CIS9557 GROUP 8 PROJECT:

ANALYZING CUSTOMER CAMPAIGNS (US PACWEST USE CASE)

*- by archana k, chani t, divya k, pooja v & tanaya n*.

Marketing is a key aspect of business success and customer targeting is pivotal to marketing. Our focus is on evaluating performance of the email campaign in the Western Pacific states of USA. Through efficient data modelling, we would determine what drives a customer in the PacWest region to stay with the insurance firm and what are the attributes which define that. As our conclusion, we would provide the firm with a business strategy to help them increase their customer base, retain the current customers and increase their Customer Lifetime Value (“CLV”).

**Data Cleaning**

**Observations:**

The PacificWest Customer Campaign data provided has information about **6394** customers and **24** attributes for each customer.

Each of these attributes are meaningful and do no contain null or empty values.

The number of customers who responded to the campaign are **904,** while those who did not respond are **5,490**.

So, we did a deeper exploratory data analysis to get insights into customers who responded. This would help us better understand the data and recognize patterns, if any, of the customers who responded to the campaign.

**Exploratory Data Analysis**

We started off by analyzing the states which have customers with highest CLV. **Oregon** and **California** stand tall in that order.

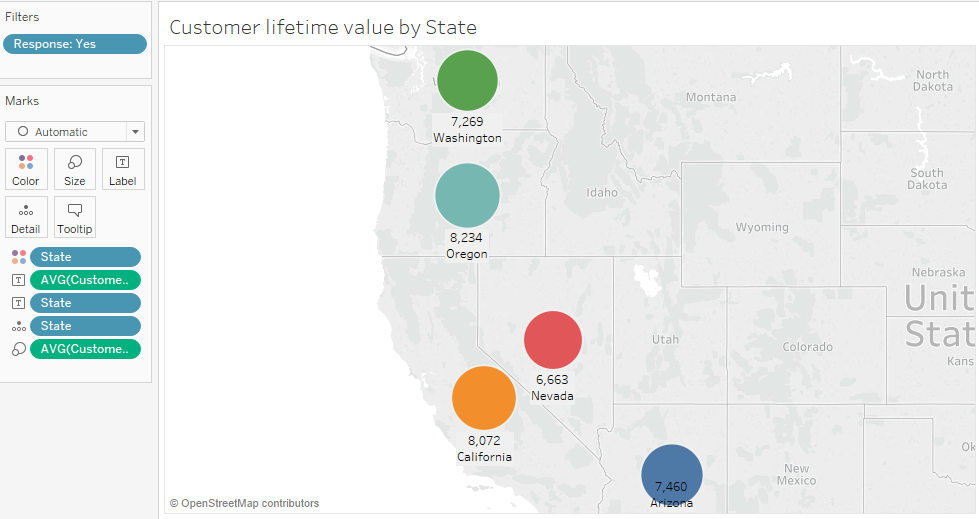


Figure 1

We then figured out that **Personal Auto** was the Policy Type which most of the customers possessed.

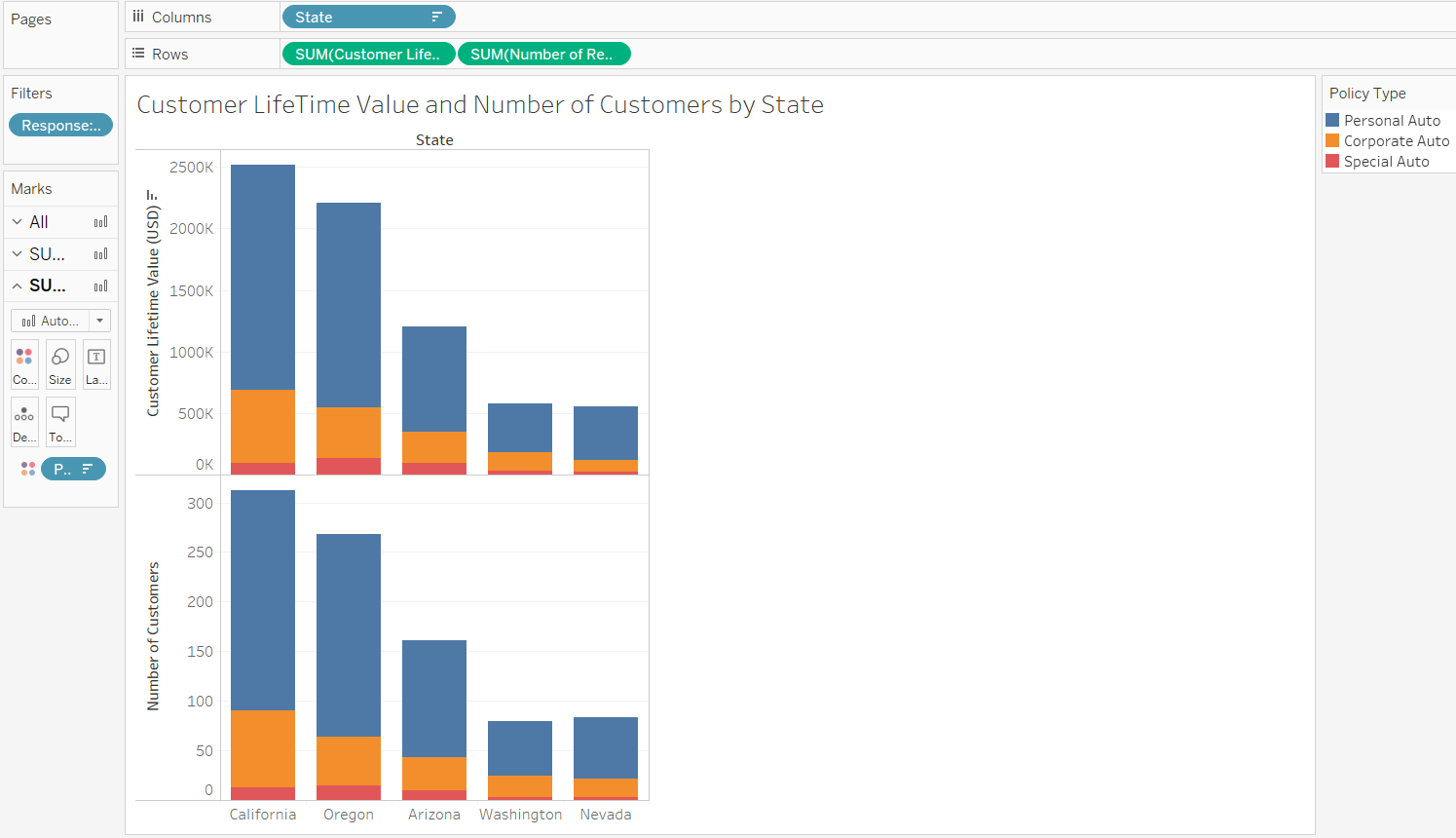


Figure 2

Further, customers who responded to the campaign, were largely those who possessed **midsize** cars with **Personal L3** type of Policy.

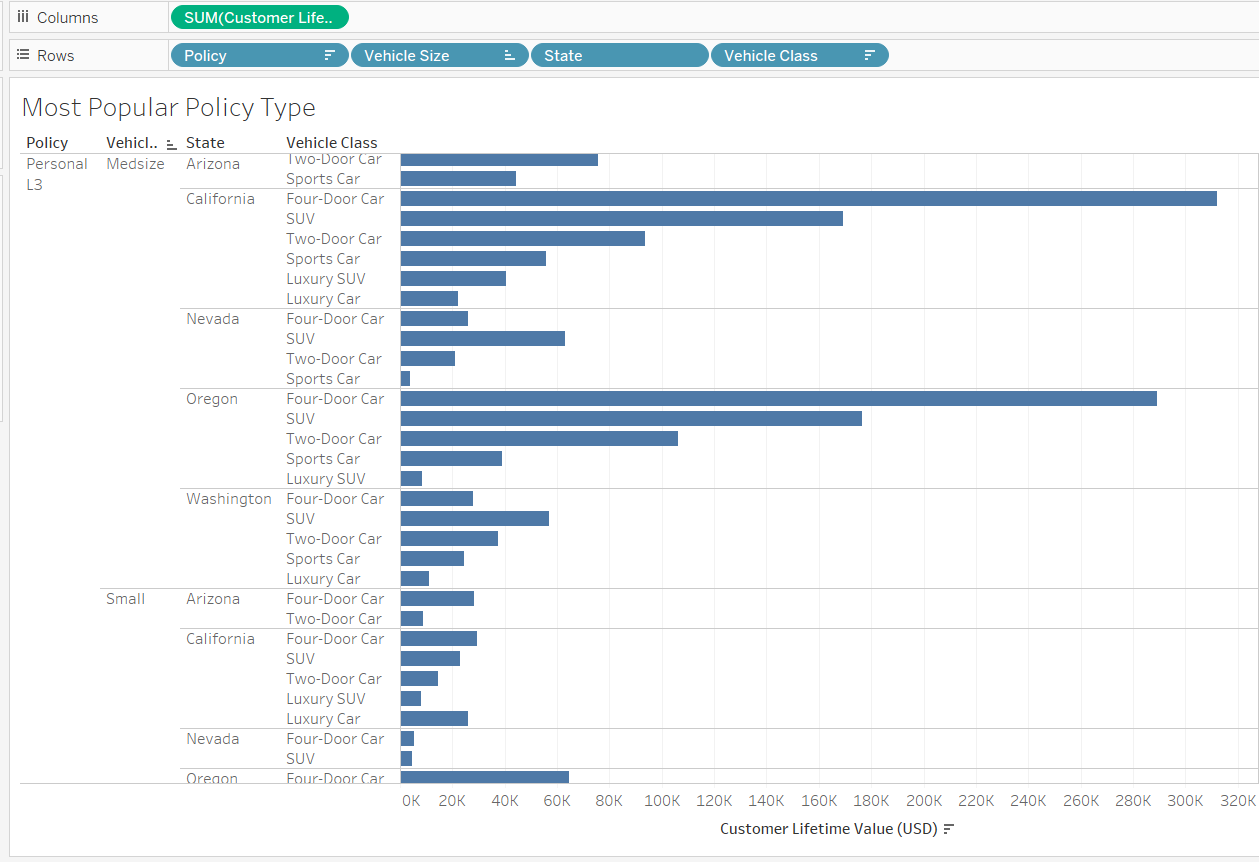


Figure 3

With this initial understanding, we started our task to gain the best f-measure and through the process, continued further exploratory data analysis through Tableau and RapidMiner Modeling.

