Washington University in St. Louis McKelvey School of Engineering

Data Analytics Boot Camp

**Final Project Self-Assessment**

Alejandra Rivera Reyes

January 27th, 2022

**Supervised ML with dependent variable “Crime Against”**

The project was conducted by three members, Becky Jones, Matt Lane, and me. At the beginning was difficult to define roles because we thought that we were not sure what type of result we wanted to present for the final project as well as we did not have a clear what kind of assignments and deliverables we needed to present with the data and first steps of content defined. The goal overall was to be aligned with what the rubric was requesting applying most of the skills learned during the bootcamp, so we did it. Our primary focused was define a realistic and achievable project; then, the common decision was to not define roles but to be focused on building the idea from scratch and setting weekly goals will be discussed on our scheduled calls.

During the first three calls, we primally worked on define the question of our project as well as completing the setup GitHub Repository where every deliverable will be hold. My contribution took a place in this piece. I was responsible to create the repo and the branches where everyone would push assignments. Although, anyone was in charge to stablish responsibilities. After all, if I have to define a role for me, “Square” would be the most suitable because of the requirements of the first segment and the flexibility I had setting up the environment where sooner, the deliverables, database and website scripts will be hosting. In other words, I was responsible about the cloud services.

*“Square: The team member in the square role will be responsible for the repository.” – 20.1.2 Build Your Foundation, Data Analytics Boot Camp*

My participation helped to set up the AWS Relational Database Service which allowed the distribution of the datasets to different interfaces such as Tableau and PostgreSQL. This step included to set up the S3 backets where the Federal Bureau of Investigation’s Crime tables from Texas data were gathered to prepare the DataFrames for Machine Learning models. With a total integration, we checked out the requirements of database connection and loading data in SQL and Tableau and applied join statements to proof ERD relationships worked. Additionally, I supported the team to review the html script and deploy the framework in Heroku which is platform that enables to run and operates applications using cloud services.

Through this step we would be able to display the exploratory data and machine learning analysis that my other team members developed. Since we decided to deploy the results and display the whole picture of the analysis, the website hosts the Tableau Public published dashboard as well as the ML results to be more in a descriptive analytical result.

Unfortunately, due to a heavy load of work at my job, I had to work for two weekends and that put me step back of the team. During this time, deliverables 2 and 3 were presented and I did not do collaborate in the visualization of the datasets in Tableau neither to model the inputs for the Supervised ML model. That did not mean that I was not aware about what kind of data we worked with, but I can say that personally I need to look for more opportunities where I can apply my skills using the diverse libraries that offered python for exploration analysis and predicting data.

Summary contribution:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Team member** | **Cloud Services** | **SQL database** | **Cleaning Data** | **Visualizations** | **Predict Analysis** | **Website Design** | **Google Slides** |
| Becky |  |  | P | P |  | P | P |
| Matt |  |  |  |  | P | P | P |
| Aly | P | P |  |  |  | P | P |

As it is shown in the table above, Becky played the role as “Circle” who created from the beginning and got in charge of creating the mockup of the database, providing the sample data, and fabricate the interactive dashboard in Tableau where all the data provided a meaning for the analyst. She created a map of Texas where the data was aggregated at the county level, bringing insights about the correlation of different variables that were chosen for the Machine Learning process. Additionally, she created a multicollinearity analysis where shows the coefficient relationship between the independent variables, finding out what were the variables with more coefficient for further analytical purposes.

On the other hand,