BIS3216 Data Mining and Knowledge Discovery

Predictive Modeling for Telco Customer Data

Semester of August 2020

1. Problem Formulation

To predict the churn rate of customers and take measures as to the effect of other inner and exterior factors. By understanding whether these factors will affect the operation and business-end of a mobile telecommunication problem, it will provide useful insights to such companies to make better decisions and changes such as implementing or upgrading their services to reduce churn rate while profit of operation increases.

2. Data Preparation

2.1 Rejected Attributes & Rationales

- PostalCode - Insufficient information to determine whether the area affects the churn rate.
- Hashcode - Rejected because it is a hash code for the postal code of a customer.
- Firstname - First name doesn't justify whether a customer will churn.
- CustomerID - Not required as it is a unique identifier for each customer.
- Rownumber - Not required as it is a unique row number for each customer.
- ChurnDate - Not used as **Churn** variable is used.
- TransactionValue - Not used as **TotalTransaction** variable is used.
- Birthday - Not used as **Age** variable is used.

2.2 Pre-processing Attributes with Methods for processing

Attribute Processed	Rationale for Processing	Methods for Processing
Transactionvalue => TotalTransaction	A customer may have many transactions during a period of time, hence, a derived value called "TotalTransaction" is created to store the customer's amount paid during that period.	SAS Studio - Using IF FIRST.variable and LAST.variable to calculate total amount for each customer.
CustomerID (termination dataset) => Terminate	The termination dataset only consists of customer ID that terminated the subscriptions; hence, a dummy variable "terminate" is created with value "1" or "0" as an indicator whether a customer has terminated the services.	SAS Studio - Using IF ELSE statement.
ChurnDate => Churn	The churn dataset consists of customer ID and churnDate, hence, a dummy variable "churn" is created with "1" or "0" as an indicator whether a customer has churn.	SAS Studio - Using IF ELSE statement.
Birthday => Age	Age is a demographic variable that can be used as an explanatory variable for the predictive model.	SAS Studio - Extract year from Birthday variable and subtract

		with the year of the dataset to get the age of each customer.
TotalTransaction	Grouping the values into groups of 4 instead of having a range of random values will lead to better model performance.	Grouping the variable using Transform Variable and Interactive Binning into groups of 4.
Age	Grouping the values into groups of 4 instead of having a range of random values will lead to better model performance.	Grouping the variable using Transform Variable and Interactive Binning into groups of 4.

2.3 Merging and Deletion

Five datasets were given for this assignment with each dataset consisting of different variables for the modelling process. After creating dummy variables on a certain dataset, all five datasets were merged by a common primary key which is customerID. Customer and Firstname mapping dataset were merged using the "firstname" variable as primary key and some values have leading blanks which have to be removed to properly merge both datasets. Once all datasets were merged, one observation was removed. SAS Studio was used to perform merging and deletion.

3. Data Modelling

3.1. Modeling Methods and Model Performance

No.	Modeling technique and Naming	Partition ratio	Partition Method	Other preparation methods applied	Model Performance
1.	Decision Tree (70,30TV)	Train: 70% Validation: 30%	Default	Connect after the Transform Variable Node, the property of Decision Tree is default.	Misclassification Rate: 0.096346
2.	Decision Tree (70,30IB)	Train: 70% Validation: 30%	Default	Connect after the Interactive Binning Node, the property of Decision Tree is default.	Misclassification Rate: 0.99668
3.	Logistic Regression (70,30TV)	Train: 70% Validation: 30%	Default	Connect after the Transform Variable Node, the property of Logistic Regression is default.	Misclassification Rate: 0.122924
4.	Logistic Regression (70,30IB)	Train: 70% Validation: 30%	Default	Connect after the Interactive Binning Node, the property of Logistic Regression is default.	Misclassification Rate: 0.122924

4. Justification and Rationale of Best Selected Model

Selected Model	Predecessor Node	Model Description	Target Variable	Selection Criterion: Valid: Misclassification Rate
Y	Tree2	Decision Tree (70,30TV)	REP_churn	0.096346
	Tree	Decision Tree (70,30IB)	REP_churn	0.099668
	Reg	Regression (70,30TV)	REP_churn	0.122924
	Reg2	Regression (70,30IB)	REP_churn	0.122924

Figure 1. Screenshot of the Model Comparison Node's Fit Statistics.