Task I: Determining Response to campaign

The business problem here, is to determine whether a customer will respond to the campaign. We have used techniques such as exploratory data analysis using Tableau and trial and error methods to determine which attributes will directly affect the response of a customer. In the following section, we have described some of our detailed analysis on understanding the data and selection of the most relevant attributes:

**Step I:** **Identify the Attributes**

From the below graph, we observed that customers only responded to Offers 1,2 and 3. The response rate was the highest for Offer 2 followed by Offer 1. Offer 4 wasn’t purchased by any customer.

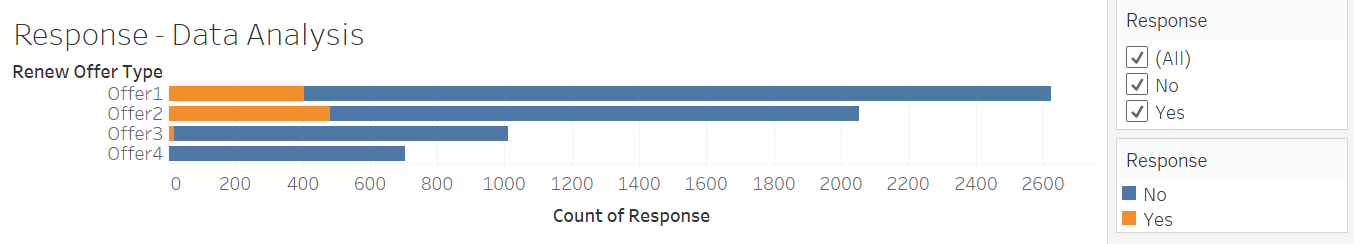


Figure 4

In the following chart, we see that the customers Education and Employment status affects the ‘Yes’ responses. Especially people who are employed tend to respond positively to the survey.

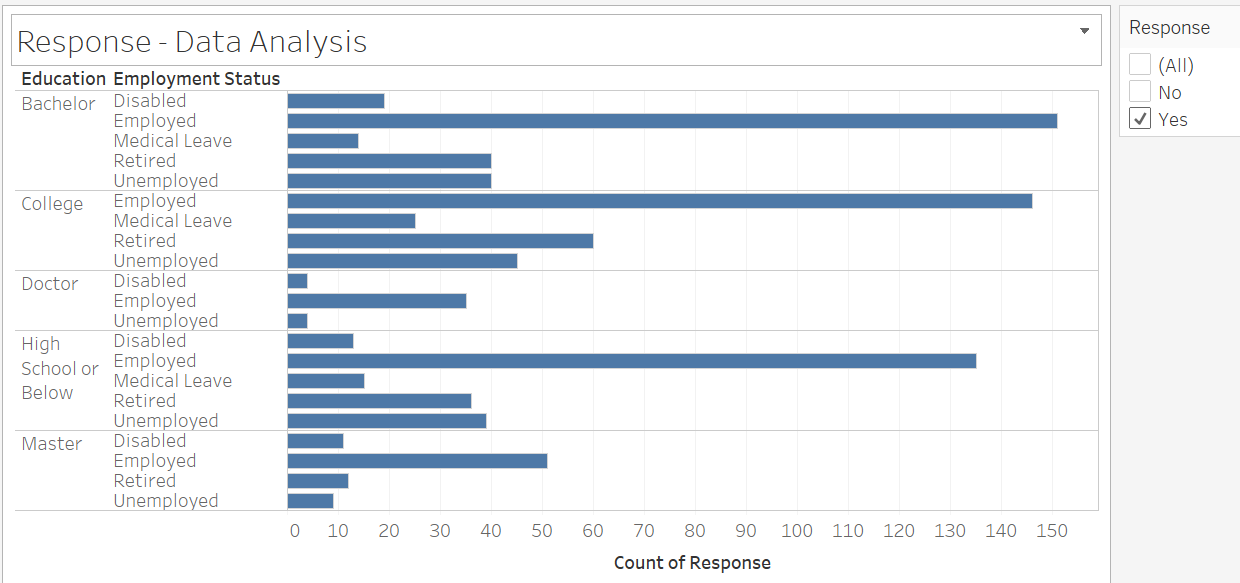


Figure 5

In the below figure, we see that Customers whose total claim amount exceeds a threshold of 1400, don’t respond to the campaign.

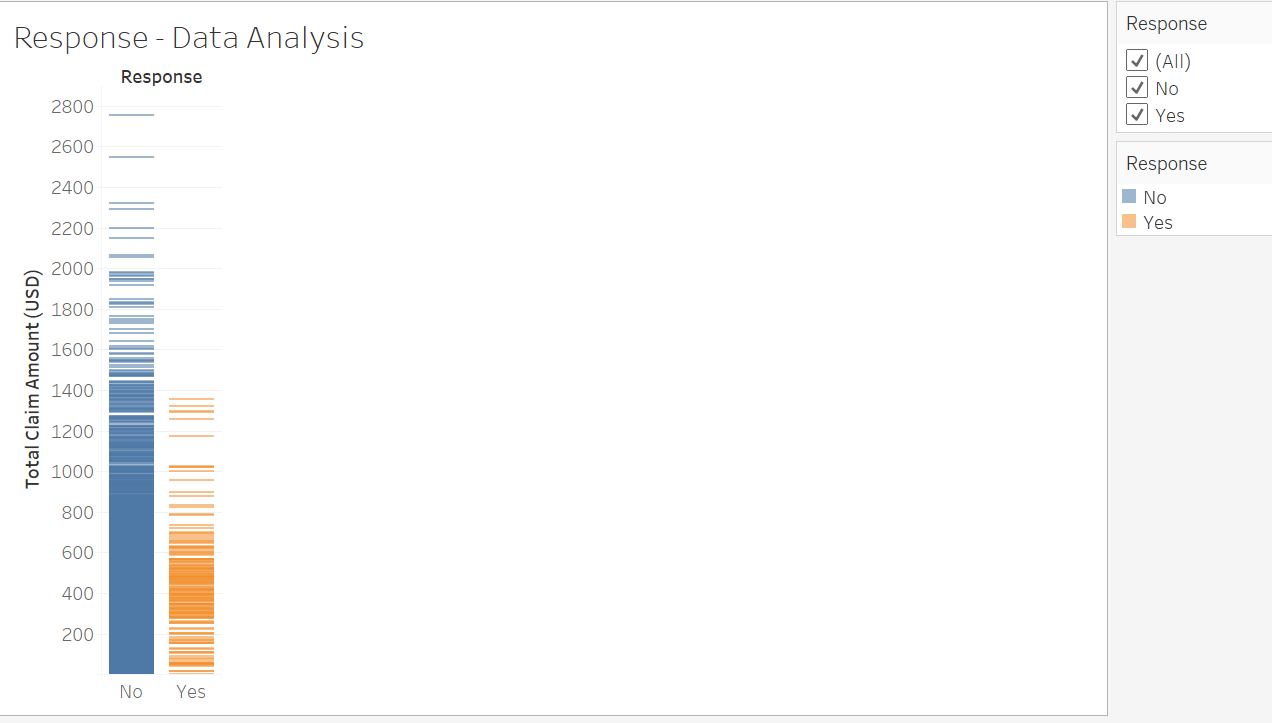


Figure 6

Customers who are ‘Married’ responded the most to the campaign. We have higher response from ‘Female’ customers than ‘Male’ customers whose marital status was ‘Divorced’. ‘Single’, male and female customers have almost similar response.

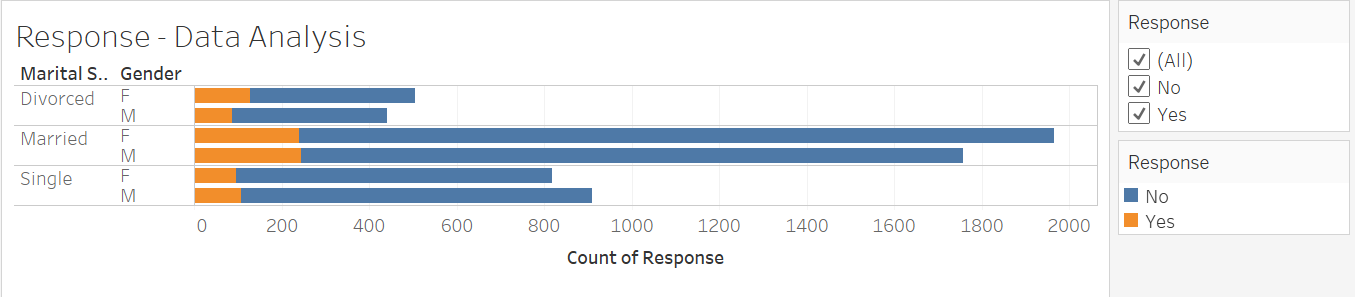


Figure 7

**Step II: Comparing the Models**

The next step was to compare different classification models and choose the model with best performance. We would use f-measure as a parameter to compare different models.

As mentioned in Step I, we selected the below attributes as they were the most important attributes in determining whether a response would be received or not.

