Customer Churn Analysis

T. Lang

DAT 650 – Advanced Data Analytics

Dr. P. Sheldon

2 April 2023

**Data Sources and Business Value**

GE Healthcare provided data to be evaluated for the identification of possible factors related to customer churn for the smartphone application and service. The data set includes 436 rows of data, with each row representing one customer. The data was collected from application information, billing, and smartphone use data. The variables reported include whether the customer remained or left the service, information on the usage of the smartphone, customer demographic information, and history of cell phone use.

GE touts their data-based decision making and utilize data in business decisions big and small (Ferguson, 2014). The company is invested in finding use for the data in commercial operations. Through analysis, data-driven business choices in sales and strategy lead GE’s business practices. The data that has been supplied by GE Healthcare is anticipated to open a greater understanding of the problem of customer churn, identifying factors that lead to customers remaining with the service, and components of the service provided that could be addressed in order to retain customers.

**CRISP-DM Application**

The six phases of the CRISP-DM methodology will be used throughout this data analytics project. With the first step being business understanding, the research on the product by GE, the target audience, the business objectives, and discussions with the various stakeholders will provide the outline for how this project will meet business objectives. An article by Shearer (2000) elaborates on the business understanding phase, indicating this is the opportunity to identify the requirements, assumptions, and constraints of the situation, as well as understand what the measure of success will be. For GE, the goal of this project is to develop a predictive model that will accurately identify whether a given customer will remain with the service or leave (churn).

The second phase is data understanding. The data that has been provided by GE Healthcare will be evaluated for the ability to meet the goal of an accurate predictive model. The data will be viewed for any gaps in factors that may be influential or that should be investigated further for relevance. Exploratory analysis will be completed within the scope of the pilot project, indicating whether the factors evaluated are appropriate for the model (Prado, 2020). Once the data is understood, it can then be prepared in the next phase. In this phase the data will be described completely, as well as evaluated for accuracy, missing values, and formatted appropriately to be used.

The final three phases are to build the model, evaluate the model, and then deploy the model. For the pilot project, the model created will identify whether the data set can be used to create an appropriate predictive model as part of the testing. Once this has been completed to the satisfaction of the objective, the predictive model can be built. Through multiple iterations of the model, with the potential to revisit phases, such as understanding the business needs or identifying issues with the data set that bring the process back to data understanding, the final model will be created. The final step will be to deploy the final predictive model.

**Purpose of Analytic Structure**

Descriptive analytics look to identify what has occurred, such as a count of a number of events, a dollar amount spent or earned, or a trend (Ghosh, 2022). This structure is useful in describing the data set without identifying any correlative relationships. Predictive analytics seek the correlations between factors to find what data points are related to the outcomes. This allows for the possibility of predicting outcomes based on the presence or absence of specific factors.

The data set from GE Healthcare can be summarized with descriptive analytics. The data points can be aggregated to show the number of customers lost in specific timeframes or the quantity of various marital statuses of customers who remain with the service. This will not allow inferences to be drawn from the data, though, which would rely on predictive analytics. A predictive model will allow for the determination of the probability of an individual who has a specific set of characteristics remaining with the smartphone service or churning. This will allow business decisions to be made around marketing, pricing, or other ways to address identified factors that significantly impact customer churn.

**Tool Selection**

Several tools are available to complete the analysis. The data will be received as a csv file. Using Excel the data will be cleaned and validated. Once this has been completed, the analysis will be run through RapidMiner. Results from the analysis will be presented in PowerBI for the purposes of clean visualization and the potential use of a dashboard feature for GE.

Excel is a readily available Microsoft product that can import and export csv files. As part of the Microsoft Office suite of applications, it is anticipated all employees of GE IT will be able to create and submit the csv file for analysis. RapidMiner is an open source data analytics tool that has numerous analysis functions (Pynam, et al, 2018). It will allow the preliminary evaluation of data through tools such as a decision tree, as well as more advanced analytics, such as logistical regression.

PowerBI is a tool that allows for comprehensive dashboards with potential for drilling down or up for further information. Once results are obtained, presentation with PowerBI may be helpful to indicate the relative significance of various factors and descriptive statistics about those factors. Presenting in this format would allow for an easier understanding by those who are in decision-making capacities for the business.

