Capstone Project: Customer Churn

Tricia M. Lang

DAT 690 – Capstone in Data Analytics

Dr. P. Sheldon

18 June 2023

**Project Summary and Analytic Plan**

**Business Understanding Phase**

GE Healthcare has had a significant number of customers leave their smartphone service. This loss in customer base has negatively affected the bottom line of the company. GE Healthcare would like to get a better handle on why their customers are leaving and potentially identify ways to remediate any issues. The question to be answered with this model is: what the likelihood is that a given customer will leave, or churn. A predictive model will be created in order to identify the customers who have a higher probability of leaving the service.

The model will help GE Healthcare to identify which factors have an impact, and the magnitude of that impact, on whether an individual will churn or not. Each of the factors identified in the model are ones that are significantly related to the outcomes of “churn” or “not churn”. The degree to which that variable is impactful is noted by the size of the coefficient within the formula. In identifying which factors are relevant, the company will be able to target advertising, improvements to the network, or features for their product in order to obtain and retain customers more effectively.

**Data Understanding Phase**

The data set has been provided by GE Healthcare in a csv file format. The variable definitions will be reviewed to identify which variables are continuous, categorical, and key identifiers. Data points that have the potential for multicollinearity will also be identified. The data will be assessed in Excel with missing or invalid data points identified. Descriptive statistics will calculate the measures of central tendency and the data points compared to those in order to identify those points that are not consistent with the data set.

**Data Preparation Phase**

The data will be prepared by thoroughly cleaning and formatting as necessary. Missing values, errors, and outliers will be handled through removal, imputation, or another appropriate transformation. The preparation of the data will create a data set that will be useful for building the model.

**Modeling Phase**

The predictive model that will be built is a linear regression using RapidMiner. The data set will be divided 70/30 for train/test sets. A linear regression model will be used to separate out the significant and irrelevant variables. The best performing models created will then be compared to determine which performs the best when using the test data set. This will be measured using an ROC/AUC, Root Mean Square Error, Adjusted R Square, and Mean Absolute Error. These measures will identify the goodness of fit of the linear regression model and the errors in the prediction.

**Evaluation Phase**

The results of the model will be evaluated for how well they address what the business needs. Will the model provide GE Healthcare the ability to predict the probability of a customer churning? Reviewing the business needs and the data the model should be able to meet requirements for acceptance by the company before moving to the next phase.

**Deployment Phase**

The completed model will be reported out with the intention of being put in production by GE Healthcare. The results of the model testing will be reported, indicating the confidence in the outcomes. The documentation for how this model is to be implemented and used by GE Healthcare will be provided. The final model will be handed over for integration in GE Healthcare’s reporting and analytics system.

**Data Understanding Phase**

**Data Selection**

GE Healthcare is looking to retain customers. They have provided a data file of customer information including demographic, interactions with the service and with customer service, and financial. The file contains 850 rows of data with 84 variables. 81 of the variables are potential predictor variables and others identifiers and the target variable. Approximately half of the predictor variables are categorical and half numerical.

The variables that would be most relevant to this would be around the performance of the product for the customers as well as the ease with which they would be able to leave for another service, even if it were at a greater cost. The product is also highly specialized, targeting healthcare providers who would benefit from a smartphone and app that would increase productivity and facilitate communication. In order to achieve their business goal, most of the variables provided have relevancy to the potential use for reducing customer churn.

**Descriptive Statistics and Correlation Analysis**

For the numerical variables, descriptive statistics of mean, median, mode, variance, standard deviation, and quartile information were calculated. Many variables were shown to be skewed, with variables having interquartile ranges driving the measure of outliers to be illogical, such as negative values for those variables where negative values are not possible. Outliers and skewed data were evaluated and not eliminated as the impact on the predictive model for a customer churning may be significant. The relationship of those who have minutes used greater than four times the standard deviation may have a significant impact on the likelihood of churning or not churning. While these numbers may be above the mean and fall within the fourth quartile, they are not clearly erroneous or irrelevant.

