Customer Churn Analysis in the Telecommunications Industry

Colton B. Horton

July 4, 2023

1 Introduction

In the telecommunications (telecoms) industry, customer *churn* represents the proportion of customers who terminate their engagement with a service provider. In this competitive market, where average annual churn rates can reach 25%, prioritizing customer retention has become imperative due to the high costs associated with acquiring new customers.

The aims of this study are to use logistic regression to identify variables that are correlated with customer churn, as well as identify variables that are correlated with non-churn in "high-risk" customers. The findings from this analysis may help telecoms companies allocate resources more efficiently to reduce overall churn rates as well as reduce churn rates in high-risk customers. This paper will review the statistical methods and models used, discuss the data cleaning process, present the data analysis procedures, and provide several recommended courses of action, along with suggestions for further analysis.

This analysis was performed using Python, and employed several Python libraries: `Pandas`, `NumPy`, `Seaborn`, `Matplotlib`, `scikit-learn`, and `SciPy`.

2 Review of Statistical Methods and Models

2.1 Logistic Regression

Logistic regression is a statistical model that generates the probability that a certain observation belongs to a particular group. This approach to binary classification makes it an ideal fit for predicting churn. Logistic regression also provides coefficients for all of the variables, making it easy to quantify the impact of each variable on the model. Additionally, lasso regularization can be added during the fitting process to identify which variables are the most significant.

2.2 Feature Selection Methods

In order to perform feature selection for this analysis, two methods were utilized: forward stepwise selection and logistic lasso regression. Forward stepwise selection involves iteratively fitting logistic regression models while adding one variable at a time based on which one maximizes a certain performance metric. For this analysis, auc scores were used as the performance metric to maximize

during forward stepwise selection. Forward stepwise selection allows variables to be selected not only for their individual contribution to the model, but their cumulative contribution as well.

After selecting the most important features for the initial logistic regression model, logistic lasso regression was used for a second model, fit to predict non-churn in "high-risk" customers. This allowed a model to be fit utilizing all of the information available while simultaneously performing feature selection. The lambda value that controls regularization strength can also be optimized for a specific performance metric using a grid search and cross fold validation.

3 Data Collection and Preprocessing

3.1 Data Source

This study uses a dataset provided by Western Governors University that contains customer and churn data for an unidentified telecoms company. There are 50 variables in the dataset, 49 independent variables, and churn as the dependent variable. A list of all variables as well as a description for each one can be found in the data dictionary provided.

3.2 Data Cleaning

These were the three goals of the data cleaning process: checking for and handling missing/duplicated values, removing variables that are not applicable to the analysis, and one-hot encoding the categorical variables in preparation for logistic regression. There was only one variable with missing values: the type of internet service a customer has. After reading the data dictionary and examining the data, it became clear that there were not any missing values. "None" is one of the potential categories for internet service, but they were being imported as NaN values in the data frame, so the missing values were replaced with "No_IS." No duplicate values were found in the data set.

There were four categories of variables that were not applicable to this analysis. The first was unique identifiers. These variables were helpful in identifying any duplicate values, but they do not provide any predictive insight. The second category was survey responses. While these may have been helpful, the goal of this analysis is to provide actionable insights and a model that can be applicable to future predictions, initiative, and analyses, regardless of the availability of survey responses. The third category was location variables; other than the state variable. The reasoning for this decision is that by including all of the location identifiers, it would add unnecessary complexity to the model. The state variable avoids that problem while still capturing regional differences that may contribute to customer churn.

The last variable removed was the job variable. The reason it was removed was similar to the location variables: it adds unnecessary complexity to the model. Additionally, the salary variable is similar to the state variable in that it reduces the potential complexity added to the model, while maintaining an important aspect of the removed variables. The impact that job titles may provide to the model will most likely be well summarized in that profession's salary. Lastly, the categorical variables were one-hot encoded, bringing the total independent variable count to 96. A list of all variables included in the analysis, pre-one-hot encoding, can be found below in the appendix.

