

Above is a data mining project that I preformed in the Spring of 2017 for the graduate course, Data Mining for Business Intelligence. The project involved data mining a large data set of a store’s purchase information. I sought to create several predictive models for if a purchase resulted in an amount greater than the average purchase.

I began this project by importing the data and preforming a preliminary exploration of the data in the StatExplore node. I then preformed all necessary data cleaning and variable transformation in the Transform Variables node. Once the data was in an acceptable state, I began to explore the associations of variables to the target variable. The target variable, in this case, was a binary variable that indicated if a purchase was greater than the average or not. I used MultiPlot, Variable Clustering, StatExplore, and Cluster to explore the associations, correlations, and patterns between all the variables. I used this information to determine which variables I would use to create a model.

Before I created the models, I partitioned the data into 40% testing and 60% validation. I then sought to create several logistic regressions, each with a different selection method of variables. I created logistic regressions using the Forward, Backwards, and Stepwise model selection methods, and then compared them using the Model Comparison Node. The Stepwise Model Selection produced the best results. I also created a Decision Tree and Neural Network models. I then chose my best logistic regression, and compared it to the Decision Tree and Neural Network, using the Model Comparison Node.

I wrote a more in depth report on the procedure and results; this can be found in the Report file.