Scenario: EDA - Survival Data

Christine D Baxter

Keller Graduate School of Management—DeVry University

Professor Cynthia Stephen

Applications of Business Analytics II (BIAM 510)

April 3, 2021

Chart 1 Censored Status Proportions for Customer Tenure by Class/Category

Table

Description automatically generated with medium confidenceText

Description automatically generatedA picture containing text, receipt, screenshot

Description automatically generatedTable

Description automatically generated with low confidenceChart, bar chart, treemap chart

Description automatically generated

The graphs above, in Chart 1, focus on displaying the percentage of customers censored (1) or uncensored (0) by Rate Plan, Market Segments/Channels, and Cancel Type. For this data analysis, all customers/subscribers who were still active (or had an unknown cancel date) on the date of data retrieval were categorized as ‘Censored’. This strategy was implemented since their stop dates are an unknown point in the future.

A few items to note from this analysis:

* 81.5% of customers (censored and uncensored) fall within the Bottom Rate Plan (upper left graph)
* 50-55% of the market’s and channel’s customer base belong to Gotham (upper right graph) and Dealer (lower left graph) and both have a similar number of customers who cancelled their relationship with our company
* Both the Smallville market (upper right graph) and the Store channel (lower left graph) have a larger percentage (70-75%) of active customers
* 37% of customers are voluntarily canceling their subscriptions versus the 16% that are being involuntarily canceled (lower right graph)

Chart 2 Distribution of R10\_monthly\_fee by Censored Status

Chart

Description automatically generated

For Chart 2 above, the numeric monthly fee variable was rescaled using a Log 10 transform, creating the R10\_monthly\_fee variable. As shown in the graph, 50% of the monthly fees are approximately $40.00 or less. Please note that the base 10 logarithm of 40 is 1.6 or simply stated as 101.6. Lastly, the interquartile range of 1.4 ($25) to 1.7 ($50) covers 80% or less of the monthly fees.

Chart 3 Distribution of R10\_TenureMonths by Censored Status

Chart

Description automatically generated

Looking at Chart 3, you will notice that 50% of the customers have 20 tenure months or less. In addition, for the uncensored (0) status type, the interquartile range of 0.6 (4 months) to 1.6 (40 months) account for 80% or less of the tenure months. Lastly, for the censored (1) status type, the interquartile range of 0.8 (6 months) to 1.8 (63 months) account for 80% or less of the tenure months which also means that 20% are outside of the specified range.

Chart 4 Performance Evaluation Table

A picture containing text, receipt, screenshot

Description automatically generated

Chart 5 Decision Tree Diagram

Diagram

Description automatically generated

The performance evaluation table (Chart 4) has a root node error of 0.47285 and uses the MARKET, MONTHLY\_FEE, and TENUREMONTHS variables. As you move down the complexity parameter (CP) column the values move closer to zero which can signify the risk of over-fitting. In addition, within the Relative Error (rel error) column, each row in the table does not have a significant increase when compared to their Cross-validation Error column (xerror) counterpart decreasing the possibility of model instability. Lastly, this figure illustrates a minimal split value equal to 20, a minimum bucket of 7, a maximum depth of 30 and a complexity value of 0.01.

As noted in Chart 5, the decision tree diagram contains four levels and answers three questions as it breaks down the data to a leaf level. The CENSORSTATUS variable is displayed as a ‘1’ (censored) or a ‘0’ (not censored) at the top of each shaded box. The root node (or top box) contains 100% of the training data and the first split in the tree is on tenure months. Customers with less than 15 tenure months take one path whereas those with greater than or equal to 15 tenure months take another. The remaining questions/subsequent splits continue to concentrate the classes.

# References

R Core Team (2021). R: A language and environment for statistical  
computing. R Foundation for Statistical Computing, Vienna, Austria.   
URL https://www.R-project.org/.

Berry, G.S.L.M. J. (2010). *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*. [devry]. Retrieved from https://devry.vitalsource.com/#/books/9781118275603/