**Data Science Program Final Project**

**Executive Summary**

The data science program leads up to a final project that will call on students to use the skills they’ve learned. Students are assigned partners and told to choose a topic and dataset to study.

This document explains the purpose and scope for Nazrath and Sarah’s project.

**Business Objectives**

Showcase the skills that Sarah and Nazrath have acquired through the Data Science program. They will be using R, Python, Tableau and other programs to wrangle, analyze, and visualize the “COMPAS Recidivism Racial Bias” dataset made available by ProPublica on Kaggle.

At the end of the project, Nazrath and Sarah will give a detailed presentation via Zoom to WozU faculty, other students and potential employers.

**Background**

This project is an opportunity for the students to put everything they’ve learned about Data Science to work in a practical case study.

Sarah and Nazrath have chosen the “COMPAS Recidivism Racial Bias” dataset because they are both interested in real-world data that is both topical and impactful. They hope to be able to identify the least biased model for the COMPAS algorithm to use if there is one.

They also hope to be able to identify the age group most likely to receive high COMPAS scores.

**Scope**

Sarah and Nazrath will be using R, Python, Tableau and other programs to complete this project. They will be thoughtful and intentional choosing the tools of their interest or that may aid in job placement.

**Functional requirements**

Data Wrangling: The COMPAS dataset should be cleaned for analysis. Any unusable columns or rows with missing data should be removed. The data types for each column should be converted to whatever type it needs to be for the chosen analysis.

Data Analysis: Nazrath and Sarah will examine the dataset and determine their variable types. They will have a good understanding of what each column contains and how the values are measured. They will determine the evaluation questions they wish to answer with their data analysis. Then they will identify and map out their plan to execute their chosen analysis.

Data Visualization: Once Sarah and Nazrath have completed their comprehensive analysis of the dataset, they will move on to creating a visual presentation of their findings. They may use Tableau to create their graphs and then Powerpoint to create a slide presentation.

Presentation: Nazrath and Sarah will work with WozU to schedule a time to present their findings via Zoom. The presentation should be kept around 20 minutes in length and will be conducted in a clear and concise manner. They should be dressed professionally for this presentation.

**Personnel requirements**

Sarah and Nazrath are the two developers. They will take turns being the scrum master and will report their progress to their instructor (Product Owner.) They will need to stay organized and stick to a schedule for this project to succeed. They will touch base at least once a day via Zoom or Slack for a virtual stand-up meeting where they will review what they accomplished “yesterday”, what they plan to do “today” and what obstacles are in their way. This daily stand-up will also include checking in on and updating the project Trello board.

Once a week, they will meet with their instructor via Zoom. They should be prepared with any questions they were unable to answer on their own and seek guidance for the next steps they’re planning to take.

They may also consult with their mentor.

**Delivery schedule**

Week 1: Download data files from Kaggle and import raw dataset into preferred software to begin data wrangling. Begin to educate ourselves on aspects of the justice system that pertain to the data. Set up Github.

Week 2: Study the dataset and ask questions. What are some possible correlations? Is the data normally distributed? What are some predictive models we can make from it? Visualize the data to see if there are any potentially significant early findings.

Week 3: Modeling/Optimization (Independent Chi-Square) and Machine Learning (Random Forest).

Week 4: Review and validate findings from the previous week and draw insights/conclusions.

Week 5: Compile findings in a PowerPoint slideshow. Run through the presentation several times in front of different people to ensure that it is clear and logical. Fine tune the style and layout of the presentation to make sure it is pleasing to the eye and easy to interpret.

Week 6: Make final adjustments to the Power Point presentation. Nazrath and Sarah should practice presenting several times with just each other and at least once more with their instructor prior to their final presentation recording.

**Other requirements**

All programs used should be free of charge, though Nazrath and Sarah may decide to use a paid service should the need arise and they determine it to be worthwhile.

**Assumptions**

The software programs and platforms Sarah and Nazrath use should be available, up-to-date and fully functional. The COMPAS dataset the team is relying on is expected to be complete and accurate.

**Limitations**

Nazrath and/or Sarah live in different time-zones which could cause scheduling issues during this six-week period that could lead to project delays. The instructor or mentor may have scheduled or unscheduled time-off that could cause delays in the project. Sarah and/or Nazrath may experience roadblocks in their work, which may push back the completion date.

**Risks**

The risks that may arise include the current COVID-19 situation, family emergencies, natural disasters, power outages, or broken software/hardware. Nazrath and Sarah are both eager to complete the program so there should be no motivation issues and therefore the chances of submitting an incomplete project remains low. The instructor and mentor are helpful and generally available when needed so there is no concern of unresponsiveness from them.