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 Pacific West 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.

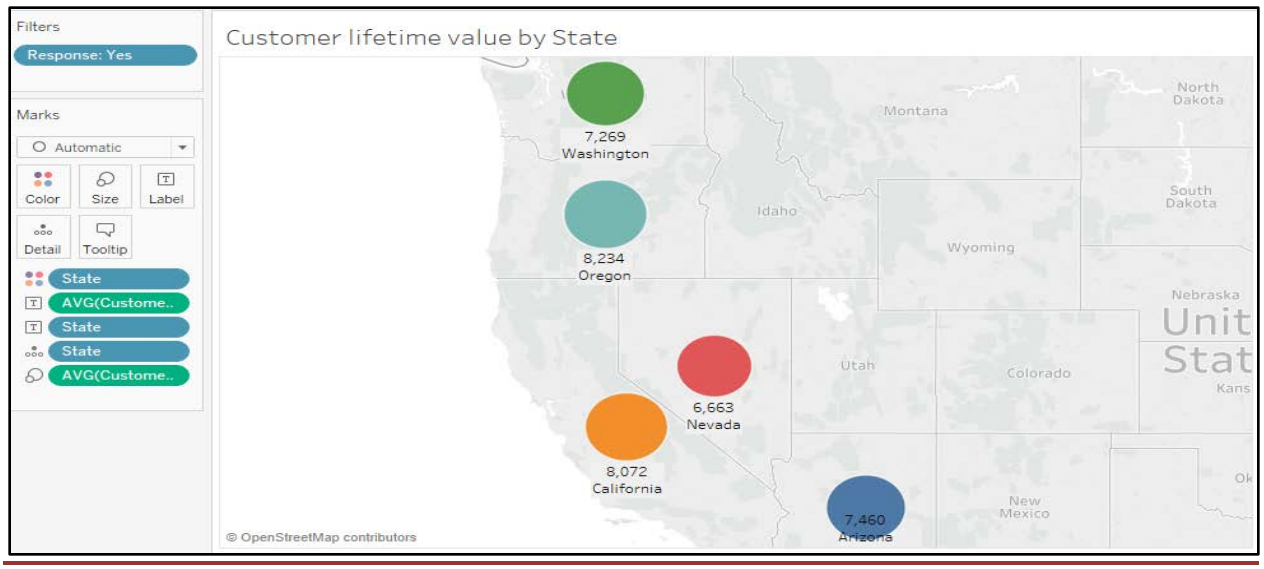


Figure 1

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.

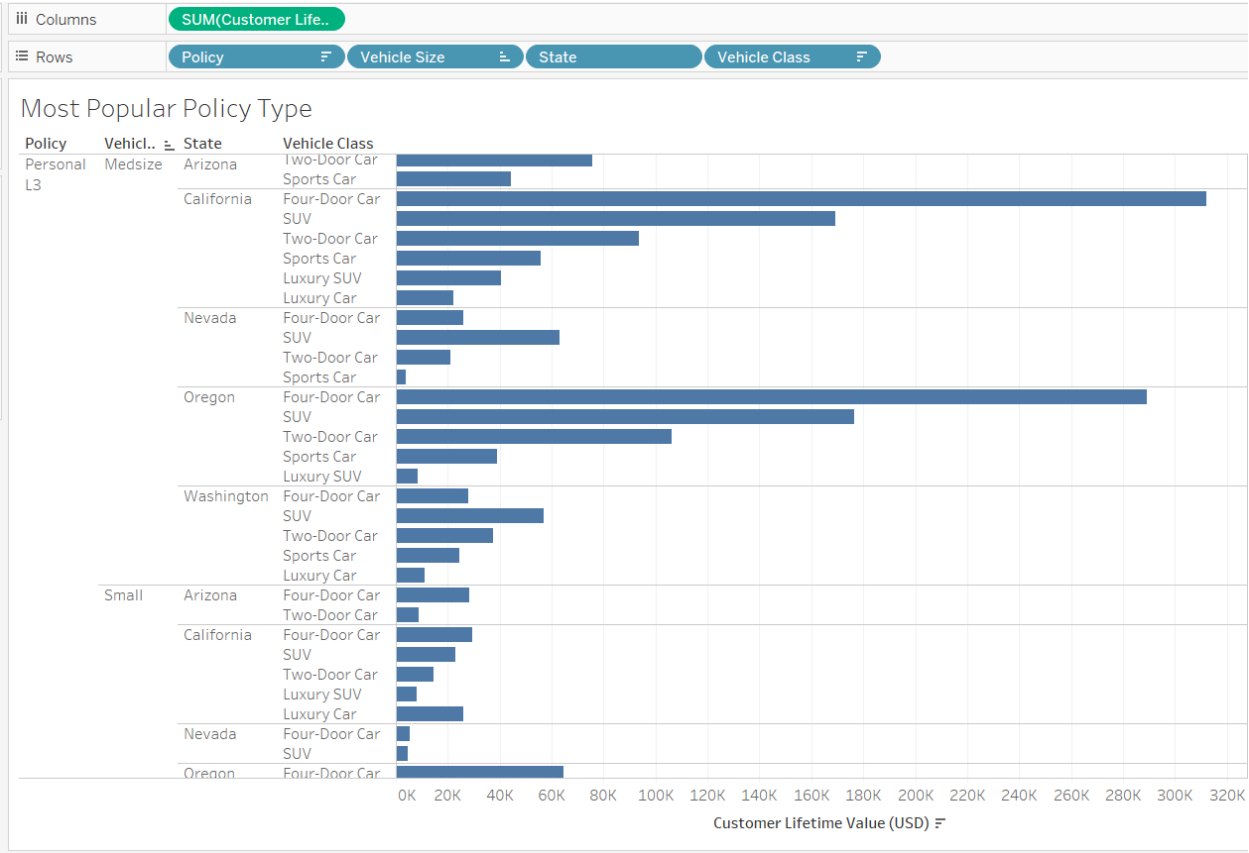


Figure 3

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.

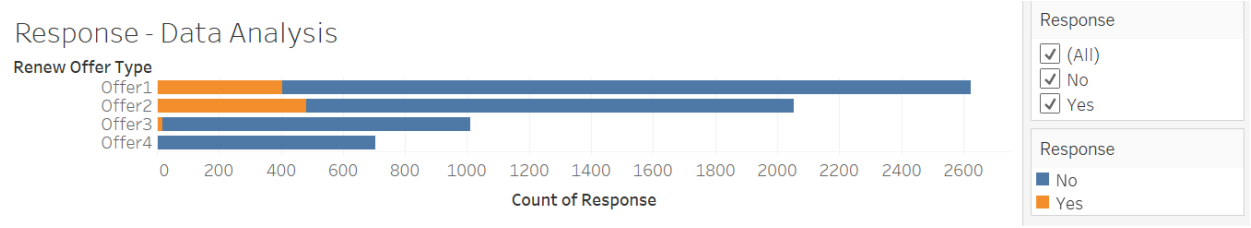


Figure 4

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.

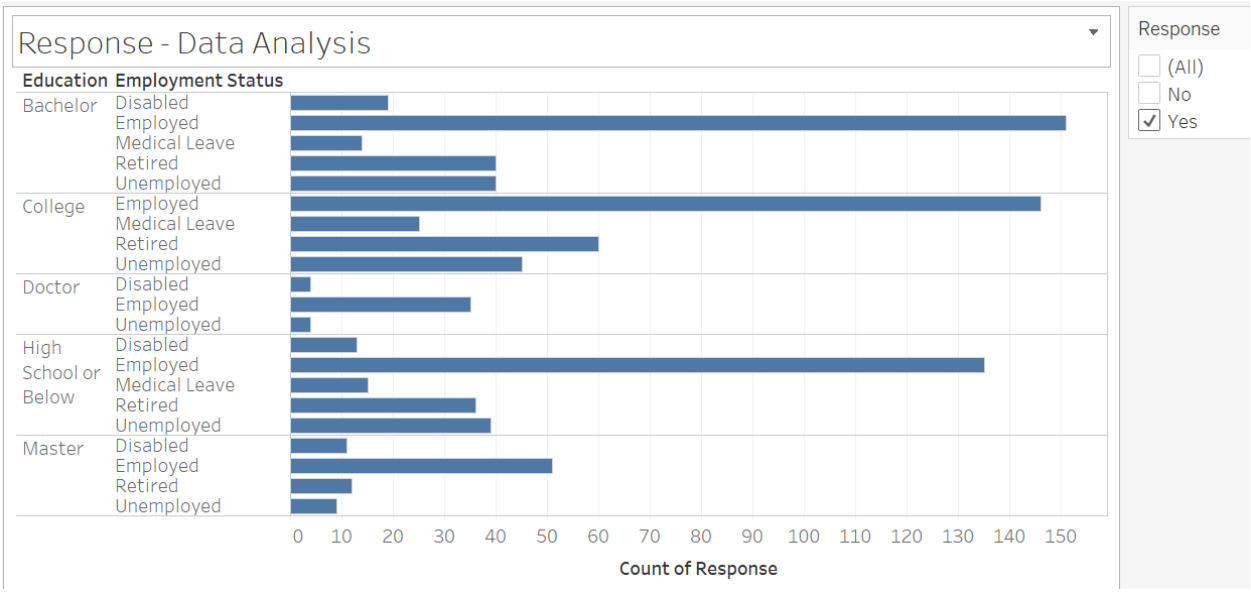


Figure 5

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

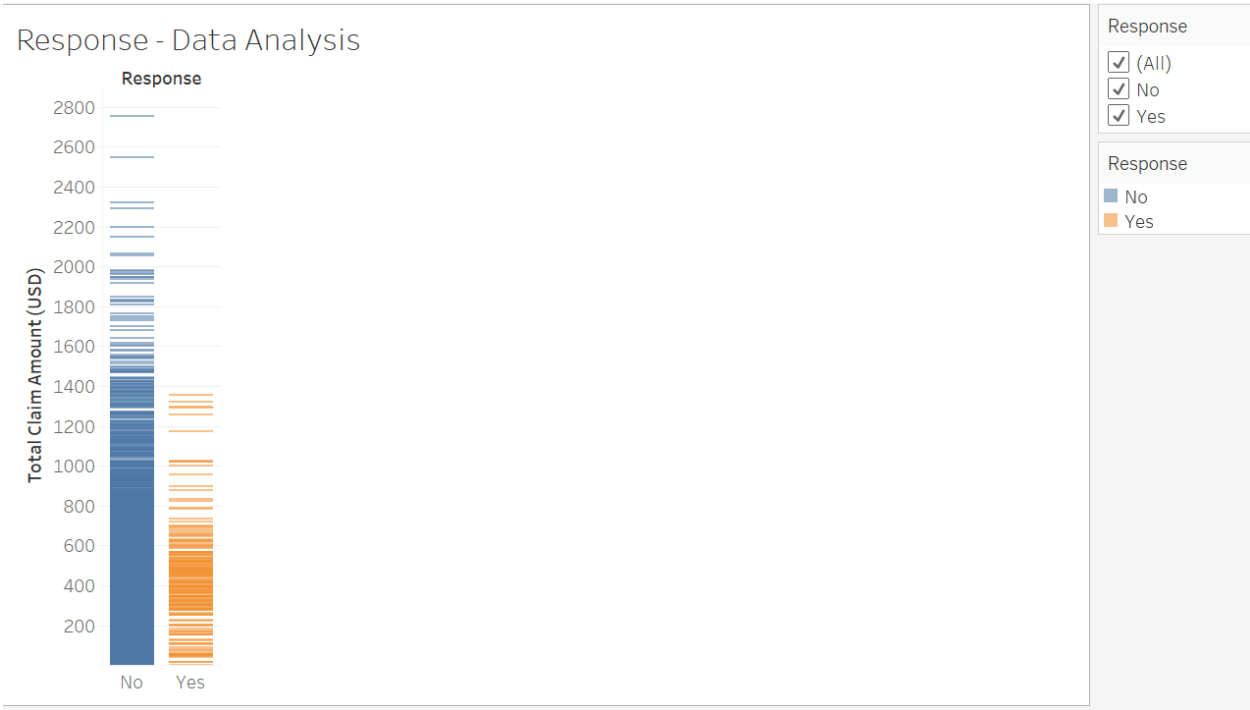


Figure 6

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.

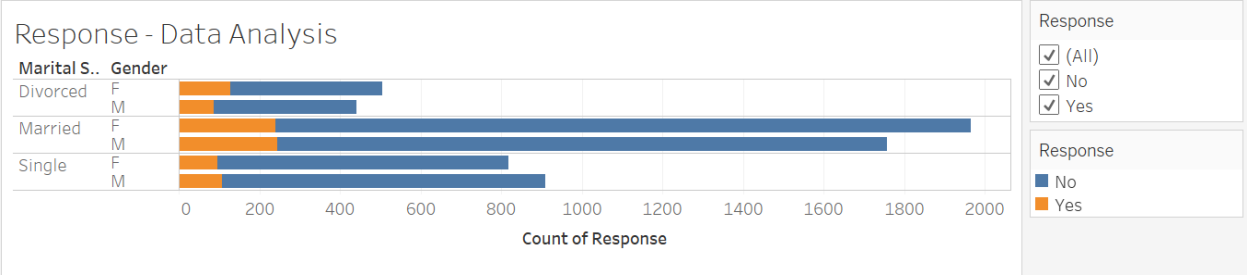


Figure 7

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:

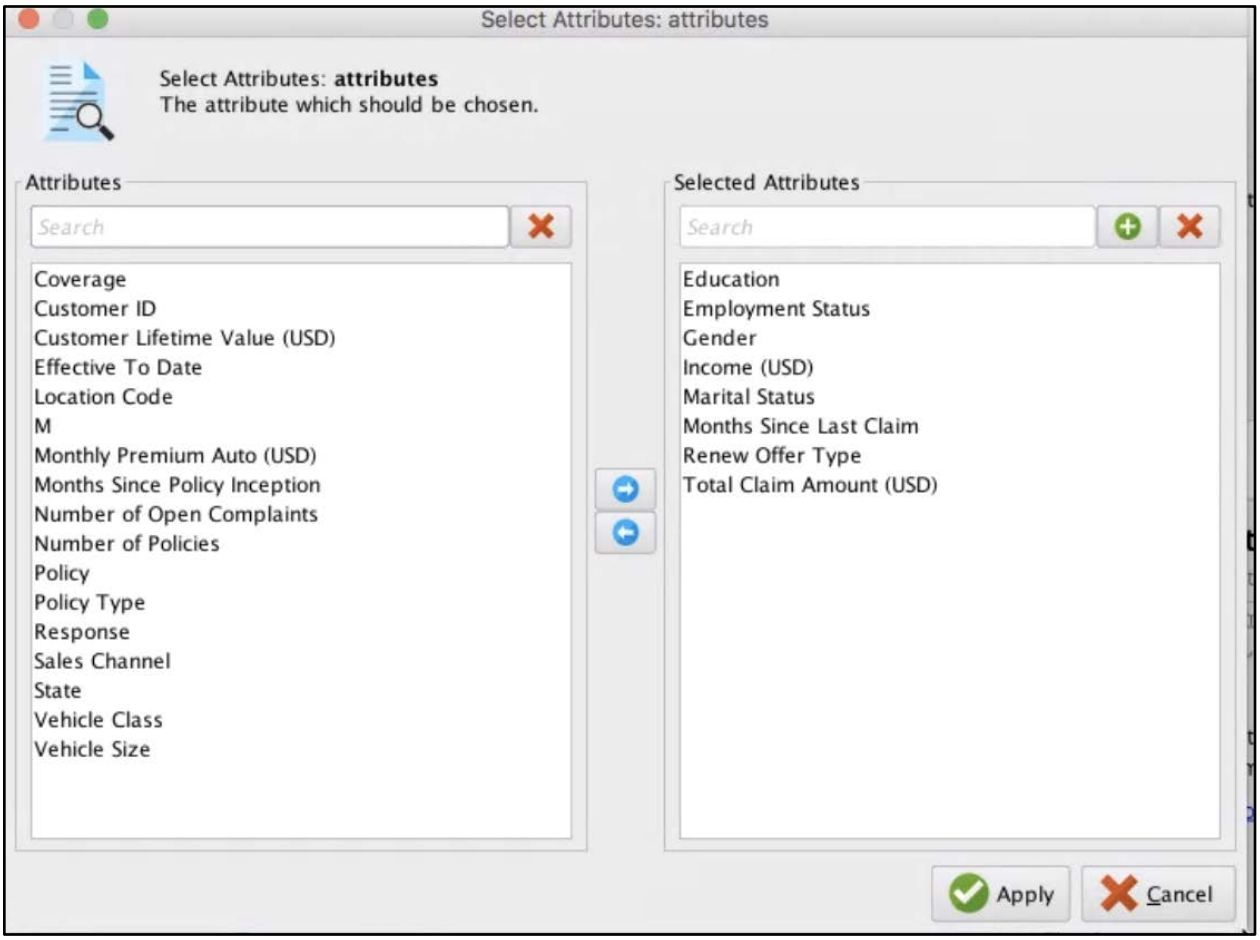


Figure 8

Comparing multiple Models:

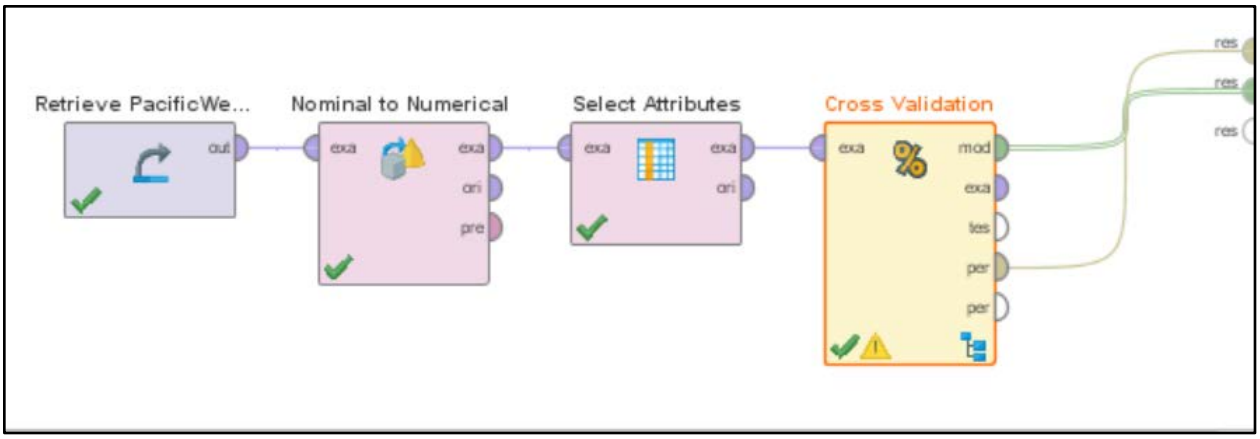


Figure 9

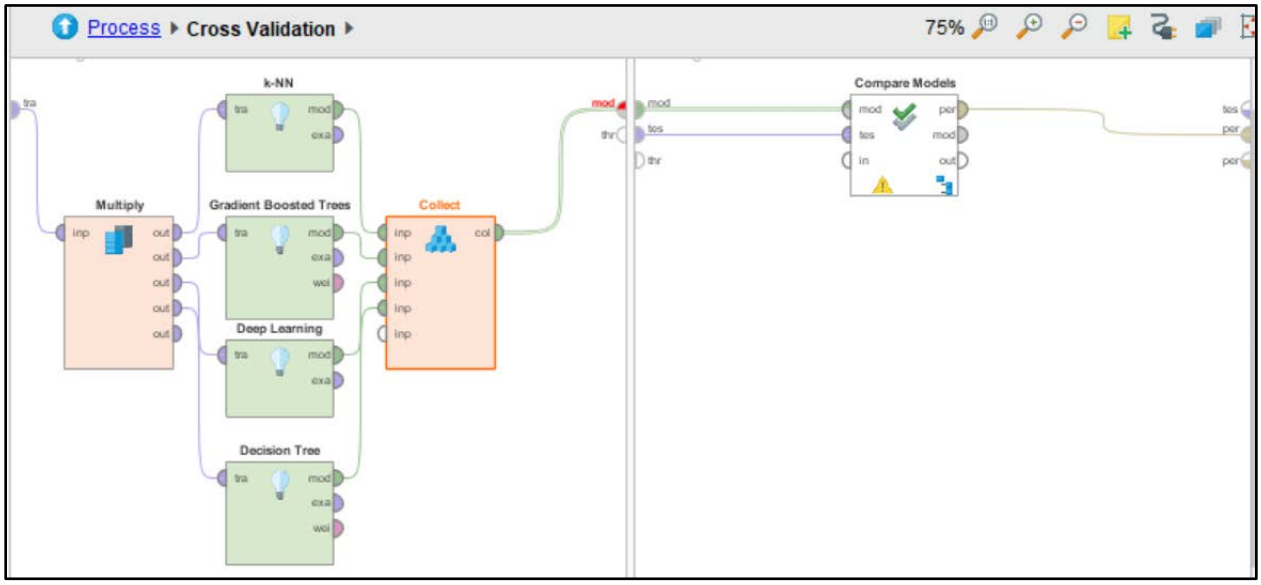


Figure 10

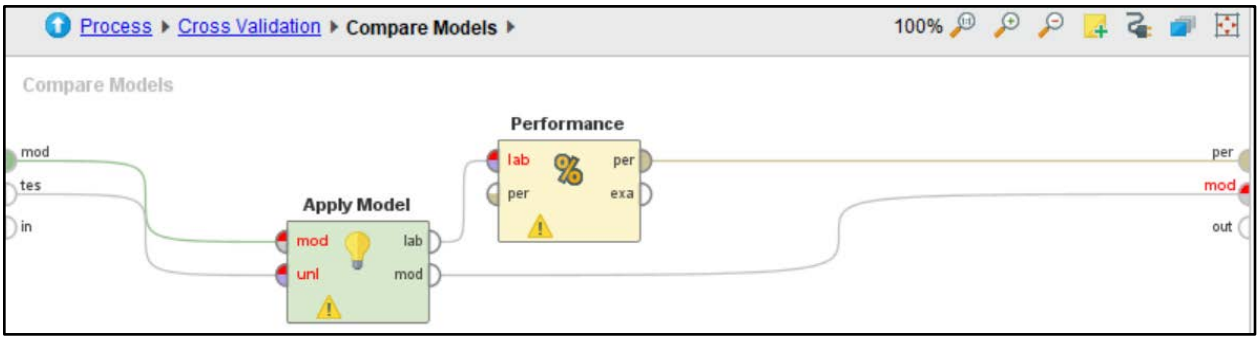


Figure 11

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:

ExampleSet (640 examples, 0 special attributes, 7 regular attributes) Filter (640 / 640 examples):

Row No.	Model Name	Date of testing	Location of Model	Criterion	Value	Standard D...	Variance
1	3_Decision Tree	May 11, 2018 9:31:37 PM EDT	//Local Repository/...	accuracy	0.859	?	?
2	3_Decision Tree	May 11, 2018 9:31:37 PM EDT	//Local Repository/...	precision	?	?	?
3	3_Decision Tree	May 11, 2018 9:31:37 PM EDT	//Local Repository/...	recall	0	?	?
4	3_Decision Tree	May 11, 2018 9:31:37 PM EDT	//Local Repository/...	f_measure	?	?	?
5	2_Deep Learning	May 11, 2018 9:31:37 PM EDT	//Local Repository/...	accuracy	0.789	?	?
6	2_Deep Learning	May 11, 2018 9:31:37 PM EDT	//Local Repository/...	precision	0.357	?	?
7	2_Deep Learning	May 11, 2018 9:31:37 PM EDT	//Local Repository/...	recall	0.622	?	?
8	2_Deep Learning	May 11, 2018 9:31:37 PM EDT	//Local Repository/...	f_measure	0.453	?	?
9	1_Gradient Boosted Trees	May 11, 2018 9:31:37 PM EDT	//Local Repository/...	accuracy	0.842	?	?
10	1_Gradient Boosted Trees	May 11, 2018 9:31:37 PM EDT	//Local Repository/...	precision	0.434	?	?
11	1_Gradient Boosted Trees	May 11, 2018 9:31:37 PM EDT	//Local Repository/...	recall	0.400	?	?
12	1_Gradient Boosted Trees	May 11, 2018 9:31:37 PM EDT	//Local Repository/...	f_measure	0.416	?	?
13	0_k-NN	May 11, 2018 9:31:37 PM EDT	//Local Repository/...	accuracy	0.975	?	?
14	0_k-NN	May 11, 2018 9:31:37 PM EDT	//Local Repository/...	precision	0.849	?	?
15	0_k-NN	May 11, 2018 9:31:37 PM EDT	//Local Repository/...	recall	1	?	?
16	0_k-NN	May 11, 2018 9:31:37 PM EDT	//Local Repository/...	f_measure	0.918	?	?

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.

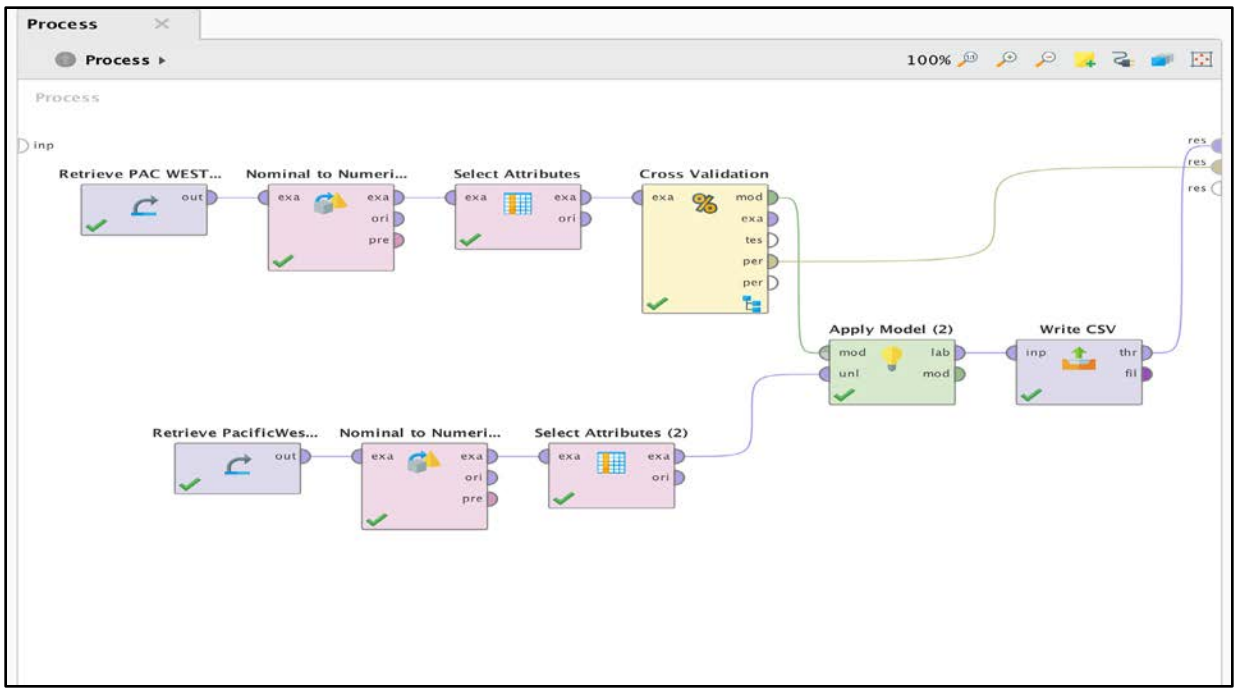


Figure 13

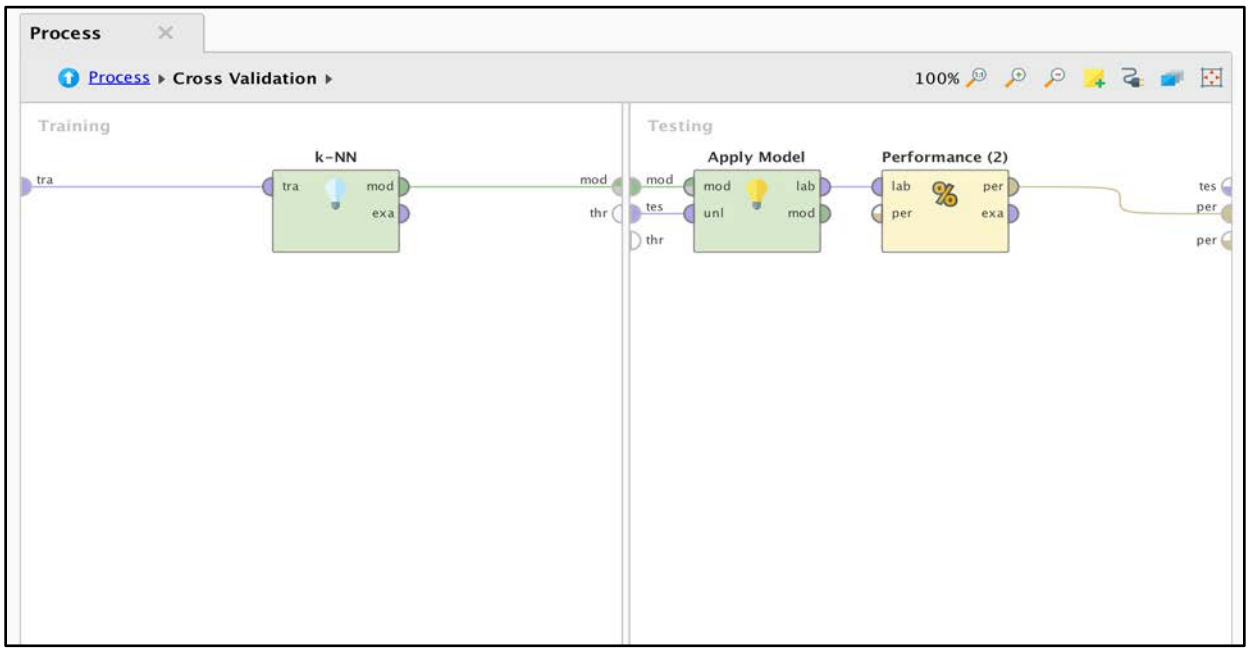


Figure 14

Below is the output of the task I: Classification

PACWEST-KNN Task 1 Submission - Grou

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.

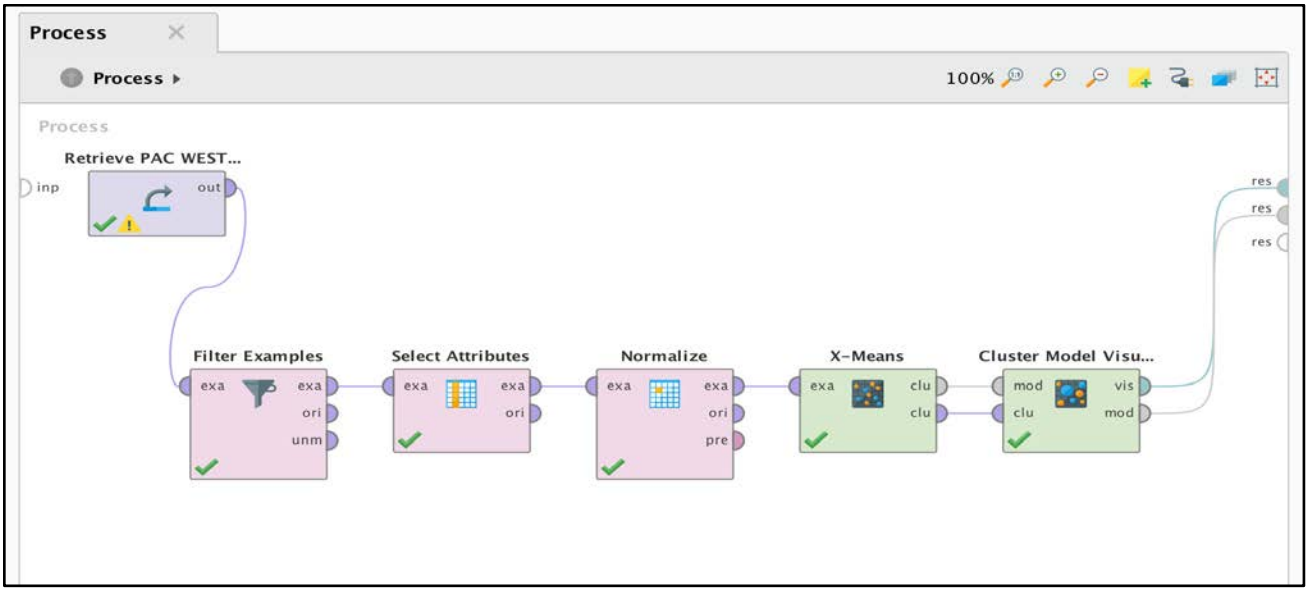


Figure 15

Step II: Customer Segmentation using X-Means

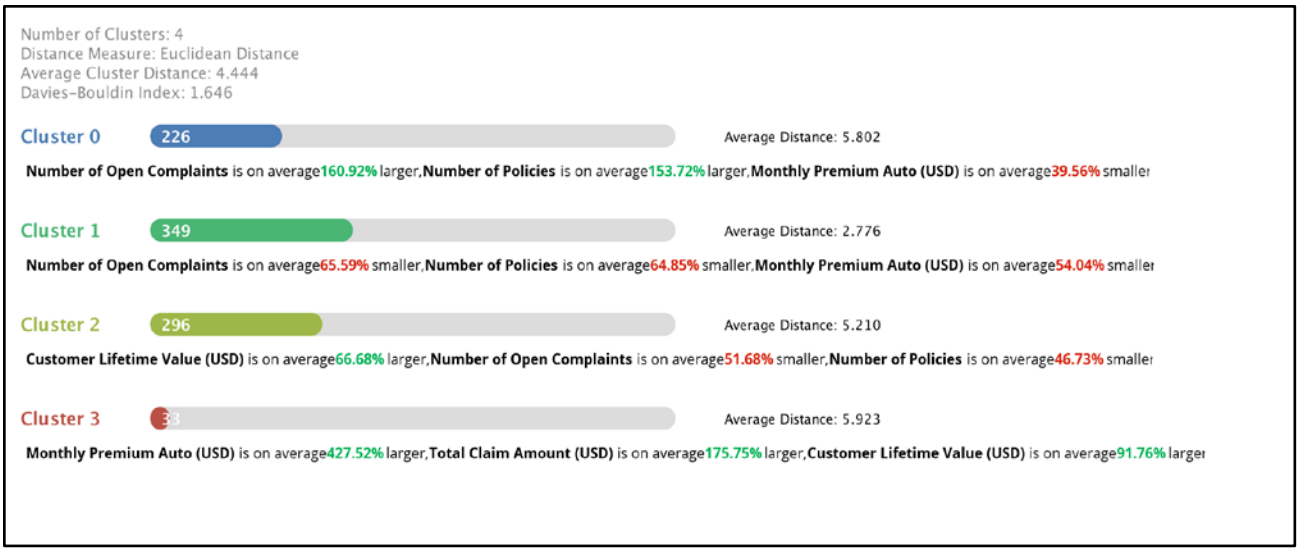


Figure 16

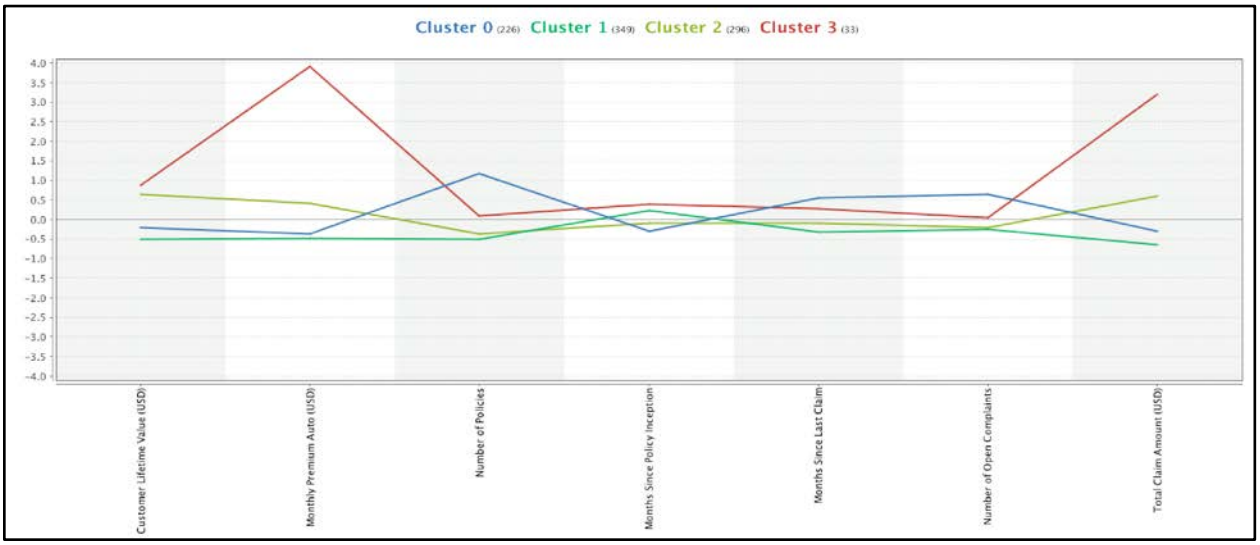


Figure 17

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.

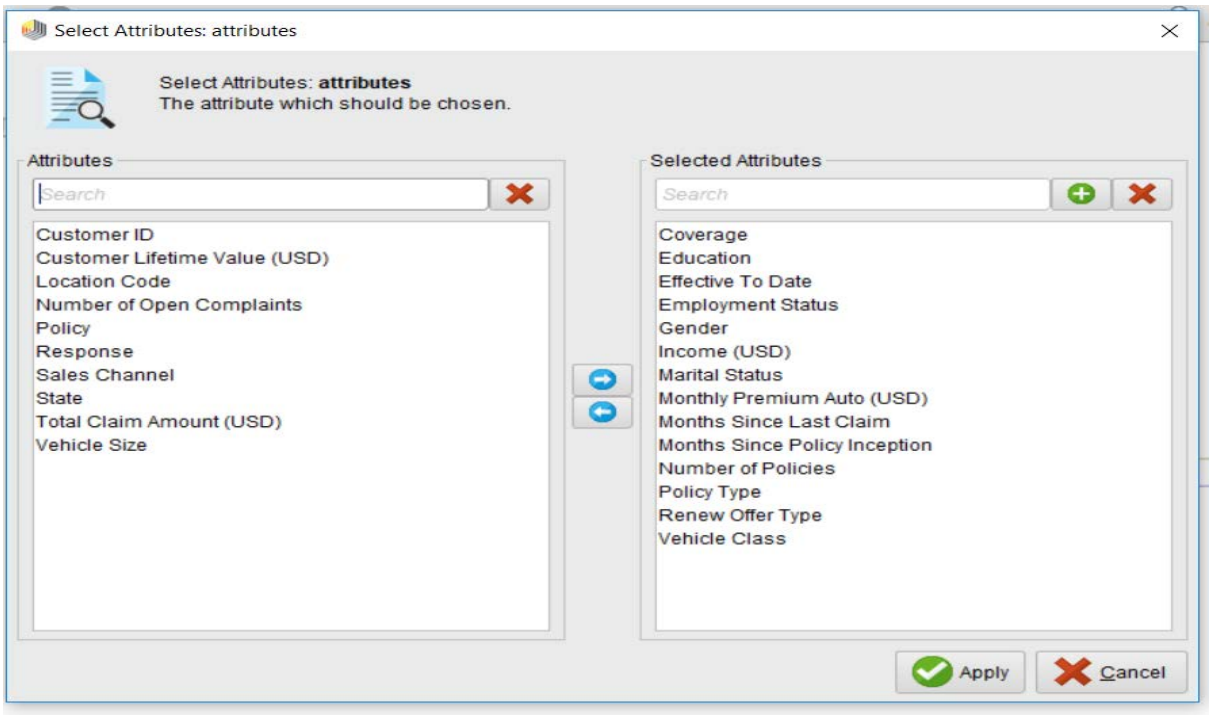


Figure 18

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

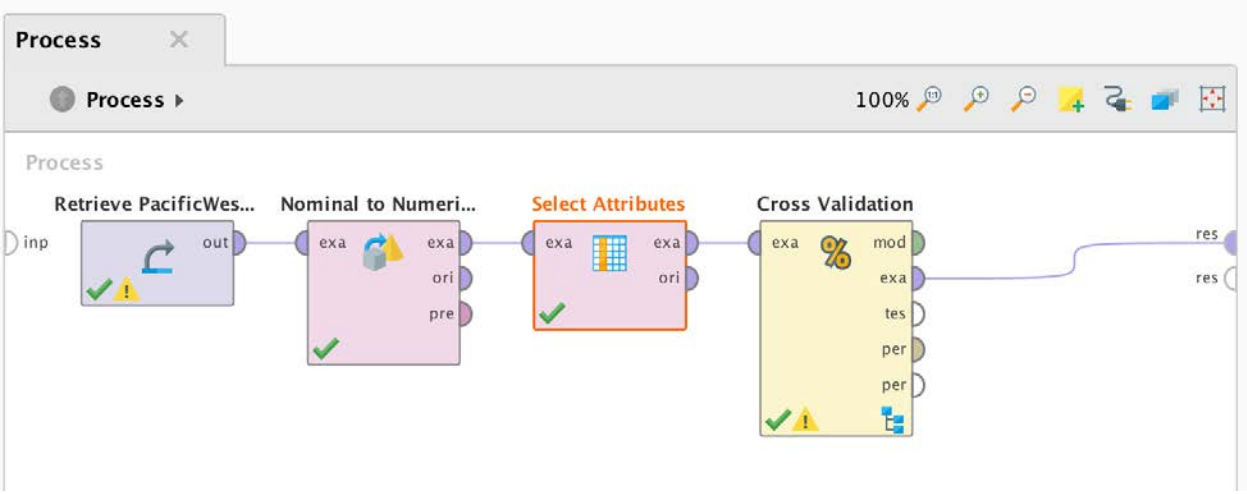


Figure 19

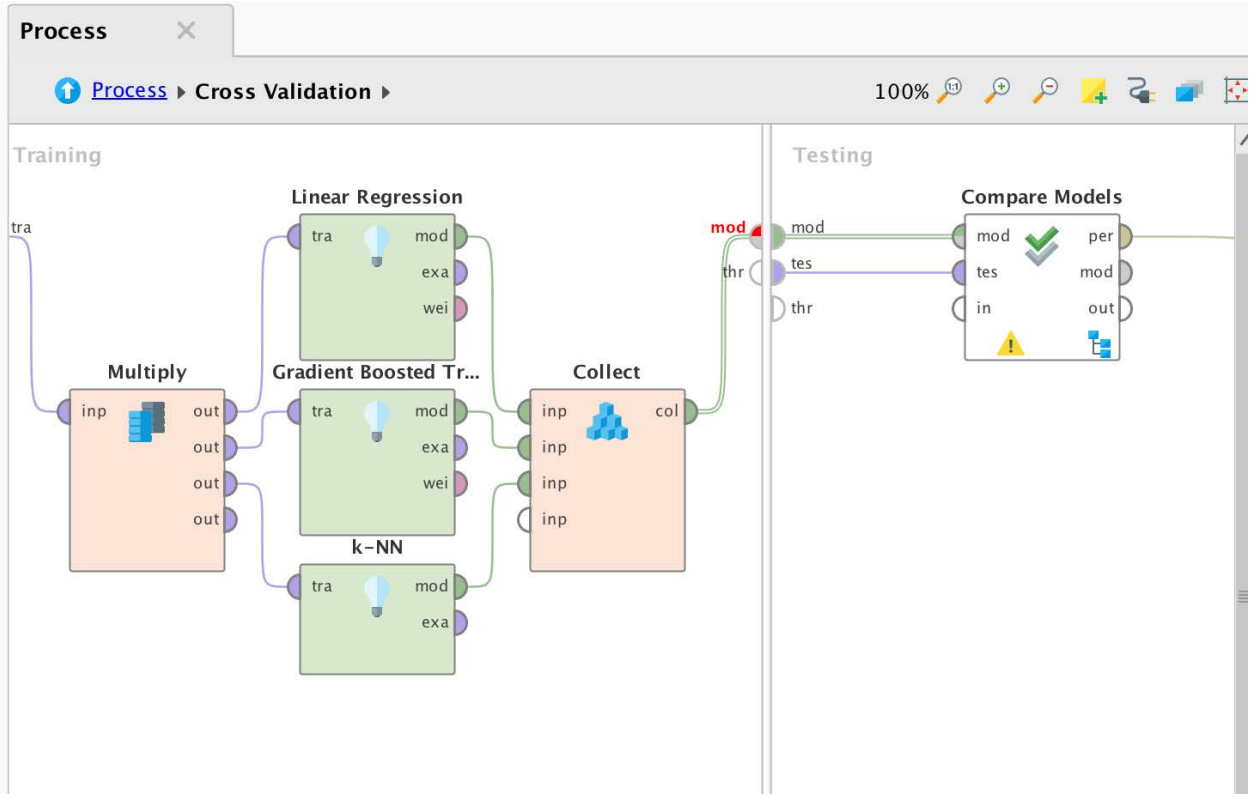


Figure 20

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

ExampleSet (213 examples, 0 special attributes, 7 regular attributes) Filter (213 / 213 examples): all

Row No.	Model Name	Date of testing	Location of...	Criterion	Value	Standard D...	Variance
1	2_k-NN	May 13, 2018 1:52:14 AM EDT	//Project_Pa...	root_mean_...	8509.312	0	-63145447...
2	1_Gradient Boosted Tr...	May 13, 2018 1:52:14 AM EDT	//Project_Pa...	root_mean_...	3577.581	0	-10692474...
3	0_Linear Regression	May 13, 2018 1:52:14 AM EDT	//Project_Pa...	root_mean_...	5937.197	0	-30874746...
4	2_k-NN	May 13, 2018 1:52:16 AM EDT	//Project_Pa...	root_mean_...	9158.842	0	-74386627...
5	1_Gradient Boosted Tr...	May 13, 2018 1:52:16 AM EDT	//Project_Pa...	root_mean_...	3959.557	0	-13094138...
6	0_Linear Regression	May 13, 2018 1:52:16 AM EDT	//Project_Pa...	root_mean_...	6591.819	0	-37817728...

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:

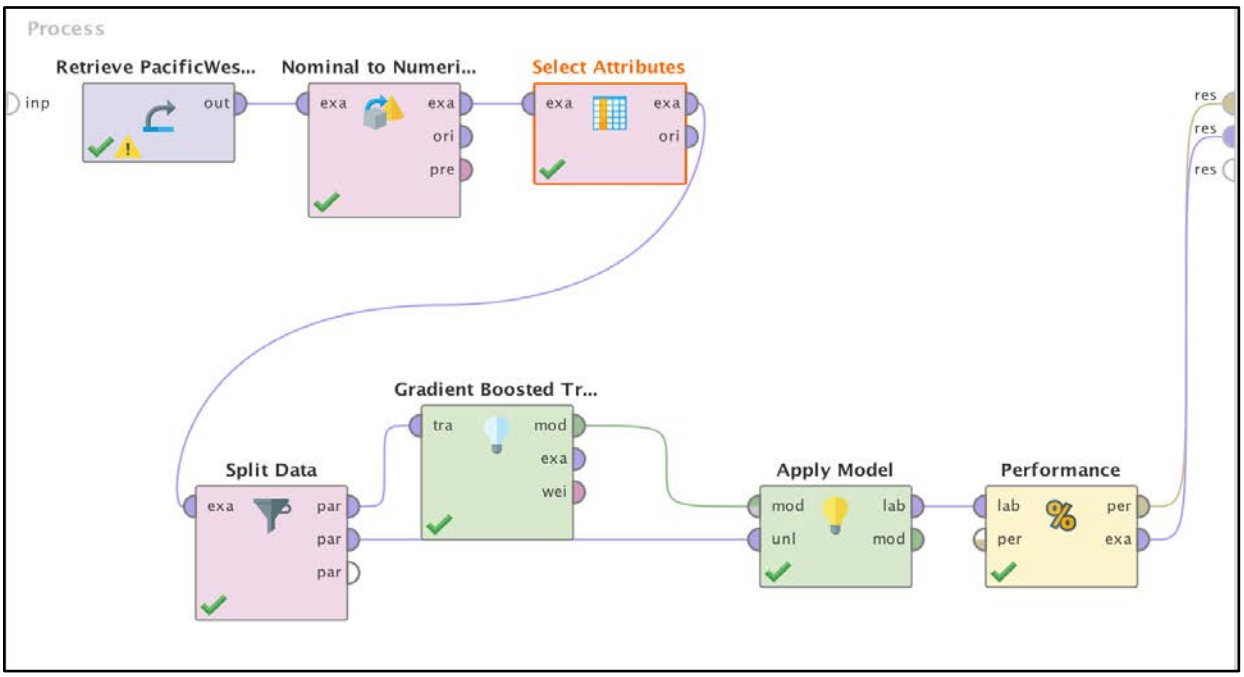
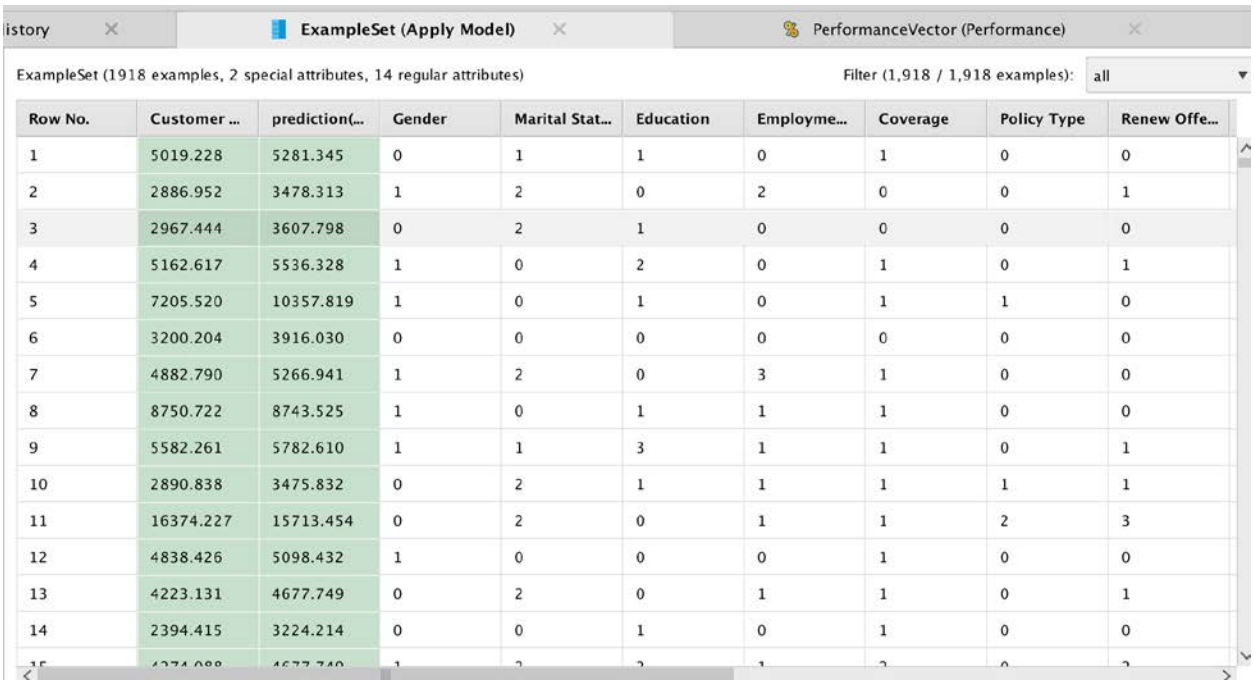


Figure 22

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.



istory | ExampleSet (Apply Model) | PerformanceVector (Performance)

ExampleSet (1918 examples, 2 special attributes, 14 regular attributes) Filter (1,918 / 1,918 examples): all

Row No.	Customer ...	prediction(...	Gender	Marital Stat...	Education	Employme...	Coverage	Policy Type	Renew Offe...
1	5019.228	5281.345	0	1	1	0	1	0	0
2	2886.952	3478.313	1	2	0	2	0	0	1
3	2967.444	3607.798	0	2	1	0	0	0	0
4	5162.617	5536.328	1	0	2	0	1	0	1
5	7205.520	10357.819	1	0	1	0	1	1	0
6	3200.204	3916.030	0	0	0	0	0	0	0
7	4882.790	5266.941	1	2	0	3	1	0	0
8	8750.722	8743.525	1	0	1	1	1	0	0
9	5582.261	5782.610	1	1	3	1	1	0	1
10	2890.838	3475.832	0	2	1	1	1	1	1
11	16374.227	15713.454	0	2	0	1	1	2	3
12	4838.426	5098.432	1	0	0	0	1	0	0
13	4223.131	4677.749	0	2	0	1	1	0	1
14	2394.415	3224.214	0	0	1	0	1	0	0

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:

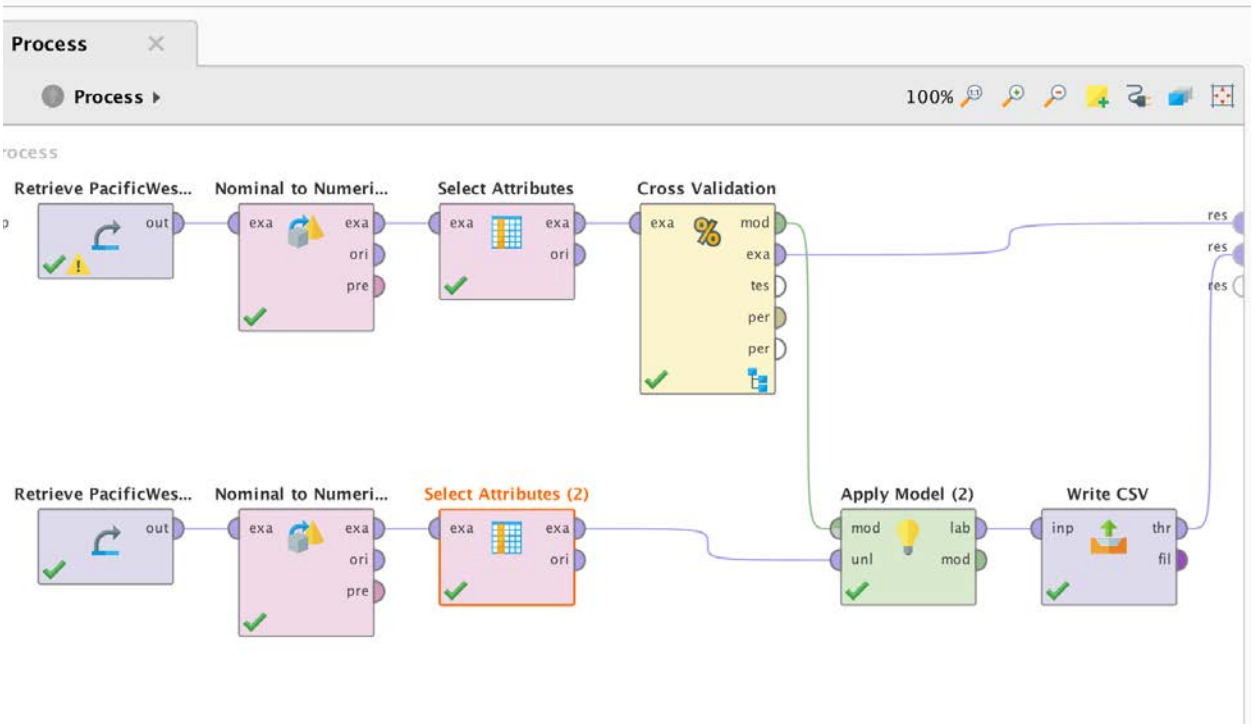


Figure 25

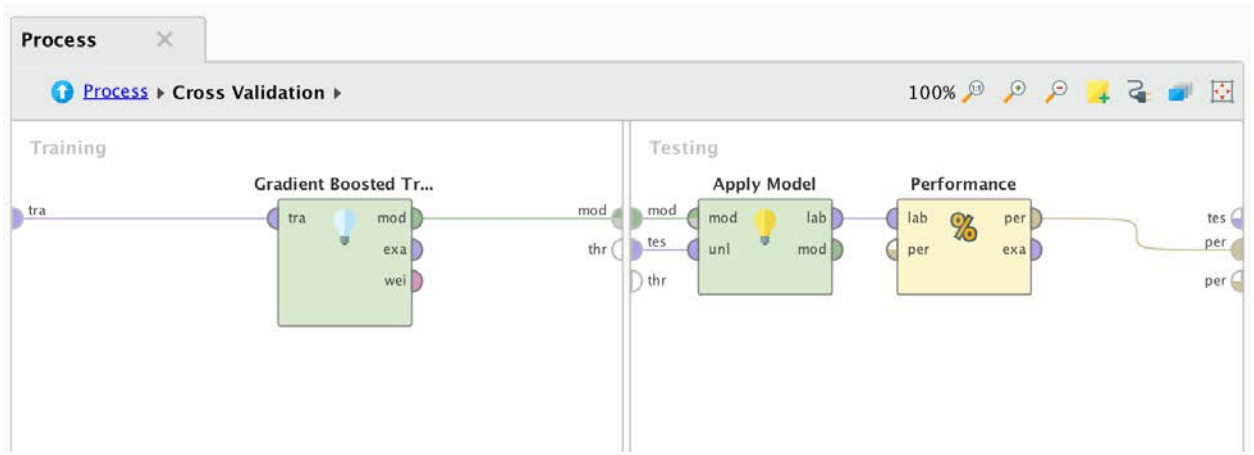
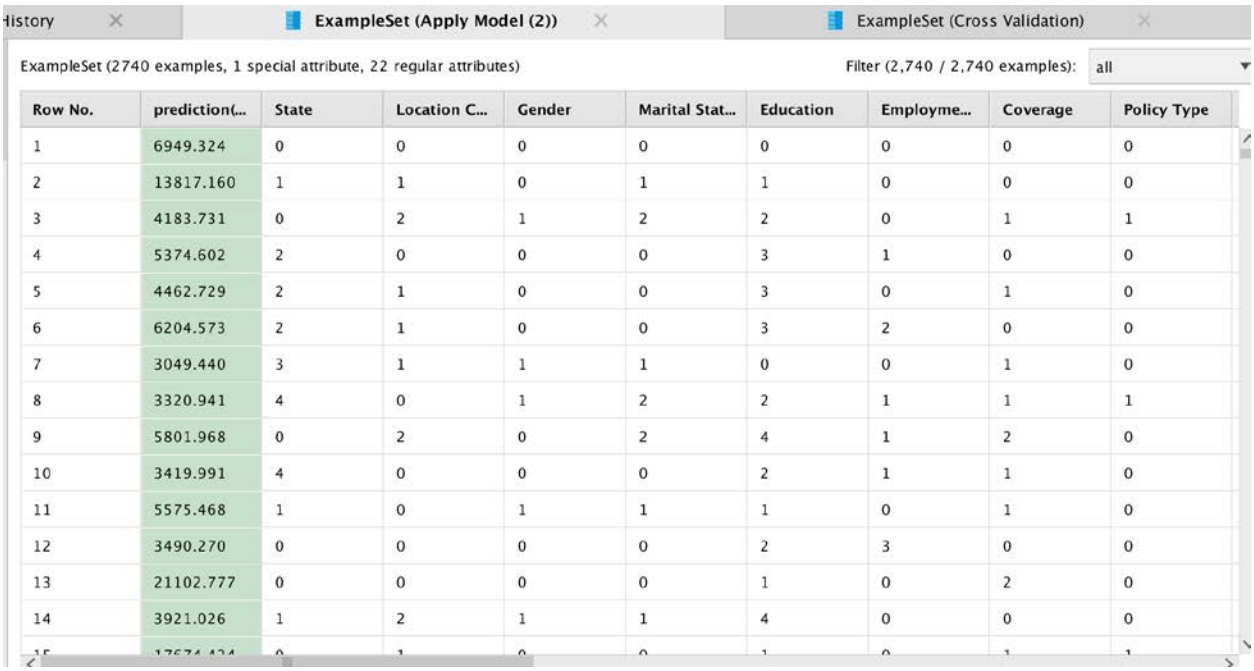


Figure 26

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



History | ExampleSet (Apply Model (2)) | ExampleSet (Cross Validation)

ExampleSet (2740 examples, 1 special attribute, 22 regular attributes) Filter (2,740 / 2,740 examples): all

Row No.	prediction(...	State	Location C...	Gender	Marital Stat...	Education	Employme...	Coverage	Policy Type
1	6949.324	0	0	0	0	0	0	0	0
2	13817.160	1	1	0	1	1	0	0	0
3	4183.731	0	2	1	2	2	0	1	1
4	5374.602	2	0	0	0	3	1	0	0
5	4462.729	2	1	0	0	3	0	1	0
6	6204.573	2	1	0	0	3	2	0	0
7	3049.440	3	1	1	1	0	0	1	0
8	3320.941	4	0	1	2	2	1	1	1
9	5801.968	0	2	0	2	4	1	2	0
10	3419.991	4	0	0	0	2	1	1	0
11	5575.468	1	0	1	1	1	0	1	0
12	3490.270	0	0	0	0	2	3	0	0
13	21102.777	0	0	0	0	1	0	2	0
14	3921.026	1	2	1	1	4	0	0	0

Figure 27

regression._updated.csv

Business Strategies:

I. 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.

II. 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

III. 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.

IV. The target customers should be those who possess mid-sized cars in Oregon and California regions.

V. 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.