**Selected Attributes:**

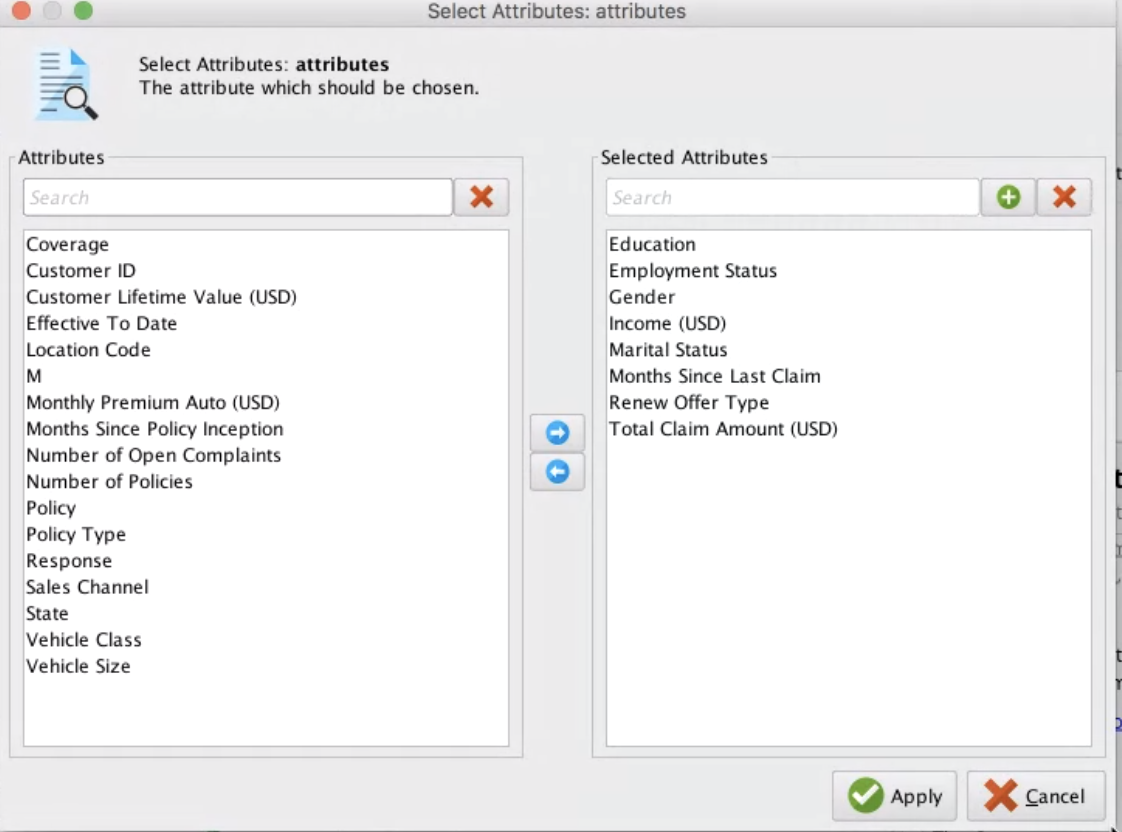


Figure 8

**Comparing multiple Models:**

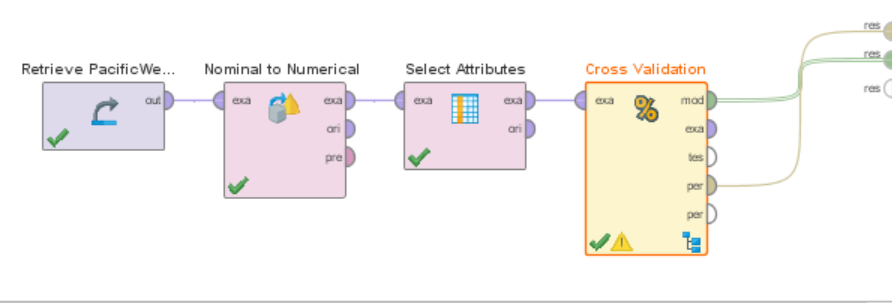


Figure 9

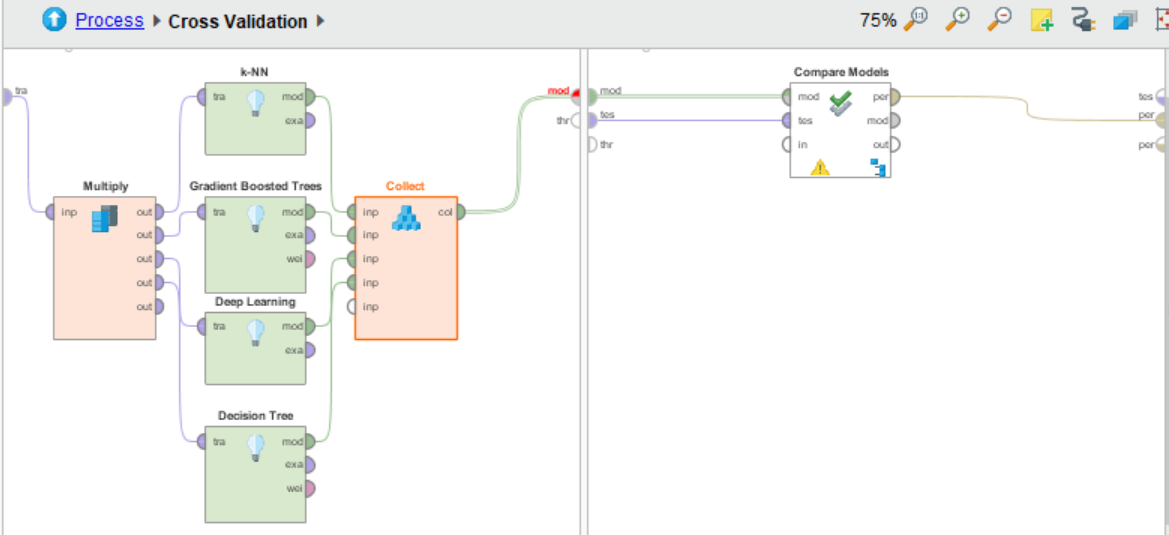


Figure 10

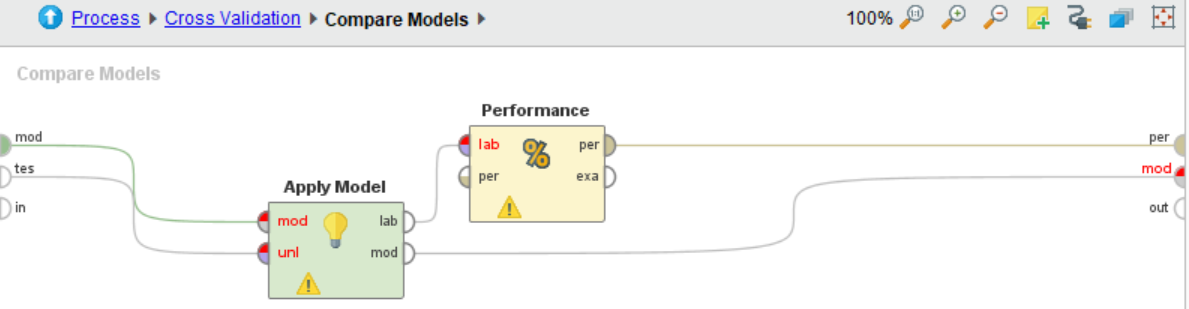


Figure 11

**Selection of the Model with highest F-Measure:**

Out of the above 4 models, we selected k-NN since it had the best performance (i.e. the highest f-measure amongst all models). The following table describes the performance of each model based on accuracy, precision, recall and f-measure:

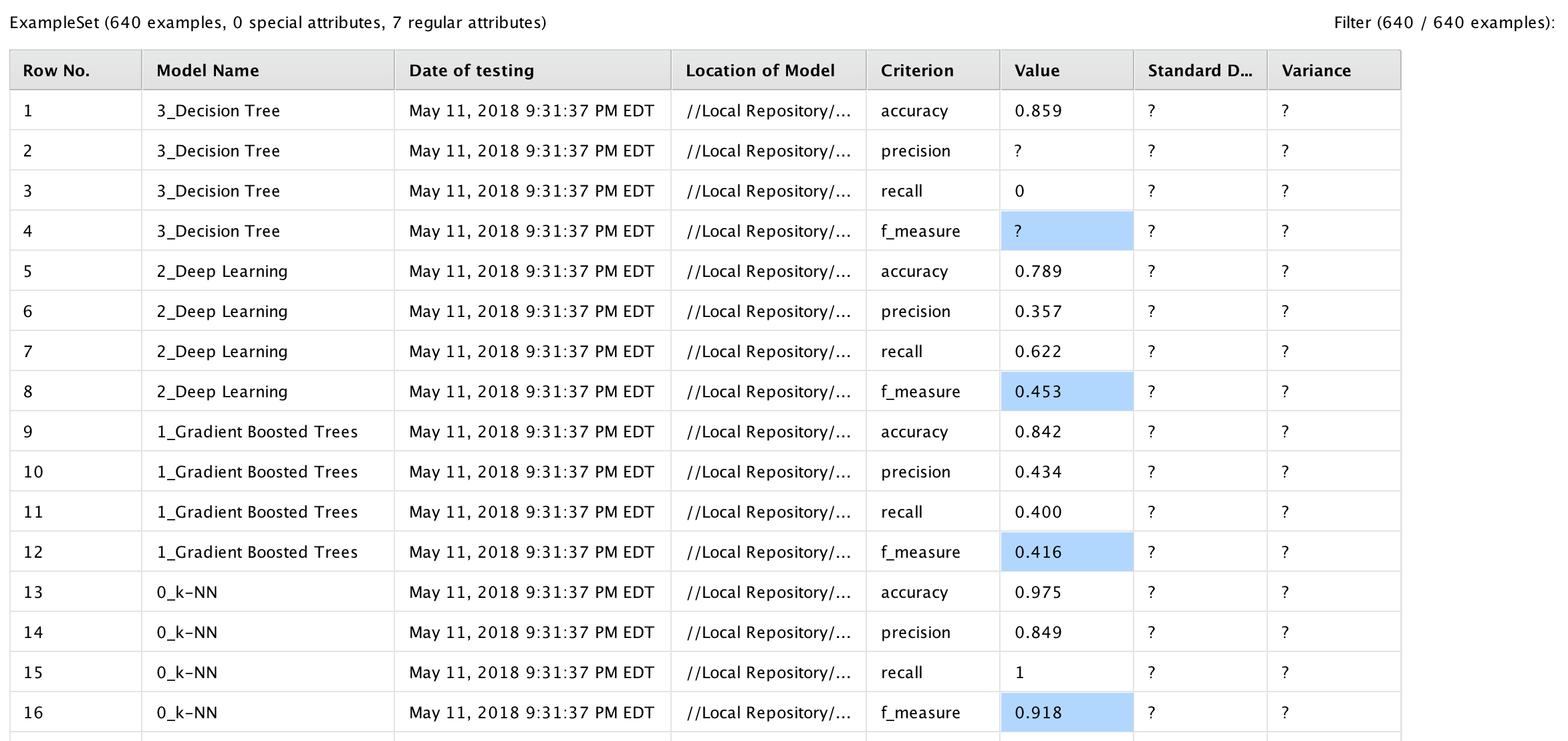


Figure 12

*Note:* We did not get any precision and f-measure value for Decision Tree because it is unable to predict ‘Yes’ responses. It is only able to predict the ‘No’ responses.

