Title: Crime Trend Analysis in the City of Austin using ‘Crime Reports’ Data Part II

Introduction:

***(same** as the EDA report project and identify clearly what your outcome variable is and what are your potential predictors)*

We continued investigating the “Crime Reports” dataset from data.austintexas.gov, which holds compiled data on crime incidents in Austin, Texas under the jurisdiction of the Austin Police Department. It is a collection of incidents responded to and filed written reports by APD from 2003 to the present, with weekly updates. The dataset focuses on the highest-level offense, meaning only the most severe offense is included in the dataset when multiple offenses are associated with one incident. Limitations include that varying data sources can yield different results due to methodological disparities, and replicating the research on other dates may yield distinct results as the database is updated on a weekly basis.

The “Crime Reports” dataset drew our attention because we are both residents of Austin and students who prioritize safety. This dataset provides us an opportunity to engage with real-world data, specifically the incidents that occur in our city, and allows us to explore crime patterns and trends. Through this dataset, we can identify locations that require special attention due to notable crime trends while gaining overall insight into crime patterns and public safety in relation to their geographical areas in Austin.

The outcome variables of our interest are “Days Till Clearance,” “Family Violence,” “Day or Night.” The potential predictors are “Clearance Stat,” “Council District,” and “Days Till Clearance.” We created a new categorical variable “Day or Night” for analysis.

Below are descriptions of our variables:

Days Till Clearance: The number of days it took for the case to be cleared (date cleared - date reported)
Family Violence: Incident involves family violence? (Y = yes, N = no)
Day or Night: Crime occurred during the day or night? (Day = between 7:00 AM to 8:00 PM, otherwise- Night)
Clearance Stat: How/whether crime was solve (N = not cleared, C = cleared by arrest, O = cleared by exception)
Council District: Austin city council district where the incident occurred

What are some of the questions that our exploratory data analysis will answer?

1. Is the clearance status a reliable predictor of how many days it takes for a case to be cleared in Austin?
2. Does the council district of where the incident occurred serve as a reliable predictor of family violence incidents in Austin?

3. To what extent does the number of days it takes for a case to be cleared predict when the incident occurred - day or night?

What trends / relationships do we expect to observe?

1. We expect that the clearance status alone is not a reliable predictor for days till clearance because we hypothesize that days it takes to clear an incident is impacted by other factors such as location type and highest offense description, in addition to clearance status.
2. We expect that the council district would be a good predictor of family violence incidents in Austin because we hypothesize that each council district has communities with distinct socio-economic status, population densities, or policies that influence the occurrence of family violence incidents.
3. We expect that the days its takes to clear the case cannot predict if the incident occured at day or night. We hypothesize that due to complexity of crime patterns and variable highest offense descriptions, days till clearance alone would not be able to reliably predict the timing of crime records.

Methods:

- Might be the same as the EDA project, unless you need to use your dataset(s) in a slightly different shape which you should document. Make improvements if you have received any suggestions about this section!

Exploratory Data Analysis:

- Same as the EDA project. Make improvements if you have received any suggestions about this section!

Classification and Prediction:

(Fit at least 1 model (linear regression, logistic regression, kNearest Neighbors, or decision tree) to predict an outcome (numeric or categorical), making sure to use a model that makes sense depending on the type of outcome, and based on at least 2 predictors in your dataset. If a team project, each group member fits a different model. o First, fit the model to the entire dataset and then use it to get predictions for all observations. ➢ If you are predicting a numeric outcome: calculate the value of the RMSE. ➢ If you are predicting a categorical outcome: build a ROC curve and calculate the value of AUC. o Second, perform 5-fold cross-validation with the same model. Report the average performance of the model across your k folds. If a team projects, each group performs cross validation for their own model. Discuss the results in a paragraph. How well does your classifier predict new observations? Are there any potential signs of overfitting? If a team project, compare the performance of the different models.)

1. When examining the relationship between predictor and outcome variables and assessing the RMSE values for both the initial linear regression model and the 5-Fold cross-validation model, elevated RMSE values suggest that the predictor yields inaccurate predictions with a high error percentage. This observation holds true for both the original model (RMSE: 370.3987) and the 5-Fold cross-validation model (RMSE: 319.5789). The adjusted R^2 value further supports this inference, registering at

-0.004062356. The low R^2 value indicates that the model does not align well with the dataset.

Moreover, potential signs of overfitting emerge, notably in the 5-Fold cross-validation, where substantial variations in RMSE values across iterations suggest that the model may struggle to generalize to new data and make reliable predictions for future observations.

2. In examining the association between the predictor variable, council district, and the outcome variable of family violence, we assessed the AUC values through logistic regression. Subsequently, we employed a cross-validation model to scrutinize potential issues such as underfitting or overfitting. The initial evaluation revealed an AUC value of 0.5894737 for the logistic regression model, and the cross-validation model yielded an AUC value of 0.5729167. These results collectively indicate suboptimal predictive performance of the classifier in the context of family violence.
   Furthermore, the presence of overfitting was once again noted, as evidenced by considerable fluctuations in AUC values across iterations of the cross-validation process. This model exhibits similarities with the previous one, showing subpar predictor performance and signs of overfitting.

