Module 2: Prediction Complaints against Chicago Police Officers from 2015 to 2016

Seokhyun Yoon 301349313, Jason Dang 301400032, Tyler Oh 301320847

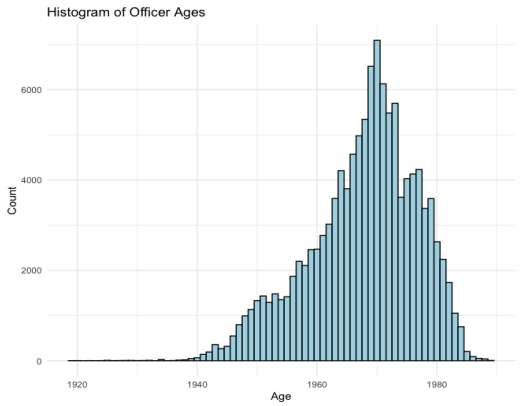
Dec 6, 2023

STAT 440 D100

Data:

For this analysis, we were provided a comprehensive collection of datasets. The ‘Accused Dataset’ contained 119367 observations and 11 different variables. These variables were both categorical, temporal, and numeric. The ‘Awards Dataset’ had 699912 observations and 13 variables. The variables in this dataset were categorical and temporal. The ‘Complaints Dataset’ had 125776 observations and 10 variables which were also categorical and temporal. The ‘Salary Dataset’ had 212508 observations and 10 variables. The types of variables in this dataset ranged from categorical, temporal and numeric. Lastly, the ‘Xte Dataset’ contained 33118 observations and 2 variables both of which were categorical.

Data Exploration: The data exploration step helped us gain valuable insight towards which features we should construct. This histogram showed the count of the officer’s age and suggested a large number of middle-aged officers in the workforce. Therefore, we believed that these officers were more likely to have more years of service. This histogram helped us identify that a feature such as the number of years worked would have been important to capture in our model.



Preprocessing: Our preprocessing step focused on filling in missing dates in the “start_date” column. This preprocessing step was crucial because we needed these dates to calculate how many years each officer worked between 2015 and 2016. We converted whitespaces in the date columns to NAs for both ‘start_date’ and ‘appointed_date’. To impute the missing dates, we merged all the datasets into a new dataset called “TEMP”. Afterwards, we were able to fill in the missing dates in “start_date” using dates from “appointed_date” as an approximation.

Constructing Features: We believed that if police officers had been accused at least once before, they would be more likely to be accused in the years 2015 and 2016. Therefore, variables Y1112 and Y1314 were created to capture instances of previous complaints against police officers within specific time frames. Variables A1112, A1314, and A1516 were created to represent the number of awards received by officers during different periods. The salary-related attributes, S1112, S1314, and S1516, were formulated to capture average salary data over distinct time spans. We believed that certain salary ranges, such as salaries below the average, would be correlated to complaints. Furthermore, the creation of P1112, P1314, and P1516 features denoted the probationary status within given periods. This would further capture officers that are likely to be accused. D1112 and D1314 were also constructed to represent the number of complaints in the same department by beat number during different time intervals. Lastly, W1516 was created to capture the years worked of each officer in 2015-2016.

Problem: To address missing data, we replaced any NA values in 'A1112', 'A1314', 'A1516' and 'P1112', 'P1314', 'P1516' with 0. For salary, the average salaries for 2011-2012, 2013-2014, and 2015-2016 were calculated. We also replaced NA values in ('S1112', 'S1314', 'S1516') with the respective calculated averages. Concerning distinct complaints, the 'complaints' and 'accused' merged dataset was employed to calculate the number of distinct complaints for each beat. The mean of the number of complaints was calculated to impute the NA values in ('D1112', 'D1314'). During the merging process, we encountered identical names, same first name and same last name, with conflicting information. Therefore to address this issue, we set the ‘multiple = first’ in our left_join operation. This ensured only one entry.