**Step III: Applying the Model to Scoring dataset**

We applied the K-NN model to the scoring data set to determine the responses.

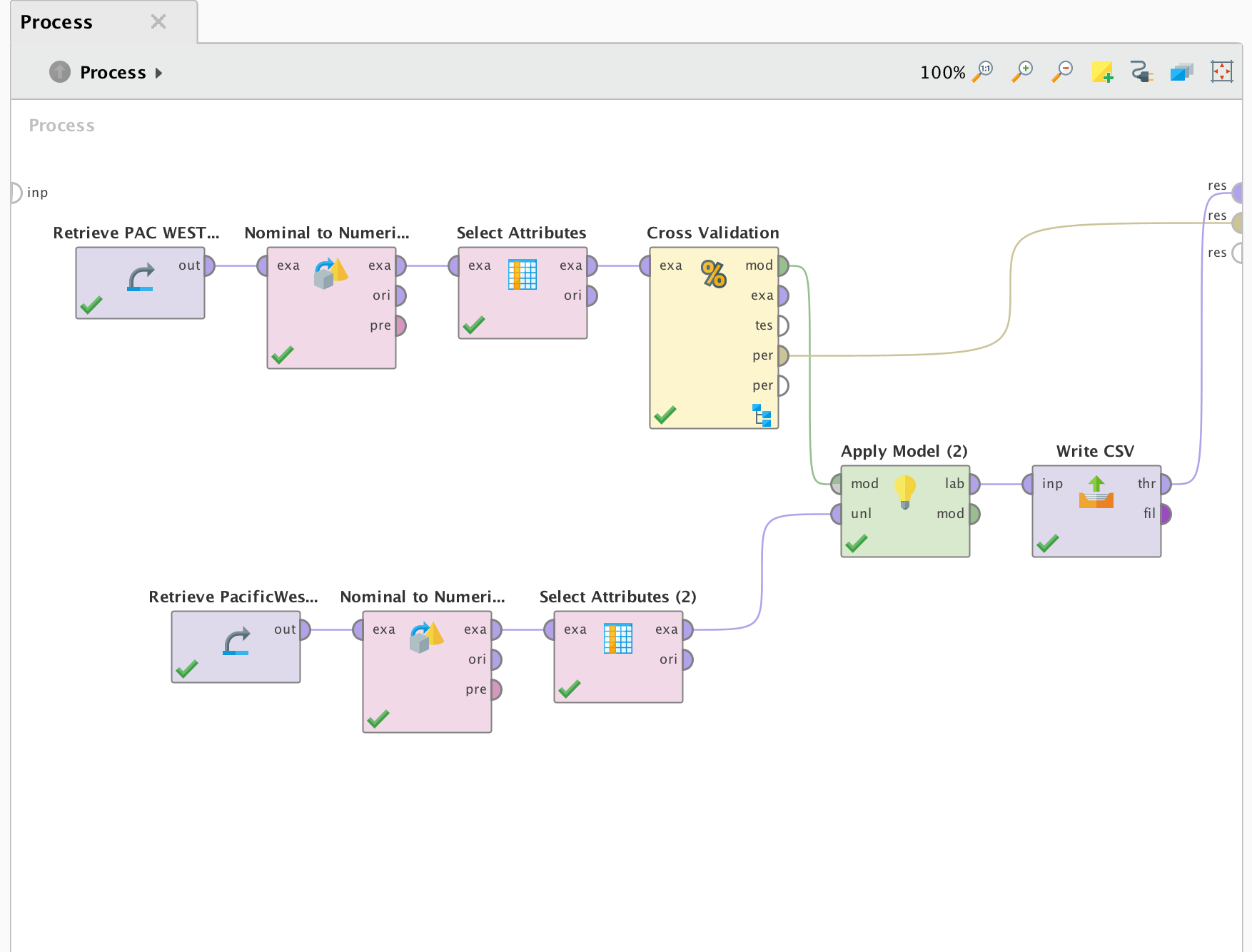


Figure 13

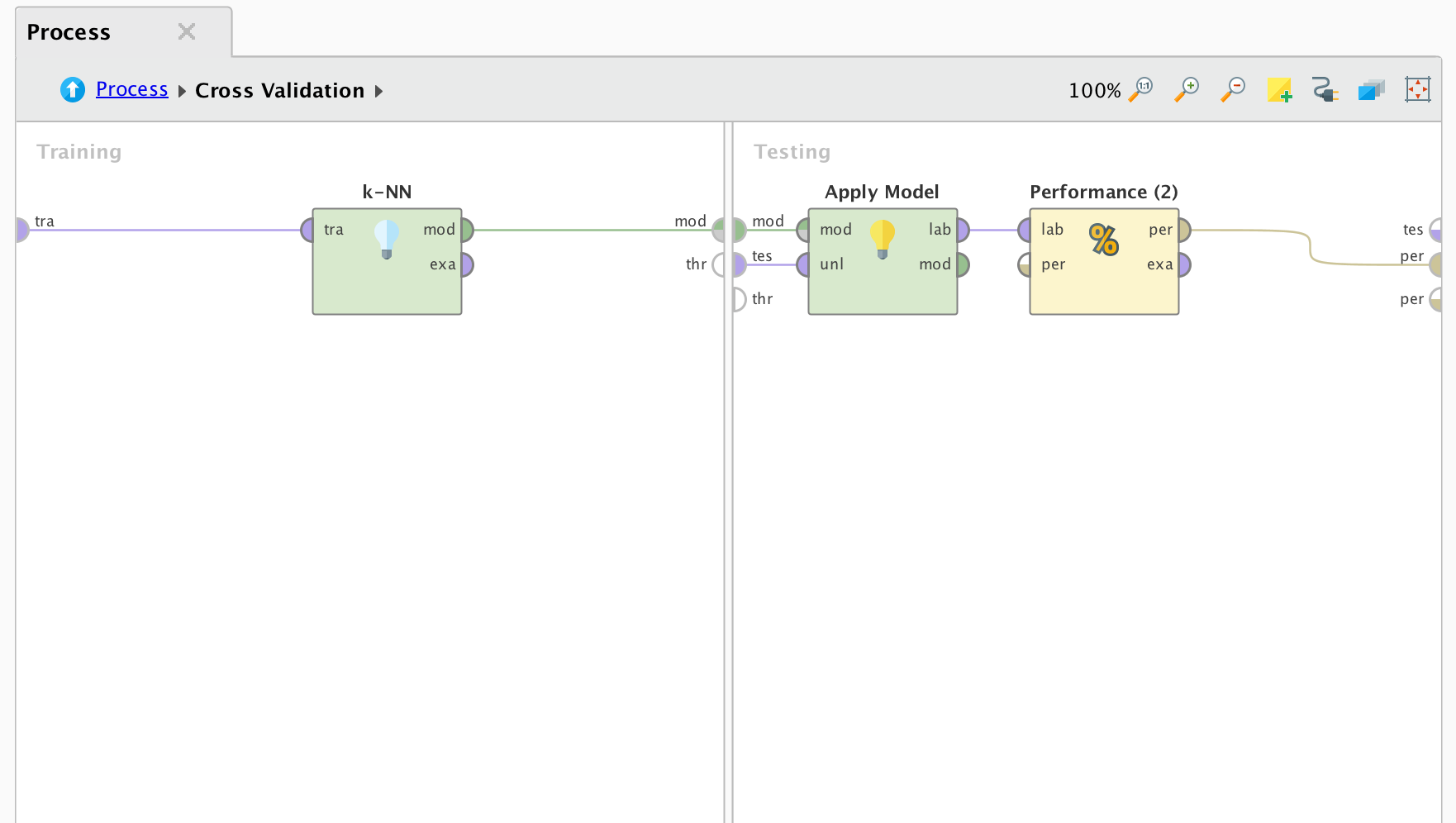


Figure 14

**Below is the output of the task I: Classification**

# 

# Task II: Determining characteristics of customers that responded to campaign

**Step I: Clustering**

To determine the characteristics of customers who responded to our campaign we used the X-Means clustering operator in RapidMiner.

X-Means is a clustering algorithm which determines the correct number of centroids based on a heuristic approach. It begins with a minimum set of centroids and then iteratively exploits if using more centroids makes sense according to the data.

