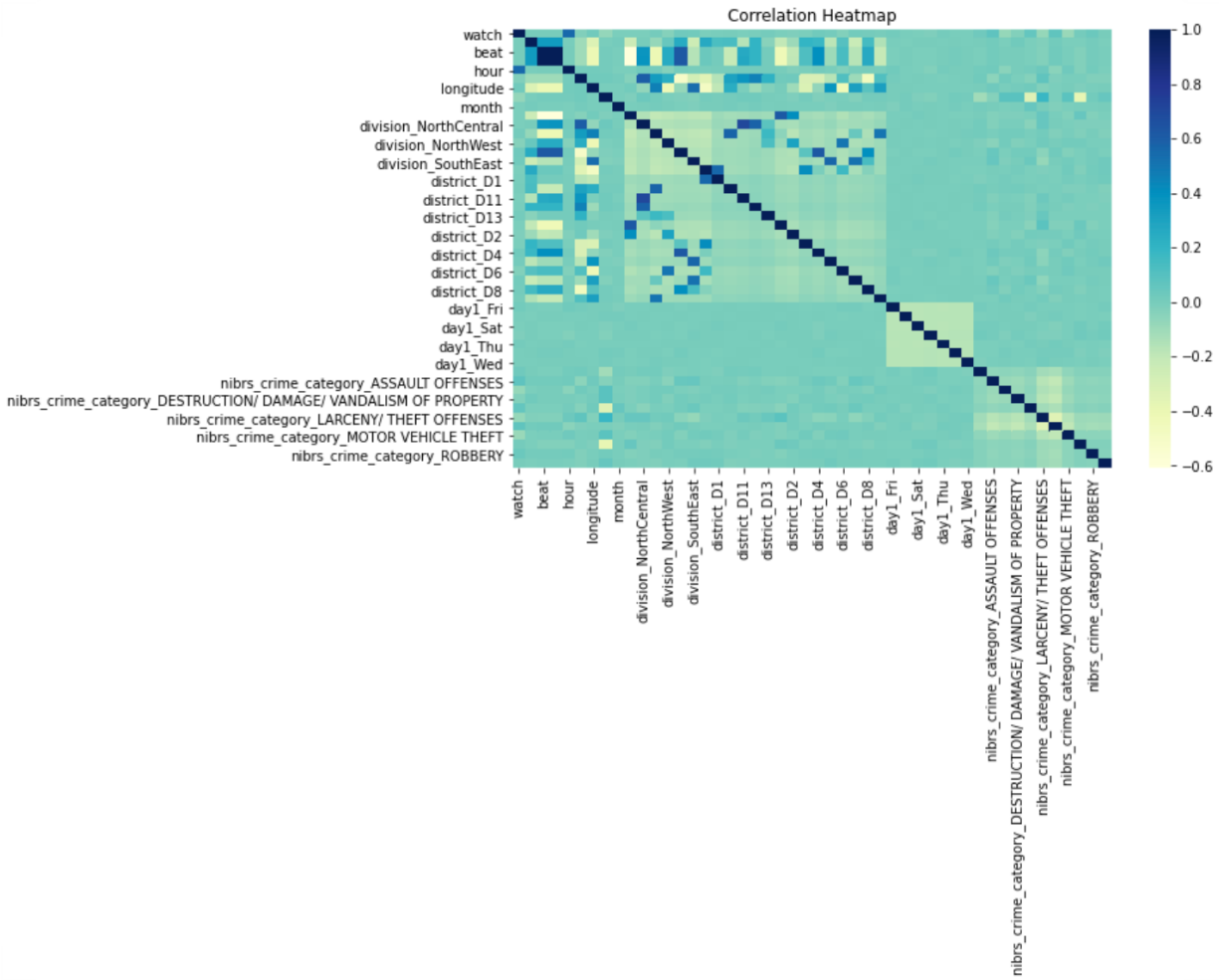
The status of crime incidents became our target, and we used feature engineering to prepare our data features. We dropped insignificant columns and insignificant categories within the nibrs_crime_category column. This column provides the kind of crime as reported by the NIBRS classification system that police units use. We only used the columns that were coded to be classified within this classification system and took out all of those categories not listed here. These other values were an insignificant portion of the overall dataset at only 4% of the total data.

