**Data Science Program Final Project**

**Executive Summary**

Companies can profit exponentially by the small act of a customer when he or she leaves a review on an item(s) purchased from them. Simultaneously, these comments can also cause more harm than good for the company. The consumer of course would be profiting, being able to save their own hard earned money trusting the experience of another. But how can both companies and consumers trust these reviews? Are they done with the right intention or with personal resentment?

This document will be uncovering the usability of these comments for a women’s clothing company that will be disclosed for privacy reasons. This document explores the context of reviews left on items using sentiment analyses to decipher and recode comments into a good and useful, otherwise “positive” comment, and the other as a “negative” comment. Using Python for sentiment analyses and data testing, R studio for data analyzation and wrangling, jupyter notebook and a multitude of statistical analyses testing to detect, if any, the relations between certain key variables and what their outcomes are.

**Business Objectives**

This document will aid in showcasing the effectiveness of reviews left on products and thus assisting this clothing company with knowledge of what products receive a higher review rate from customers and what products do not. This way companies can have the knowledge and take the initiative to either improve or remove said items.

I will be able to confidently discuss the matters of products in layman’s terms for the company executives in an attempt for the company to rebrand itself using positive comments.

**Background**

To showcase the skills I’ve learned and the programming languages I am efficient in I wanted to use a dataset of clothing reviews to see how I can break data and mold it to find answers to evaluations that may have not been there before.

Using my own skepticism and critiques of reviews left on products from my favorite online stores I wanted to see how I could implicate data science into this subject.

**Scope**

-

**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, Jane and Jessica should consider sub-setting the dataset in a proper manner, meaning the subset should be a random selection of the data. The datatypes for each column should also be converted to a usable format for the needed analysis.

Data Analysis: Jane and Jessica 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 Jane and Jessica 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 Power Point slideshow.

Presentation: Working with school leaders, Jane and Jessica 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**

I am the only developer. I will need to give myself structured time to work on parts of the project. Using the Kanboard and my repository I will update proficiently.

I will be the Scrum Master every week but I will still assign a day to review the work I’ve done and the progress I’ve made to keep up with the delivery schedule. I will also use this time to determine if anything on my part in scheduling, time, analyses, or research needs to be adjusted.

Once a week, I will meet with my instructor. I will be prepared to ask questions and seek guidance for the next steps.

**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 is 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 Power Point 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 Power Point presentation. Jane and Jessica 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**

Data from the same company in a different year/ time frame could’ve helped aid in more analyses of popularity trends.

All data/programs used are free.

**Assumptions**

I can load Natural Language Processing Tool Kit and Hugging Face and launch it effectively.

**Limitations**

Because this project is being done solo, if I am not conscious with my time management and using the time effectively I can delay the structure of my schedule leaving me with less time to complete the project.

My laptop, which the instructor was witness to, performs slowly at times so if I don’t use my resources well and save my continued progress I risk losing work.

**Risks**

My laptop, which the instructor was witness to, performs slowly at times so if I don’t use my resources well and save my continued progress I risk losing work.