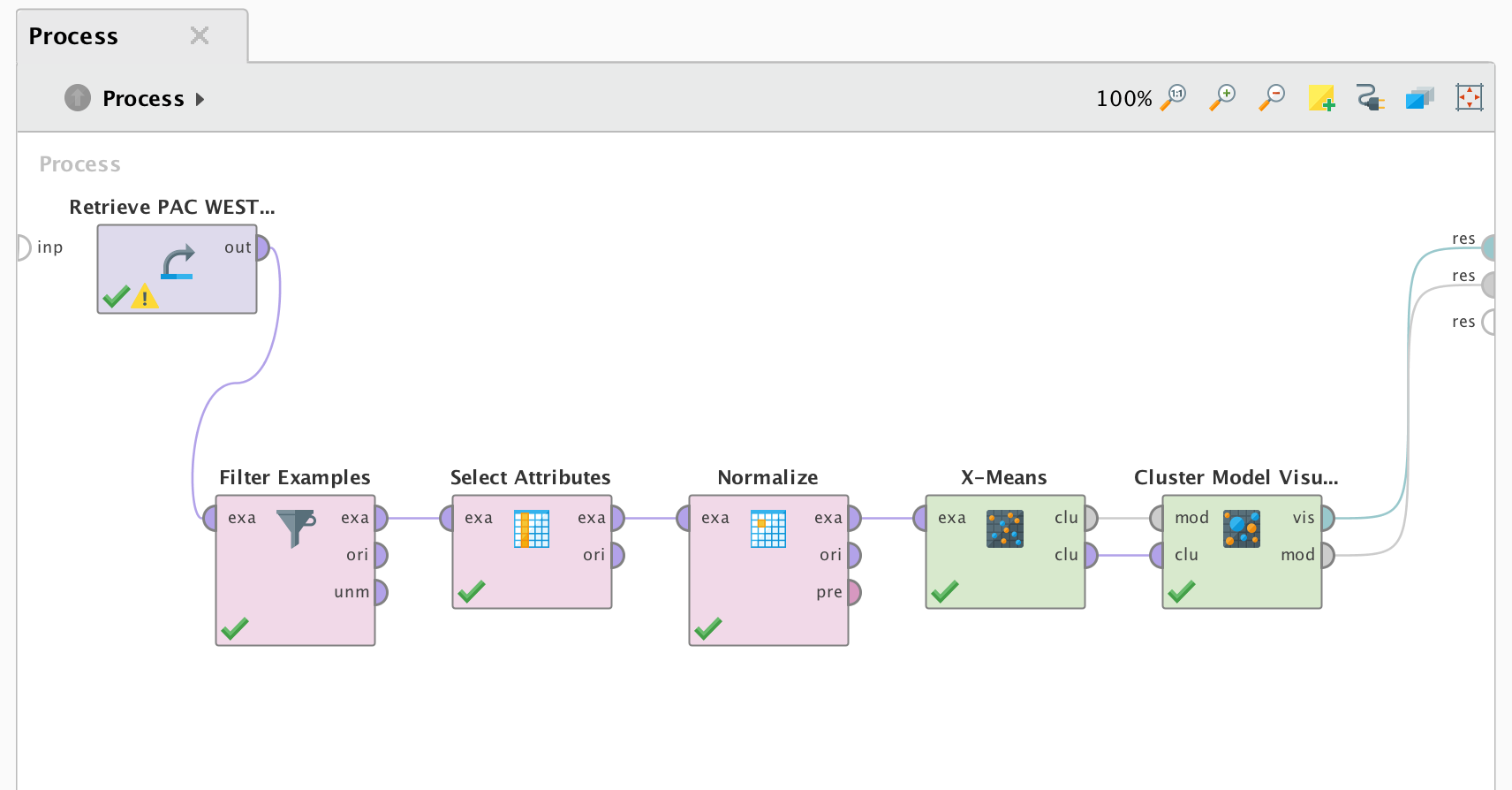


Figure 15

**Step II: Customer Segmentation using X-Means**

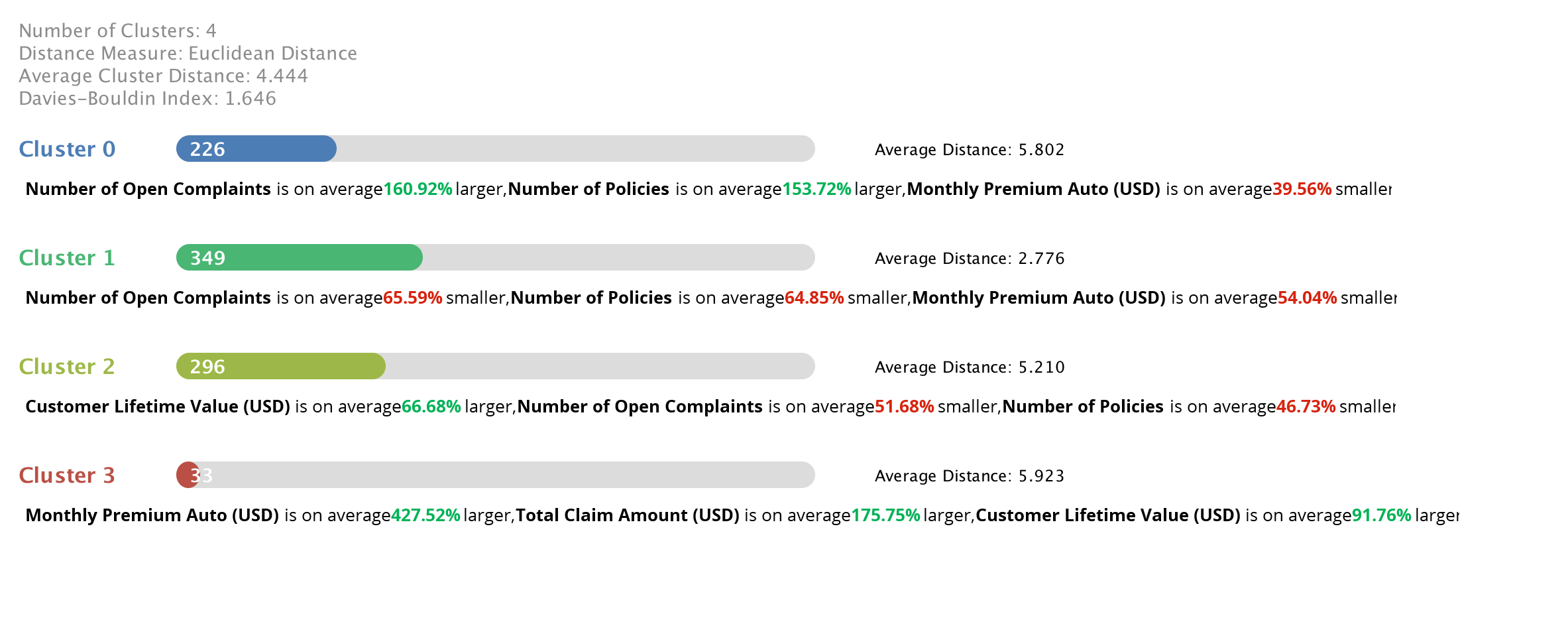


Figure 16

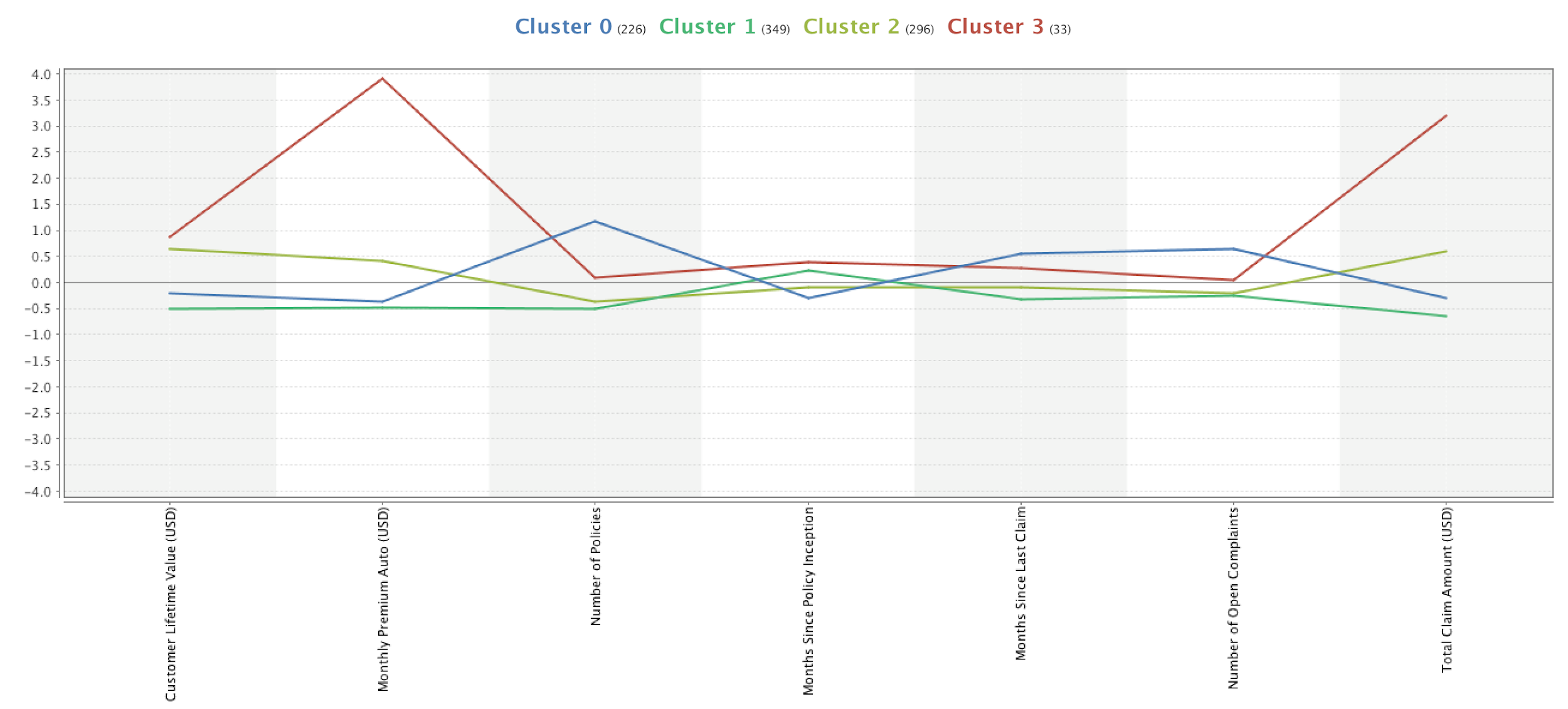


Figure 17

**Step III: Analysis of Customer Characteristics**

Based on above analysis we were able to identify the below 4 categories in which our customers belonged:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Clusters | Cluster Name (Label) | Number of Customers | Months since inception | Monthly Premium | Total Claim Amount | Customer Lifetime Value | Number of Policies | Number of Open Complaints |
| Cluster 3 | Platinum Class - High Value | 33 | Premium Customers | 428% **↑** | 176% **↑** | 92% ↑ | Average | Average |
| Cluster 2 | Gold Class - Medium Value | 296 | Loyal Customers | Slightly Higher than average | Higher than average | 67% ↑ | 47% ↓ | 52% ↓ |
| Cluster 1 | Silver Class - High Potential | 349 | Regular Customers | 54% ↓ | Lower than average | Lower than average | 65% ↓ | 66% ↓ |
| Cluster 0 | Basic Class - Mid-Low Potential | 226 | New customers | 40% ↓ | Slightly Lower than average | Slightly Lower than average | 154% ↑ | 161%↑ |

# Task III: customer lifetime value

**Step I: Identify the attributes**

The attributes were selected based on the higher correlation. While applying the models only those attributes that reduced the “root mean square error” values significantly were identified and used in the model comparison process.

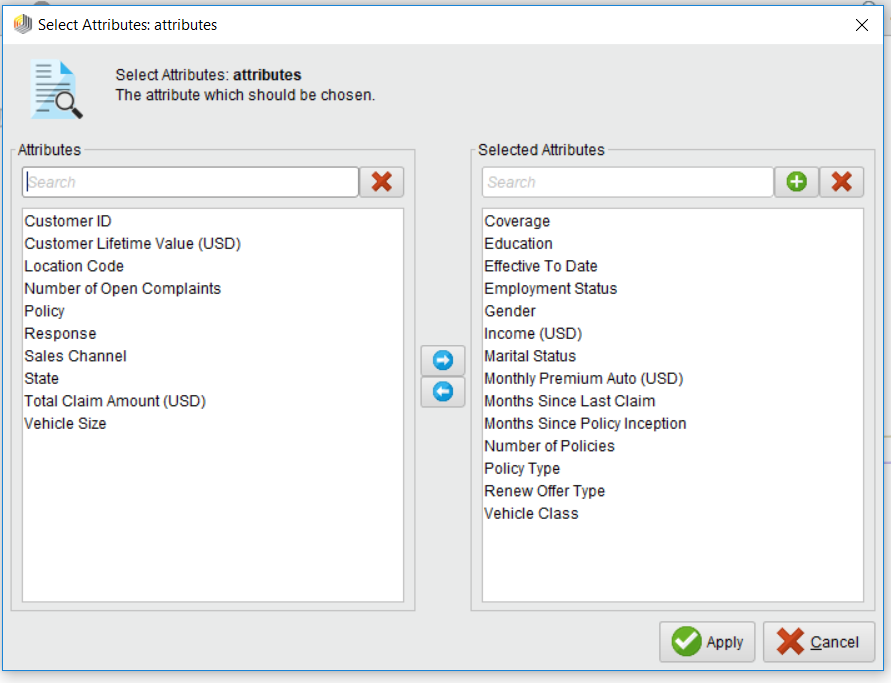


Figure 18

**Step II: Comparing the models**

The next step was to compare different regression models and choose the model with best performance. We used root mean square error as a parameter to compare different models.

As mentioned in Step I, we selected the above attributes as they were the most important attributes in determining the customer lifetime value.

The models used for comparison were:

1. Linear Regression
2. K-NN
3. Gradient Boosted Tree

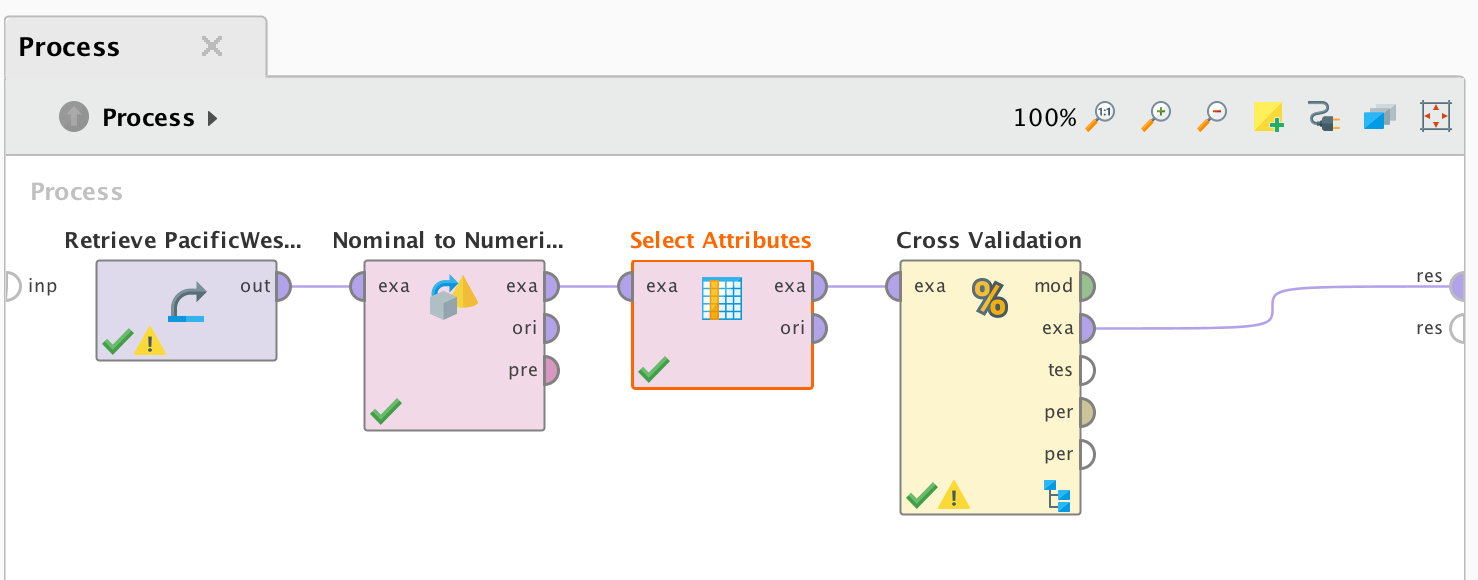


Figure 19

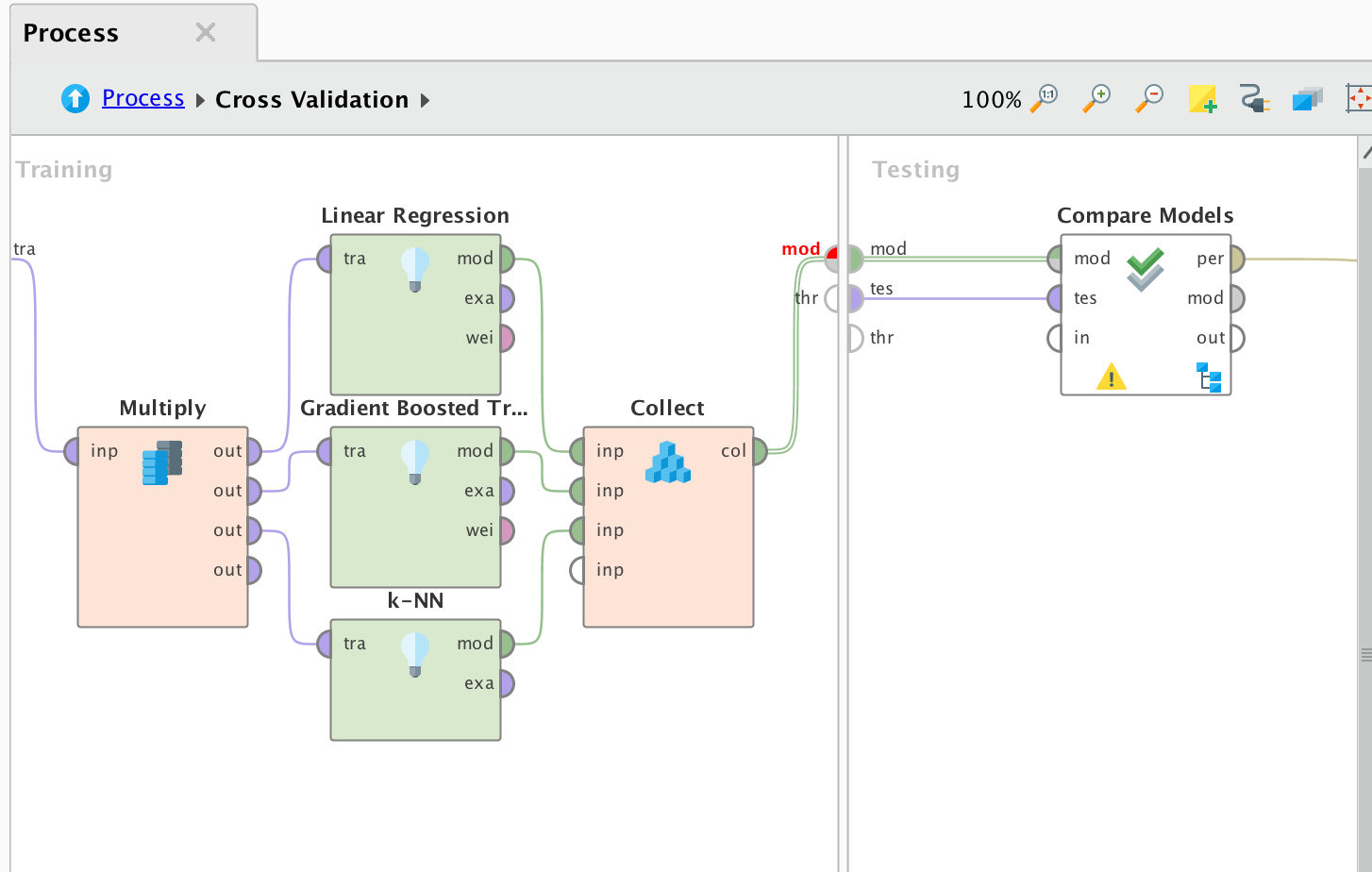
****

Figure 20

We compared the RMSE values for the above mentioned three models and select Gradient Boosted Trees since it had the lowest RMSE value.

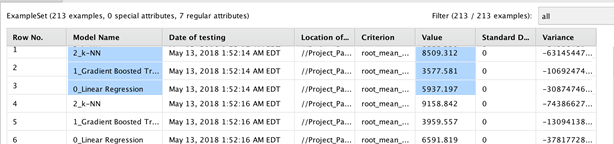
****

Figure 21

**Step III: Apply the Model**

**Applying the Model on training dataset**

In this task, we have developed an estimator for determining the Customer Lifetime Value(CLV).

