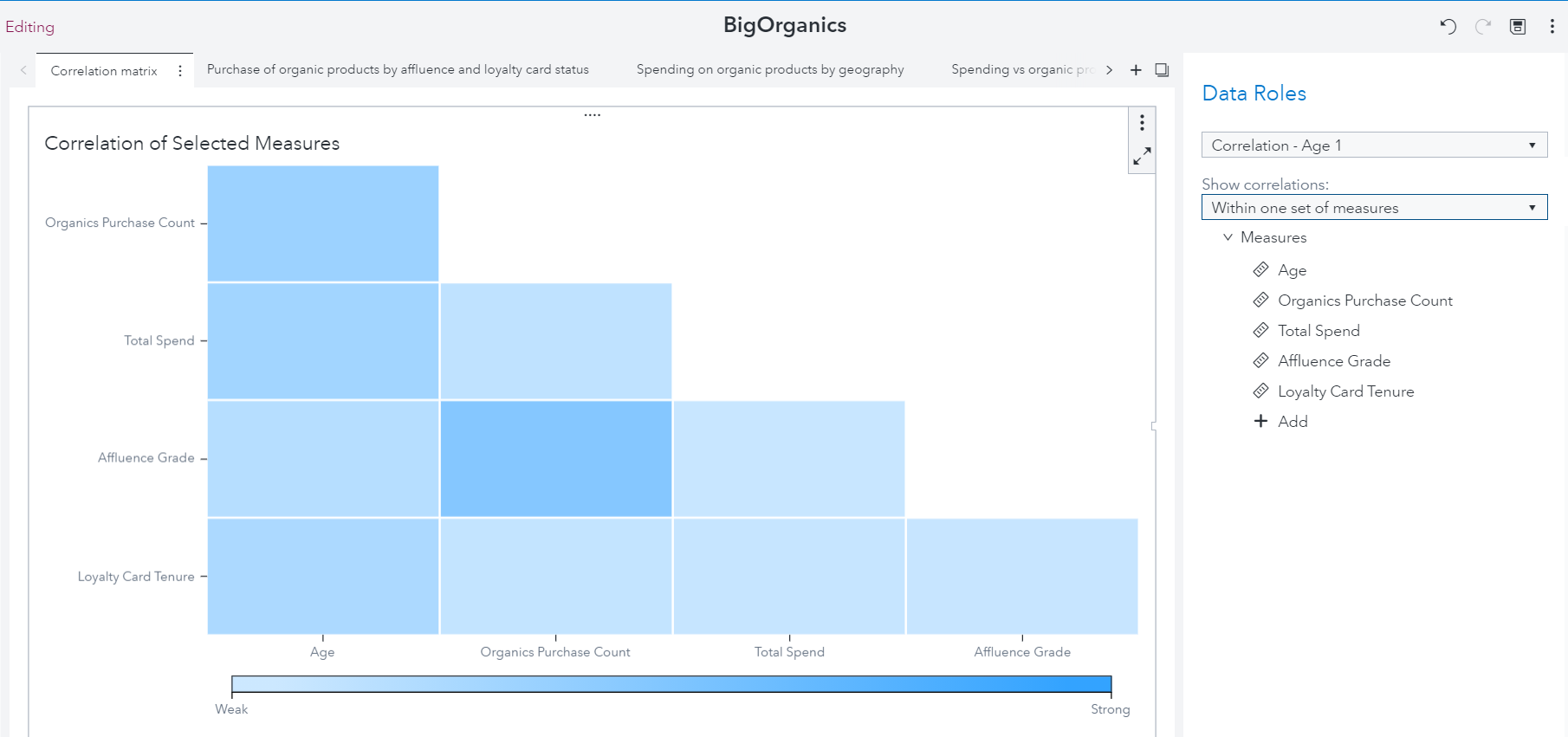
**SAS VIYA- MINI PROJECT**

**NAME: Ritika CHAKRAVARTY ORGANISATION: DSTI**

**DATE: 31st May,2021**

**DATA VISUALISATION**

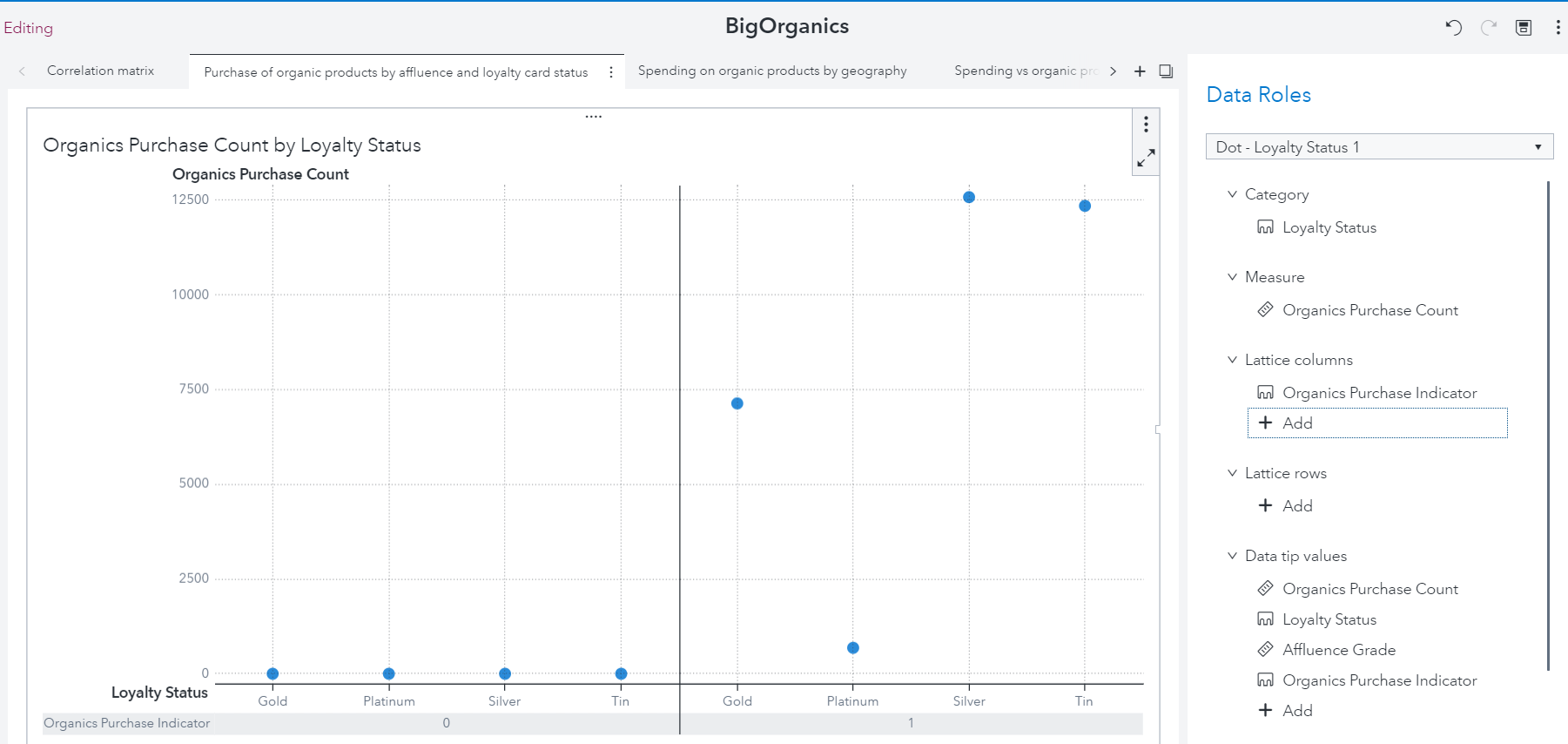
**Finding variable correlations:**

****

In this section, I try to understand the correlation between the variables. I use the Age, Organics Purchase Count, Total Spend, Affluence Grade and Loyalty Card Status in my analysis.