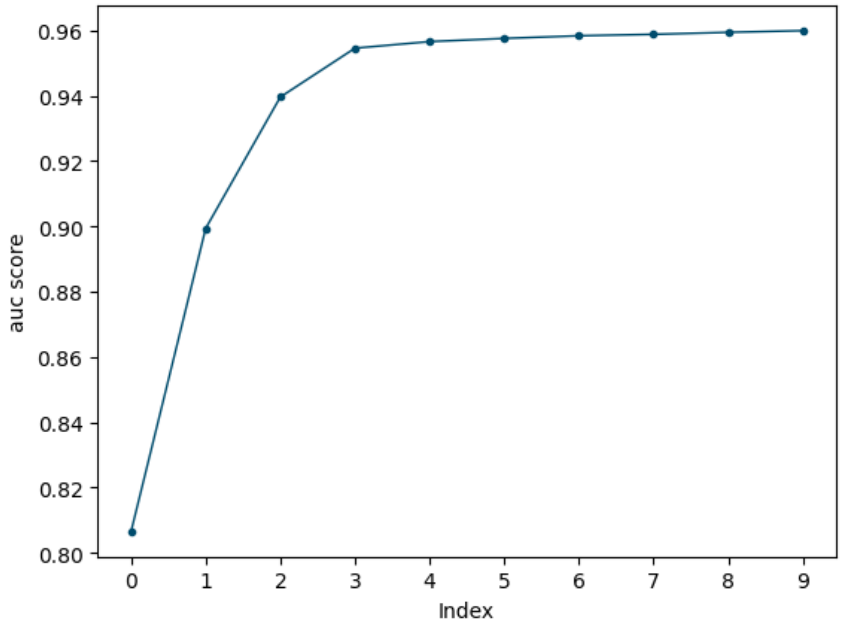


Figure 1: auc Scores by Index

4 Data Analysis Procedures

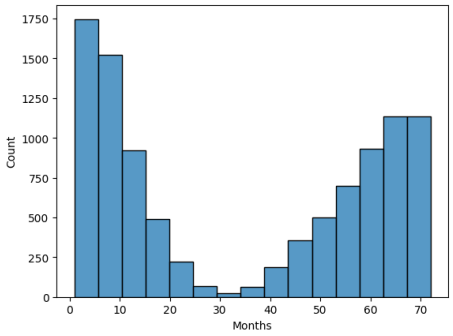
4.1 Forward Stepwise Selection

To set up the forward stepwise variable selection, 3 lists were created: a candidate variables list containing all of the independent variables, and 2 other empty lists to add the most significant variables along with their associated auc scores. Two functions were defined: one that fits a logistic regression model using the most significant variables with one additional candidate variable in order to provide the auc score associated with that candidate variable, and another function that iterates through the candidate variables, appends it to the current most significant variables, and compares it's auc score to the other candidate variables, then providing the next best variable to append to the significant variables list. Lastly, a loop was created that iterates through those functions 10 times to obtain the 10 most significant variables, along with their associated auc scores.

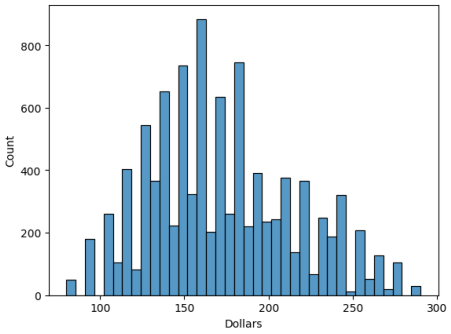
Plotting the auc scores reveals that the auc scores level off after the fourth variable (Figure 1). The four most significant variables are the customer's tenure measured in months, their monthly charge measured in dollars, whether the customer has a monthly contract, and whether the customer has fiber optic internet services.

4.2 Significant Variable Visualization

Univariate visualizations show that tenure has a bimodal distribution, monthly charge is normally distributed (Figure 2), and both monthly contract and fiber optic internet are roughly evenly distributed (Figure 3).