Methods:

Feature Engineering: In the process of feature engineering, we considered several categories of data. First, we focused on salary related attributes extracted from the salary dataset. The salary dataset contained information such as yearly salary of each police officer and probationary status indicators. This is where we extracted the information to construct our salary and probationary related features. Next, we combined accusation records from the accused and complaints datasets to examine past complaints against police officers. This involved associating accusations with officers' names and unique complaint IDs. Additionally, the awards feature we engineered was created by extracting from the awards dataset and grouping by name to sum the total number of awards for each officer. To track complaints in specific departments, we utilized complaint dates, departmental beat numbers, and officer names from the merged complaints and accused datasets. To effectively incorporate these features into our models, we applied a 2-level factorization using “as.factor” for variables such as Y1112, Y1314 to indicate accusation (1 if accused in, 0 otherwise). The same process was followed for P1112, P1314, and P1516.

Machine Learning Techniques: In this prediction, our primary methods were Random Forest and XGBoost. A significant difference between the two methods was observed: Random Forest handles take factor variables but not NA values, while XGBoost accepts NA values but is restricted to numerical variables. Consequently, we had to copy the data into two separate datasets for implementation purposes. For the optimization of hyperparameters, we conducted a computation of mean ROC accuracy score for various parameter combinations (ntree, nodesize, and mtry) using an out-of-bag estimate as well as performed a random search using the *caret* package in R.

Preventing Overfitting: In both Random Forest and XGBoost, we implemented 5-fold cross-validation to assess model performance robustly and prevent overfitting. For Random Forest, we utilized its built-in out-of-bag estimation. With XGBoost, we applied early stopping and the built in regularization parameters. Early stopping involved constantly checking the model’s performance on a separate validation set. Those methods were important in preventing overfitting, ensuring our models remained stable and capable of performing well on new, unseen data.

Results:

During our analysis, our group considered two main models: XGBoost and Random Forest. Among these models, Random Forest yielded the most accurate results. Our Random Forest model attained an AUC score of 0.9048 on the public leaderboard whereas our XGBoost model yielded an AUC score of 0.89043 on the public leaderboard. However, we also implemented a 5-fold cross-validation approach to assess both our model's predictive performance with the results shown below. Taken both these factors into consideration, this was the main motivation for our group to use Random Forest as our final model.

5-CV Fold Random Forest ROC AUC Score:	5-CV Fold XGBoost ROC AUC score:
0.9270	0.9158

As a result, our Random Forest Model was able to yield a final AUC score of 0.89864 on the private leaderboard. Our intuition to why Random Forest worked best was due to its strength in handling categorical variables and factors which was prevalent in our dataset. In contrast, XGBoost required one-hot encoding which made working with categorical variables complicated. Random forest manages categorical data efficiently by partitioning the feature space, making it adept at capturing intricate relationships within factor variables. This capability proved crucial in our analysis as the dataset contained categorical features that significantly influenced the target variable.

Next Steps: Based on our results, some potential next steps our group had taken into consideration were using more advanced data imputation techniques such as K-nearest neighbours or Multiple Imputation by Chained Equations (MICE). In regards to feature engineering, being able to implement an interaction effect between street and incident time in our model may have improved our accuracy. Being able to implement the street and incident time may have given us some insight towards particular officers who may regularly receive complaints during their shift, or route they patrol during work. However, due to the large number of levels in this categorical variable and date time format, this was a challenge we faced towards implementing this idea in our model.

Limitations: Although there were many advantages to having a large dataset spanning over many years, this was also a limitation to our model and problem to the approach. Our model did not account for changes over time such as any amendments in law, law enforcement policies, or societal attitudes. These variables can alter how incidents and complaints are reported, investigated, and resolved. For example, changes in 'use-of-force' protocols can significantly impact officer behaviour and complaint patterns. In addition, changes in societal attitudes is a large limitation because high media coverage, and social movements such as the George Floyd incident can influence the rate and nature of complaints against officers as well which our model does not account for well. For further information regarding the George Floyd incident, we provided a link:
https://www.nytimes.com/2020/05/31/us/george-floyd-investigation.html