As we can see, the Organics Purchase Count is strongly correlated to the Affluence Grade. Age has some correlation with Organics Purchase Count, Total Spend and Loyalty Card Tenure. All other variable relations are weakly correlated.

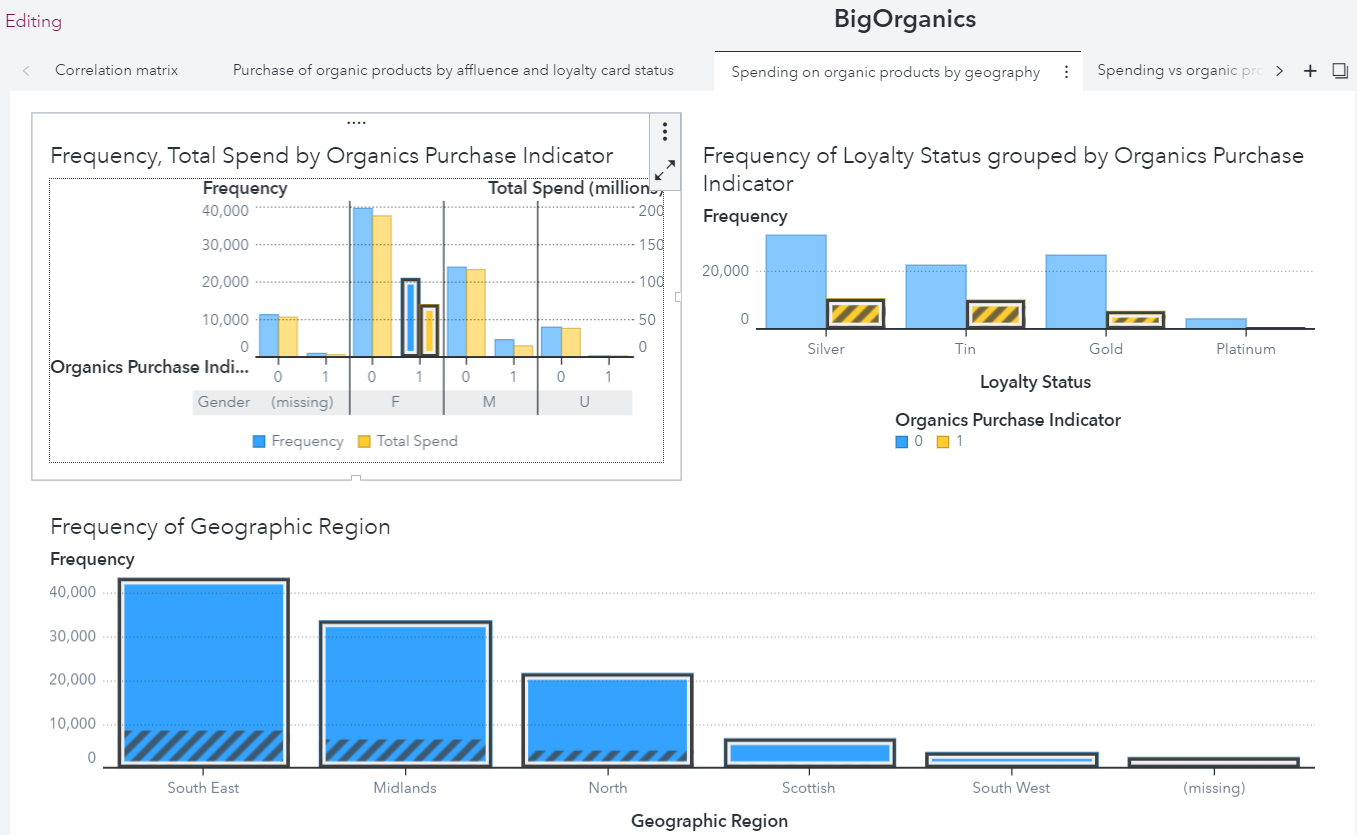
**Purchase of organic products with respect to Affluence grade:**

