Alberta “Albi” Kovatcheva and Barbra Treston

DSO110 - Final Group Project

November 28, 2021

**New York State Mega Millions:**

**Data Science Program Final Project**

**Executive Summary**

At the end of the Data Science program, students are required to complete a final project of their choice. They are given six weeks to work on the project. Oftentimes, they will be paired up with another fellow-student.This final project is an ideal way to activate and put to practical use to what the students have learned over the course of the Data Science program .

This document will explain the purpose and scope for the project proposed by Alberta “Albi” Kovatcheva and Barbra Treston.

**Business Objectives**

To showcase the skills that Albi and Barbra have acquired through the Data Science program. They will be using R, Python, Tableau and other programs to wrangle, analyze, and visualize the “Mega Millions Winning Numbers” dataset made available by t\he New York State Lottery on Data.gov.

At the end of the project, Albi and Barbra will explain their work in layman’s terms, and present their findings to the students, faculty, staff, and potential employers, along with other interested parties via Zoom.

**Background**

Albi and Barbra have chosen the “Mega Millions Winning Numbers” dataset because the lottery is something that is familiar and accessible to a wide range of people worldwide; it would be difficult to find someone who hasn’t dreamed of hitting the jackpot and changing their life forever. However, it is also widely accepted that the lottery is not set up to favor the player. In the case of Mega Millions, although there is a 1 in 24 chance of winning something, the odds of choosing all 6 numbers correctly to win the jackpot is 1 in 302,575,350 - a fact that is posted openly on both the New York Lottery and Mega Millions websites. By analyzing the winning numbers data as well as complementary datasets on lottery retailers, lottery aid to local school districts, and monies recouped from the lottery winnings of public aid recipients, Albi and Barbra hope to glean insight to make actionable suggestions on how lottery players can get the best return on their investment as well as to demonstrate for the average person whether the lottery serves any societal good or whether it may be best to abstain from playing altogether.

**Scope**

Albi and Barbra will be using Python and R to complete the project. Data wrangling will be performed in Python. Data transformation and statistical analyses will be performed mainly in R. Lastly, data will be visualized in Tableau.

**Functional requirements**

Data Wrangling: The downloaded dataset should be successfully cleaned up for analyzing. Columns and unusable columns should be removed. As the dataset is fairly large, Albi and Barbra should consider sub-setting the dataset in a proper manner, meaning the subset should be a random selection of the data. The data types for each column should also be converted to a usable format for the needed analysis.

Data Analysis: Albi and Barbra will familiarize themselves with the dataset. They should have a good understanding of what each column means, and how the values are measured. They will brainstorm on questions to ask, and what they might gather from the dataset. Then, they will identify the proper functions to create models, predictions, etc.

Data Visualization: Once Albi and Barbra have a comprehensive understanding of and insight gathered from the dataset, they will work on visualizing the findings. They may decide to use Tableau or other graphing programs, and compile the visuals and texts in a PowerPoint slideshow.

Presentation: Working with school leaders, Albi and Barbra will schedule a time to present their findings via Zoom. They should be able to communicate in a clear and easy-to-understand manner. The presentation should be kept around 20 minutes. They should be dressed professionally for this occasion.

**Personnel requirements**

Albi and Barbra are the two developers. They will need to work closely for this project to succeed. They will touch base once a day via Zoom or Slack to problem-solve or to check in on work progresses. Once a week, they will review the past week's workload and plan out the next week. They will take turns being the scrum master, and report their progress to their instructor (Product Owner.)

Once a week, they will meet with their instructor. They should be prepared to ask questions and seek guidance for the next steps.

They may also consult with their coding mentor.

**Delivery schedule**

Week 1: Import dataset into preferred software to begin data wrangling. Any unnecessary columns should be removed. Educate ourselves on breast cancer. 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 interesting findings.

Week 3: Modeling/Optimization (Combined Stepwise - Forward and Backward Selection) and Machine Learning (Random Forest.)

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

Week 5: Compile findings into a PowerPoint slideshow. Go over it with their instructor and friend/family member to ensure that the presentation is clear and logical. Work on the style and layout of the presentation so it is delightful on the eyes.

Week 6: Make final touches to the PowerPoint presentation. Albi and Barbra should not attempt to come up with a brand-new analysis. There will not be enough time to verify their findings. They should practice presenting at least a couple times with the two of them, and at least once with their instructor.

**Other requirements**

All programs used should be free of charge. Though Albi and Barbra may decide to use a paid service, such as a more advanced version of Tableau.

**Assumptions**

The software programs and platforms Albi and Barbra use should be available, up-to-date, and not broken.

**Limitations**

If something should come up for Albi and Barbra during this six-week period, the project may be delayed. If the instructor or mentor has scheduled or unscheduled time-off, the project may be delayed as well. Albi and Barbra may experience a roadblock in their work, which may push back the completion date.

**Risks**

The risks that may arise are such as natural disasters, power outages, family emergencies or broken software/hardware. Albi and Barbra are eager to complete the program so there should be no motivation issues. The instructor and mentor are phenomenal so there is no concern of no help from them. The risk of this project being incomplete is minimal. They will be successful in completing this project!