**Additional Data Fields**

Consideration may be made for adding data fields to the data set, if available, should there be factors that seem to be impactful and in proximity to other factors. The GE Healthcare smartphones are designed for the medical professional and it may be helpful to know the setting the individual works in, such as hospital, urgent care, family practice, etc. Additionally, the role of the individual in the field may be relevant, such as whether the individual is a doctor, nurse, or a tech, with the possibility of further classifications, such as emergency room, cardiac care, or ICU. These factors may help narrow the focus of target customers.

**Ethical Implications**

It is suggested by Satz that to develop a model with ethical guidelines “will take collaboration between programmers, statisticians, legal scholars, and philosophers” (n.d.). Though including all of those individuals in determining the ethical considerations of the GE predictive model may be beyond the scope of this project, the need for considering legal and statistical needs is important.

Within the dataset provided by GE are variables such as credit rating, ages of household members, and prizm code, which is aa code for the PRIZM NE “urbanization model” (Claritas, 2004). The prizm code is noted to be capable of identifying profiles of customers based on segmenting the areas where they live, lending insight into the shopping habits, routines, and connections individuals in those segments have based on location.

Categorizing people in the way of the prizm code may present ethical issues, given there are potential assumptions about individuals who live in different areas of varying population density. Coupled with other information, such as credit score and age, the potential for profiling individuals based on unintended factors, such as race, may increase, as systemic racial bias issues arise with certain assumptions and increase the potential for discrimination (Campisi & Lupini, 2021). The authors continue by saying that errors in credit reports may negatively impact the credit scores of Black consumers disproportionately. It may be beneficial to run the predictive model including and excluding the prizm code and credit score to identify whether there is an impact on the outcome.

**Ethical Recommendations**

The ethical implications of the model using factors that may highlight a bias for or against specific categories of individuals based on race, gender, age, or other protected factors, can be significant. Should a resolution for the problem of churn be to offer individuals within specific geographic areas a discount, this may lend to discriminatory practices. Data points that may be more relevant to identify the individuals to be targeted may include the cost of housing, rather than homeownership, which is also disproportionately biased against Black and Hispanic individuals (Campisi & Lupini, 2021).

Age separated into categories rather than a continuous variable may also identify whether individuals who are early-career professionals, mid-career professionals, senior-career professionals, or retired, may be a better indicator than age without a potential for bias based on age. Legal consultation may be advisable as well, due to the potential for implications legally for perceived bias.

**Model Creation**

**Applicability**

Descriptive statistics about the data set will provide a better understanding of the data. Measures of central tendency and variance will show whether the data are normally distributed or if there is a skew. Skewed data may present issues with regression modeling as the tails can be treated as outliers (Sharma & Whitfield, 2022), which would lead to the model performing poorly. This is also pointed out by Weiss (2004), who noted, “Noise is a known factor that usually affects models performance. In imbalanced domains, noisy data has a greater impact on the least represented examples.” Data that is skewed would need to be evaluated further for potential data transformation should a regression model be implemented so as to reduce the impact of the skewness.

A random forest model is a collection of decision trees from randomly selected data within the set. This algorithm can be executed over large data sets quickly and effectively. A random forest model is the most efficient way to manage having imbalanced data sets. Kirasich, et al. (2018) compared random forest models, both weighted and balanced, and found both performed well when evaluating skewed data sets. One study comparing the random forest and logistic regression found them to be comparable for data sets with fewer than 1000 observations, but the random forest performed better in situations with a greater number of noise variables (Chen, et al., 2004). Other studies, such as that by Wålinder (2014), showed no significant difference between logistic regressions and random forest models. The ease of executing a random forest model on a data set, coupled with the management of potentially skewed data without a need for transformation indicates this would be an effective pilot model for the GE data set.

**Value**

Analysis of customer churn will provide significant value to GE. Understanding the reasons for customers remaining or leaving the service provides many benefits to the company, including those pointed out in a blog post by Ramchadndra (2022): improving profits, improving customer experience, the ability to customize products and services, and increasing customer retention. Through predictive modeling, the predictor variables of customer churn can be identified, providing insights into the customers who remain and those who leave the GE service plan.

