**Walmart’s Women’s Clothing Reviews Data Analysis - Report**

By Esmeralda Hernandez

I worked with Walmart data on women’s clothing reviews with the objective of evaluating the connection that Age, Rating, Department Name, and Class Name had to Recommended IND and Positive Feedback Count. In this dataset, the fields Age, Rating, Recommended IND, and Positive Feedback Count are related to the customer where Age is the customer’s age, Rating is the value from 1 to 5 given by the customer to a product, Recommended IND is whether or not a customer would recommend the product to others, and Positive Feedback Count represents the number of other customers who found the review given to be positive. The fields Department Name and Class Name both pertain to categorizing the product that the customer had reviewed.

My initial hypothesis was that customer ratings of products would be the most influential factor in predicting the amount of positive feedback and whether a customer would recommend the product. To test this hypothesis, I cleaned and processed the data by removing null values and creating dummy variables for the fields in the data that I was going to use in my predictive models. Next, I used Random Forest Classifier for predicting Recommended IND and Random Forest Regressor for predicting Positive Feedback Count. By using this predictive modeling tool, I was able to perform feature importance on both models to identify the importance or weight that each independent variable held in predicting Recommended IND and Positive Feedback Count.

I was able to develop a predictive model for Recommended IND using Random Forest Classifier that had a high accuracy score of 92.2%. While high accuracy is great to have achieved, I noticed that my model tends to produce more false positives (210) than false negatives (154). Given the context of the data, I find that this is a weakness in my model as it means that it is more likely to predict that a customer had recommended the given product to others when in reality they did not. This can potentially lead to Walmart deciding to buy more inventory for a given product believing that the larger number of customer recommendations would lead to more sales when in reality the demand for that product was overestimated. Ratings seem to have a stronger connection to Recommended IND as its score exceeds the second largest score by about 78.5% meaning that the ratings given by a customer are the dominating factor that predicts whether or not a customer would recommend a product.

When creating a predictive model for Positive Feedback Count, I was able to attain a root mean error of about 5.97 meaning that the average difference between the predicted and cultural values is very low. As a result, my model will likely produce highly accurate Positive Feedback Counts. By using feature importance, I was able to find that while the difference between the importance of the feature variables was not as large as for my other model, it is clear that the feature variable that holds more importance above the others was Age (50.8%) with Rating (21.1%) following it. This means that the age ranges of customers play a crucial factor in predicting whether others find the feedback given as positive.

While only part of my hypothesis proved to be correct, I was still surprised by the distribution of the feature importance for both my predictive models. With more time and data, I would have liked to explore creating a model for predicting Recommended IND with other features to see if Ranking continues to be the dominant feature with the highest importance score. Additionally, I would have liked how the Positive Feedback Count amount shifts with varying age ranges to better understand the customer behavior of Walmart customers who buy women’s clothing.