After this, we label-encoded our target of status and the month column. In addition, we dummified our string columns of “division,” “district,” “day1”, and “nibrs_crime_category.” After dummying our data, we ended up with the feature categories of:

- Watch
- RA
- Beat
- Division
- Sector
- District
- Month
- Day
- Crime Category
- Hour
- Longitude
- Latitude

Correlations

The label and one-hot encoding we performed on our features provided a lot of additional columns for our dataset to investigate our correlations. We took a subset of the correlations with the largest correlating factors as displayed in the Correlation Heatmap.



We removed the feature categories with low correlations, leaving us with the categories of:

- Watch
- Hour
- Division
- District
- Month
- Day
- Crime Category
- Hour
- Latitude

These are the finalized feature categories we ran through our Machine Learning model.

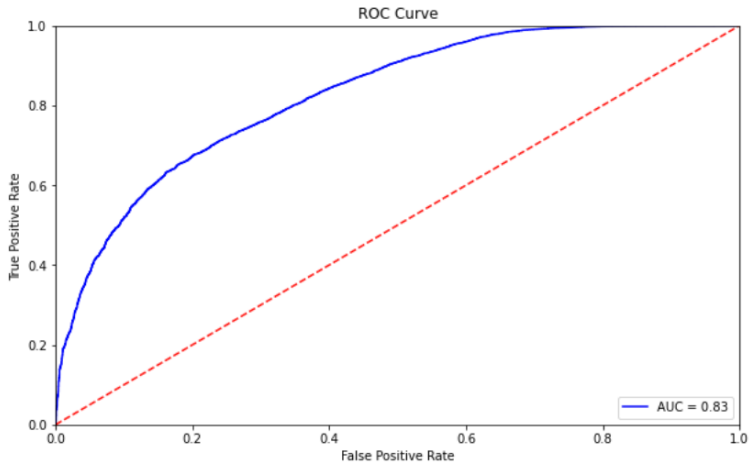
Machine Learning Models

We ran our data using the classification models, but we saw that our model was overfitting the data. The model was overlooking our second class of target status of "Closed/Cleared." To counter this, we decided to use two machine learning models. The first model determines the probability of whether or not our target status ended in suspended or not suspended. Our second model then took the not suspended incidents to predict the probability that the incident either ended in arrest or ended in the incident being closed or cleared.

Model 1

The X-treme Gradient Boost Classifier was the model that we chose to predict the probability of our data for Model 1. It did not overfit the data and it had the highest accuracy, F-1, and AUC scores.