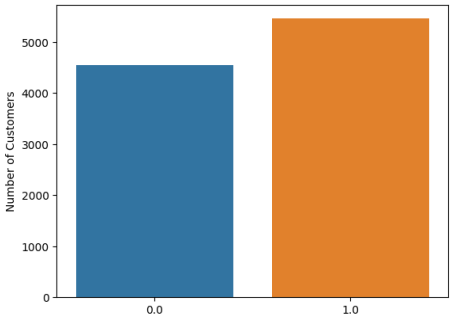


(a) Tenure Distribution

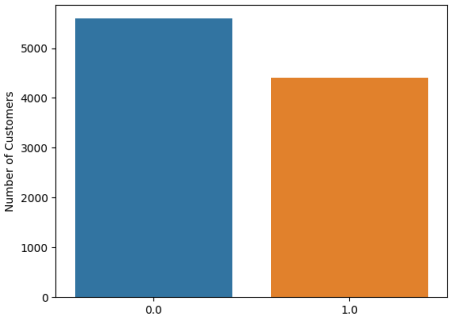


(b) Monthly Charge Distribution

Figure 2: Quantitative Univariate Visualizations

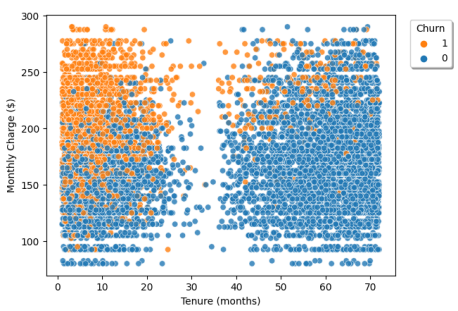


(a) Month-to-Month Distribution

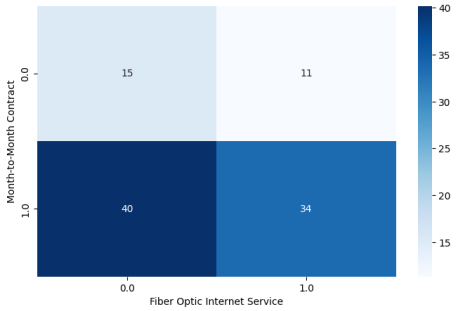


(b) Fiber Optic Internet Service Distribution

Figure 3: Categorical Univariate Visualizations



(a) Churn Tenure and Monthly Charge



(b) Churn Monthly Contract and Fiber Optic

Figure 4: Quantitative Univariate Visualizations

Bivariate visualizations were created for the quantitative and categorical variables. Each visualization includes churn through color coding or as the values in the heat map. The visualizations show that customers with shorter tenure and higher monthly charge have higher churn rates. Additionally, customers with a monthly contract and without fiber optic internet services have higher churn rates (Figure 4).

4.3 Logistic Regression Model

In order to quantify how well these variables can predict churn, a logistic regression model was fit using 10-fold cross-validation, stratifying on churn. 3 performance metrics were gathered from each cross-fold: accuracy, sensitivity, and specificity. The lists for these values were then used to generate a 95% confidence interval for each performance metric. The mean metric and the associated confidence interval were then plotted (Figure 5). The performance metric plots indicate a well-fitting model, but the model had a little more difficulty with the false positive rate, as seen in the reduced sensitivity. In order to obtain coefficients and predictions, a follow-up model was fit on all of the data. These coefficients confirm the trends identified in the bivariate statistics and will be provided in the results section.

4.4 Identifying High-Risk customers

The reduced sensitivity performance was an interesting finding and prompted an additional goal for this analysis. The original goal for this analysis was to simply find the variables associated with churn, and use that information to provide recommended courses of action. However, due to the relatively large number of people who were predicted to churn, but did not, the additional goal of identifying any potential variables that are correlated with non-churn in high-risk customers was added. "High-risk" has been defined in this study to be the group of customers who were predicted to churn. The data set was reduced to the high-risk customers for the rest of the analysis procedures.