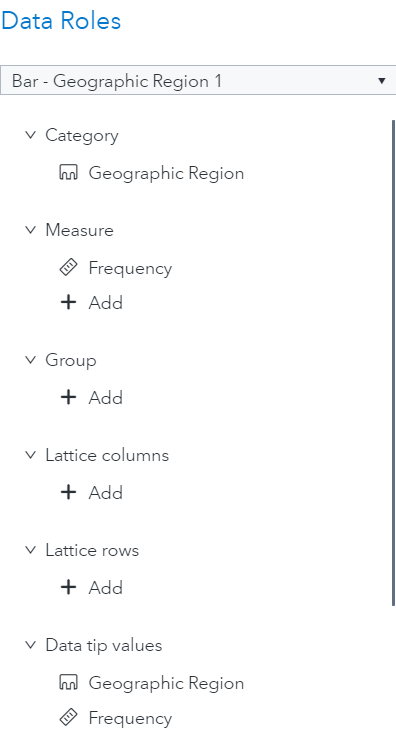
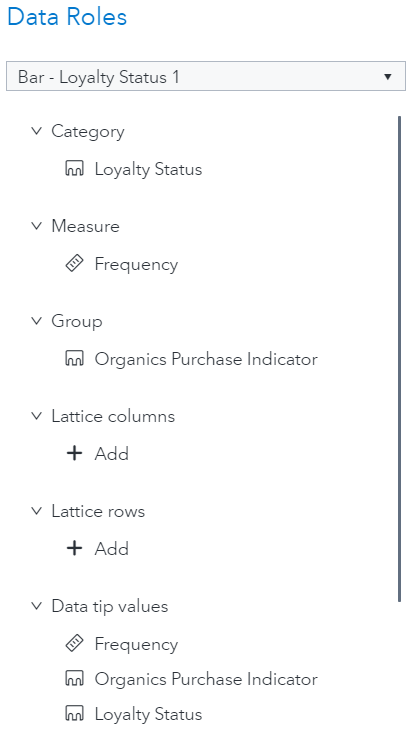
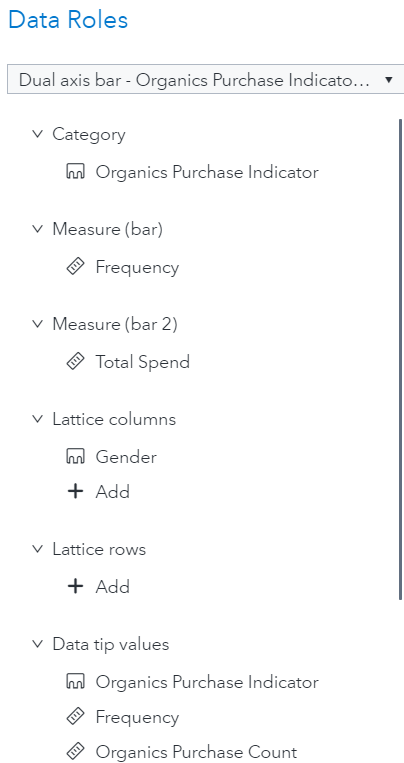
****

I now analyze how the purchase of organic products is influenced by the Loyalty Status of customers.

We see that all loyalty card customers purchase organic products. However, even as loyalty status increases, the purchase of organic products does not always increase. This trend is seen between the Gold and Platinum status customers.