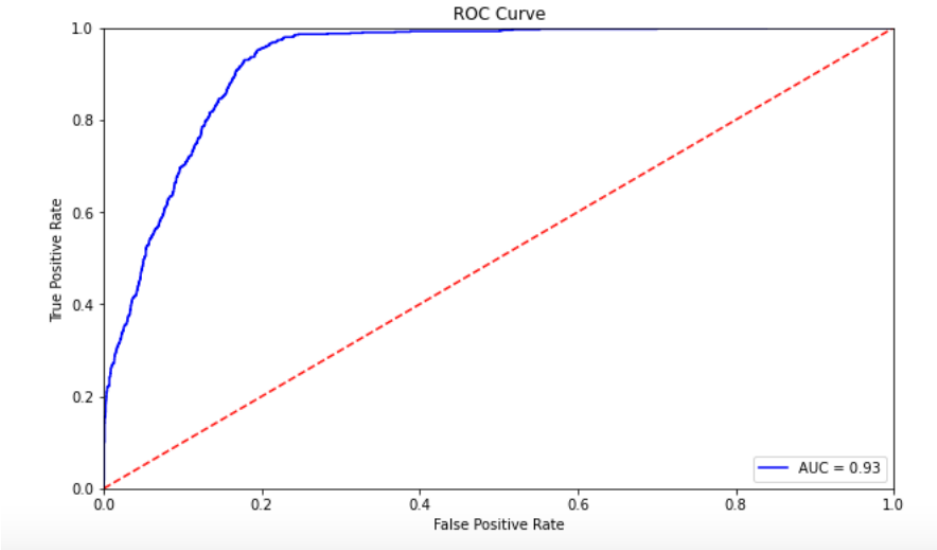
```
XG Boost Classifier:
Classification Report - In Sample
              precision    recall  f1-score   support

           0       0.86      0.37      0.52     17380
           1       0.87      0.99      0.93     77042

    accuracy                           0.87     94422
   macro avg       0.87      0.68      0.72     94422
weighted avg       0.87      0.87      0.85     94422


Confusion Matrix - In Sample
[[ 6452 10928]
 [ 1032 76010]]


Classification Report - Out Sample
              precision    recall  f1-score   support

           0       0.81      0.34      0.48      4345
           1       0.87      0.98      0.92     19261

    accuracy                           0.86     23606
   macro avg       0.84      0.66      0.70     23606
weighted avg       0.86      0.86      0.84     23606


Confusion Matrix - Out Sample
[[ 1491  2854]
 [  339 18922]]
```



Model 2

We also chose to use The X-treme Gradient Boost Classifier to predict the probability of Model 2. It did not overfit the data and it had the highest accuracy, F-1, and AUC scores.

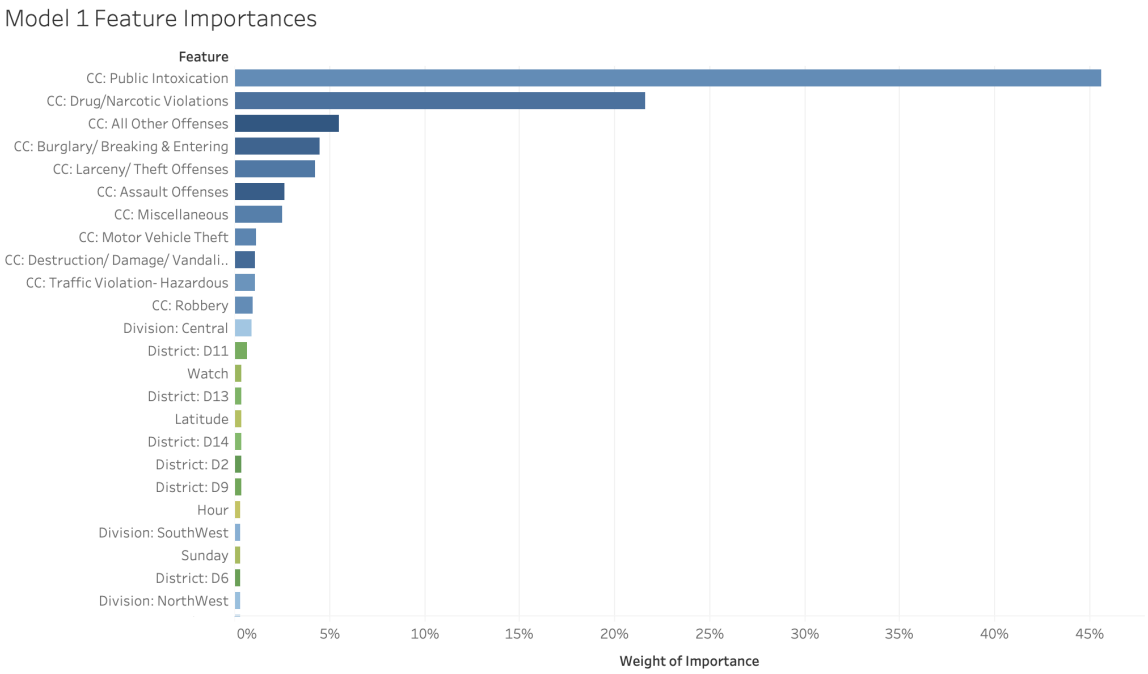
```
XG Boost Classifier:
Classification Report - In Sample
              precision    recall  f1-score   support

           0       0.95      0.97      0.96     14237
           1       0.87      0.77      0.82      3143

    accuracy                           0.94     17380
   macro avg       0.91      0.87      0.89     17380
weighted avg       0.94      0.94      0.94     17380


Confusion Matrix - In Sample
[[13879   358]
 [  713  2430]]


Classification Report - Out Sample
              precision    recall  f1-score   support

           0       0.91      0.93      0.92      3559
           1       0.64      0.58      0.61       786

    accuracy                           0.87      4345
   macro avg       0.78      0.75      0.76      4345
weighted avg       0.86      0.87      0.86      4345


Confusion Matrix - Out Sample
[[3304  255]
 [ 329  457]]
```