Other values that were not included are those that were duplicative. The credit rating, Prizm codes, marital status, and occupations were all broken down by specific ranges and also accounted for in a combined variable. Variables of CREDIT\_RATING for the ranked credit rating, PRZM\_NUM for Prizm code – unified, and OCC for Occupation – Unified, were kept while the individual subdivisions of each, such as “Occupation – professional, Occupation – clerical”, etc. were eliminated.

In performing a correlation analysis within Excel, there were several variables that showed high correlations. With a cutoff of +/- 0.80, seven interactions were noted to be correlated, with results shown in Figure 1. The relationships between some of these variables are clear, such as RETCALLS, the number of calls to the retention center, and RETCALL, whether a call was made to the retention center or not. The categorical variable of RETCALL can be omitted as this is a duplication of capturing the data around calls to the retention center. If no calls were made, the value for RETCALLS as well as RETCALL are 0.

Figure 1. Correlations between Predictor Variables



MAILORD, identifying those who have purchased via mail order, and MAILRES, those who respond to mail order solicitations, are also highly correlated. None of those who have purchased via mail order are identified as not responding to mail order solicitations. Only 12 individuals are noted to respond to mail order solicitations but not purchase by mail order. The variable MAILORD was therefore retained and MAILRES removed as being duplicative.

Other variables, such as OWNRENT, value for homeownership is missing, and CREDITCD, the individual possesses a credit card, are less obviously related. Given the definition of OWNRENT, and there not being other variables identified capturing homeownership or renting status, this data point was removed as it did not appear relevant or reliable as a predictor variable. Other variables with missing data were handled with the value of 0, such as handset price and income.

The values of DROPVCE, number of dropped voice calls, and DROPBLK, number of blocked voice calls, were noted as being highly correlated with a correlation coefficient of 0.84137. Both of these variables are continuous number variables. There is no clear relationship between dropped and blocked calls. Both of these variables were maintained in the data set.

The variables PHONES and MODELS correspond with the number of phones that have been issued and the number of different models. These were very highly correlated. A combined variable that divided the number of phones by the number of different models was created to capture this, named PHOMO. This provided a more normal distribution of the data points without losing the integrity of the data.

**Results of Preliminary Analysis**

The preliminary analysis included the evaluation of the descriptive statistics for all numeric variables. The mean, median, mode, range, and standard deviation were evaluated along with the quartiles, interquartile ranges, and specific percentiles for some data where the skew appeared to be greatest.

The mean monthly minutes of use (MOU) was one variable that showed great variance. The range was from 0 to 3656.25 and the mean 540.89. The standard deviation of this variable was 542.88 and the 3rd quartile calculated to be 766.5. This would indicate that 25% of the data falls between 766.5 and 3656.25. Visualizing this skew can be seen in the figure below.

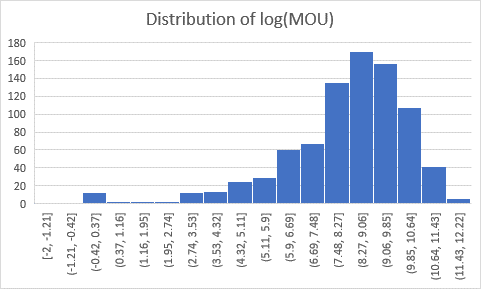
Figure 2. Distribution of MOU data points.

A picture containing text, screenshot, font, number

Description automatically generated

The 99th percentile was with a result of 2547.92, which includes 8 data points. A log transformation with base 2 shows the points of 0 as being outliers, as illustrated in Figure 3. This distribution, however, is significantly more normalized.

Figure 3. Distribution of log transformation of MOU.



**Data Preparation Phase**

The data selected for the predictive model excludes some variables based on the preliminary statistics. There were numerous values for AGE1 (the age of the primary household member) that were 0, which is not a logical option. The corresponding AGE2 values for the second household member were also 0 for each of these selections. It is unclear whether this indicates that the person lives alone or if the data were unavailable. For this reason, AGE2 was eliminated from the calculations as the data integrity issue was significant.

