****

**Project Proposal – Women’s E-Commerce Clothing Review Analysis**

Course: ALY 6020 Predictive Analytics

**Instructor: Dr. Marco Montes De Oca**

Submitted By,

Ashlesha Kshirsagar

Shruti Avinash Pawar

29 May 2020

**Introduction:**

Recent increase in the base of ecommerce vendors has made the women clothing ecommerce industry highly competitive. Thus, customer reviews received on the websites for the products are used for text mining and building ML models. Our objective is to work on such data to build classification models from features extracted using the review text. We have therefore selected a data set from Kaggle known as Women’s Ecommerce Clothing reviews. <https://www.kaggle.com/nicapotato/womens-ecommerce-clothing-reviews>

**Description of the data:**

The dataset has 23486 rows and 10 feature columns. The features are variables like Clothing ID, the age of the reviewer, the title of the review, the rating given by the reviewer, numerical variable of recommendation, the overall feedback of the product (displayed as a cumulative positive), the division, the department and the name of the product.

**Methodology:**

The tool that will be used to work on this analysis is python. First, we will introspect the data by analyzing the descriptive statistics summary. Further, we will carry out data cleaning and EDA to better understand the variables. The outcome variable based on which classification algorithms will be built is ‘Recommended IND’. It is a binary variable where 0 means the product is not recommended and 1 means the product is recommended by the reviewer. We will be creating a dataframe extracting features from the text using nltk package from python. Once our final dataframe is ready, we plan to build classification models like logistic regression, Random forest, KNN and Naïve Bayes in order to classify the product as recommendable or non-recommendable. Further, we will be using confusion matrix to determine the model accuracy, precision and recall score and choose the best fitting model with high accuracy and precision. Once we select the best model amongst 4 mentioned models, we plan to focus on improving the accuracy of the selected model. The current plan is to focus on the features that can be dropped based on the significance however we also look forward to exploring model tuning that will be learned over the course.

**Link to the dataset:** <https://www.kaggle.com/nicapotato/womens-ecommerce-clothing-reviews>