Feature Importances

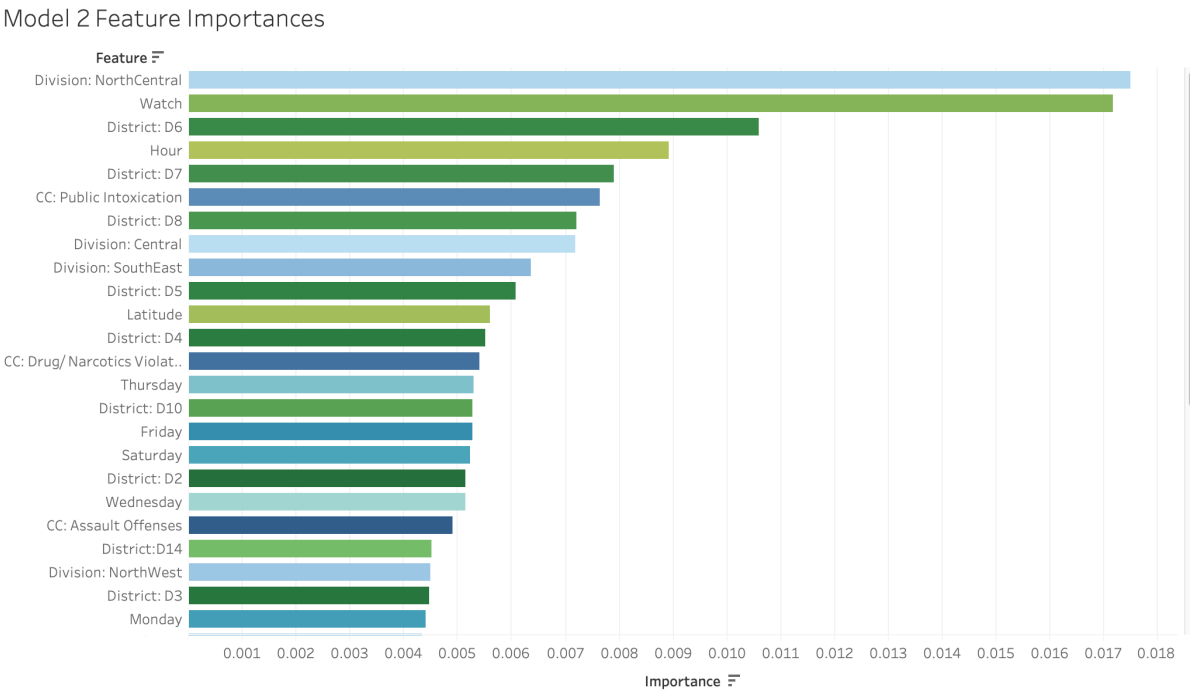
We also investigated the feature importances of each of these models to determine what, if any features played a greater role in predicting status outcomes.