**Spending on organic products by loyalty status and geography:**

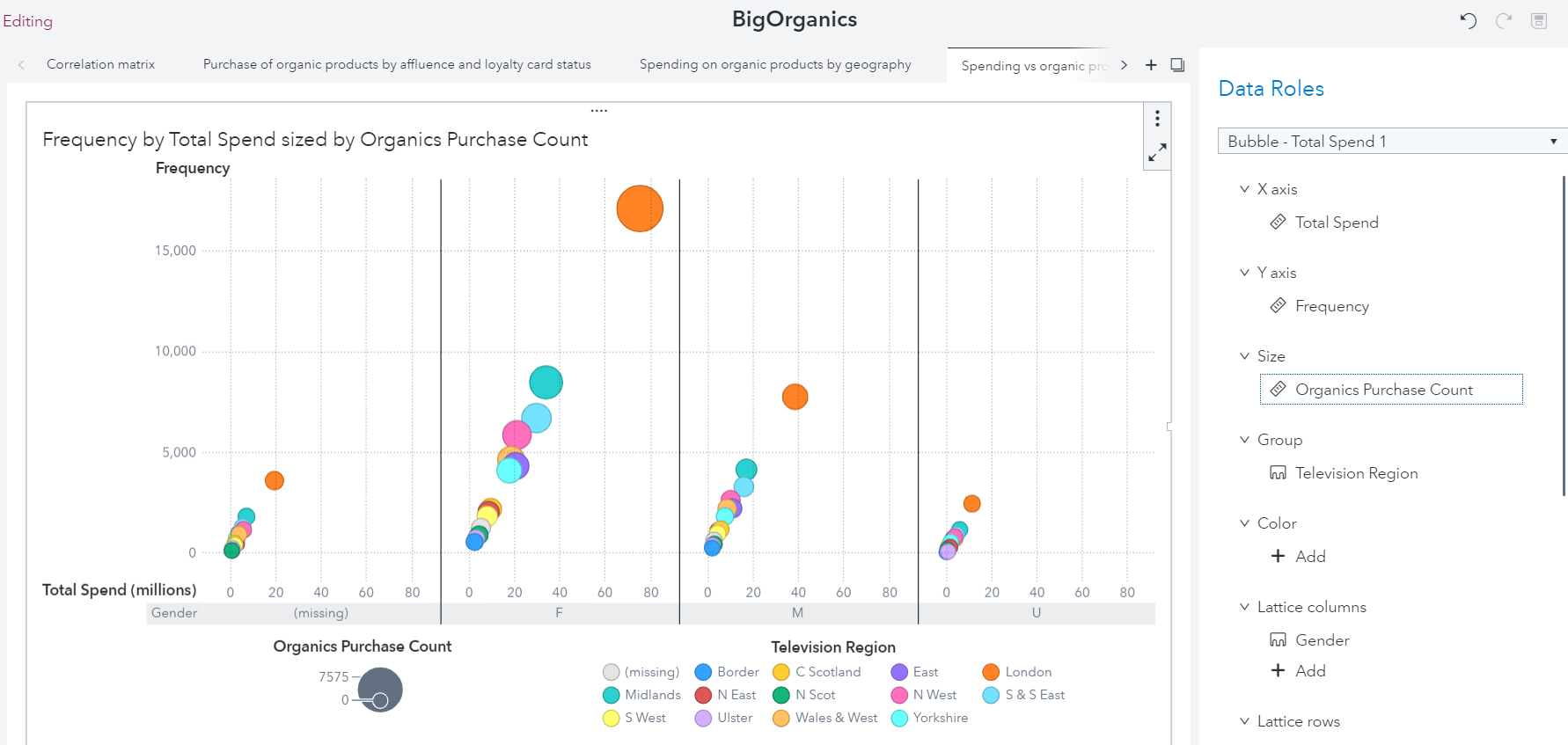


****

Now, I analyze the purchase habits based on Gender, Loyalty Status and Geographic region. There are a significant number of missing values, so the result may not be as accurate as we want it to be. All charts are inter-linked for this analysis.

The graphs shows that women in general (especially in the South East, Midlands and North) buy more organic products than other genders and across all Loyalty statuses. Surprisingly, a similar trend is followed when it comes to not buying organic products.

**Spending vs organic products purchased:**

****

For this graph, I have analyzed the purchase of organic products based on the television region and classified by gender.

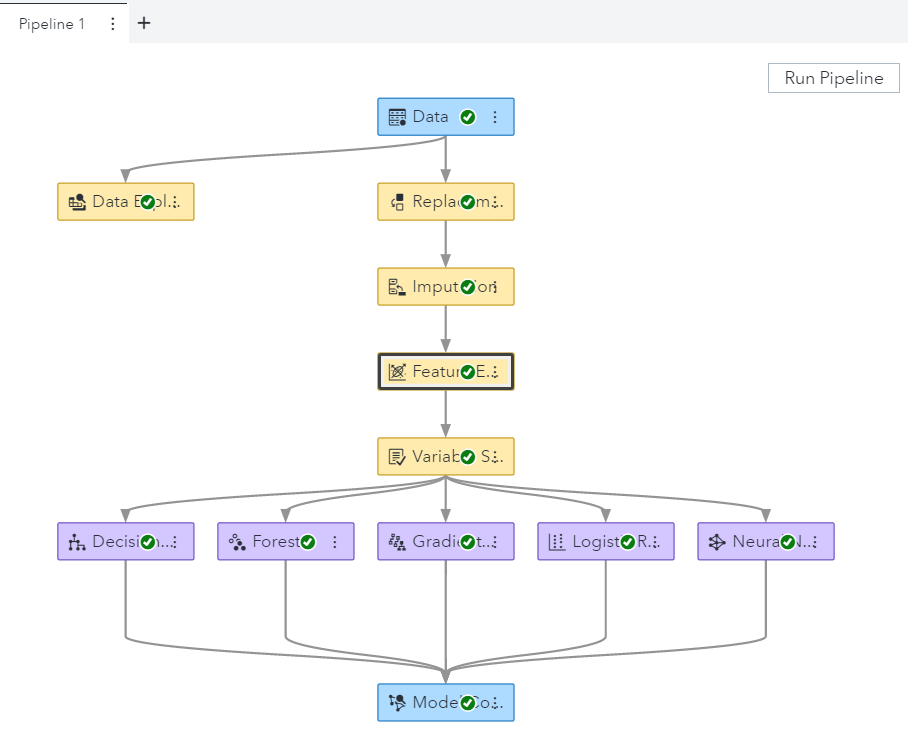
Women in London are the largest buyers of organic products compared to other genders and regions, while the men in Northern Scotland buy the least number of organic products across all other genders and regions.

