Walmart Women’s Clothing E-commerce Dataset

The objective of this project/code is to which model yields a more accurate predictions on unseen or new data in the future, by training and splitting the data set for testing with the available dataset. The dataset is made up of different types of data such as integer, string variable, ordinal integer, binary variable, and names. There are a total of 10 feature variables in this data set, such as age, title, review text etc. However, the main goal for this project is to determine whether these 4 factors/predictor variables (Age, Class Name, Department Class and Rating) have a connection or correlation to two target variables (Recommended IND and Positive Feedback Count).

I started the project by conducting some basic exploratory analysis. Plotting graphs to show the distribution between each of these individual predictor variables, to determine which is the get more information regarding these predictors, to see if there were any imbalance in the data sets or just a general feel for the data. The graphs indicated that there was an imbalance in the recommendation (Recommended IND) column. I suspect that there is a class imbalance in this dataset that can skew the models results. More plots on correlation between these 4 predictors against the target variables also yielded interesting finding. For example, the general trend for Rating and Recommended IND showed that higher rating produces more recommendation. An odd discovery was that a rating score of 3 has about the same number of recommendations, despite there being slightly more no recommendations for that rating score.

For data cleaning, I started by checking and eliminating and missing values. Then, I checked the number of unique values for the categorical name columns and created dummy variables for preprocessing purposes. Additionally, I removed the other unnecessary columns, like review text that is not essentials to the main objective. One thing that I learned from this process is the importance of data cleaning and preprocessing steps. I initially used the label encoder method, but I found that these models yielded weird and extreme results. I realized that it doesn’t make sense to encode class names and department names as integer values, but instead it makes more sense for each of the variables to me classified as binary values for a more accurate prediction. Another problem that I encountered was during classifying the X and y variables. I kept on getting could not convert strings to float error even after preprocessing the data set. For X, I just listed the variables described on the assignment page without realizing that I have 12 additional new columns from one hot encoding. To overcome this, I just dropped the 2 target variable, along with the original class name and department name column.

For the first part, the models (LogReg and SVC), yielded high accuracy score of 93%-94%, with a high precision and recall score (above 80s avg). This shows that these models are good for identifying true positive/negative and avoiding false positive/negative. High f-1 score also indicated good model a balanced model performance. For 2nd part, I ran 3 models (LinReg, KNnReg, and DTR). Both DTR and KNnReg shows similar r2 scores of .5s avg, while the LinReg yieldd the highest r2 score of 0.65. I think that this is because LinReg are good at capturing linear relationships between predictors and target variables. As I suspected in the EDA section, the data was predominantly filled with high/positive recommendation and positive review score. Overfitting is one of the main problems with this model. The other two models yielded a fairly good performance, with low MAE and RMSE score indicating that the predictions are like the actual values on average.

Overall, I think that these models performed well in helping me achieve my main objective of this project. If I had more access to more data and more time, I would say that I want to explore a more unbiased data. In the sense that they would include a more balanced data set in the recommended IND and positive feedback count columns. It may be the case that generally Walmart’s women e-commerce store is doing well, but I suspect that they are not including some of the negative recommendation or low positive feedback count in the data set. Also, I would like to explore more parameters for my KNN model. I tried using the elbow method to find the optimal k for this model, but it was taking too long, so I ended up with the Davies Bouldin method. Finding the right optimal k is crucial to the model’s performance. The wrong k number can lead to underfitting or overfitting.