Model 1



The feature importances of Model 1 when determining whether or not an incident ends in suspension is heavily influenced by various crime categories overall.

Model 2



The feature importances of Models 2, in which we determine whether an incident ends in an arrest or a clearance, is heavily weighted by the Crime Category: Miscellaneous at approximately 78%. Therefore, we have excluded that category to show the other contributing factors and their weights respectively.

The feature importances of the models show how we can fine tune the input features to better predict the status of crime incidents.

Data Tables

We included an interactive datatable that shows the data we used to create our machine learning model. The user can filter this data by inputting an hour, selecting a day, division, watch, district, and crime category.

Conclusion

We determined that crime occurs throughout Dallas with a concentration of crime incidents occurring in South Dallas and Central Dallas. We accepted our alternative hypothesis and determined that South Dallas experiences the most crime out of each division in our dataset. We also discovered that April had the lowest percentage of crime incidents out of every month in 2020.

After investigating what type of crime occurs the most, we failed to reject our null hypothesis and determined that the most common crimes that occur throughout Dallas are miscellaneous crimes, larceny and theft, followed by assault crimes. We also determined that most crimes occur on Friday's throughout the year, but this can change depending on month

and crime committed. When investigating what time of day crimes occur it depends on what type of crime was committed and the status outcome of the crime itself.

Recommendations

With the new Dallas Police Chief beginning his position in February, we recommend he dive into these crime trends that we analyzed. Based on our analysis, we recommend that the Dallas police force have more officers on duty in the month of August and allow officers vacation time in April. We also recommend providing more officers in patrolling Central and South Dallas areas due to the increase of percentage of crime incidents in those areas.

Limitations

When dealing with live data we often have to deal with discrepancies in our data due to human error. This prevented us from being able to include every incident documented in 2020 and limits our overall dataset.

We also were only able to use a subset of data for our visualizations due to Tableau limits on file size for documents being uploaded. This prevents our visualizations from being able to show our entire dataset and rather a small percentage of the incidents that occured in 2020.

To continue, we were only able to analyze the data for the year 2020. This limits our ability to look at trends over time and could also have an impact on our trained machine learning model.

Furthermore, the dataset included a vast category titleted NIBRS crime category. This category was very insightful in providing information about the crime committed, but there were two very large categories titled “Miscellaneous” and “All Other Crimes”. Neither of these categories included descriptions on what crimes they included, impacting our overall understanding of the data.

Future Work

To further investigate the vast amount of crime data it would be vital to look into trends overtime. This would allow us to compare crime incidents in between years and see if overall crime has trended down or up over the last several years.

If provided more time we would also like to create a mapbox map using leaflet in order to provide different filters for the user. This would allow the user to see where the majority of crime incidents occur closest to them, the type of crime incidents, the hours that these crime incidents occur, as well as a variety of other factors.