Predictive modeling of customer churn has been studied previously. Of note, Li & Li (2019) have completed a literature review of various methods to address customer churn. Evaluating various modeling methods, including logistic regression, XGBoost algorithm, and a hybrid of the two. Upon evaluation of the accuracy, precision, and recall, the hybrid model was identified as more accurate than logistic regression alone.

For GE, preliminary results from a pilot evaluation of the data with a random forest model will demonstrate if there is potential for the data set to be effective in a predictive model. Once complete, further data preparation and evaluation with a hybrid logistic regression-XGBoost algorithm is anticipated to provide a model allowing GE to predict customer churn.

**Pilot Plan**

To establish the potential for predicting customer churn a random forest will be run on the dataset provided by GE Healthcare. Following cleansing of the data, the variables with potential multicollinearity will be removed and the random forest model run. An initial random forest will be run to identify a general assessment of accuracy and precision through evaluating the Area under the Curve (AUC) as well as the error matrix. Additionally, variable curation will begin with the identification of important variables through the Mean Decrease Gini.

Iterations of this process will continue by removing and changing the variables that are evaluated in the random forest, with measures from the AUC and error matrix indicating whether the recall, precision, accuracy, and specificity indicate the variables are impactful on the outcome. Guidance for which variables to include or exclude will be based on which are scoring high on the Mean Decrease Gini. Additional consideration will be made for variables that are relevant to answering the overall question: what variables predict a customer’s likelihood of churning?

**Pilot Test**

A random forest was run with the intent of identifying relevant factors leading to customer churn. Using Rattle and R, the data set was loaded and evaluated for missing variables or unexpected values. Missing values were few and imputed as the error rate decreased in doing so compared to removing those rows of data entirely. Variables that were identified as correlated were ignored to avoid overfitting due to multicollinearity. These variables included overlap with credit, occupation, and notifications of missing data for other variables.

The initial random forest created had a significantly lower Area Under the Curve (AUC) than was reasonable at .5107. The most important variable identified was mean monthly minutes of use, followed by change in minutes of use and number of days the individual had the equipment. These variables were then used as building blocks for the next iteration.

Variables about the service that were unable to be altered were excluded, such as dropped calls, as the cell tower reach and reliability is not in the control of GE Healthcare. A random forest with these factors removed were improved. Identifying the most important variables and reducing the random forest showed some improvement to the model.

Running the random forest with only the 10 most important identified variables increased the AUC to 0.6141, but the overall error rate increased to 51.2%. This indicates that with these ten variables, the model was able to better differentiate positive and negative outcomes, but greater inaccuracy with predicting the negative outcomes consistently. Diversity of variables that have impact on the outcome of churn as well as not churn may be a reason for this.

The random forest run with a set of the 22 variables identified as most important resulted in an AUC of 0.61 and an error rate of 39.88%. This random forest was built creating 500 trees and using 6 variables at the decision points and the data set being split 70/30/0. The variables of importance were identified with several having significantly greater impact on the decision tree, such as whether the customer has purchased from mail order, the number of days they have owned the current equipment, and the percentage change in minutes. The mean decrease in Gini, a measure of the importance of the variable within the random forest (Louppe, et al, 2013), was higher for these variables. The percent change in minutes and number of days with the current equipment had the highest values at 12.69 and 13.42 respectively.

This model shows there is significant potential for further exploration of the variables related to customer churn. The variables identified may inform the company on areas that may need improvement that ultimately contribute to customer churn. Further predictive modeling may help GE Healthcare focus efforts on specific areas that could reduce customer churn overall.

**References**

Campisi, N., & Lupini, C. (2021, February 26). *From inherent racial bias to incorrect data-the problems with current credit scoring models*. Forbes. Retrieved February 14, 2023, from https://www.forbes.com/advisor/credit-cards/from-inherent-racial-bias-to-incorrect-data-the-problems-with-current-credit-scoring-models/

Chen, C., Liaw, A., & Breiman, L. (2004). *Using random forest to learn imbalanced data*. University of California Berkley. Retrieved February 28, 2023, from https://statistics.berkeley.edu/sites/default/files/tech-reports/666.pdf

Claritas. (2004). *PRIZM NE Methodology Summary*. UVM.edu. Retrieved February 14, 2023, from https://www.uvm.edu/rsenr/gradgis/advanced/prizm\_method.pdf