The values for AGE1 that were 0 were substituted with value of the median and mode, 43, which was very close to the mean of 42 prior to the imputation. Histograms of the distributions pre- and post- substitution can be seen in the figures below.

Figure 4. AGE1 Distribution from raw data set

Chart, histogram

Description automatically generated

Figure 5. AGE1 distribution after substituting median/mode for values of 0.Chart, histogram

Description automatically generated

Additionally, several variables were inconsistent, such as negative values for the variable EQPDAYS, which measures the number of days a customer has had their equipment. This value was substituted with the median of 317, which had no impact on the mean or median, as well as the quartiles. The mean, 382, is noted to fall in the fourth quartile, indicating this data is skewed even with the value of -1, with little to no impact of substituting a value that is the median (317) and very close to the mode (316).

The data points of NEWCELLY and NEWCELLN were combined into a third variable, NEWBIE, that captured whether the cell phone owner was known to be a new cell phone user (NEWCELLY=1), known to not be a new cell phone user (NEWCELLN=1), or unknown (NEWCELLY=0 & NEWCELLN=0). The categorical values for this are 0 = no, 1 = yes, 2 = unknown. The NEWCELLY and NEWCELLN variables were then removed from the data set.

Interaction effects were also examined, initially assessing factors that may appear to be connected, such as income and credit rating or profession and income. None of the interactions explored showed any significant impact on the outcome variable.

**Modeling Phase**

**Artifacts**

Graphs and tables explaining preliminary results from the model. Confusion matrix or other outputs. (3-5 sentences)

Running the model for a logistic regression with the variables identified in the data cleaning and preparation phase created a model with an accuracy of 70.60% and an AUC of 0.6939.

Figure 1. First Iteration Results

A picture containing text, screenshot, line, font

Description automatically generated

A screenshot of a computer program

Description automatically generated with low confidence

CM: Confusion Matrix (Row labels: Actual class; Column labels: Predicted class):

0 1 Error Rate

0 384 200 0.3425 200 / 584

1 95 168 0.3612 95 / 263

Totals 479 368 0.3483 295 / 847

A screenshot of a computer code

Description automatically generated with low confidence

A screenshot of a computer

Description automatically generated with medium confidence

Weight by Gini Index and AttributeWeights (Logistic Regression)

A picture containing text, screenshot, font, number

Description automatically generated A picture containing text, screenshot, number, font

Description automatically generated

Comparing the Weights by Gini Index and Attribute Weights, several variables showed on both lists as having weights of 0. These variables, CREDITCD, MARRY, MAILFLAG, PCOWN, RV, TRUCK, MCYCLE, and NEWBIE, were removed for the next iteration. The impact was a small drop in the AUC and no change in accuracy.

Figure 2. Second Iteration Results

A screenshot of a computer

Description automatically generated with low confidence

A screenshot of a computer

Description automatically generated with medium confidence

Comparing weights of Gini Index and Attributes again led to several more variables identified as having minimal impact on the model, as shown below. The next variables removed were REFER, CREDITAD, CHILDREN, UNIQSUBS, and MAILORD.

Figure 3. Comparison of Weight by Gini Index and AttributeWeights (Logistic Regression)

A picture containing text, screenshot, font, number

Description automatically generated A picture containing text, screenshot, font, number

Description automatically generated

This resulted in a poorer accuracy and AUC, so all removed variables were placed back in the model, resulting in the AUC of 0.704 and an accuracy of 70.60% +/- 3.23%. This model includes 55 predictors and 8 iterations to arrive at this model. The RapidMiner configuration for this model is shown below.

Figure 4. RapidMiner Model

A screenshot of a computer

Description automatically generated with low confidence

**Data Quality**

After initial modeling, the variables for INCOME as well as SETPRC (price of the phone) were deleted. Both had extensive values of 0, indicative of the data points missing. Additionally, as INCOME was a categorical variable there were no clear indications of the values for income ranges of the categories, rendering this variable uninterpretable. Returning to the stakeholders to inquire about this and whether it might be relevant would be warranted.