3. Examining the final outcome and predictor variables, we conducted an evaluation using the same methodology applied to the previous model—utilizing logistic regression and a 5-Fold cross-validation approach. The initial logistic regression model produced an AUC value of 0.5584591, while the 5-Fold cross-validation model yielded an AUC value of 0.4320038. These results once again highlight inadequate predictive performance from the classifier, consistent with findings from prior models.

   However, in contrast to earlier iterations, the AUC values observed across each fold iteration suggest a potential issue of underfitting when compared to the preceding models that exhibited indications of overfitting. This suggests a limitation in the model's ability to accurately capture the relationship between the input and output variables in the dataset.

DISCUSSION

Putting it all together, what did you learn from your data? Answer your research question(s) in context, referring to the performance of the model(s). o Did the data match what you expected? Anything you are curious about? o Consider ethical issues (from data collection to interpretation): what impacts/implications could your results have on the community? o If you were going to share your findings with the City of Austin (and you might actually do that!), what would be the main takeaway from your exploratory data analysis? Anything they should know about the state of the dataset (e.g., inconsistency in categories, typos, …)?

Research Question 1: Is the clearance status a reliable predictor how many days it takes to clear a case in Austin?

outcome: days till clearance
predictor: clearance stat

We conclude that the clearance status is not a reliable predictor for days it takes to clear incidents in Austin. Both the initial linear regression model and the 5-Fold cross-validation model show elevated RMSE values, where the original model yielded an RMSE of 370.3987 and the 5-Fold cross-validation model had an RMSE of 319.5789. These high RMSE values suggest that the clearance status alone is not a reliable predictor as there is a high error percentage. Moreover, the adjusted R^2 value of -0.004062356 suggests that there is a poor fit between the clearance status and days till clearance. This means that the variation in the days it takes to clear incidents is not well-captured by clearance status.

RESEARCH QUESTION 2: Does the council district of where the incident occurred serve as a reliable predictor of family violence incidents in Austin?

outcome: family violence
predictor: council district

Based on our results, we conclude that the council district is a suboptimal predictor of family violence incidents in Austin. Upon quantifying the performance for the logistic regression model, the initial corresponding AUC was 0.5894737, showing a moderate level of prediction performance of family violence. The subsequent cross-validation model yielded a lower AUC value of 0.5729167. Collectively, the performance of the logistic regression classifier's prediction is suboptimal. The results did not meet our expectation as both initial evaluation and cross validation AUC values fall short of predicting family violence incidents in Austin based on council district. The model does not capture the relationship between the council district and family violence optimally.

RESEARCH QUESTION 3: To what extent does the days it takes to clear the case predict when the incident occured- day or night?

outcome: day_or_night
predictor: days_till_clearance

The results show that The initial logistic regression model and the 5-Fold cross -validation model both yielded a poor AUC value of 0.5584591 and 0.4320038, respectively. The values are far off from the AUC value of 1, meaning that our model struggles to discriminate and predict day and night incidentss by incident clearance time. Unlike the previous models, this model shows indications of underfitting and has challenges of capturing comlex relationship between the predictor and output variable.

Therefore, days to clear a case alone is not a robust predictor of whether the incidents occur at day or night. In conclusion, the results reveal various limitations of how well our model predicts the outcome variables.

Was our expectations met?

1. The results align with our expectations as we predicted that clearance status alone would not be a reliable predictor for how long it takes to clear crime incidents in Austin. We are curious about whether other variables along with clearance status could enhance the predictive reliability of days till clearance in Austin.
2. The results do not meet our expectations as we predicted that council district would be a good predictor of family violence incidents in Austin, We are curious to see if further exploration using other variables such as “Time of Day” can enhance the accuracy of the model and address the signs of overfitting.
3. The results does support our initial expectations as we predicted that due to complex crime patterns, the days it takes to clear a case would not be a reliable predictor of when the incident occured (day or night) in Austin. We are curious to see which combination of multiple predictors would best enhance the predictive performance of our model.

ETHICAL CONCERNS- what impacts/implications could our results have on our community?

It is important to note that the dataset is updated on a weekly basis, and therefore, when conducting the same study on a different day, the implications and interpretations for the same study may be variable. This means that our study is only an evolving representation of the incidents that occur in Austin. Our results may help the Austin community to better understand various factors that can be used to predict trends related to crime incidents. City officials and policy makes may refer to the results when integrating new policies and establishing new resource allocations for public safety.

Reflection, acknowledgments, and references:

- Reflect on the process of conducting this project. What was challenging, what have you learned from the process itself? Include acknowledgments for any help received: who helped you with the project? Thank TAs, instructors, data owners… If a team project, that is where you report the contribution of each member: who did what). Include references: dataset link(s), a citation for background context).