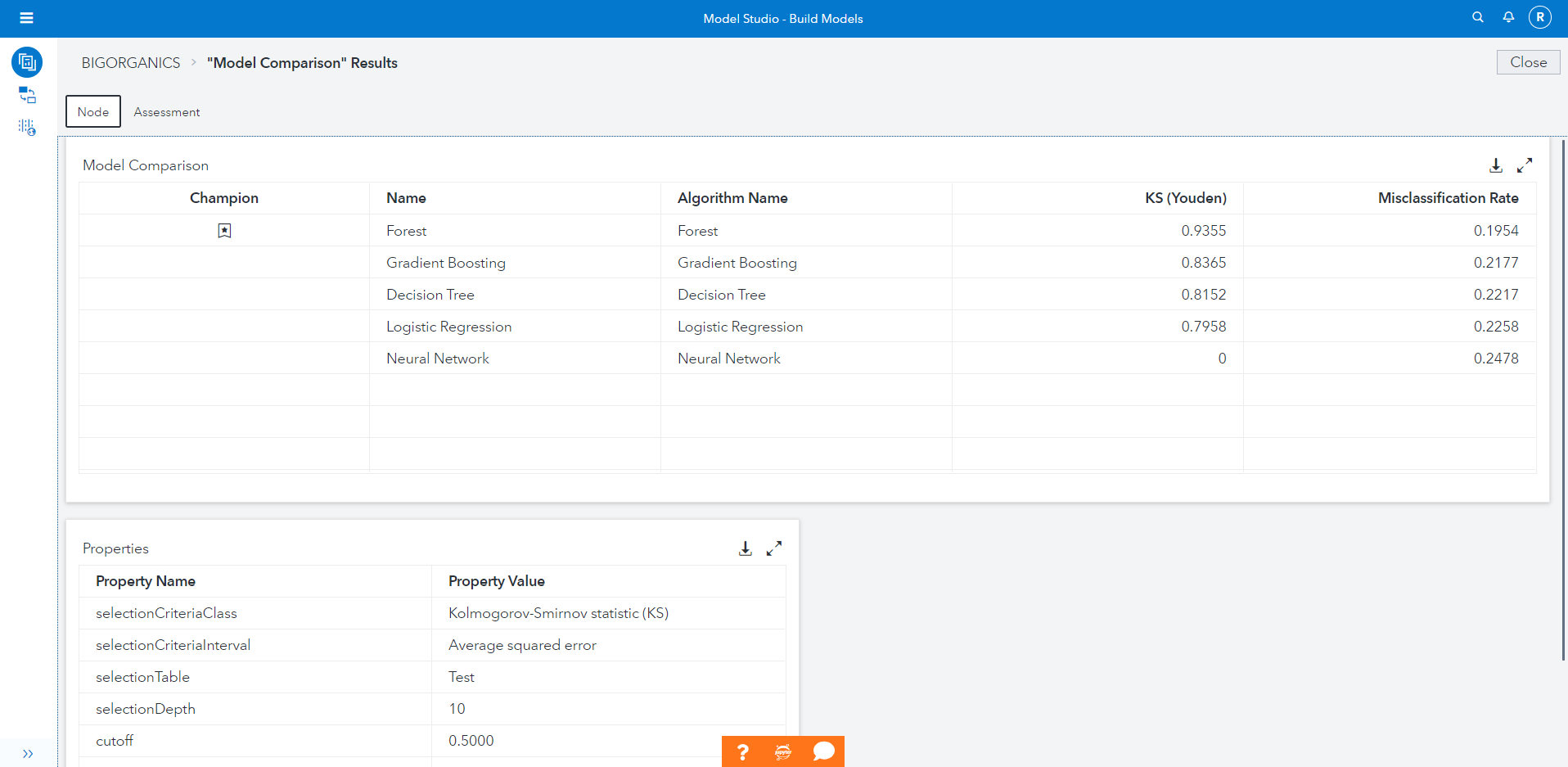
**DATA PIPELINES:**

In this section, I created two types of data pipelines to understand which method had the lowest misclassification rate and the highest KS (Youden) values. The models I have used for comparison are Decision Trees, Forest, Gradient Boosting, Logistic Regression and Neural Networks.

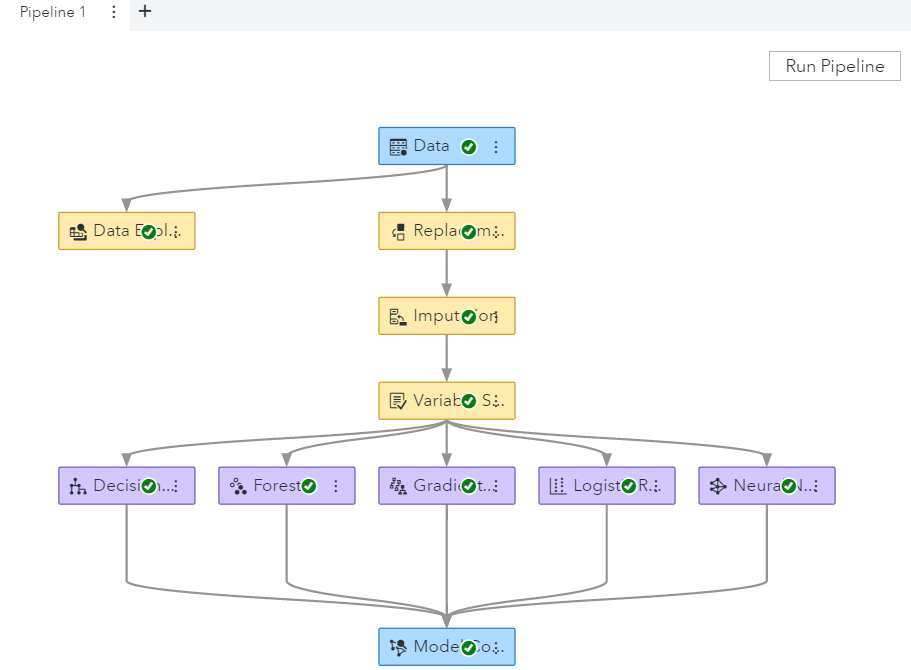
The first data pipeline has the feature extraction node and the second one does not. We will find the best model for our dataset.

