Project Proposal

November 8, 2022

Team Member Names: Madeline Witters

Project Title: Predicting Customer Churn and Identifying Attributes of At-Risk Customers

Problem Statement

One of the most important metrics of success in any business is customer retention. A business's customer churn rate can have very significant financial impacts that effect the company in a multitude of ways. A Harvard Business Review article recently stated that merely increasing a company's customer retention rates by 5% can increase profits upwards of 25%; conversely, the cost of obtaining a new customer can be anywhere from 5 to 25 times as expensive as retaining an existing one.

It therefore follows that if a company is able to predict which customers are at risk of leaving, they can use this information to better position themselves in a variety of ways, such as: creating an intervention plan for customers at risk of leaving, calculating potential loss of revenue in the next quarter, or simply better understanding their customer demographic and various market segments.

In this project, there are two central research questions that I will aim to answer:

1. Can I create a model that will predict customer churn with a reasonable accuracy rate?
    - Furthermore, does one model type outperform another in predicting customer churn?
2. What features/customer attributes (present in the dataset) are most important in predicting whether a customer will churn?
    - Additionally, what does interpretation of these attributes reveal (for example, are younger customers more at risk of churning)?

Data Source

For this project I am using the "Bank Customer Churn" dataset, sourced from Kaggle. The dataset has 10,000 data points and 12 variables. Each row represents an individual customer. Variables are both categorical and quantitative, and include demographic variables (age, gender, country), as well as industry-specific variables (balance, credit_card, etc.). The response variable "churn" is categorical, containing a 1 if the client left the bank and a 0 if they remained a customer.

Methodology

My methodology will consist of three distinct parts:

- Exploratory Data Analysis
- Variable Selection

- Modeling

Exploratory Data Analysis

Exploratory data analysis will consist of any necessary data cleaning (addressing missing data, removing unnecessary/not relevant variables, one hot encoding, scaling/standardization, etc.), outlier detection, and examination for potential multicollinearity between predictor variables. As part of this step, I also plan to do some basic visual analysis of the distribution of predictor variables via boxplots, histograms, or density plots. Finally, I will partition the dataset into a train/test split.

Variable Selection

For variable selection, I plan to use Lasso or Elastic Net to identify the most relevant features, as I anticipate some features will not contain predictive power. I will select the best λ parameter for Lasso/Elastic Net via cross-validation on the training dataset. Depending on the results from my analysis, I may also examine the "most important" features from a Random Forest model (fit to all variables), and then re-fit my final model(s) to those chosen variables (more on this below).

Modeling

I plan to build and compare two classification models: a logistic regression model and a random forest.

Logistic regression lends itself well to this dataset in two ways: it can be easily used for binary classification/prediction, and it is also very interpretable (meaning we can extract information about which attributes have the most predictive power). This is especially important for business-related data, as oftentimes ones needs to explain the results from a model to non-technical stakeholders. Furthermore, the fact that the response data can be represented in terms of a probability is particularly appealing. While I will use cross-validation to identify the optimal threshold for classification accuracy, one can imagine a scenario where a business may prefer to set their own threshold and/or have the model directly report the probability that a customer may leave, rather than a binary indicator.

Second, I plan to contrast the logistic regression model with a random forest model. While random forests do not have the interpretability of logistic regression, and are known as being a bit more of a "black box" model, they are also known for their predictive prowess. I think it will be interesting to contrast the performance of these two models and see which achieves better predictive performance. As mentioned in the variable selection section, I also plan to examine the "most important" features from the Random Forest model and compare those to the variables chosen by Lasso/Elastic Net. If different subsets of variables are identified, I may build 4 models total (logistic regression fit to Lasso/Elastic Net variables, logistic regression fit to Random Forest most important, Random Forest fit to Lasso/Elastic Net, Random forest fit to most important), and compare and contrast performance.

Finally, if I have time, I may also build and compare a third classification model, such as a support vector machine or k-nearest neighbors model. If I choose to build this additional model, I will of course discuss and compare its performance with that of the logistic regression and random forest models. I will also discuss pros/cons of using each model and when it may be appropriate to use one above another.

Evaluation and Final Results

I will evaluate and compare the models in a variety of ways. The primary evaluation method will be by examining the classification error rate for the test data. Related to this, I'll likely construct a confusion matrix and/or AUC-ROC curve. I may also consider some additional metrics by which to compare model performance, such as a pseudo R^2 value.

Second, I will also provide an interpretation of results and discussion with respect to the chosen variables. This is especially relevant for this project, as businesses are often hoping to gain insight into which particular customer segments are leaving. Specifically, I will examine and interpret the coefficients included in the final logistic regression model, and explain how the coefficients could be translated into a business insight (e.g. younger customers are more at risk of leaving, given the other predictors in the model).

I will finally discuss how one could use the results from my project to design their own study, as well as how the project could be improved. This will include discussion of additional models one could test, different methods for variable selection, variables that could be added to the dataset, and more.

Citations

- "Bank Customer Churn", https://www.kaggle.com/datasets/gauravtopre/bank-customer-churn-dataset
- "The Value of Keeping the Right Customers", Amy Gallo, https://hbr.org/2014/10/the-value-of-keeping-the-right-customers