Missing values for Age1 were imputed with the average, which was also very close to the mode and median values. This change had no impact on the model and the contribution to the model by Age1 remained relevant but not skewed.

**Data Structure**

The data structures were attributed to each specific variable. In this data set, the outcome variable of CHURNDEP is a binomial categorical variable. Other categorical variables are also binomial, such as CHILDREN and TRAVEL, where 1 is yes and 0 is no. The values for categorical variables MARRY (marital status), OCC (occupation), and PRZM\_NUM (neighborhood type) are polynomial categorical values, where each number relates to a category of an answer, such as 0 is no, 1 is yes, and 2 is unknown. Numeric values for other variables, such as REVENUE and combined variables like PHOMO (phones divided by models), are also coded as such in the RapidMiner model.

**Evaluation Phase**

**Model Evaluation**

The purpose of the model is to accurately predict whether a customer would churn given the measures available. This model has done acceptably well in performing this task. With the accuracy of 70.60% the model was able to predict the accurate outcome at a much greater percentage than by chance alone. The AUC had a value of 0.704, demonstrating the model was able to distinguish the true from false positives adequately.

**Areas of Concern**

The data points that were regarded as more impactful on the model were logical, such as those related to the quality of the service or the utilization of the services. The variables that had less of an impact on churn were also logical, such as marital status or having children. Only the variable of occupation, with the category of professional specifically having minimal impact, was surprising. Given the product is geared towards individuals who are in the medical profession, it would have seemed that professionals would be less likely to churn than non-professionals, and as such the category of occupation as professional was presumed to have a greater impact on a customer not churning.

**Fit of Model**

Throughout the model creation, the statistics that were easily apparent for evaluation included the confusion matrix, AUC, and accuracy measures for the model. For individual variables, the mean decrease gini and attribute weight were used as measures of impact for the model. Through each iteration, the confusion matrix statistics and the AUC were evaluated to determine whether subsequent iterations improved or reduced the performance of the model. Individual variables were identified as being more or less impactful.

The model was determined to not be overfit as none of the variables were overly weighted than others. Additional measures were taken to reduce the correlation between the variables. A linear regression for the predictor variables to identify the direct relationship with the outcome variable was also evaluated in the data preparation phase.

**Model Results**

The predictor variables were able to give a reasonably accurate prediction for the outcome variable. The data points were normalized and weights by Gini Index were applied. Further evaluation of whether the model can be improved with having a combined variable may be examined.

The model was able to predict true positives 60.84% of the time and true negatives 72.6% of the time. The MSE was calculated to be 0.1889 and RMSE 0.4346. Both of these values indicate the data points are not far from the model’s regression line. This would indicate the model is satisfactory in predicting the outcomes based on the values of the variables.

Figure 5. Model Evaluation Statistics

A screenshot of a computer

Description automatically generated with medium confidence

Appendix A

# DAT 690 Production Turnover Report Template

**Date:** June 11, 2023

**Business Department:** Sales/Marketing

**Project Name:** GE Healthcare Customer Churn Predictive Model

**Project Description:** Predictive model for the identification of customers who are likely to churn.

**Model Baselines:** Expected outcomes would approximate the testing and verification results:

Accuracy: 67.98%- training; 78.67%- verification

Precision: 44.44%- training; 33.33%- verification

AUC: 0.582- training; 0.558- verification

**Model Performance:** Modifications to variables are not anticipated at the time of initial deployment. The execution of the model is not anticipated to take longer than 15 minutes for a dataset of approximately 50,000 rows of data.

**Comments:** The outputs the business users will be using predictions of churn to identify those who are likely to leave the service.