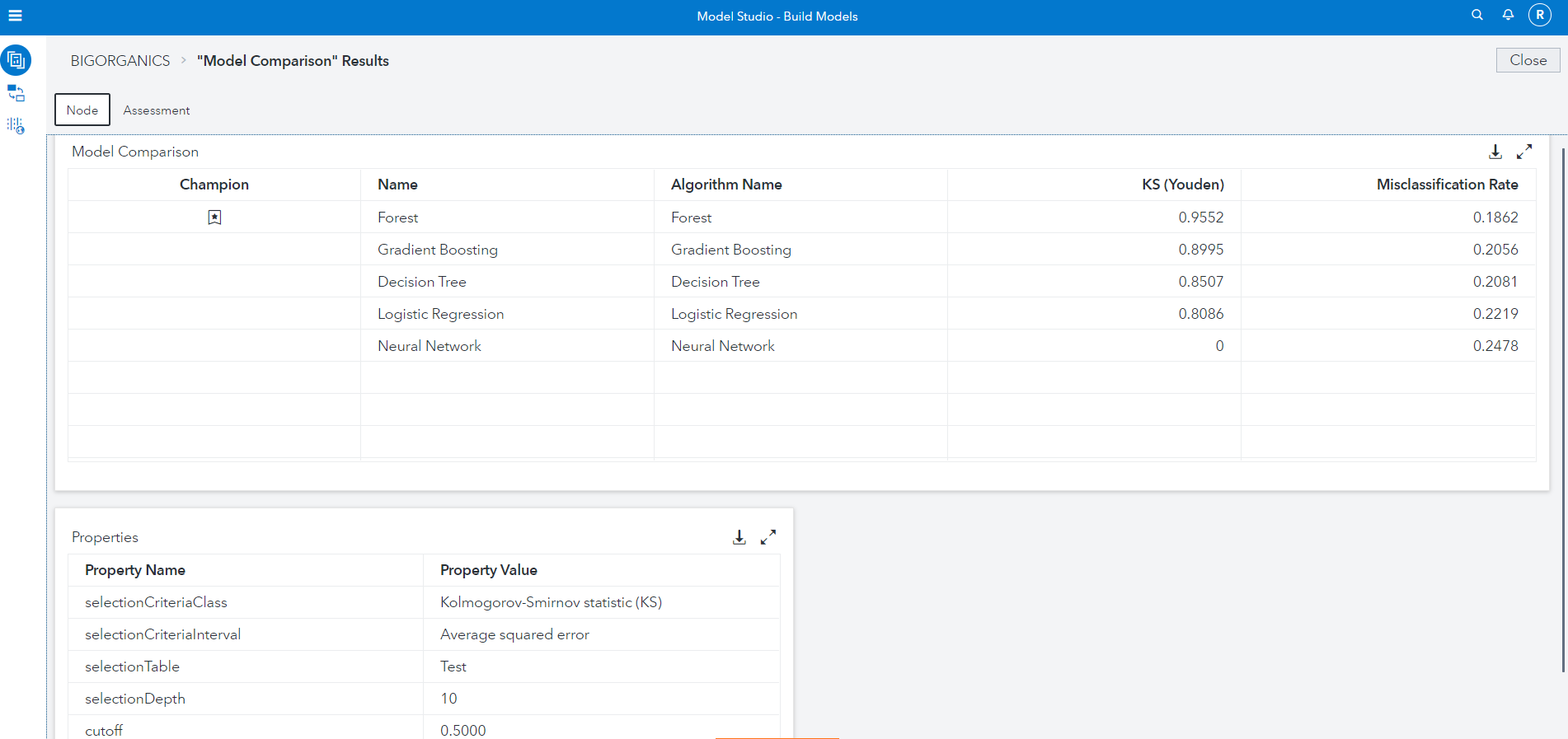
**With Feature Extraction:**





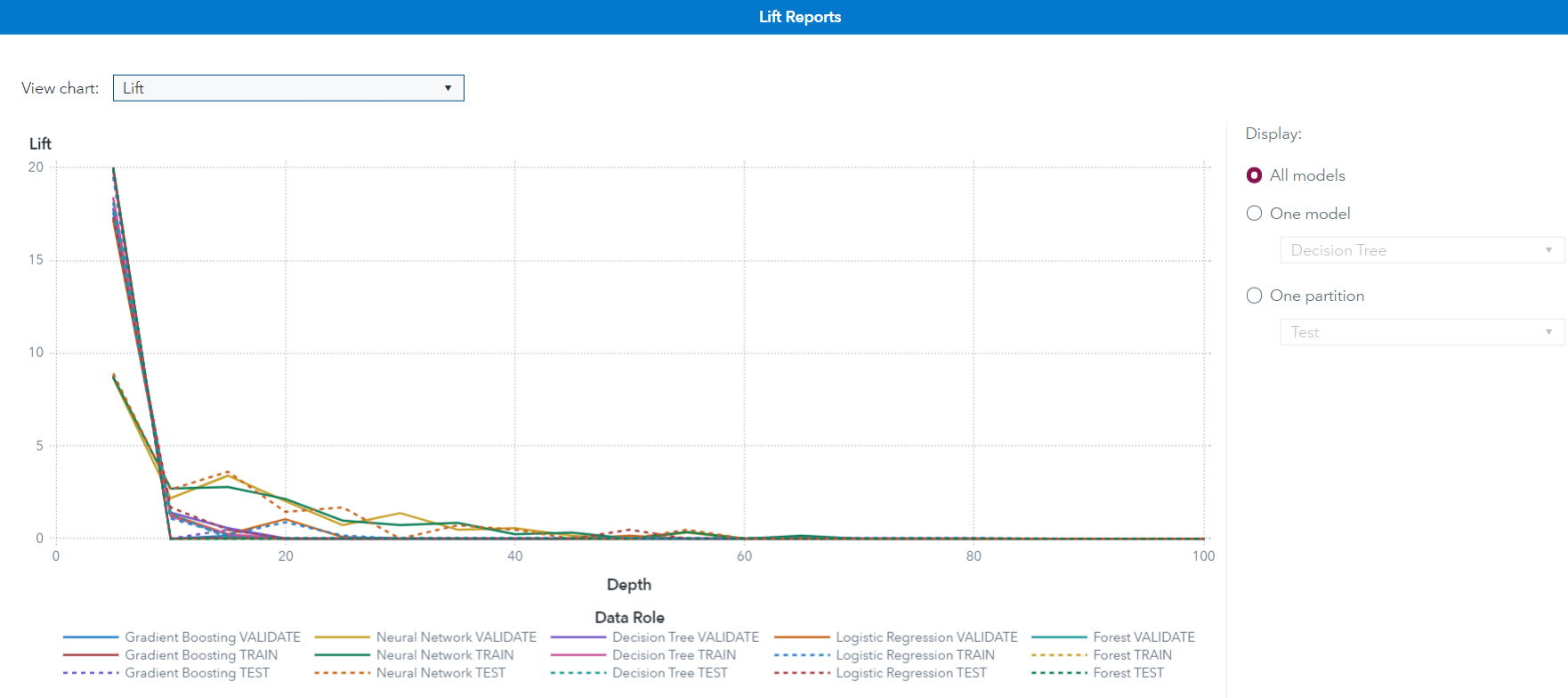
**Without feature extraction:**





We can see that the KS (Youden) score is higher in the model without the feature extraction and the misclassification error is also lower here. Therefore, for the rest of my analysis, I will proceed with the pipeline with no feature extraction.

