1. Define Project

The Project I chose was to work on the Walmart data set in which the main problem I wanted to solve was whether Age, Rating, Department, and Class Name have some type of correlation with the Recommend IND using Classification. The second problem I want to know is if those 4 predictor variables will develop a response or correlation with the Positive Feedback Count Column using Regression. I will run a Classifier to classify the recommended IND which is a binary variable under a confusion matrix which I will define under True positive and true negative and vice versa under false negative and false positive. This will allow me to dissect the data and understand which matrix is the best result while running the accuracy score seeing how accurate my data is so far. Allowing me to not only see how accurate the test and train are but also see if there are any correlations between the two response variables.

1. Solutions

The way I decided to handle the solution to my problem was by running a KNN classifier and regressor, but the challenge I faced was finding the classifier and regressor that would fit the data set. This was due to it containing words and text, so I thought I had to add a preprocessor, but this wasn’t the case I still added an NLTK (‘punkt’) and (‘stopwords’) for any word outputs. I dropped the other columns so that I could get rid of the noise in the data set. After, this I decided to run the KNN classifier and then regressor to find the correlation between the two response variables.

1. Results

As for my results, I was able to Extract a confusion matrix which returned a matrix with an accuracy score of 21.76% which two things resulted from the model. This is because the model needed more data or the Recommend IND didn’t have a strong correlation with the data. This in return made me suggest that the predictor variables weak correlation with the recommended IND, and people’s only item recommendation was only if the customer had a good rating. I would later test to see if the ratings would correlate with the response variable. My second result was that the positive feedback was close to zero but not exactly zero. This returned an RMSE of 6.21. This wasn’t the best correlation but it was close to see if the Positive feedback count was affected by the predictor variables. This led me to believe that ratings did positively affect people’s sentiments. Below are the results of my Problem.

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated