|  |
| --- |
| CS 699 Project Report |
| [Telecom users dataset] |

|  |
| --- |
| Summer 2021 | By: Yung-Chieh Yang |

****

**Statement of Data Mining Goal**

Analyzing which Classification Algorithms and Attributes Selection Methods can predict the best user churn.

**Detailed Description of the Dataset**

This dataset has **22 attributes** and **5986 tuples**.

1. ID - ID of the tuple
2. customerID – customer id
3. gender – client gender (male/female)
4. SeniorCitizen – is the client retired (Yes, No)
5. Partner - is the client married (Yes, No)
6. Dependents – is the client has dependents (Yes, No)
7. tenure - how many months a person has been a client of the company
8. PhoneService - is the telephone service connected (Yes, No)
9. MultipleLines - are multiple phone lines connected (Yes, No, No phone service)
10. InternetService - client's Internet service provider (DSL, Fiber optic, No)
11. OnlineSecurity - is the online security service connected (Yes, No, No internet service)
12. OnlineBackup - is the online backup service activated (Yes, No, No internet service)
13. DeviceProtection - does the client have equipment insurance (Yes, No, No internet service)
14. TechSupport - is the technical support service connected (Yes, No, No internet service)
15. StreamingTV - is the streaming TV service connected (Yes, No, No internet service)
16. StreamingMovies - is the streaming cinema service activated (Yes, No, No internet service)
17. Contract - type of customer contract (Month-to-month, One year, Two year)
18. PaperlessBilling - whether the client uses paperless billing (Yes, No)
19. PaymentMethod - payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))
20. MonthlyCharges - current monthly payment
21. TotalCharges - the total amount that the client paid for the services for the entire time
22. Churn - whether there was a churn (Yes or No)

Any business wants to maximize the number of customers. In order to achieve this goal, it is important not only to try to attract newcomers, but also to retain existing ones. Compared with attracting new customers, the cost of retaining customers will reduce the company's costs. In addition, new customers may have weak interest in business services and it is difficult to work with them, while old customers already have the data needed to interact with the service.

Therefore, by predicting churn, we can react in a timely manner and try to keep customers who want to leave. Based on the service data used by the customer, we can offer him a special discount in an attempt to change his decision to leave the operator. This will make the retention task easier to implement than the task of attracting new users, and we don't know anything about it.

Provide you with data sets from telecommunications companies. This data contains information about nearly 6,000 users, their demographic characteristics, the services they use, the length of time they use the service of the operator, the payment method and the payment amount.

The task is to analyze the data and predict user churn (to determine who will and will not renew the contract).

**Description of Data Mining Tools**

Weka and JMP Pro

Test classification model using 10-fold cross-validation in Weka; Five attribution selection methods in Weka.

Build classification models in JMP Pro.

**Description of Classification Algorithms**

1. Naïve Bayes Classification

2. K-nearest Neighbors Classification

3. Neural Network Classification

4. Partition Model Classification (Decision Tree)

5. Boosted Tree Classification

**Description of Attribute Selection Methods**

1. BestFirst + CfsSubsetEval

2. Ranker + CorrelationAttributeEval

3. Ranker + GainRatioAttributeEval

4. Ranker + InfoGainAttributeEval

5. Ranker + ReliefAttributeEval

**Set of Attributes Selected by Each Attribute Selection Method**

1. tenure, OnlineSecurity, TechSupport, Contract, PaymendMethod (BestFirst + CfsSubsetEval)

2. tenure, Contract, OnlineSecurity, TechSupport, InternetService (Ranker + CorrelationAttributeEval)

3. Contract, OnlineSecurity, TechSupport, InternetService, OnlineBackup (Ranker + GainRatioAttributeEval)

4. Contract, tenure, OnlineSecurity, TechSupport, InternetService (Ranker + InfoGainAttributeEval)

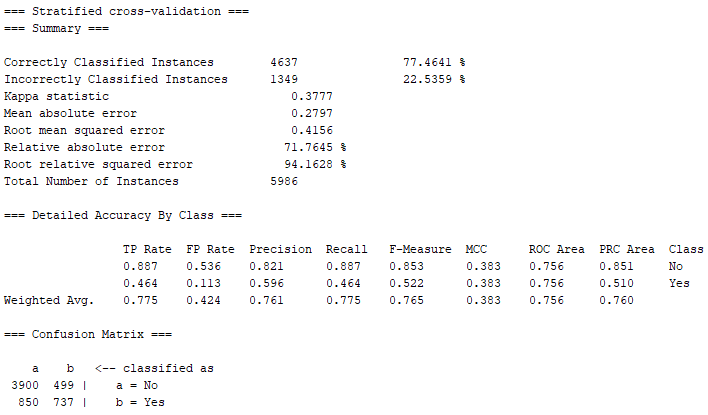
5. Contract, PaymentMethod, tenure, InternetService, MultipleLines (Ranker + ReliefAttributeEval)

**Detailed Description of Data Mining Procedure**

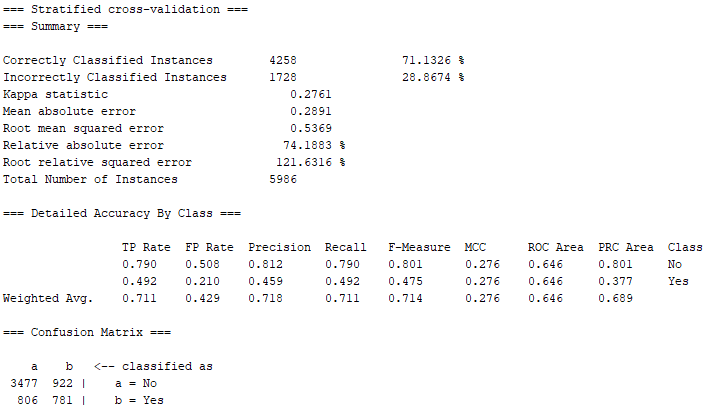
1. Preprocessing: remove **2** attributes – **ID** and **customerID** when building classification models

2. Test 5 classification models using 10-fold cross-validation in Weka

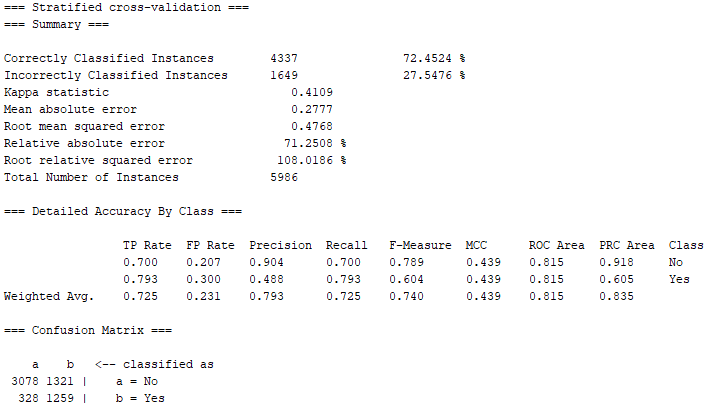
**J48/Partition Model Classification**



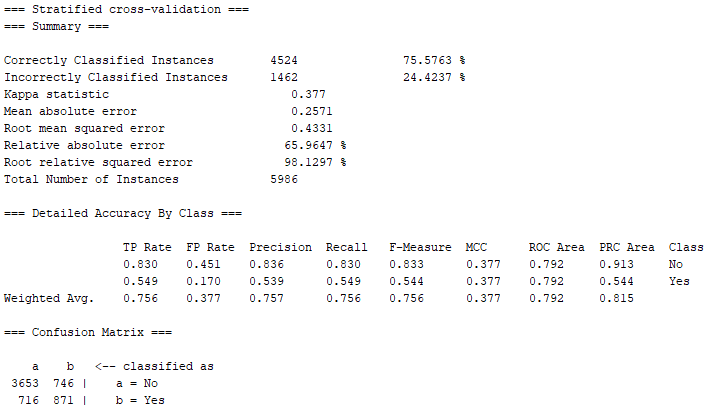
**K-nearest Neighbors Classification**

****

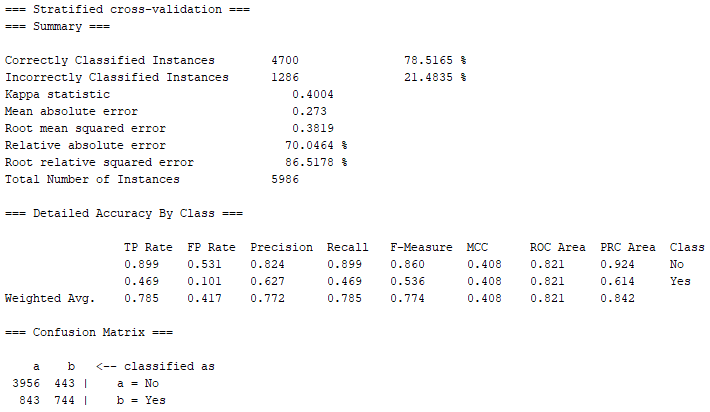
**Naïve Bayes Classification**

****

**Neural Network Classification**

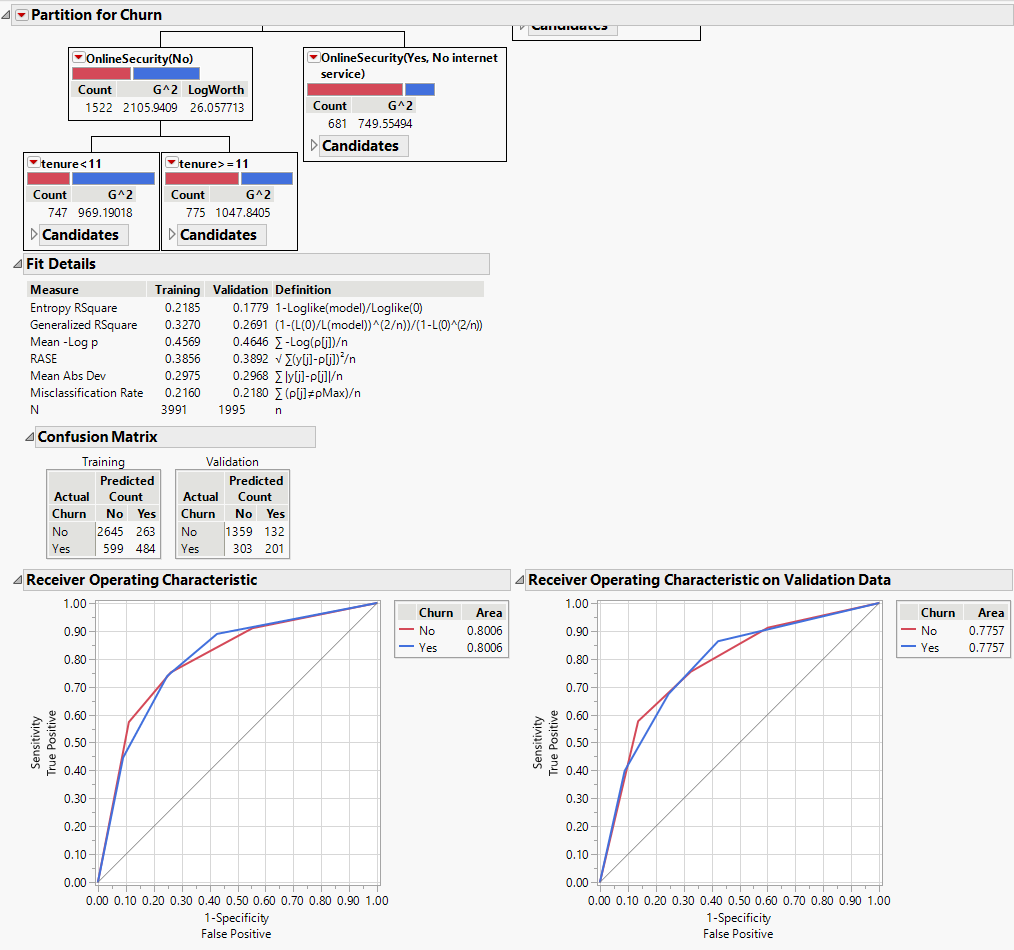
****

**Boosted Tree Classification**

****

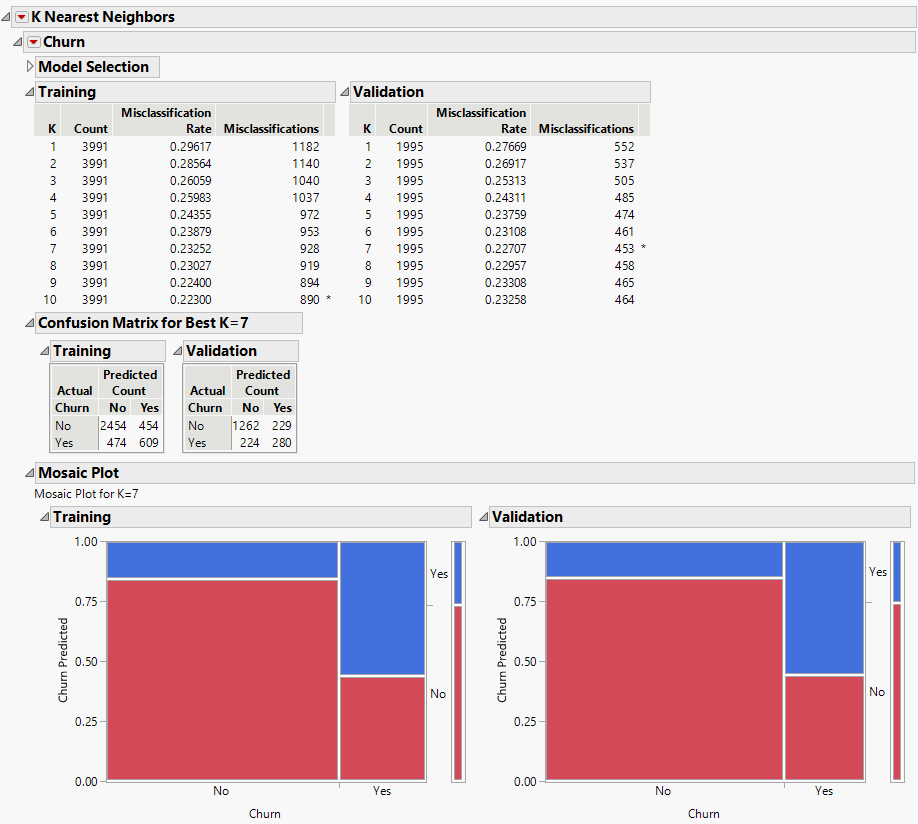
3. Build 5 classification models in JMP Pro

**J48/Partition Model Classification**



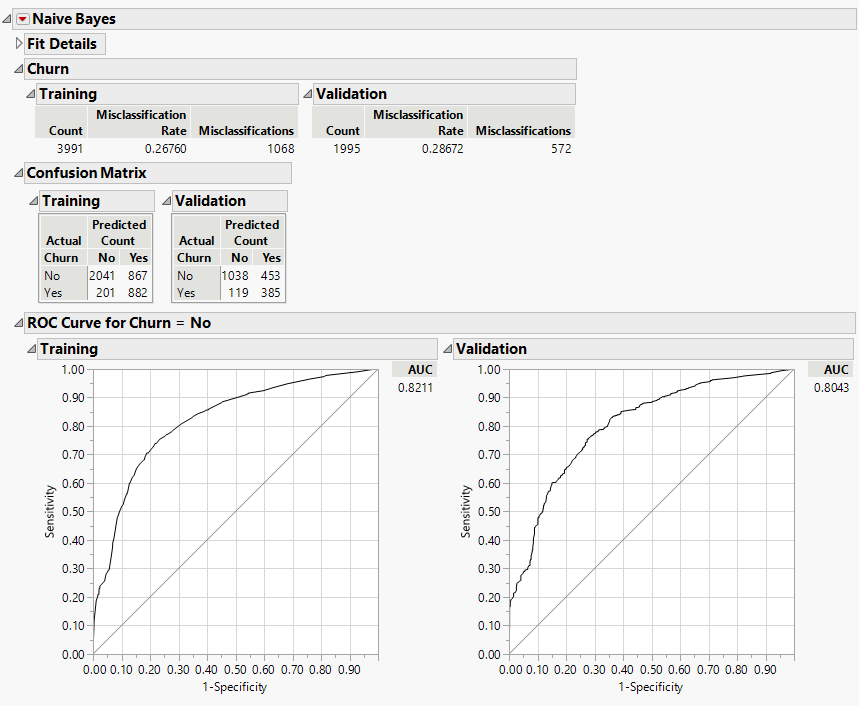
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TP Rate** | **FP Rate** | **Precision** | **Recall** | **F-Measure** | **MCC** | **Class** |
| 0.8177 | 0.3964 | 0.9115 | 0.8177 | 0.8620 | 0.3615 | No |
| 0.6036 | 0.1823 | 0.3988 | 0.6036 | 0.4803 | 0.3615 | Yes |

**K-nearest Neighbors Classification**



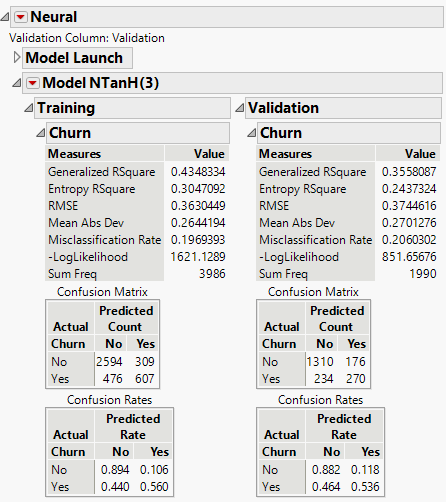
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TP Rate** | **FP Rate** | **Precision** | **Recall** | **F-Measure** | **MCC** | **Class** |
| 0.8493 | 0.4499 | 0.8464 | 0.8493 | 0.8478 | 0.4007 | No |
| 0.5501 | 0.1507 | 0.5556 | 0.5501 | 0.5528 | 0.4007 | Yes |

**Naïve Bayes Classification**



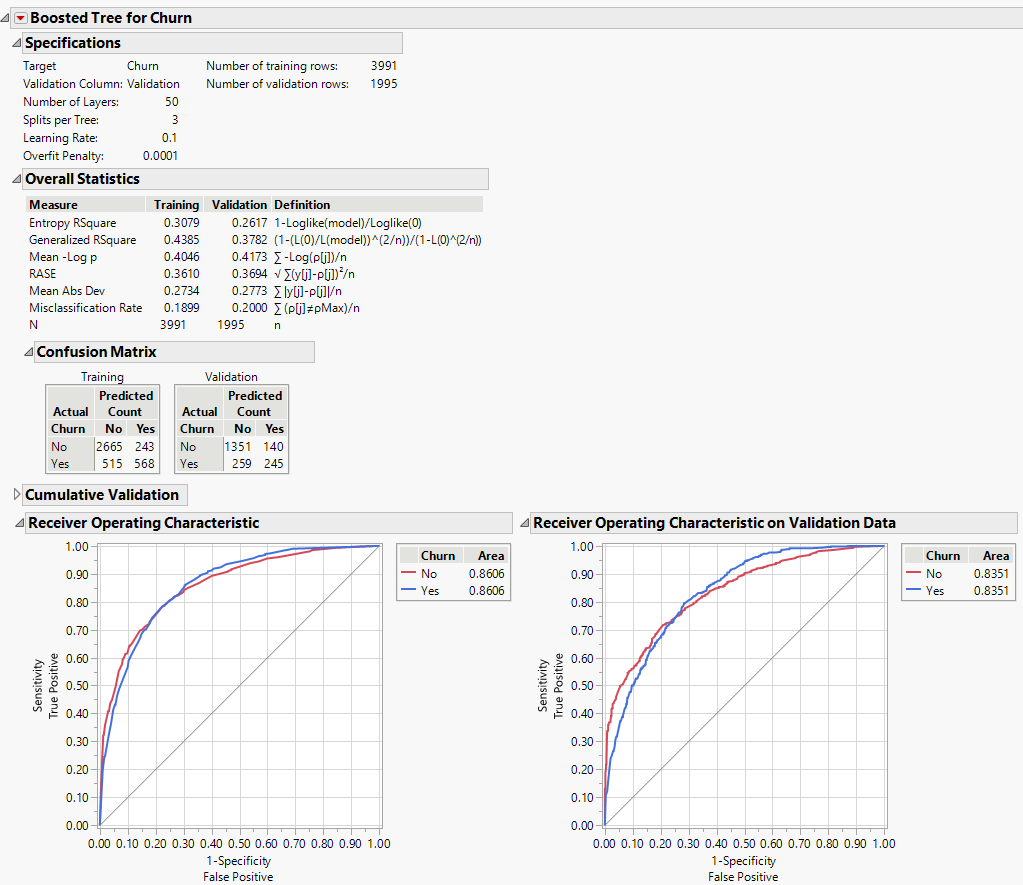
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TP Rate** | **FP Rate** | **Precision** | **Recall** | **F-Measure** | **MCC** | **Class** |
| 0.8971 | 0.5406 | 0.6962 | 0.8971 | 0.7840 | 0.4050 | No |
| 0.4594 | 0.1029 | 0.7639 | 0.4594 | 0.5738 | 0.4050 | Yes |

**Neural Network Classification**



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TP Rate** | **FP Rate** | **Precision** | **Recall** | **F-Measure** | **MCC** | **Class** |
| 0.8484 | 0.3946 | 0.8816 | 0.8484 | 0.8647 | 0.4352 | No |
| 0.6054 | 0.1516 | 0.5357 | 0.6054 | 0.5684 | 0.4352 | Yes |

**Boosted Tree Classification**



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TP Rate** | **FP Rate** | **Precision** | **Recall** | **F-Measure** | **MCC** | **Class** |
| 0.8391 | 0.3636 | 0.9061 | 0.8391 | 0.8713 | 0.4319 | No |
| 0.6364 | 0.1609 | 0.4861 | 0.6364 | 0.5512 | 0.4319 | Yes |

4. Select 5 attribution selection methods in Weka

5. Select 5 sets of attributes by each attribute selection method and prepare five reduced training datasets

6. Add an extra attribute called “Validation”, split the preprocessed dataset into a training dataset and a test dataset. Training dataset has **3991 tuples (66%)**, test dataset has **1995 tuples (33%)**. Total with **5986 tuples**

7. From each reduced training dataset, build five classification models and test them on the corresponding reduced test dataset

8. Build and test 25 classification models (Shown in Data Mining Result and Evaluation)

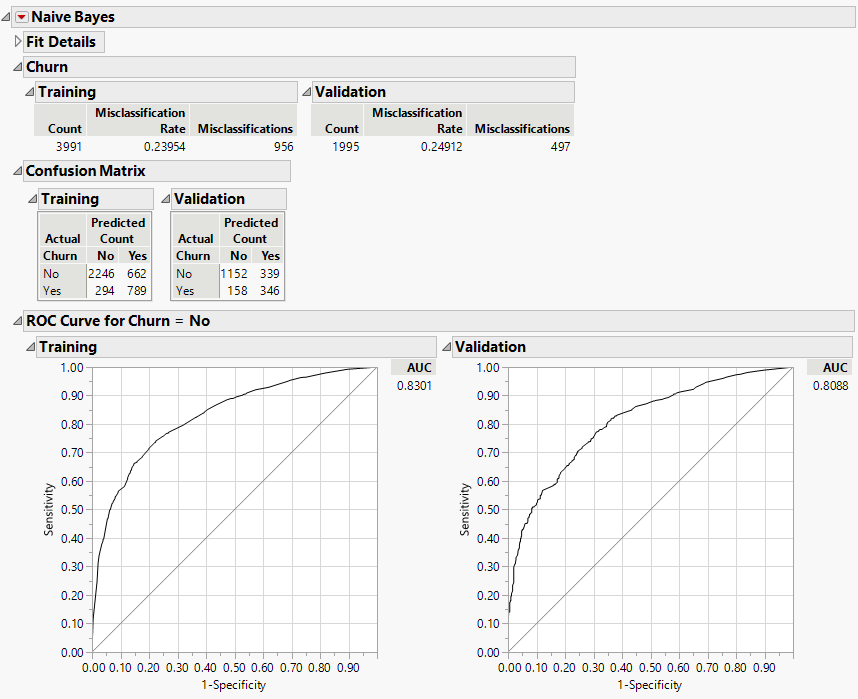
9. Compare and select the best model

**Data Mining Result and Evaluation**

**1. BestFirst + CfsSubsetEval**

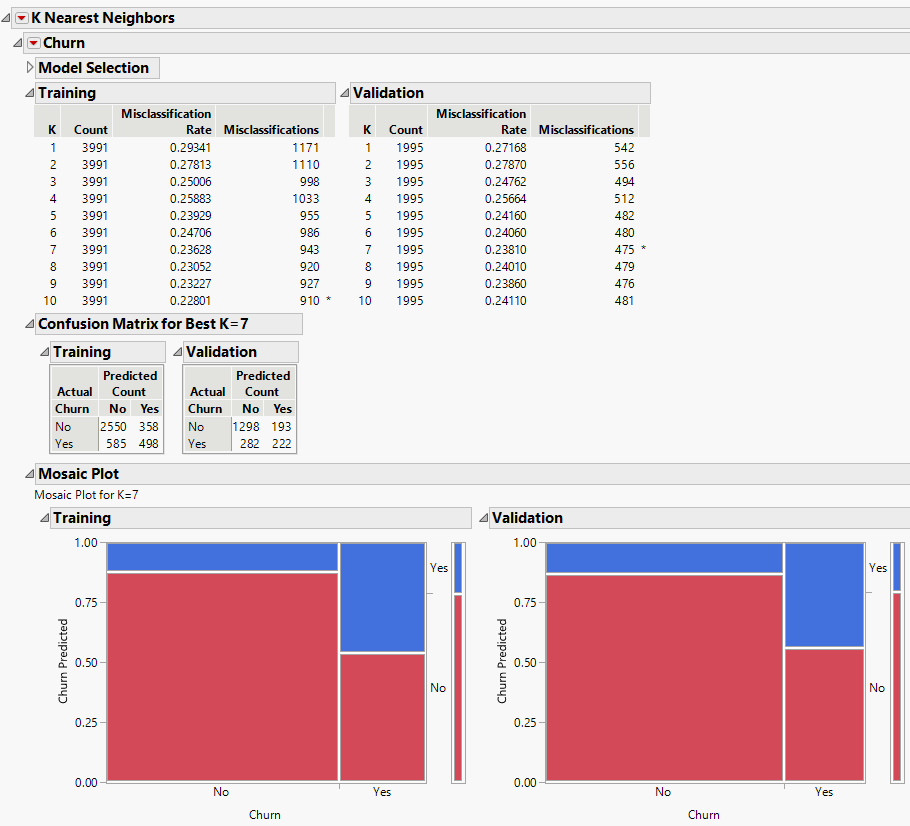
**tenure, OnlineSecurity, TechSupport, Contract, PaymendMethod**

**(1) Naïve Bayes Classification**

****

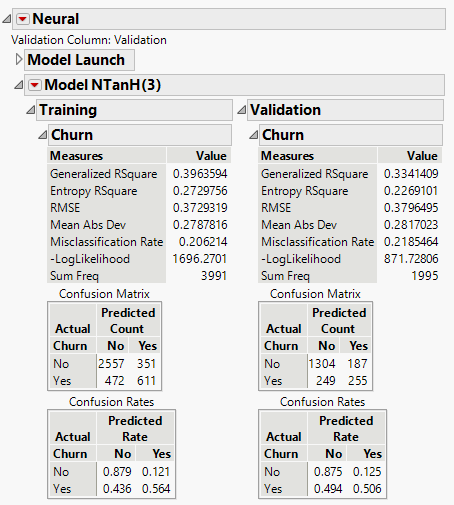
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TP Rate** | **FP Rate** | **Precision** | **Recall** | **F-Measure** | **MCC** | **Class** |
| 0.8794 | 0.4949 | 0.7726 | 0.8794 | 0.8226 | 0.4202 | No |
| 0.5051 | 0.1206 | 0.6865 | 0.5051 | 0.5820 | 0.4202 | Yes |

**(2) K-nearest Neighbors Classification**

****

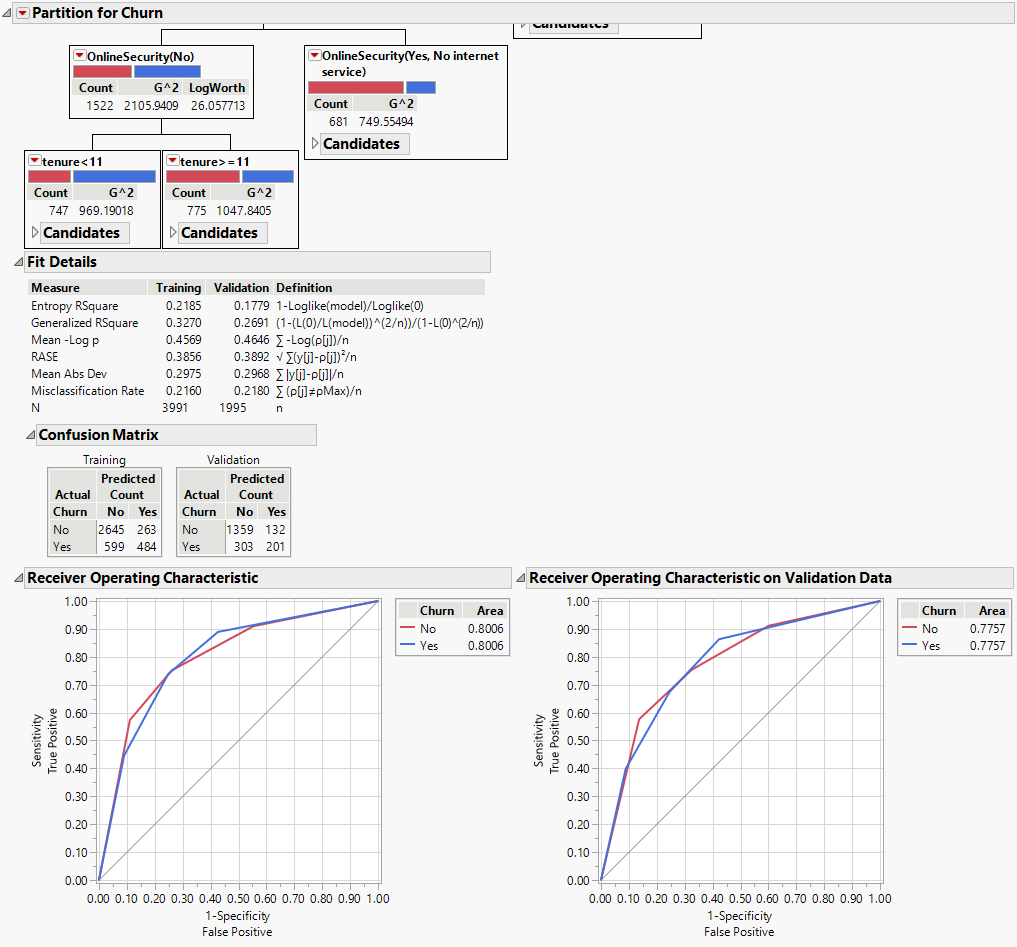
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TP Rate** | **FP Rate** | **Precision** | **Recall** | **F-Measure** | **MCC** | **Class** |
| 0.8215 | 0.4651 | 0.8706 | 0.8215 | 0.8453 | 0.3330 | No |
| 0.5349 | 0.1785 | 0.4405 | 0.5349 | 0.4831 | 0.3330 | Yes |

**(3) Neural Network Classification**

****

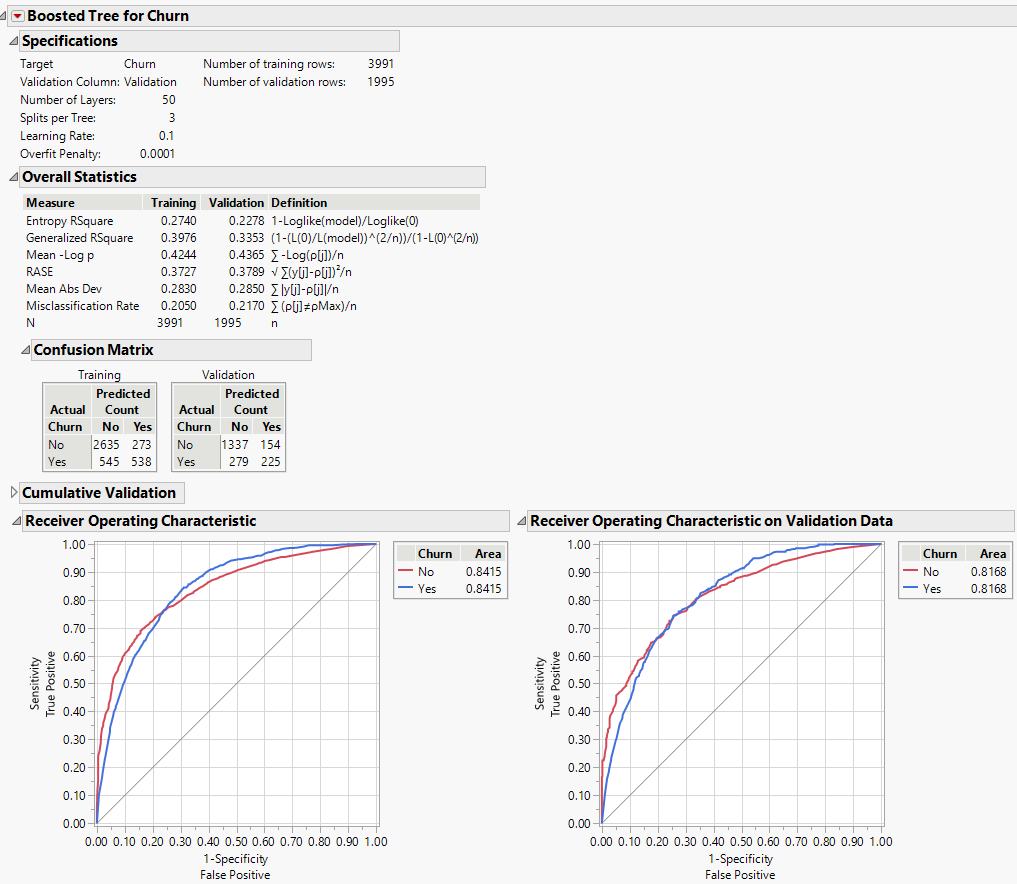
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TP Rate** | **FP Rate** | **Precision** | **Recall** | **F-Measure** | **MCC** | **Class** |
| 0.8397 | 0.4231 | 0.8746 | 0.8397 | 0.8568 | 0.3982 | No |
| 0.5769 | 0.1603 | 0.5060 | 0.5769 | 0.5391 | 0.3982 | Yes |

**(4) Partition Model Classification (Decision Tree)**

****

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TP Rate** | **FP Rate** | **Precision** | **Recall** | **F-Measure** | **MCC** | **Class** |
| 0.8177 | 0.3964 | 0.9115 | 0.8177 | 0.8620 | 0.3615 | No |
| 0.6036 | 0.1823 | 0.3988 | 0.6036 | 0.4803 | 0.3615 | Yes |

**(5) Boosted Tree Classification**

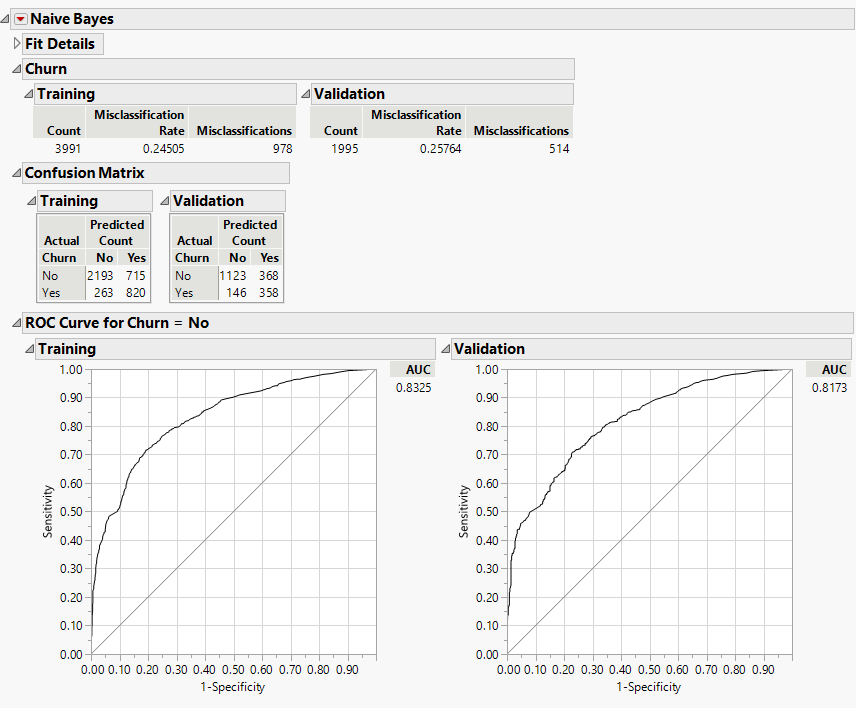
****

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TP Rate** | **FP Rate** | **Precision** | **Recall** | **F-Measure** | **MCC** | **Class** |
| 0.8274 | 0.4063 | 0.8967 | 0.8274 | 0.8606 | 0.3801 | No |
| 0.5937 | 0.1726 | 0.4464 | 0.5937 | 0.5096 | 0.3801 | Yes |

**2. Ranker + CorrelationAttributeEval**

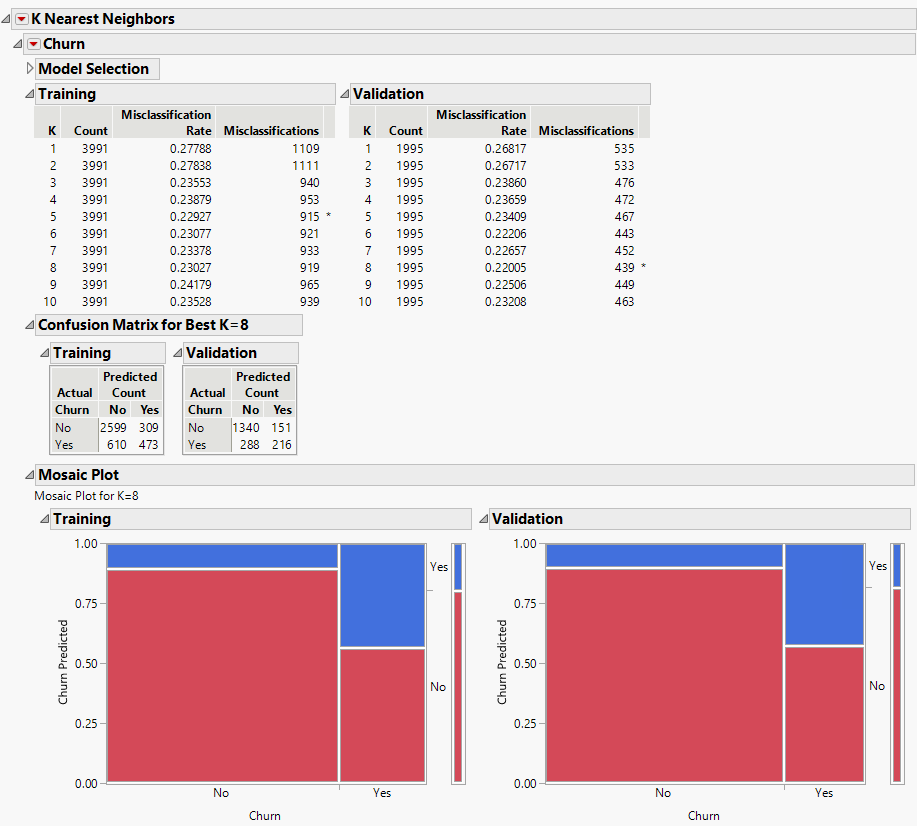
**tenure, Contract, OnlineSecurity, TechSupport, InternetService**

**(1) Naïve Bayes Classification**

****

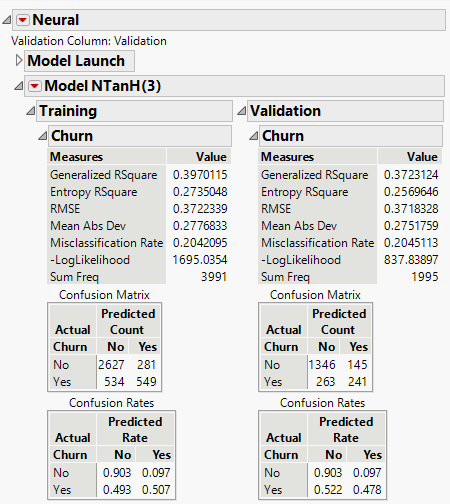
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TP Rate** | **FP Rate** | **Precision** | **Recall** | **F-Measure** | **MCC** | **Class** |
| 0.8849 | 0.5069 | 0.7532 | 0.8849 | 0.8138 | 0.4186 | No |
| 0.4931 | 0.1151 | 0.7103 | 0.4931 | 0.5821 | 0.4186 | Yes |

**(2) K-nearest Neighbors Classification**

****

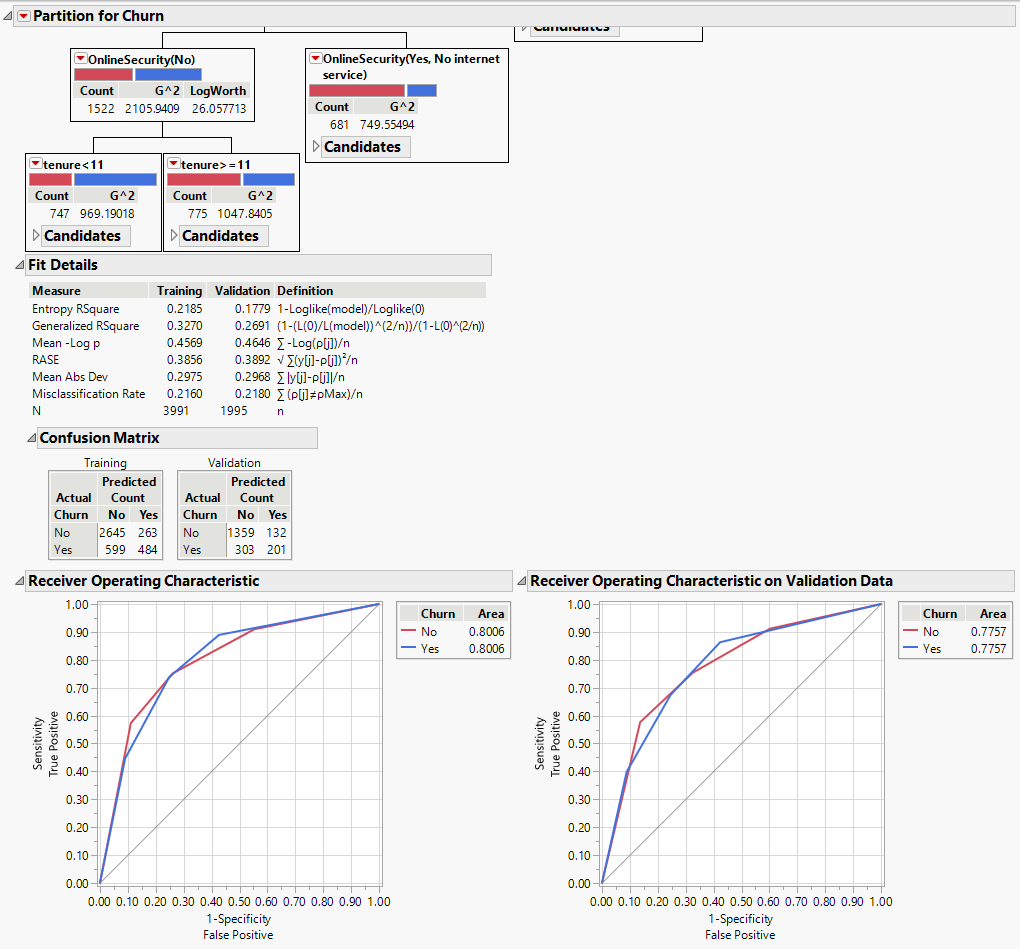
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TP Rate** | **FP Rate** | **Precision** | **Recall** | **F-Measure** | **MCC** | **Class** |
| 0.8231 | 0.4114 | 0.8987 | 0.8231 | 0.8592 | 0.3671 | No |
| 0.5886 | 0.1769 | 0.4286 | 0.5886 | 0.4960 | 0.3671 | Yes |

**(3) Neural Network Classification**

****

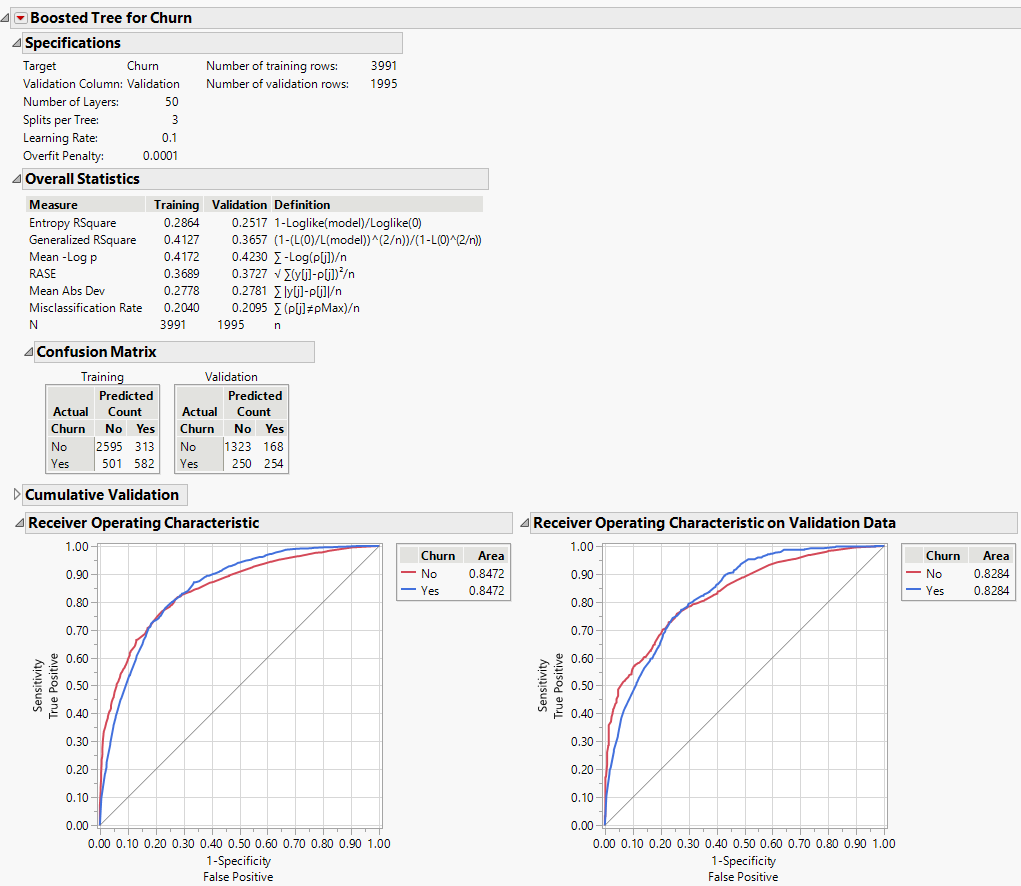
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TP Rate** | **FP Rate** | **Precision** | **Recall** | **F-Measure** | **MCC** | **Class** |
| 0.8365 | 0.3756 | 0.9027 | 0.8365 | 0.8684 | 0.4190 | No |
| 0.6244 | 0.1635 | 0.4782 | 0.6244 | 0.5416 | 0.4190 | Yes |

**(4) Partition Model Classification (Decision Tree)**

****

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TP Rate** | **FP Rate** | **Precision** | **Recall** | **F-Measure** | **MCC** | **Class** |
| 0.8177 | 0.3964 | 0.9115 | 0.8177 | 0.8620 | 0.3615 | No |
| 0.6036 | 0.1823 | 0.3988 | 0.6036 | 0.4803 | 0.3615 | Yes |

**(5) Boosted Tree Classification**

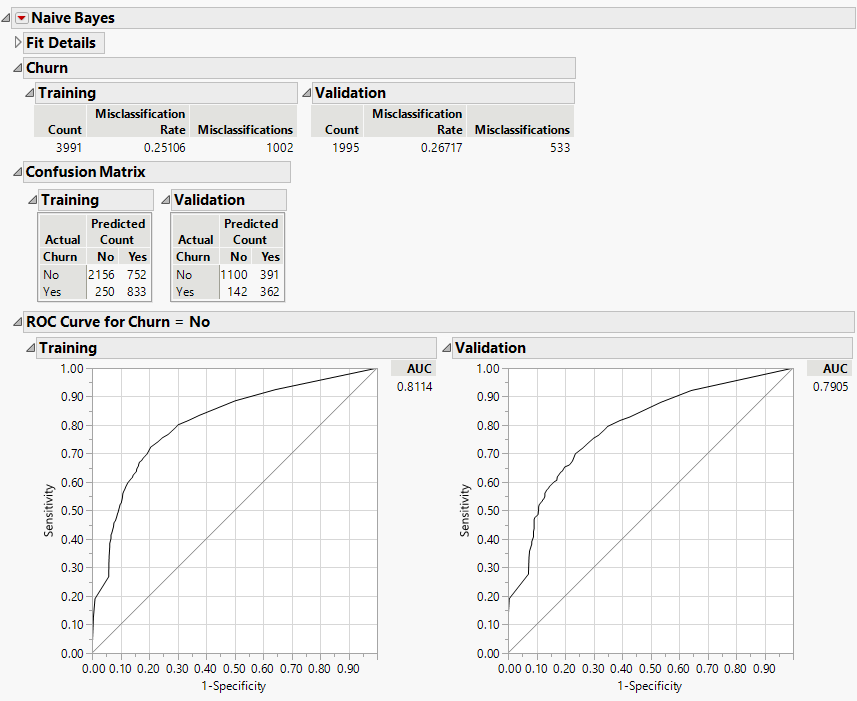
****

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TP Rate** | **FP Rate** | **Precision** | **Recall** | **F-Measure** | **MCC** | **Class** |
| 0.8411 | 0.3981 | 0.8873 | 0.8411 | 0.8636 | 0.4163 | No |
| 0.6019 | 0.1589 | 0.5040 | 0.6019 | 0.5486 | 0.4163 | Yes |

**3. Ranker + GainRatioAttributeEval**

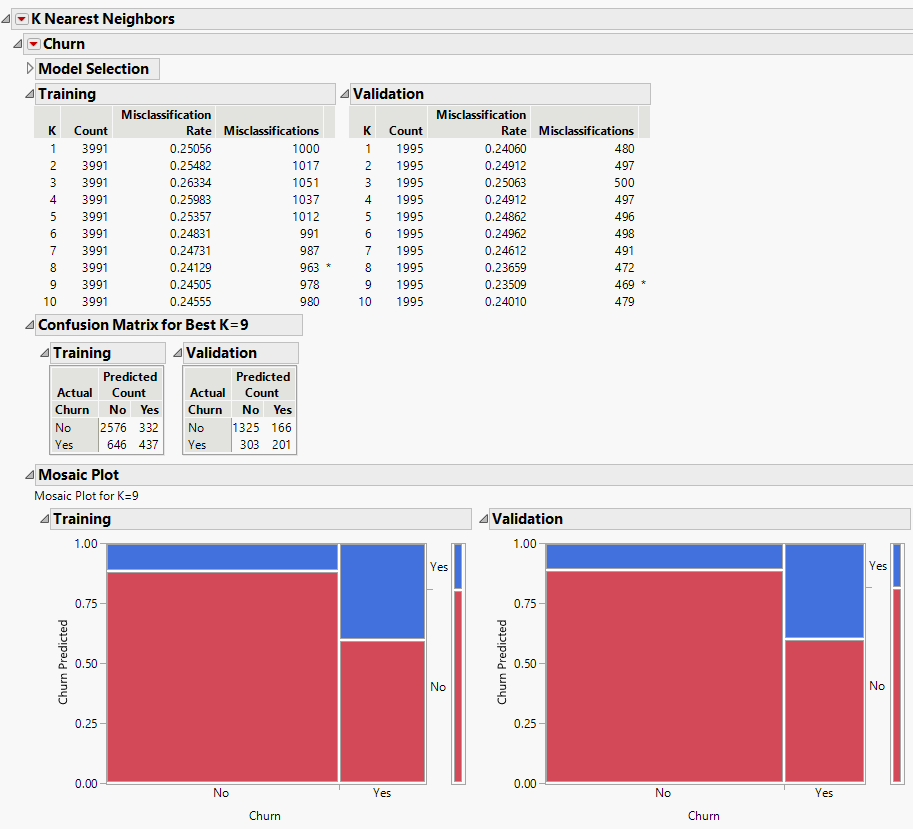
**Contract, OnlineSecurity, TechSupport, InternetService, OnlineBackup**

**(1) Naïve Bayes Classification**

****

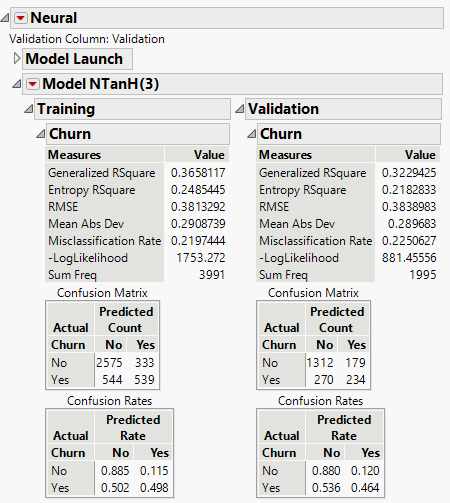
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TP Rate** | **FP Rate** | **Precision** | **Recall** | **F-Measure** | **MCC** | **Class** |
| 0.8857 | 0.5193 | 0.7378 | 0.8857 | 0.8050 | 0.4088 | No |
| 0.4807 | 0.1143 | 0.7183 | 0.4807 | 0.5760 | 0.4088 | Yes |

**(2) K-nearest Neighbors Classification**

****

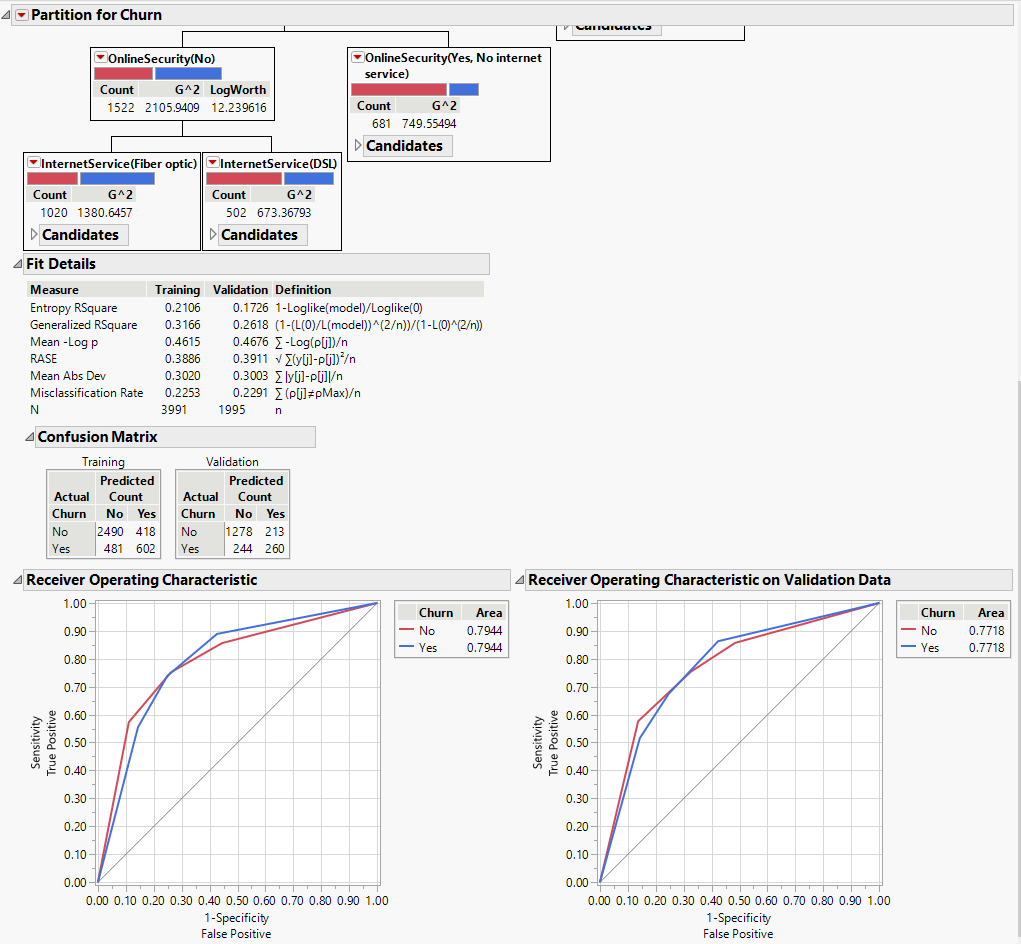
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TP Rate** | **FP Rate** | **Precision** | **Recall** | **F-Measure** | **MCC** | **Class** |
| 0.8139 | 0.4523 | 0.8887 | 0.8139 | 0.8496 | 0.3224 | No |
| 0.5477 | 0.1861 | 0.3988 | 0.5477 | 0.4615 | 0.3224 | Yes |

**(3) Neural Network Classification**

****

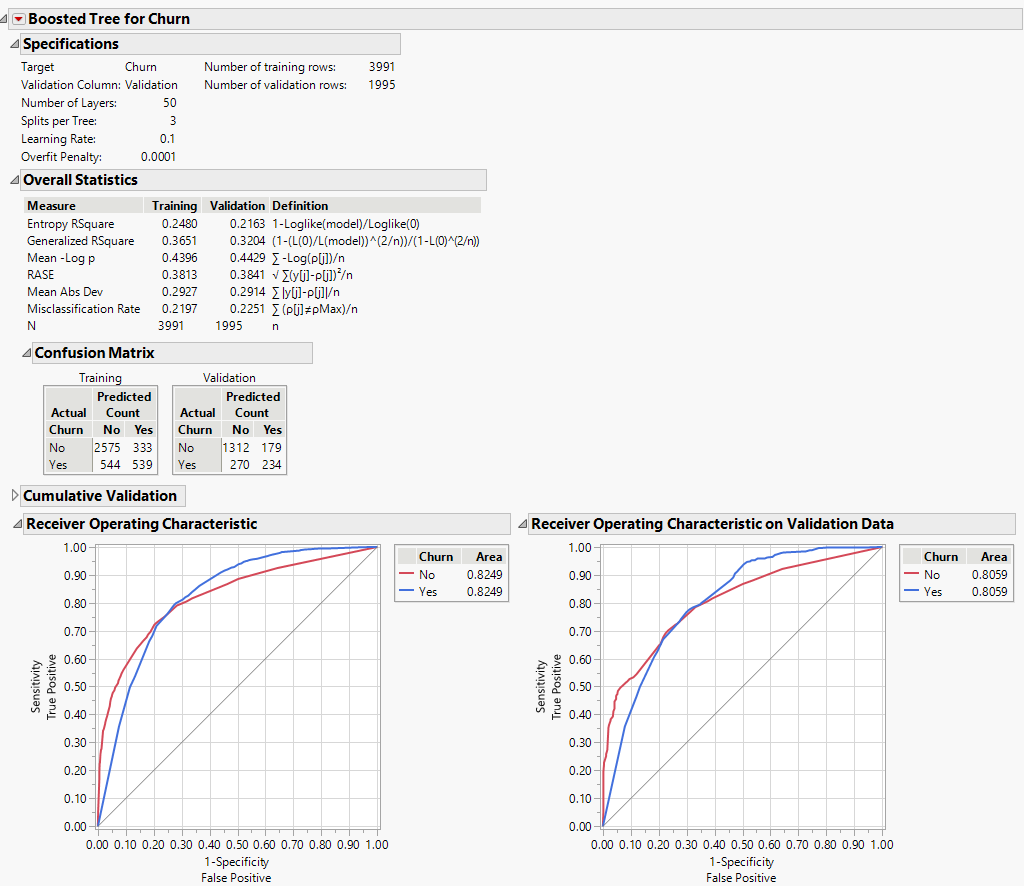
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TP Rate** | **FP Rate** | **Precision** | **Recall** | **F-Measure** | **MCC** | **Class** |
| 0.8293 | 0.4334 | 0.8799 | 0.8293 | 0.8539 | 0.3692 | No |
| 0.5666 | 0.1707 | 0.4643 | 0.5666 | 0.5104 | 0.3692 | Yes |

**(4) Partition Model Classification (Decision Tree)**

****

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TP Rate** | **FP Rate** | **Precision** | **Recall** | **F-Measure** | **MCC** | **Class** |
| 0.8397 | 0.4503 | 0.8571 | 0.8397 | 0.8483 | 0.3811 | No |
| 0.5497 | 0.1603 | 0.5159 | 0.5497 | 0.5322 | 0.3811 | Yes |

**(5) Boosted Tree Classification**

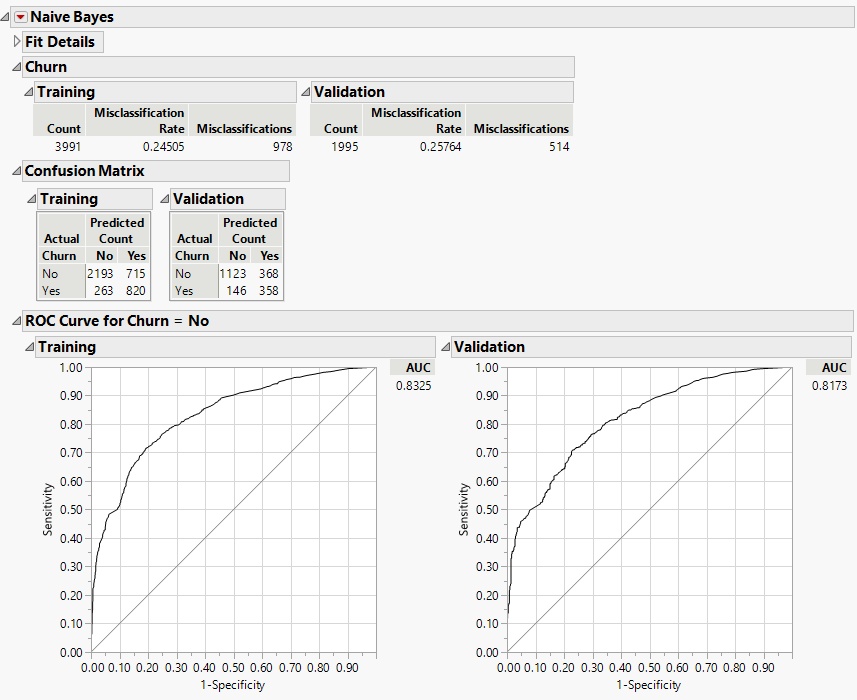
****

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TP Rate** | **FP Rate** | **Precision** | **Recall** | **F-Measure** | **MCC** | **Class** |
| 0.8293 | 0.4334 | 0.8799 | 0.8293 | 0.8539 | 0.3692 | No |
| 0.5666 | 0.1707 | 0.4643 | 0.5666 | 0.5104 | 0.3692 | Yes |

**4. Ranker + InfoGainAttributeEval**

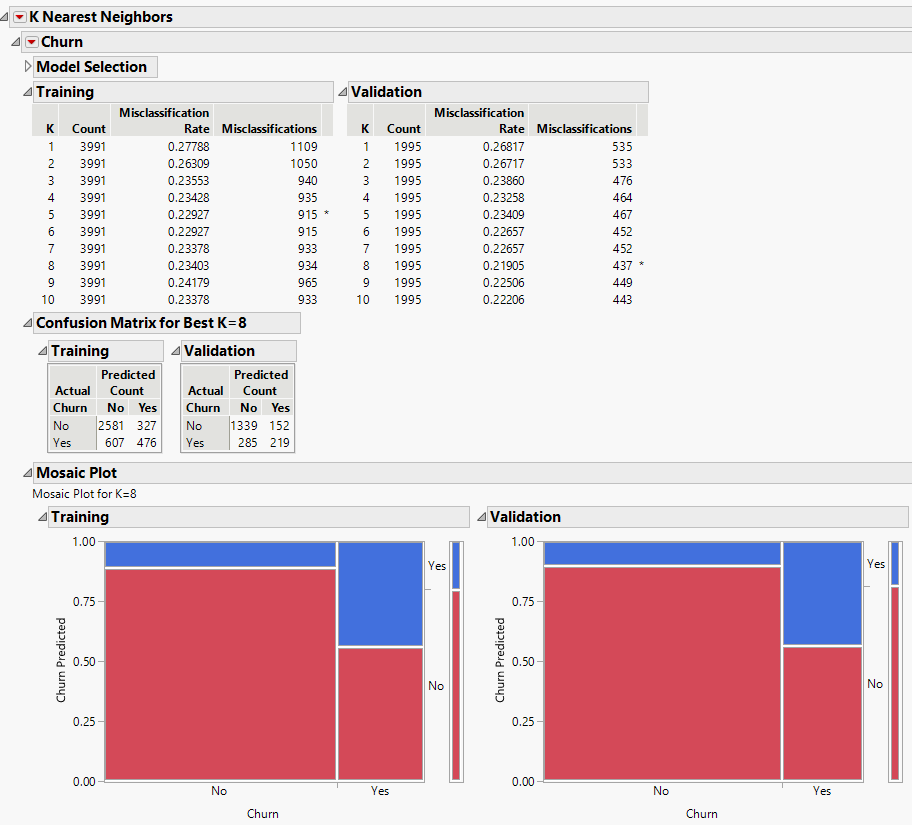
**Contract, tenure, OnlineSecurity, TechSupport, InternetService**

**(1) Naïve Bayes Classification**

****

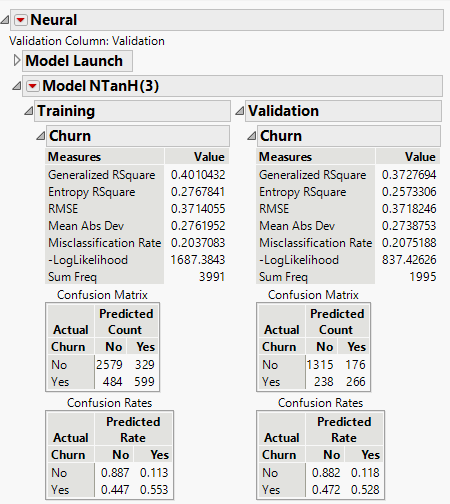
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TP Rate** | **FP Rate** | **Precision** | **Recall** | **F-Measure** | **MCC** | **Class** |
| 0.8849 | 0.5069 | 0.7532 | 0.8849 | 0.8138 | 0.4186 | No |
| 0.4931 | 0.1151 | 0.7103 | 0.4931 | 0.5821 | 0.4186 | Yes |

**(2) K-nearest Neighbors Classification**

****

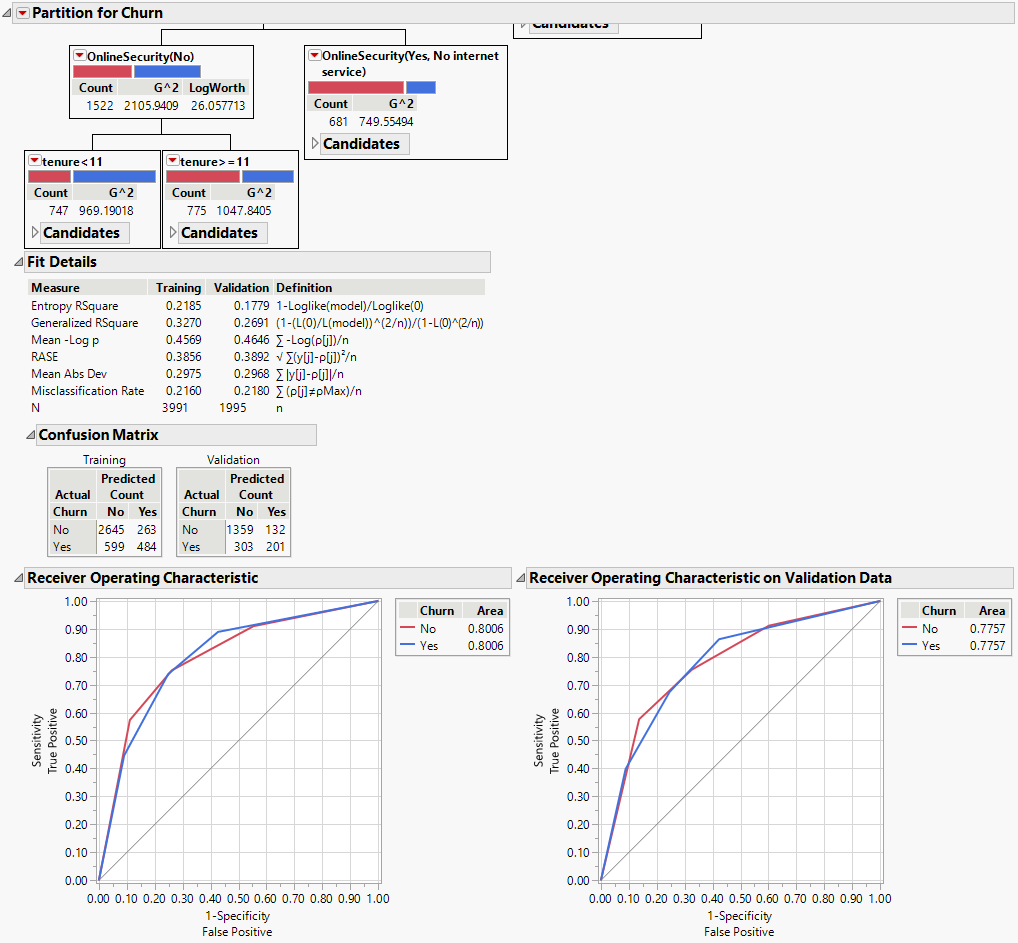
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TP Rate** | **FP Rate** | **Precision** | **Recall** | **F-Measure** | **MCC** | **Class** |
| 0.8245 | 0.4097 | 0.8981 | 0.8245 | 0.8597 | 0.3714 | No |
| 0.5903 | 0.1755 | 0.4345 | 0.5903 | 0.5006 | 0.3714 | Yes |

**(3) Neural Network Classification**

****

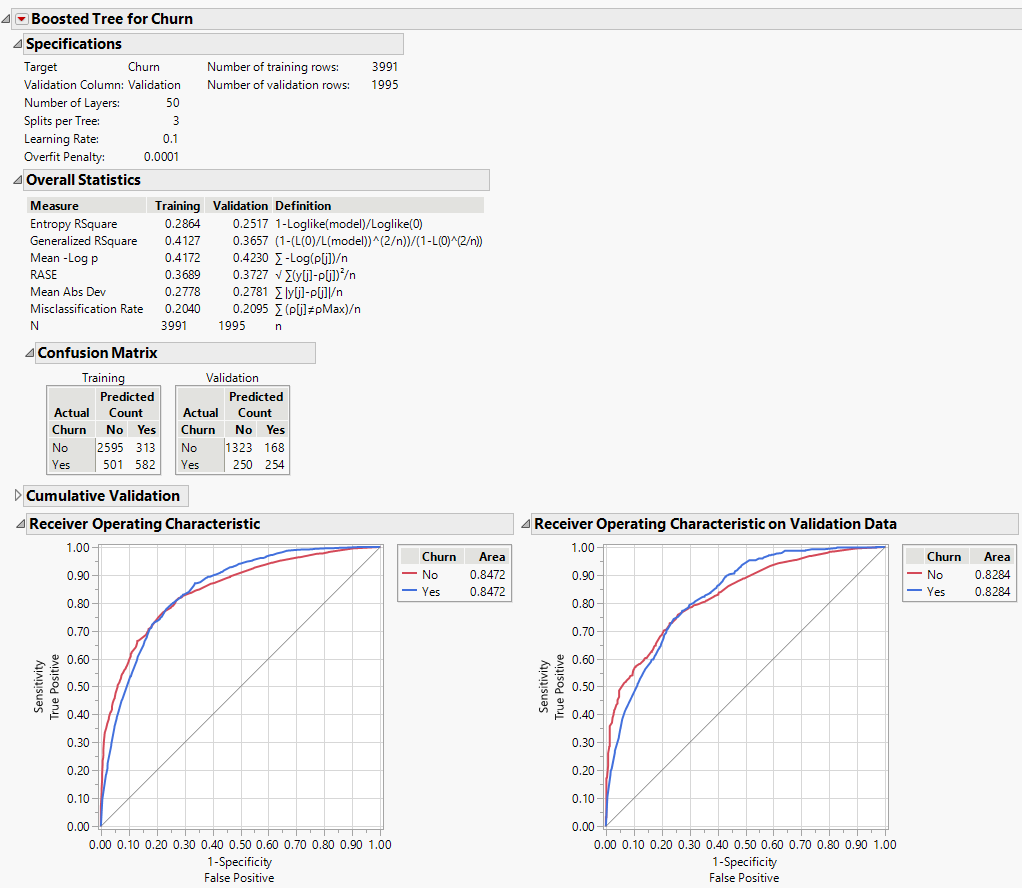
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TP Rate** | **FP Rate** | **Precision** | **Recall** | **F-Measure** | **MCC** | **Class** |
| 0.8467 | 0.3982 | 0.8820 | 0.8467 | 0.8640 | 0.4287 | No |
| 0.6018 | 0.1533 | 0.5278 | 0.6018 | 0.5624 | 0.4287 | Yes |

**(4) Partition Model Classification (Decision Tree)**

****

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TP Rate** | **FP Rate** | **Precision** | **Recall** | **F-Measure** | **MCC** | **Class** |
| 0.8177 | 0.3964 | 0.9115 | 0.8177 | 0.8620 | 0.3615 | No |
| 0.6036 | 0.1823 | 0.3988 | 0.6036 | 0.4803 | 0.3615 | Yes |

**(5) Boosted Tree Classification**

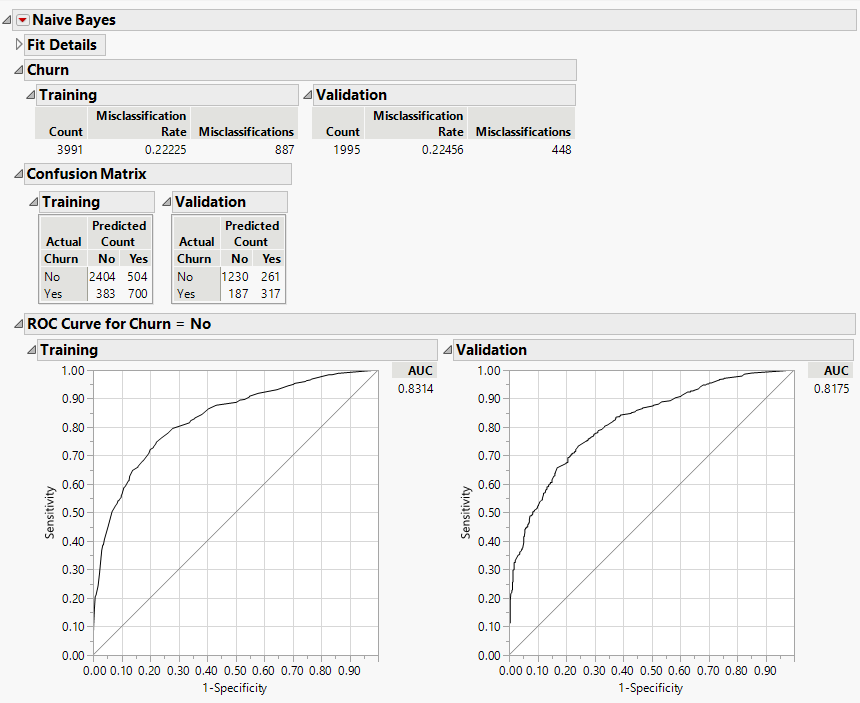
****

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TP Rate** | **FP Rate** | **Precision** | **Recall** | **F-Measure** | **MCC** | **Class** |
| 0.8411 | 0.3981 | 0.8873 | 0.8411 | 0.8636 | 0.4163 | No |
| 0.6019 | 0.1589 | 0.5040 | 0.6019 | 0.4163 | 0.4163 | Yes |

**5. Ranker + ReliefAttributeEval**

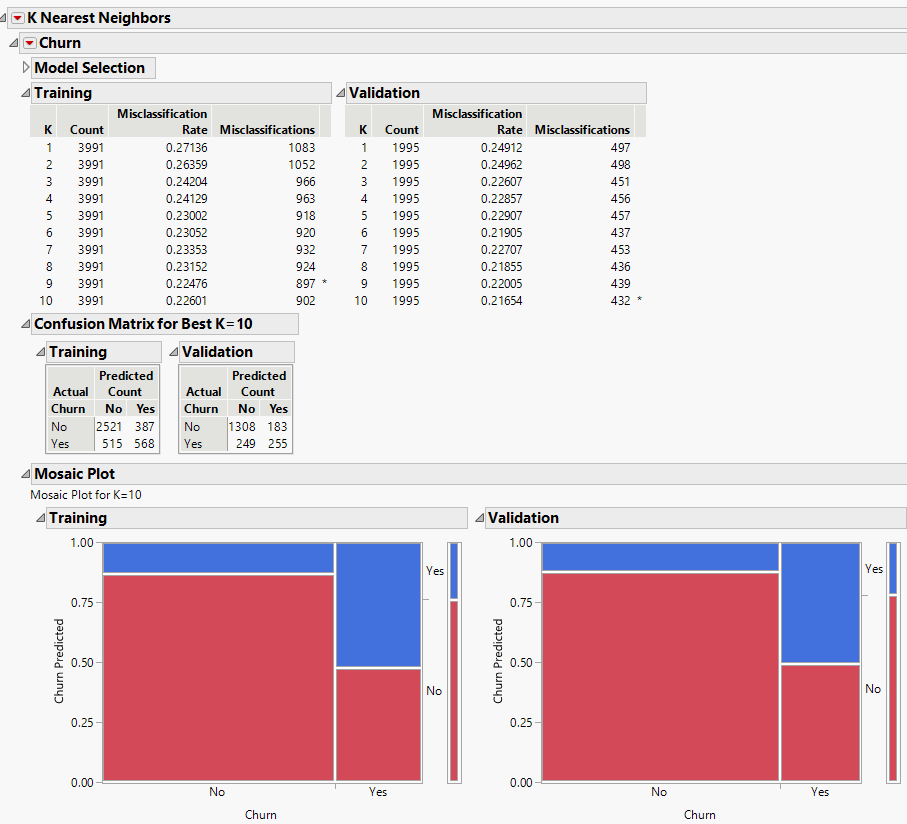
**Contract, PaymentMethod, tenure, InternetService, MultipleLines**

