**Name: Saivijay Panuganti**

**Date: November 20, 2022**

Dr. Barber

Data606: Data Science Capstone

**Personal Contribution Write-up**

To Grader,

Good day, my name is **Saivijay**. I am a member of Baltimore Arrest Data machine learning capstone project group. The following are my personal contributions to the capstone project.

* Development of business question
  + Initially my team members and I have brainstormed several ideas on what we want to do for the project, all of us have put forward our ideas and discussed them together and decided to go forward with the Baltimore Arrest Data.
  + Then we discussed as a team to come up with the business, every teammate gave their inputs we have decided to go ahead with the idea of finding out what factors are causing arrests in different districts of Baltimore and check for any bias, so that effective policing efforts can be made and What model can we use to in our analysis which gives the best accuracy and takes least time?
  + Overall, I can say it was a team effort and a lot of meetings/discussions that went into developing our business question and I have contributed my share in doing so.
* Contribution to Data Acquisition
  + Data Acquisition is done by one of my teammates and all of us assisted him in doing it.
* Contribution to Data Access
  + For the easy access of the data to all the members of the team we have uploaded the data to the GitHub and provided access to all of us.
* Contribution to EDA
  + I did EDA for our data to find what columns are there in the data we acquired, what are the datatypes of the columns and are they relevant to our model.
  + I deleted the columns which are not relevant to our business question and checked for any missing values in the data and how to manage the missing values.
  + Statistical analysis of the data is mainly done by my other teammate, but I have assisted him throughout his analysis when he needed.
  + I did the bulk of the visualizations for our project; I used the insights gained from our EDA and visualized important aspects so that they can be easily interpreted by the users.
* Contribution to Data Cleaning
  + Data cleaning is an important step before modelling the data. I did my fair share of Data Cleaning in the project, like to change the data type of the column to correct format, to split the columns to be meaningful, to handle the missing values to name a few.
* Contribution to Data Preprocessing
  + I basically took the task of doing the SparkML for our project, so I did the data preprocessing required for the pyspark models.
  + I did the steps that need to be performed before injecting our data into the model, like split the data into train and test data, perform One hot encoding, String indexer, Vector assembler and separate independent variables from dependent variables.
* Contribution to Data Modelling
  + As I mentioned above, I took up the task of doing the SparkML, our goal in doing pyspark model is to take the model with highest accuracy from sklearn and tune the parameters of the model using pyspark to get better accuracy.
  + I did the Logistic Regression using pyspark using pipeline and different parameters to get the best result.
  + I also used k-fold Cross validation and Multiclass Classification Evaluator to validate our model and get the accuracy of the model; This model turned out to be the best performing model of all the models used in the project by almost doubling its accuracy from the previous models.
  + I also did the best parameters checking for the model, to find the best parameters as our model takes a lot of time about 52 minutes to execute, as it checks the cross validation on different parameters; We can significantly reduce the execution time by using only the best parameters for the model.
* Contribution to Project structure
* When doing a project in groups, it is easy for the project to go out of track and head in some other direction which we have not intended to do, my job is to stop this from happening and guide and explain the team when we are not going according to the plan of the project.
* One of my important tasks is to find out where we are lacking in our project, like basically do the task of a tester by testing the project to see of it is functioning as we planned and discuss with the team and the concerned teammate if found some abnormality in the code and figured out possible solutions of rectify them.
* Also, the project needs to have a flow and each concept is boxed in its space and remove any unwanted code that doesn’t provide much input to our project, I have helped in structuring the project so that it can be deliverable.
* Contribution to Research Paper
* We have divided the research paper into four sections and each team member we take the responsibility of doing research and present it to the team and complete the assigned section of the paper.
* The sections which I have done for our research paper are Data Cleaning, EDA, and Dataset Basics.
* Contribution to Group planning and organization
  + Our group plannings and meetings have gone smoothly throughout our project even though our team has 4 members. In every meeting and planning one of the team members used to take charge of the meeting based on the content we are going to discuss in that meeting and others would provide their inputs to the teammate leading the meeting.
  + I also had my opportunities to lead our meeting sessions when I had something to present to the team, we decide on it beforehand on who will present their case that meeting to avoid confusion.
  + I held meeting with the team regarding the EDA, visualizations, pyspark model and the structuring of our project, whenever I had something to present to the team.
  + All the teammates joined the team meetings without fail on the timing we have decided together making it easy for everyone.

Please note that what was listed here is a surface level look into my contributions to the project, as all our group members are involved in all the aspects of the project from decision making to modeling and helped each other throughout the project timeline. Thank you for your time and have a blessed day.

Regards,

**Saivijay**