Appendix B

A picture containing text, diagram, screenshot, line

Description automatically generated

1. **Modeling Process Flowchart**
2. **Data Cleaning**
3. The data will be provided by GE Healthcare. The csv file will be imported into Excel. Save the data sheet as a file name that indicates it is a working document (i.e.: DataSet-Working.csv)
4. Each of the variables and the type of data will be identified. Variables will be noted as categorical or numerical. Categorical variables will be nominal, such as marital status, or ordinal, such as a ranked order. Numerical variables will be continuous or discrete.
5. Missing values will be identified within the csv file.
   1. Select all data.
   2. From Home tab, go to Edit, Go To Special

Graphical user interface, application

Description automatically generated

* 1. Select “blanks”

Graphical user interface

Description automatically generated

* 1. Click on highlighter to identify all empty cells
  2. If no missing values, continue to step 6. Otherwise, continue to step 4.

1. For each of the missing data points, identify whether there are numerous missing data points or if this is a rare event for both the column and the row.
2. Address missing data and document in what way they were addressed.
   1. Rare events may be managed by ignoring or imputing the data with the mean or mode.
   2. Frequently occurring empty cells may warrant removing that row or column.
3. Calculate the descriptive statistics, primarily measures of central tendency, for each variable. This would include range (min/max), mean, median, mode, variance, and standard deviation for numerical variables and the mode for categorical variables.
4. Identify any errors or invalid data points, such as negative numbers when not possible (i.e. age) or characters for expected numbers.
5. Identify any data errors for categorical variables, such as numbers that do not fall within the range of data points. If none, continue to step 11. Otherwise, continue to step 9.
6. Address data errors and document way in which they were addressed.
   1. Identify whether handling of the invalid data points would best be done by removing, modifying, or transforming.
   2. Ignoring these data points is not an option.
7. Return to step 6.
8. Identify any outlier data points.
   1. Note data points that are greater than or less than one standard deviation of the mean.
   2. Determine quartiles in Excel using quartile function.
      1. Sort data column smallest to largest, expanding set to keep all data together.
      2. Determine quartiles using quartile function (=quartile.inc(range, quartile #))
      3. Calculate IQR (Q3-Q1)
      4. Calculate upper and lower limits (Q1 - (1.5\*IQR)), (Q3 + (1.5\*IQR))
      5. Highlight cells that do not fall in the range of the LL to UL, these are the outliers.

Graphical user interface, application, table, Excel

Description automatically generated

1. If no outliers, continue to step 13. Otherwise, examine these outliers to determine if they are data errors to be handled or if they should be removed. Return to step 6.
2. Save the data set. Once saved, Save As with a file name indicating it is a clean data set (i.e. DataSet-clean.csv).
3. Delete the calculations such as quartiles and measures of central tendency.
4. File is now ready to be uploaded into RapidMiner for modeling.
5. **Data Modeling**
6. Load cleaned data set into RapidMiner.
   1. Format columns so as to mark the outcome variable as a label and any keys as an id variable.
   2. Check that the variables are loaded as the appropriate variable type (binomial, polynomial, integer, etc.).

A screenshot of a computer

Description automatically generated with medium confidence

1. Once loaded, visually inspect the description of the data points loaded for any overlooked missing variables or blank cells.
2. Review the distributions of the data points.
3. Check for collinearity through a correlation matrix.
   1. As this was done in the previous data cleaning step, verify there are no correlations out of range previously determined. For example, in this instance +/- 0.80.

A screenshot of a computer

Description automatically generated with low confidence

A screenshot of a computer

Description automatically generated with low confidence

1. Run model for the complete data set for generalized linear model.
2. Note the p-values for variables in this model.
3. If there are no p-values < 0.05, return to step 1b. Otherwise continue to step 8.
4. De-select all attributes for the model other than the lowest p-value variable and run the model.
5. If there are no more variables, continue to step 11. Otherwise, add the variable with the next lowest p-value and run the model.
6. If the p-value of the most recently added variable is > 0.05, if it is not, leave it in the model. Return to step 9. If there are no variables in the model with a p-value > 0.05, return to step 9.
7. Assess the goodness of fit of the model for predicting the outcome. If model is satisfactory, continue to step 12, otherwise return to 1b.
   1. Error matrix for quality of predictions.
   2. AUC/ROC for ability to reliably predict outcome.
8. Save the model.
9. Deploy model.