The best model was selected by using the Model Comparison Node in SAS Enterprise Miner. The selected model is a decision tree after "age" and "totalTransaction" was transformed using the Transform Variable node with 0.096% of misclassification rate, followed by a decision tree linking from the Interactive Binning node and two logistic regression models. The testing parameters used for the decision trees are default. The performance of decision tree node with Transform Variable node is slightly better than decision tree node with Interactive Binning node because the "age" variable was rejected as it did not exceed the cut-off value of Gini Statistics. Hence, the "age" variable was rejected.

5. Interpretation of Best Selected Model

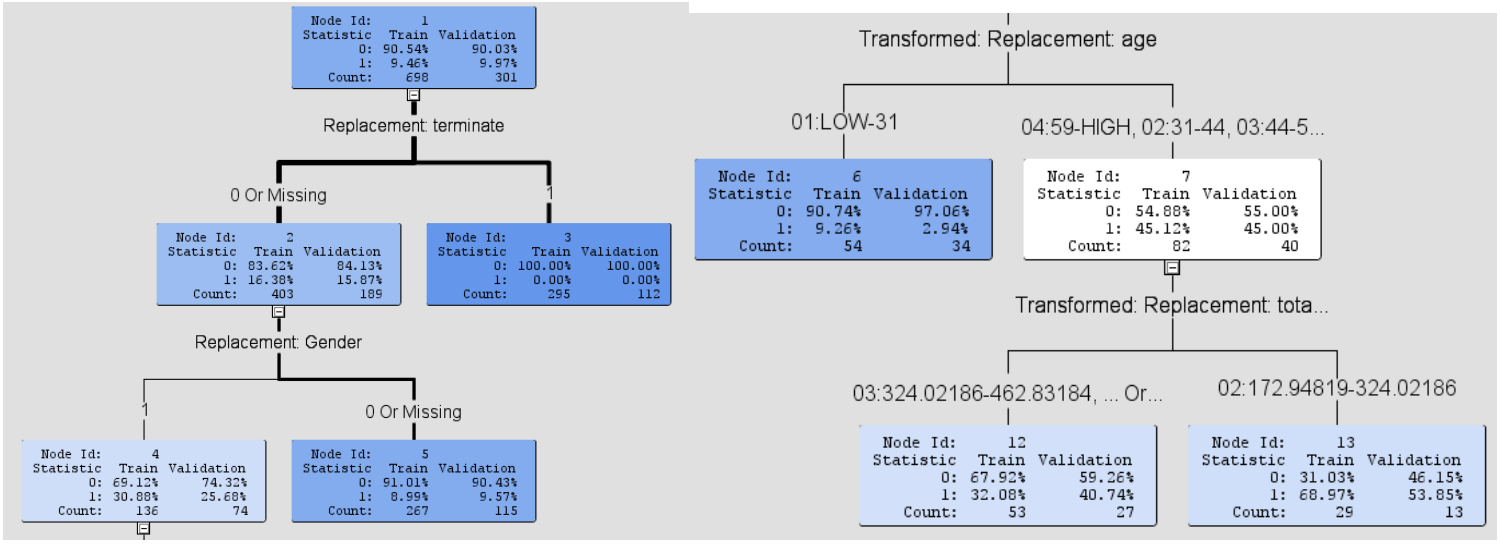

Figure 2. Decision Tree.

The result of this the best-selected model is performed using a decision tree as seen in *Figure 1*. Based on *Figure 2*, the decision tree has a total of 8 leaves with 4 splits, there are few interpretations to take note:

- **Termination (or expiry) of subscription affecting the churn rate**
  - **Results:** The first subset of the first split, corresponding to cases with termination=0, has a higher concentration of target customer churn rate=1. The second subset of the first split, corresponding

to cases with termination=1, has the maximum concentration of target customers churn rate=0. This model assigns to all cases in the left branch a predicted churn rate value equal to 0.1638 (16.38%) and all cases in the right branch a predicted churn rate value equal to 0 (0%).