For estimating the Customer Lifetime Value, we have used Gradient Boosted Tree model as described in step II. The process used for estimating the Customer Lifetime Value is as shown below:

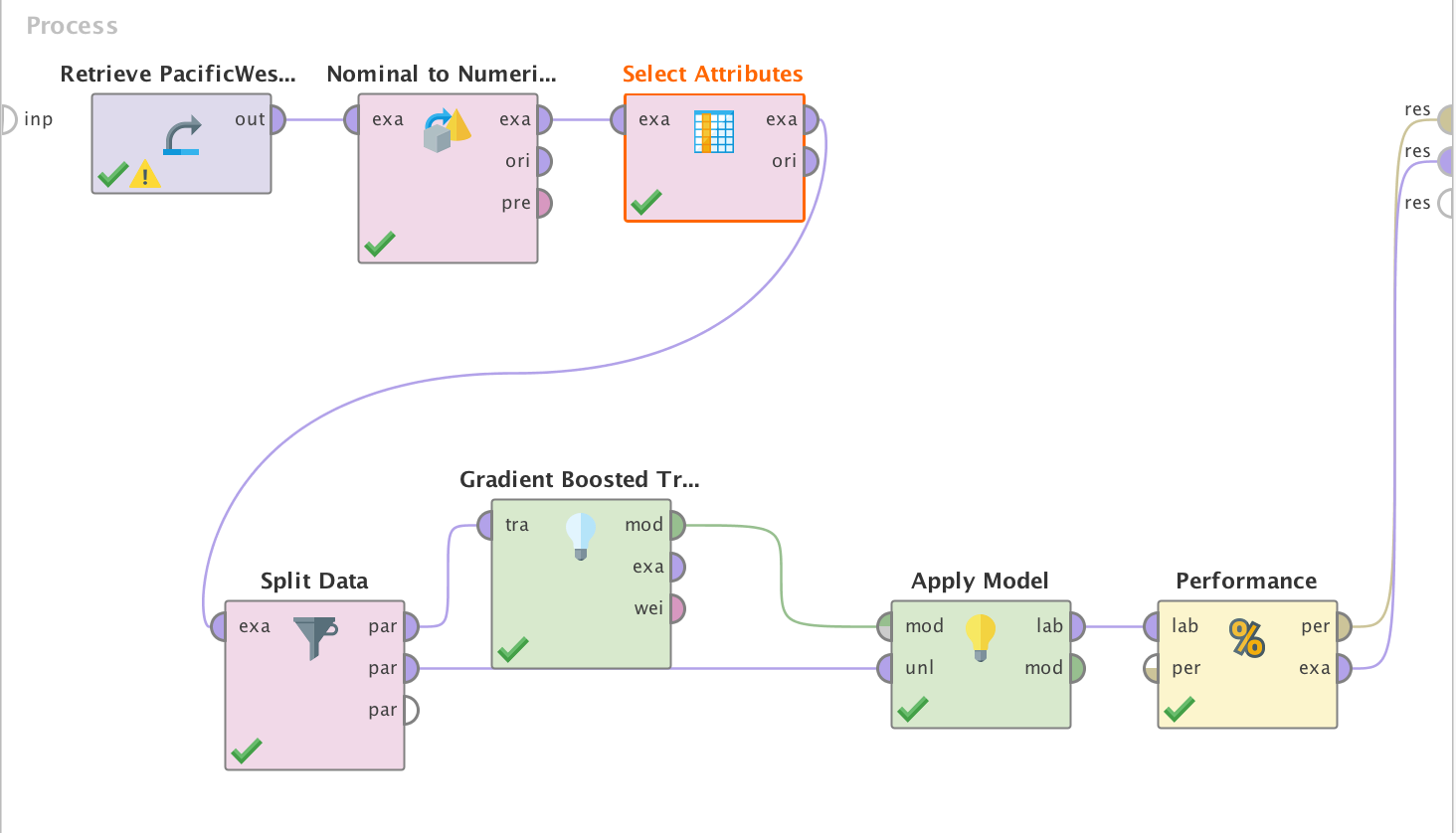


Figure 22

From the given model, we used correlation matrix in order to identify factors that drive Customer Lifetime value. The below screenshot displays the result from applying the regression and predicting the customer lifetime value.

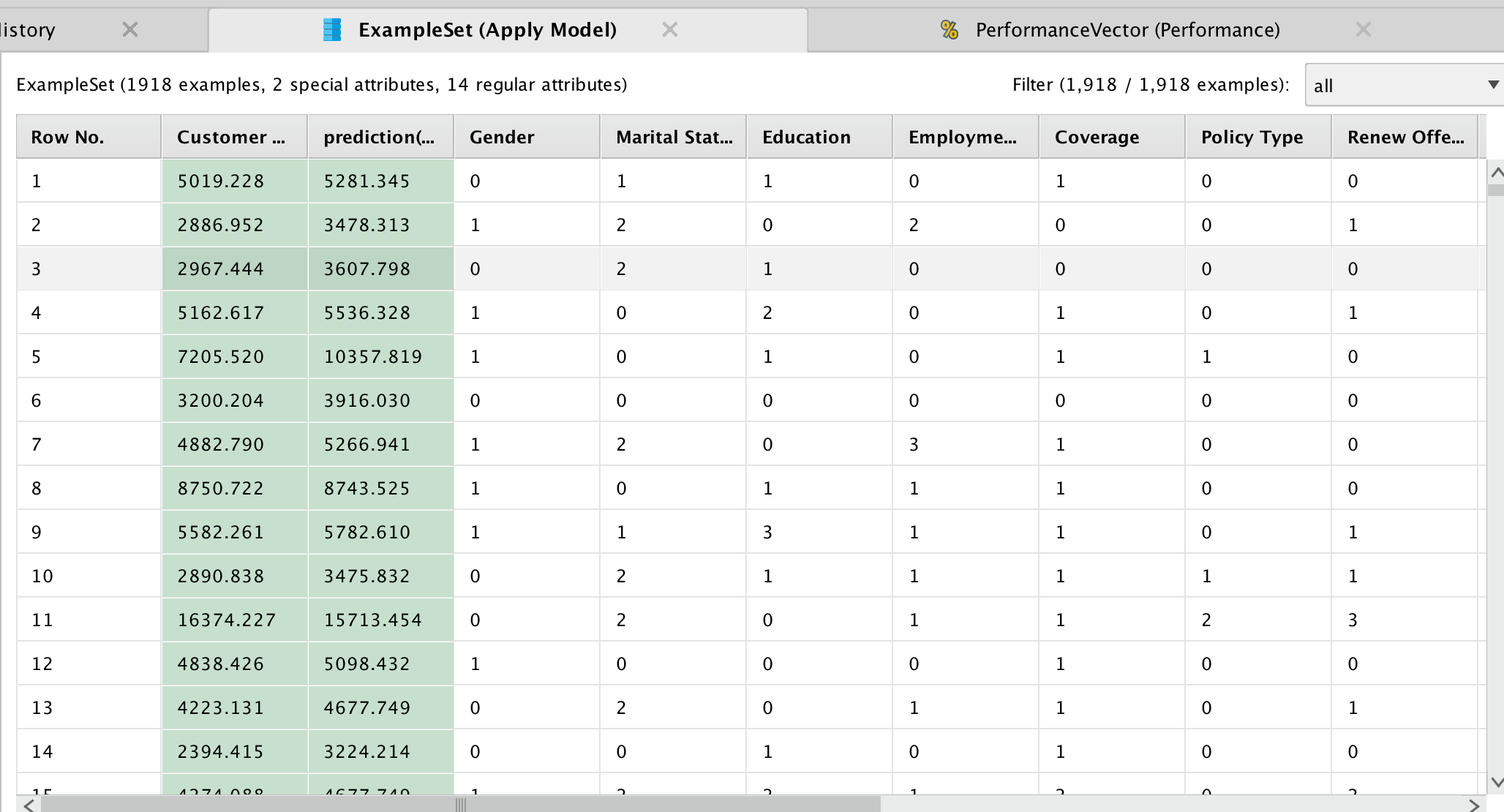


Figure 23

After running the Gradient Boosted model several times and removing variables that does not contribute to the Customer Lifetime value we achieved the root mean square error value of around **3774.720**

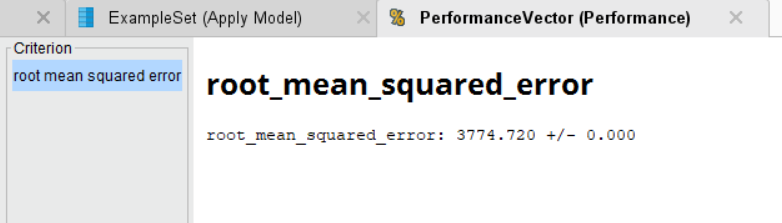


Figure 24

**Applying the model on scoring dataset:**

We have applied the gradient boosted tree on the pacific west scoring dataset.