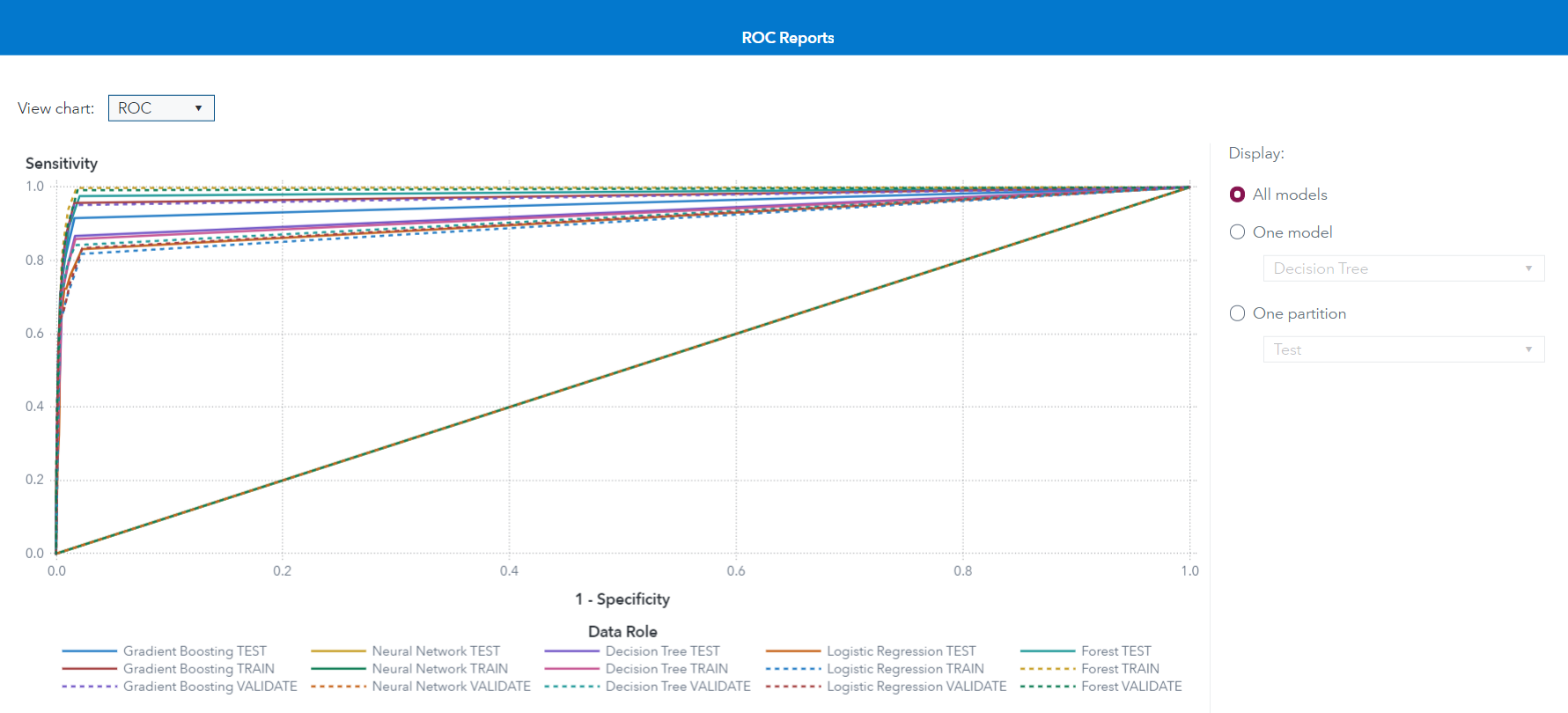
**Lift Reports:**



The gradient boosting and neural network models perform slightly better than the forest model.