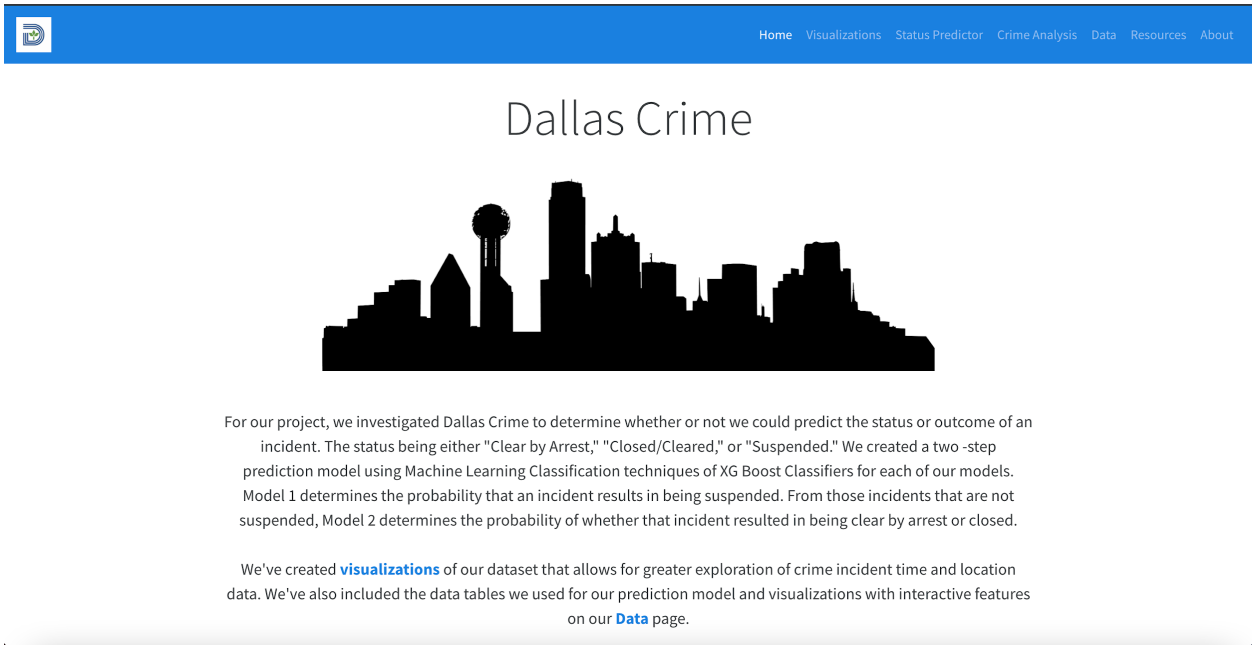
We also would like to develop more graphs and dashboards to show more insight into the other information the crime data provided such as the type of weapon used, the premise

where most crimes occur, and information on the demographics of the victims committing crimes.

It would also be interesting to investigate other types of machine learning models using different targets such as crime category or watch. This would allow us to see what else we could potentially predict and compare the different scores between each model.

Appendix

Website: https://dallas-crime-smu.herokuapp.com/



References

Data:

- https://www.dallasopendata.com/Public-Safety/Police-Incidents/qv6i-rri7

Images:

- City of Dallas Logo:
  - https://pbs.twimg.com/profile_images/726146162852405248/n1_39hmF_400x400.jpg
- Dallas Skyline:
  - http://clipart-library.com/clip-art/3-31174_skyline-clipart-dallas-skyline-dallas-clipart.htm

Inspiration:

- Gladwell, Malcolm. Talking to Strangers: What We Should Know about the People We Don't Know. 1st ed., Little, Brown & Co., 2019.
- https://dallascityhall.com/Pages/default.aspx
- https://www.dallaspolice.net/
- https://www.wfaa.com/article/news/local/new-dallas-police-chief-eddie-garcia-first-day/287-94067ec1-864e-41df-a669-6a80c8f4e2bc
- https://www.wfaa.com/article/news/local/city-of-dallas-ends-2020-with-highest-number-of-murders-in-more-than-15-years/287-d6902843-a927-4606-b1c3-019c85d5c502

Kaggle:

- https://www.kaggle.com/c/sf-crime
- https://www.kaggle.com/chasmo/sf-crime-analysis
- https://www.kaggle.com/carrie1/dallaspolicereportedincidents

NIBRS Crime Categories:

- https://www.fbi.gov/services/cjis/ucr/nibrs
- Crime Types Explanation:
  - https://www.denvergov.org/media/gis/DataCatalog/crime/pdf/NIBRS_Crime_Types.pdf
  - https://ncc.nebraska.gov/sites/ncc.nebraska.gov/files/pdf/manuals/nibrs/nibrs_man.pdf