The process used is as shown below:

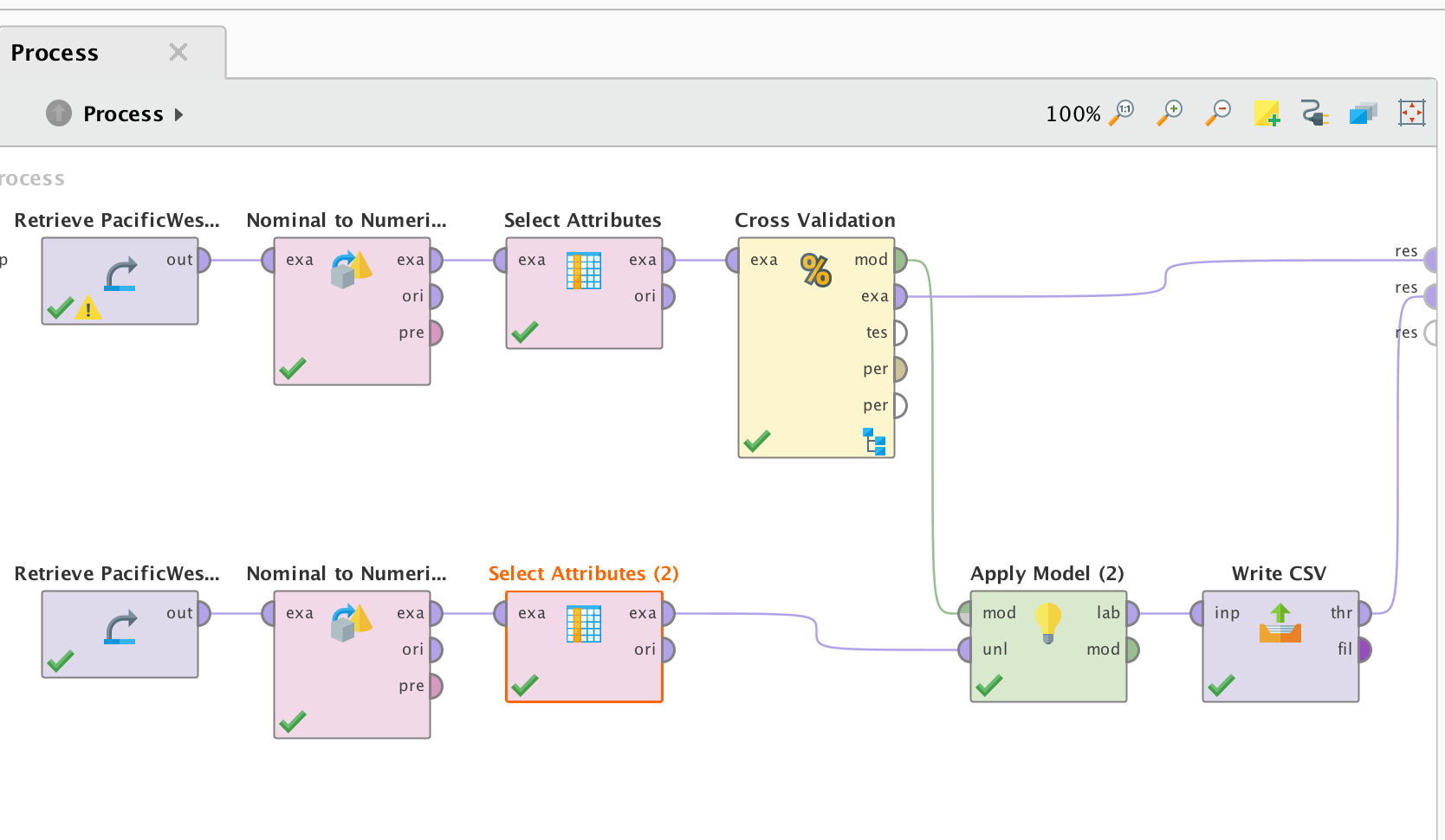


Figure 25

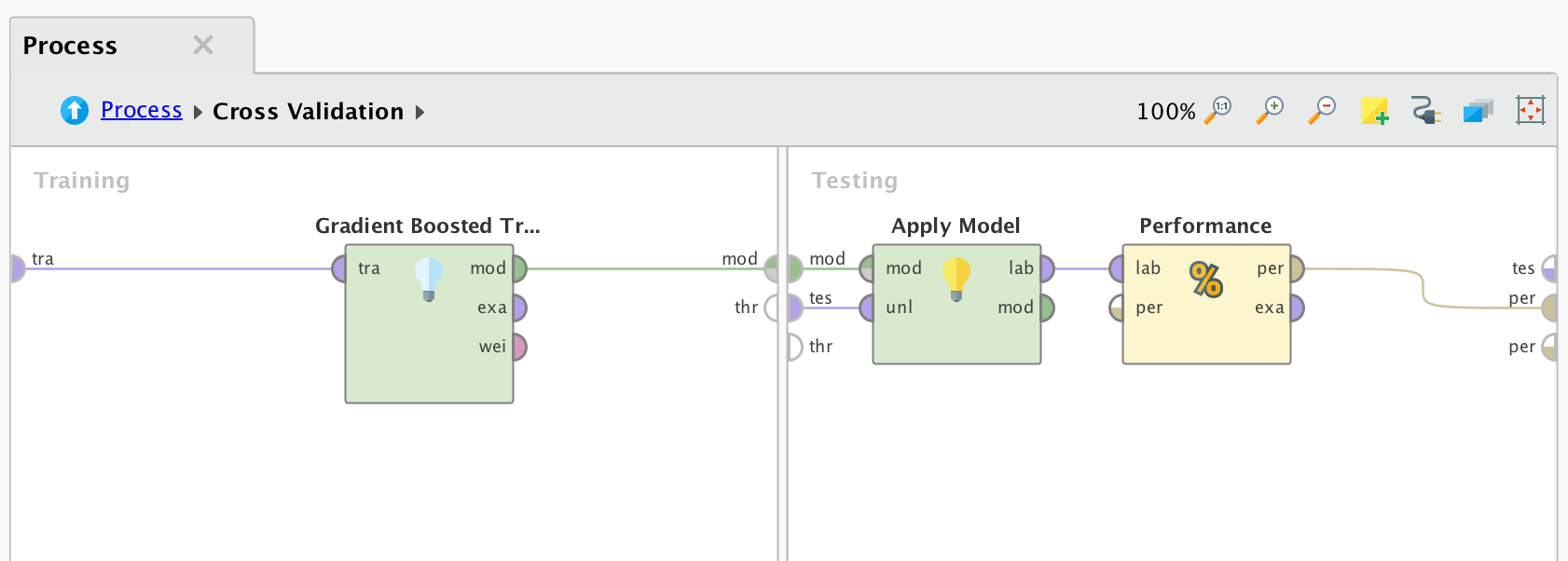


Figure 26

The predicted Customer Lifetime Value for the scoring dataset is stored in a csv and a screenshot of the result is shown below.

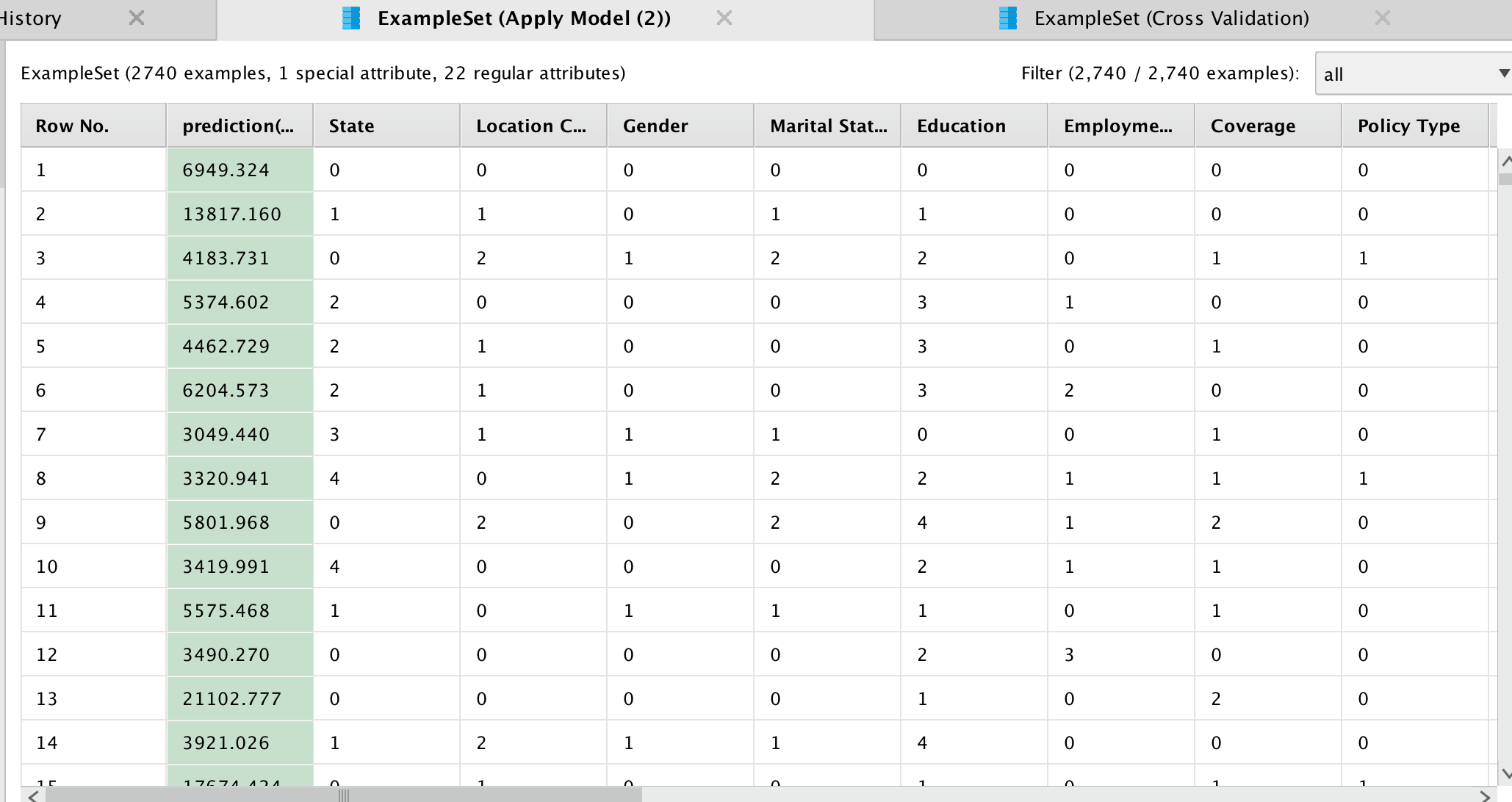


Figure 27



**Business Strategies:**

1. Based on our analysis, since Platinum and Gold Class customers already provide the firm with high customer lifetime value, it would be worthwhile to extend add-on benefits like access to a faster claim processing and increased coverage with their existing policies.
2. Additionally, for the Silver Class customers, PacWest has the potential to extend offerings by providing bundled policies at discounted rates and extending premium benefits when more than 2 policies are purchased
3. For the Basic Class customers, since the number of complains are on the higher side, PacWest can provide extended customer support to remediate complains faster to increase customer satisfaction and will result in higher customer retention.
4. The target customers should be those who possess mid-sized cars in Oregon and California regions.
5. Also, since Offer 3 and 4 are not very popular amongst customers, the company should either restructure the benefits provided or decide to discontinue them.