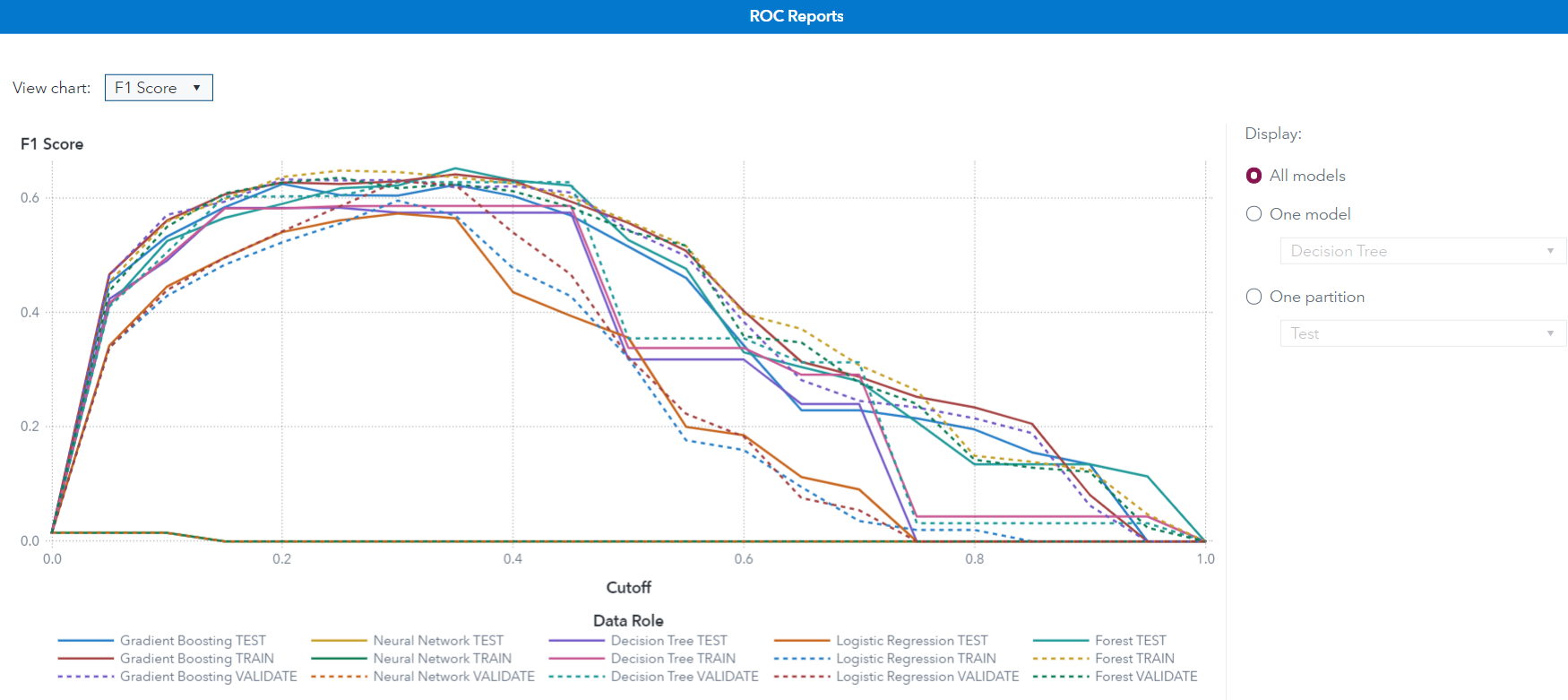
**ROC Reports:**

**ROC:**

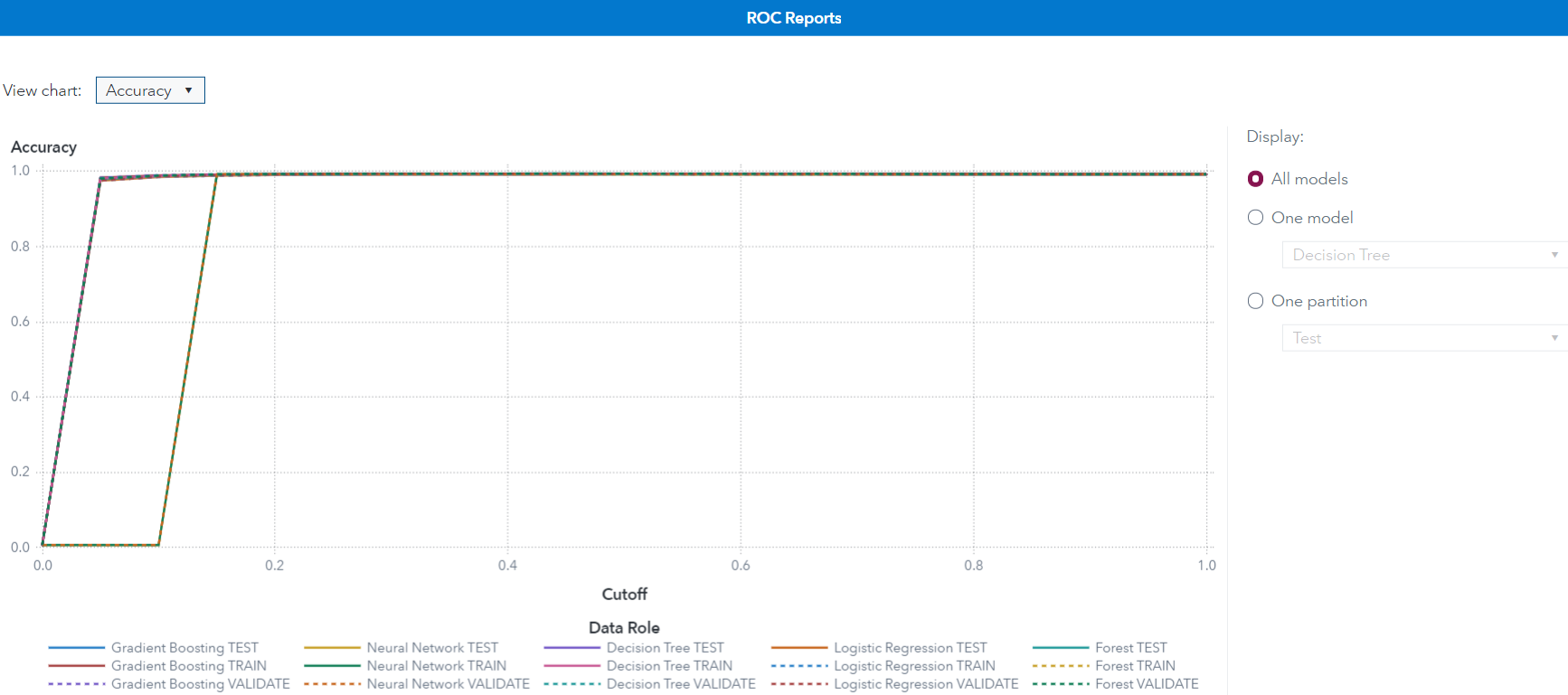


All models other than the Neural network, are a great fit for our data source. One particularly great fit is the Forest model.

**F1 Score:**



Unfortunately, the forest model also has the highest amount of misclassified data. Here, the Gradient boosting model and the Logistic regression model are the better choice, given their low classification rates.



Here too, most models perform well, other than the Neural Network model.

From the KS(Youden) score, misclassification rate and the graph analysis, we see that the Forest model is the best model fit for our current dataset.