**(1) Naïve Bayes Classification**

****

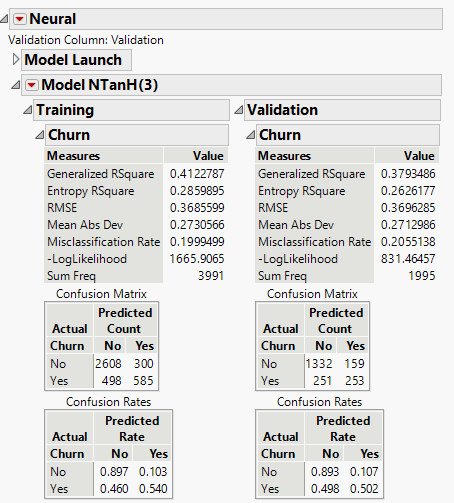
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TP Rate** | **FP Rate** | **Precision** | **Recall** | **F-Measure** | **MCC** | **Class** |
| 0.8680 | 0.4516 | 0.8249 | 0.8680 | 0.8459 | 0.4348 | No |
| 0.5484 | 0.1320 | 0.6290 | 0.5484 | 0.5860 | 0.4348 | Yes |

**(2) K-nearest Neighbors Classification**

****

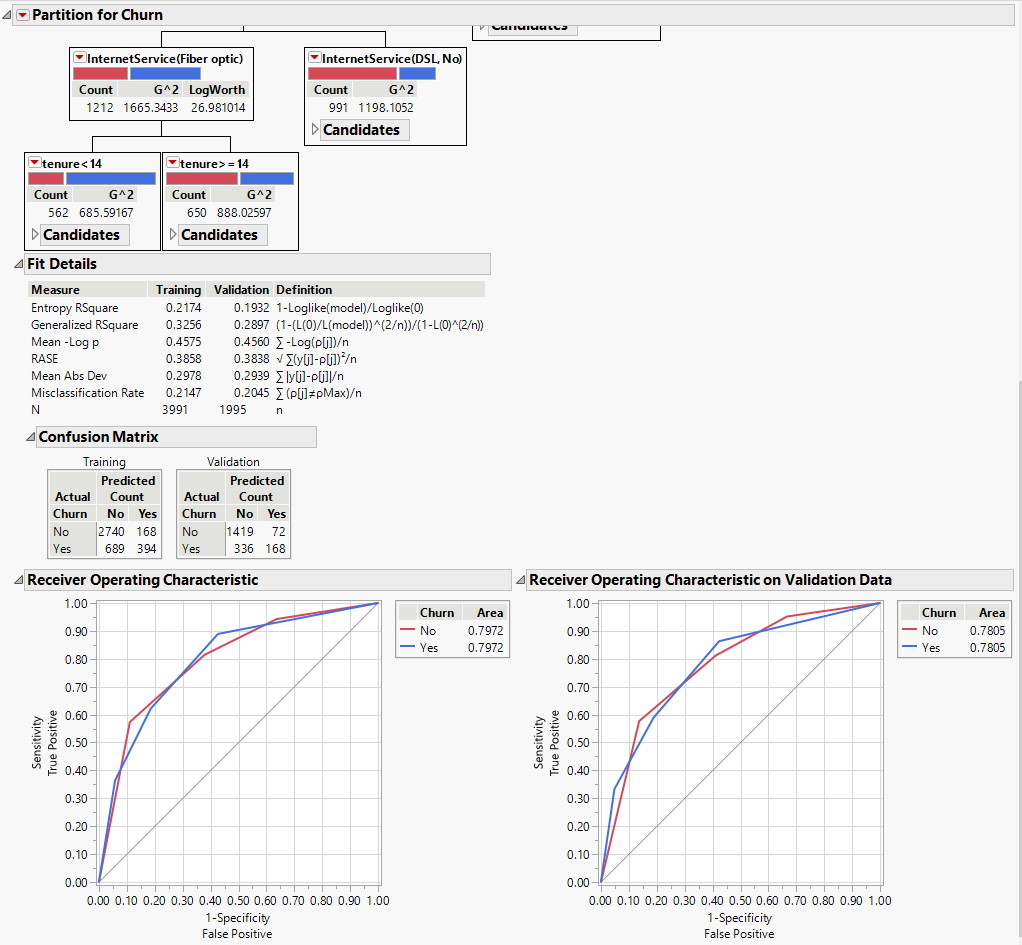
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TP Rate** | **FP Rate** | **Precision** | **Recall** | **F-Measure** | **MCC** | **Class** |
| 0.8401 | 0.4178 | 0.8773 | 0.8401 | 0.8583 | 0.4023 | No |
| 0.5822 | 0.1599 | 0.5060 | 0.5822 | 0.5414 | 0.4023 | Yes |

**(3) Neural Network Classification**

****

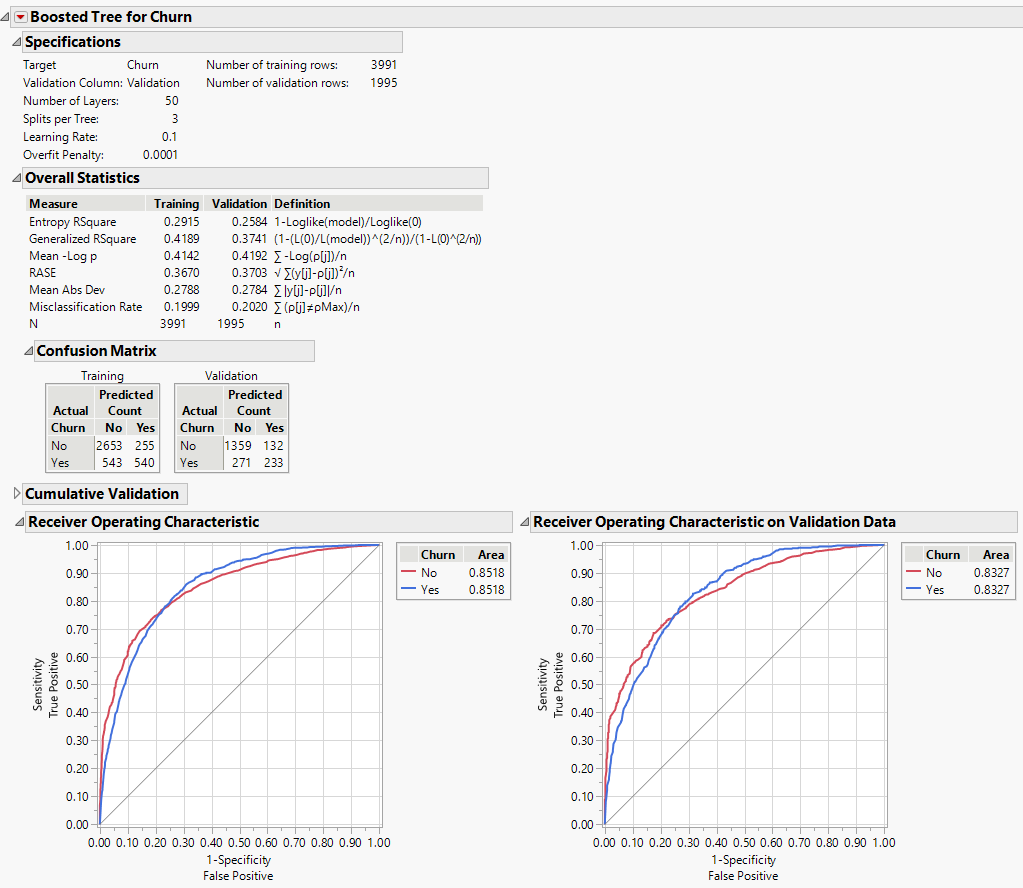
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TP Rate** | **FP Rate** | **Precision** | **Recall** | **F-Measure** | **MCC** | **Class** |
| 0.8414 | 0.3859 | 0.8934 | 0.8414 | 0.8666 | 0.4244 | No |
| 0.6141 | 0.1586 | 0.5020 | 0.6141 | 0.5524 | 0.4244 | Yes |

**(4) Partition Model Classification (Decision Tree)**

****

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TP Rate** | **FP Rate** | **Precision** | **Recall** | **F-Measure** | **MCC** | **Class** |
| 0.8085 | 0.3000 | 0.9517 | 0.8085 | 0.8743 | 0.3807 | No |
| 0.7000 | 0.1915 | 0.3333 | 0.7000 | 0.4516 | 0.3807 | Yes |

**(5) Boosted Tree Classification**

****

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TP Rate** | **FP Rate** | **Precision** | **Recall** | **F-Measure** | **MCC** | **Class** |
| 0.8337 | 0.3616 | 0.9115 | 0.8337 | 0.8709 | 0.4201 | No |
| 0.6384 | 0.1663 | 0.4623 | 0.6384 | 0.5362 | 0.4201 | Yes |

**Discussion and Conclusion**

The higher the correlation between true and predicted values, the better the prediction. Rechristened Matthews Correlation Coefficient (MCC) when applied to classifiers.

When using **Naïve Bayes Classification**, **Ranker + ReliefAttributeEval** has better prediction.

When using **K-nearest Neighbors Classification**, **Ranker + ReliefAttributeEval** has betterprediction.

When using **Neural Network Classification**, **Ranker + InfoGainAttributeEval** has better prediction.

When using **Partition Model Classification (Decision Tree)**, **Ranker + GainRatioAttributeEval** has better prediction.

When using **Boosted Tree Classification**, **Ranker + ReliefAttributeEval** has better prediction.

For all classification, **Naïve Bayes Classification** with **Ranker + ReliefAttributeEval** has the best prediction. (Contract, PaymentMethod, tenure, InternetService, MultipleLines (Ranker + ReliefAttributeEval))

**Naïve Bayes Classification**

|  |  |
| --- | --- |
|  | **MCC** |
| Original Model | 0.4050 |
| 10-fold cross-validation | 0.439 |
| BestFirst + CfsSubsetEva | 0.4202 |
| Ranker + CorrelationAttributeEval | 0.4186 |
| Ranker + GainRatioAttributeEval | 0.4088 |
| Ranker + InfoGainAttributeEval | 0.4186 |
| **Ranker + ReliefAttributeEval** | **0.4348** |

**K-nearest Neighbors Classification**

|  |  |
| --- | --- |
|  | **MCC** |
| Original Model | 0.4007 |
| 10-fold cross-validation | 0.276 |
| BestFirst + CfsSubsetEva | 0.3330 |
| Ranker + CorrelationAttributeEval | 0.3671 |
| Ranker + GainRatioAttributeEval | 0.3224 |
| Ranker + InfoGainAttributeEval | 0.3714 |
| **Ranker + ReliefAttributeEval** | **0.4023** |

**Neural Network Classification**

|  |  |
| --- | --- |
|  | **MCC** |
| Original Model | 0.4352 |
| 10-fold cross-validation | 0.377 |
| BestFirst + CfsSubsetEva | 0.3982 |
| Ranker + CorrelationAttributeEval | 0.4190 |
| Ranker + GainRatioAttributeEval | 0.3692 |
| **Ranker + InfoGainAttributeEval** | **0.4287** |
| Ranker + ReliefAttributeEval | 0.4244 |

**Partition Model Classification (Decision Tree)**

|  |  |
| --- | --- |
|  | **MCC** |
| Original Model | 0.3615 |
| 10-fold cross-validation | 0.383 |
| BestFirst + CfsSubsetEva | 0.3615 |
| Ranker + CorrelationAttributeEval | 0.3615 |
| **Ranker + GainRatioAttributeEval** | **0.3811** |
| Ranker + InfoGainAttributeEval | 0.3615 |
| Ranker + ReliefAttributeEval | 0.3807 |

**Boosted Tree Classification**

|  |  |
| --- | --- |
|  | **MCC** |
| Original Model | 0.4319 |
| 10-fold cross-validation | 0.408 |
| BestFirst + CfsSubsetEva | 0.3801 |
| Ranker + CorrelationAttributeEval | 0.4163 |
| Ranker + GainRatioAttributeEval | 0.3692 |
| Ranker + InfoGainAttributeEval | 0.4163 |
| **Ranker + ReliefAttributeEval** | **0.4201** |