- **Interpretation:** A customer may or may not churn based on their termination (or expiry) of subscription. For example, Node 2 identifies as the bigger proportion of observation in training data and count, which can be interpreted as customers who did not terminate (or expire) their subscription is 16.38% more likely to churn. Since the objective is to find out the churn rate of customers, Node 3 identifies as the worst-performing node, shows that all customers who terminate (or expire) their subscription are not going to churn. It could also indicate that the telco company does not keep any of the customer data to know whether they churned, hence the value of 100% not churning.

- **Gender of customer affecting the churn rate**
  - **Results:** The first subset of the second split, corresponding to cases with termination=0 and gender=1, has a higher concentration of target customer churn rate=1. The second subset of the second split, corresponding to cases with termination=0 and gender=0, has a higher concentration of target customer churn rate=0. This model assigns to all cases in the left branch a predicted churn rate value equal to 0.30 (30.88%) and all cases in the right branch a predicted churn rate value equal to 0.09 (8.99%).

  - **Interpretation:** The churn rate can be affected based on the customer's gender. For example, Node 4 with less training observation shows female customers who did not terminate (or expire) their subscription is 30.88% more likely to churn as compared to Node 5 with value of 8.99% of male customers to churn. Although Node 5 consists of more training observation, it shows that customers who terminated (or expired) their subscription and more likely to churn are most likely to be female

- **Age of customer affecting the churn rate**
  - **Results:** The first subset of the third split, corresponding to cases with termination=0, gender=1, and age=group 1, has a higher concentration of target customer churn rate=0. The second subset of the third split, corresponding to cases with termination=0 and gender=1 and age=group 4,2,3, has a higher concentration of target customer churn rate=1. This model assigns to all cases in the left branch a predicted churn rate value equal to 0.09 (9.26%) and all cases in the right branch a predicted churn rate value equal to 0.45 (45.12%).

  - **Interpretation:** The churn rate can be affected based on the customer's age. For example, Node 6 shows female customers who did not terminate (or expire) their subscription and aged between 1 to 31 is 9.26% value of churning. Interestingly, Node 7 identifies as the best-performing node with the highest churn rate of 45.12%, showing female customers aged between 31 to 92 are more likely to churn.

- **Total transactions of customer affecting the churn rate**
  - **Results:** The first subset of the last split, corresponding to cases with termination=0, gender=1, age=group 4,2,3 and total transaction=group 3,1,4 has a higher concentration of target customer churn rate=0. The second subset of the last split, corresponding to cases with termination=0 and gender=1, age=group 4,2,3, and total transaction=group 2, has a higher concentration of target customer churn rate=1. This model assigns to all cases in the left branch a predicted churn rate

value equal to 0.32 (32.08%) and all cases in the right branch a predicted churn rate value equal to 0.69 (68.97%).

- **Interpretation:** The churn rate can be affected based on the customer's total transactions. For example, Node 12 shows female customers who did not terminate (or expire) their subscription, aged between 31 to 92, and made a total transaction value between €2.50 to €173 and €324 to €1119 is 32.08% more likely to churn. Although Node 13 has the highest predicted value of 0.69 (68.97%) with female customers who did not terminate (or expire) their subscription, aged between 31 to 92, and made a total transaction value between €173 to €324, the decision tree model did not choose it as the best-performing model due to its low count and training and validation observations.

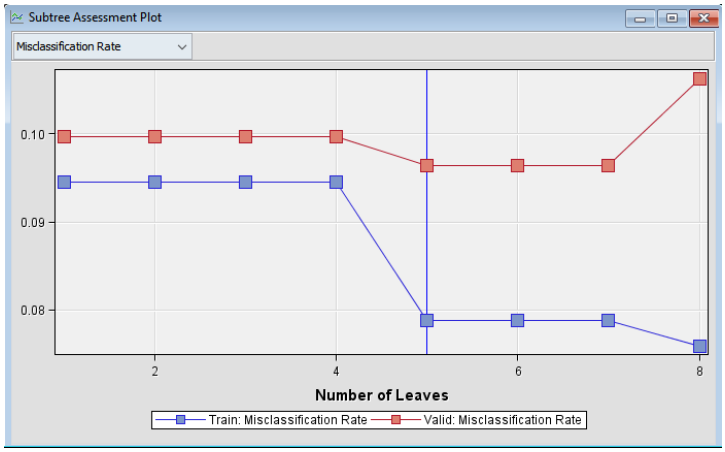

Figure 3. Subtree Assessment Plot - Misclassification Rate.