Ferguson, R. B. (2014, January 27). *GE and the Culture of Analytics*. MIT Sloan Management Review. Retrieved February 10, 2023, from https://ezproxy.snhu.edu/login?qurl=https%3A%2F%2Fwww.proquest.com%2Fscholarly-journals%2Fge-culture-analytics%2Fdocview%2F1514708633%2Fse-2%3Faccountid%3D3783

Ghosh, P. (2022, February 11). *Fundamentals of Descriptive Analytics*. DATAVERSITY. Retrieved February 10, 2023, from https://www.dataversity.net/fundamentals-descriptive-analytics/#

Kirasich, K., Smith, T., & Sadler, B. (2018). Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets . *SMU Data Science Review*, *1*(3), 9–1-9–24. Retrieved from: https://scholar.smu.edu/cgi/viewcontent.cgi?article=1041&context=datasciencereview

Li, X., & Li, Z. (2019). A hybrid prediction model for e-commerce customer churn based on logistic regression and extreme gradient boosting algorithm. *Ingénierie Des Systèmes d Information*, *24*(5), 525–530. https://doi.org/10.18280/isi.240510 Retrieved from: https://eds-s-ebscohost-com.ezproxy.snhu.edu/eds/pdfviewer/pdfviewer?vid=3&sid=076e06ab-72f1-47f5-b088-cf6bdc24f160%40redis

Louppe, G., Wehenkel, L., Sutera, A., & Geurts, P. (2013). *Understanding variable importances in forests of randomized trees - neurips*. NeurIPS. Retrieved March 10, 2023, from https://proceedings.neurips.cc/paper/2013/file/e3796ae838835da0b6f6ea37bcf8bcb7-Paper.pdf

Prado, H. M. (2020, December 12). *Predicting Churning Customers Using CRISP-DM Methodology*. Medium. Retrieved February 10, 2023, from https://hmartins-prado.medium.com/predicting-churning-customers-using-crisp-dm-methodology-d53efd57c448

Pynam, V., Spanadna, R. R., & Srikanth, K. (2018, September). An extensive study of data analysis tools (rapid miner, Weka, R tool, Knime, Orange). *International Journal of Computer Science and Engineering*, *5*(9), 4–11. https://doi.org/10.14445/23488387/ijcse-v5i9p102. Retrieved February 11, 2023, from https://www.researchgate.net/publication/331169871\_An\_Extensive\_Study\_of\_Data\_Analysis\_Tools\_Rapid\_Miner\_Weka\_R\_Tool\_Knime\_Orange

Ramchandra, S. (2022, May 31). *The importance of customer churn analysis: How to do it?* Kapture CRM. Retrieved March 2, 2023, from https://www.kapturecrm.com/blog/the-importance-of-customer-churn-analysis-how-to-do-it/

Satz, D. (n.d.). *Ethics and Data Science*. Data Science. Retrieved February 12, 2023, from https://datascience.stanford.edu/research/research-areas/ethics-and-data-science

Sharma, R., & Whitfield, B. (2022, November 8). *What's wrong with your statistical model? skewed data.* Built In. Retrieved February 26, 2023, from https://builtin.com/data-science/skewed-data

Shearer, C. (2000). The CRISP-DM Model: The New Blueprint for Data Mining. *Journal of Data Warehousing*, *5*(4), 13–22. Retrieved February 11, 2023, from https://mineracaodedados.files.wordpress.com/2012/04/the-crisp-dm-model-the-new-blueprint-for-data-mining-shearer-colin.pdf

Wålinder, A. (2014, June 10). *[PDF] evaluation of logistic regression and random forest classification based on prediction accuracy and metadata analysis: Semantic scholar*. Retrieved March 1, 2023, from https://www.diva- portal.org/smash/get/diva2:724982/FULLTEXT01.pdf

Weiss, G. M. (2004). Mining with rarity: a unifying framework. SIGKDD Explorations Newsletter, 6(1):7–19. As cited in Branco, P., Torgo, L., & Ribeiro, R. P. (2015). A Survey of Predictive Modelling under Imbalanced Distributions. *ACM Computing Surveys*, *49*(2), p. 31. https://doi.org/10.1145/2907070. Retrieved from: https://arxiv.org/pdf/1505.01658.pdf