Rebecca Jones

27 January 2022

**Summary of Project**

It has been widely reported that violent crime has been on the rise nationwide; a 5% increase between 2019 and 2020, according to FBI data. At a county level, violent crime rates are significantly impacted by the violence taking place in the country’s most dangerous cities. The intent of this project is to explore crime trends overall at the county level throughout the state of Texas. We utilized millions of reported offenses published by the FBI’s Crime Data Explorer between 2017 and 2020 to discover what variables may help us predict the likelihood that a particular type of crime will take place. We considered a variety of factors including arrestee demographics, temporal and seasonality factors, as well as considered locational characteristics of where an offense occurred. Our highest accuracy score was generated using the Random Forest Machine Learning Model; however, it is highly likely that significant overfitting occurred. Since our independent variables were mostly categorical, a better choice for Machine Learning is the Multinomial Logistic Regression which produced our second highest accuracy score. While there is certainly room for improving our machine learning models and adjusting the variables we initially considered, we have been able to gain new insights regarding overall crime trends throughout Texas.

**Self-Assessment**

Throughout the course of the project my main tasks were data exploration, processing and visualizations. In addition to these responsibilities, I also took on some extra tasks to support the team including the website, the initial database design and the updated database model and research and feedback on the machine learning algorithms to ensure we were using the appropriate models for categorical data. I coordinated and hosted all of our zoom meetings outside of class time as well as made myself available for one-on-one support to both of my teammates. My greatest personal challenge was being patient and trusting that my teammates would follow through on their parts. Often times I would complete the other parts of the segment deliverables and hope we didn’t need to use my work; however, there were a few times where having the extra work I did was all that got done and we had to use it to stay on track. Operating like this created a lot of extra work for me and if I could do it again, I would have communicated my concerns to our instructors early on.

My contributions to each of the segment deliverables are as follows:

* Segment 1
  + Initial project design and data availability research
  + Development of the initial logical data model for the postgres database
  + Data processing to prepare for machine learning and data ingest into the postgres database
* Segment 2
  + Data exploration and visualization utilizing Tableau
  + Creation and publication of the project’s Tableau dashboard
* Segment 3
  + Website design including template selection, html and css scripts, and consolidation of team inputs
* Segment 4
  + Prepare Google Slides for presentation
  + Present on Data Exploration, Processing and Visualization Steps

**Team Assessment**

In addition to class time, we met every Sunday for an hour to review our progress and make sure we had everything we needed for the segment deliverables. We each took turns sharing our screens and walking the other team members through exactly what we did. This live review was incredibly helpful and we caught a couple of potentially major errors by doing this. The first was discovering that many of the machine learning models we learned do not lend themselves to mostly categorical data. The second, was that the visualization of arrest rate was inaccurate because we had combined arrests for multiple years together and the formula is only per year. In addition to our zoom meetings, we posted status updates on our group slack chat when necessary. Communicating with one another was one of the team’s greatest strengths. Lastly, rather than relying on a single team member to manage our main GitHub branch and Google Slides we all worked together to make sure everything was up to date.

Time management was challenging, especially when some deliverables for the week depended on someone else to get their part done first. This often resulted in a duplication effort and the need to develop temporary, work-arounds especially in regards to keeping our database updated. Data exploration, processing and visualization was a very iterative process and as the person tasked with these steps, I should have also probably managed the database or at least shared in that responsibility. The actual website design was fun but very time intensive and if we could do it again, I would recommend one person start on that during at least Segment 2. I also would have preferred if all the team members took time to dig into the data and really understand it rather than rely on me to explain it to them—I’m confident they would have caught and/or helped to prevent some of the initial mistakes I made when processing the data. We could have actually outlined the data processing steps together and potentially even split up that work based on blocks of code. My recommendation for future students would be to consider how data professionals work in the real world and mirror that. Have team members focused on the front end, back end, and data visualizations but share the analysis and machine learning responsibilities.