Looking at the plot for training data, the maximal 8-leaf tree is generated. It appears that the optimal leaf is the fifth-leaf tree, which generates a lower misclassification rate than any other leaf. The plot on training data seems to indicate that the fifth-leaf tree is preferred for assigning predictions to cases rather than selecting the eighth-leaf tree, where overfitting occurred.

6. Teammates Evaluation

Contribution / Member	Initiative, Ideas & Discussion	Predictive Modelling Process & Tasks	Final Report Writing	Group Presentation Content
Jarrod Tham Kuok Yew (Student ID: 16034753)				
Chia Mun Choon (Student ID: 16074536)	5	5	5	5

Appendix A

Screenshot of the Modelling Process Diagram

File Import
FINALCLEANDATA_TRAIN
Save Data
StatExplore
Metadata
Data Partition (70,30)
Replacement
Transform Variables
Interactive Binning
Decision Tree (70,30TV)
Regression (70,30TV)
Regression (70,30IB)
Decision Tree (70,30IB)

Appendix B

Decision Tree: Node Rules

```
*------------------------------------------------------------*
 Node = 3
*------------------------------------------------------------*
if Replacement: terminate IS ONE OF: 1
then
 Tree Node Identifier   = 3
 Number of Observations = 295
 Predicted: REP_churn=1 = 0.00
 Predicted: REP_churn=0 = 1.00

*------------------------------------------------------------*
 Node = 5
*------------------------------------------------------------*
if Replacement: terminate IS ONE OF: 0 or MISSING
AND Replacement: Gender IS ONE OF: 0 or MISSING
then
 Tree Node Identifier   = 5
 Number of Observations = 267
 Predicted: REP_churn=1 = 0.09
 Predicted: REP_churn=0 = 0.91

*------------------------------------------------------------*
 Node = 6
*------------------------------------------------------------*
if Transformed: Replacement: age IS ONE OF: 01:LOW-31
AND Replacement: terminate IS ONE OF: 0 or MISSING
AND Replacement: Gender IS ONE OF: 1
then
 Tree Node Identifier   = 6
 Number of Observations = 54
 Predicted: REP_churn=1 = 0.09
 Predicted: REP_churn=0 = 0.91
```

```
*-------------------------------------------------------------*
 Node = 12
*-------------------------------------------------------------*
if Transformed: Replacement: totalTransaction IS ONE OF: 03:324.02186-462.83184, 01:LOW-172.94819, 04:462.83184-HIGH or MISSING
AND Transformed: Replacement: age IS ONE OF: 04:59-HIGH, 02:31-44, 03:44-59 or MISSING
AND Replacement: terminate IS ONE OF: 0 or MISSING
AND Replacement: Gender IS ONE OF: 1
then
 Tree Node Identifier   = 12
 Number of Observations = 53
 Predicted: REP_churn=1 = 0.32
 Predicted: REP_churn=0 = 0.68

*-------------------------------------------------------------*
 Node = 13
*-------------------------------------------------------------*
if Transformed: Replacement: totalTransaction IS ONE OF: 02:172.94819-324.02186
AND Transformed: Replacement: age IS ONE OF: 04:59-HIGH, 02:31-44, 03:44-59 or MISSING
AND Replacement: terminate IS ONE OF: 0 or MISSING
AND Replacement: Gender IS ONE OF: 1
then
 Tree Node Identifier   = 13
 Number of Observations = 29
 Predicted: REP_churn=1 = 0.69
 Predicted: REP_churn=0 = 0.31
```