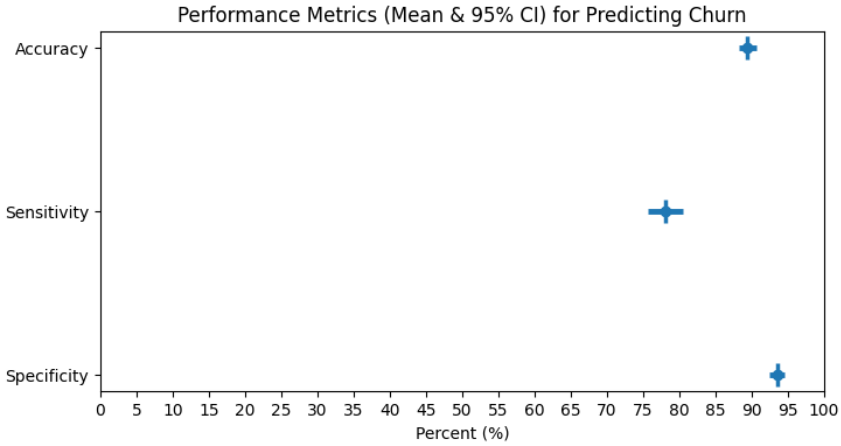


Figure 5: Mean and 95% Confidence Intervals for Logistic Regression Performance Metrics

4.5 Logistic Lasso Regression

The high-risk data frame contained 2,069 customers who churned, along with 477 customers predicted to churn but did not. The target variable, "non-churn," was derived as the inverse of churn. Additionally, the four variables used in the original logistic regression were removed to explore new variable relationships, resulting in a data set containing 93 variables for this logistic regression. Lasso regularization was employed to identify significant variables more effectively. To maximize the effectiveness of regularization, a grid search function with 5-fold cross-validation was used, stratifying on "non-churn", to determine the optimal lambda value that maximizes the sensitivity of the model. The pipeline was constructed with the logistic regression function and a standard scaler to normalize the quantitative variables for each fold.

After identifying the optimal lambda value, summary statistics for this model were obtained by conducting a 5-fold cross-validation using the optimized lambda value. The same pipeline in the grid search was used. Accuracy, sensitivity, and specificity were collected, and 80% confidence intervals were calculated. These metrics were then plotted (Figure 6).

A subsequent model was then fit using all of the data in the data set. This model was used to generate predictions for each observation in the data set and obtain the coefficients for each variable. Despite the average accuracy of the model being high, the sensitivity is very poor, indicating that the model does not fit the data very well. Additionally, there were no variables who's coefficient were reduced to 0. The regularization strength could be increased, but this would forfeit sensitivity when it already performs very poorly.

5 Results

5.1 Feature Selection Results

In the feature selection process, forward stepwise selection and logistic lasso regression was applied to identify the most significant variables for our logistic regression models. The forward stepwise

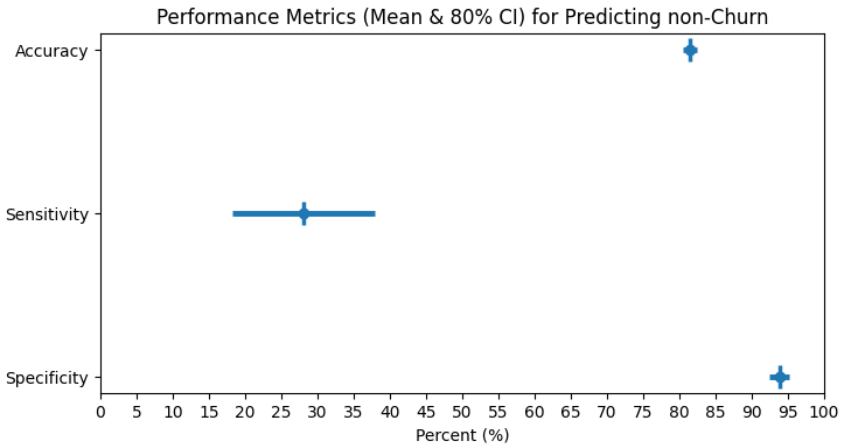


Figure 6: Mean and 80% Confidence Intervals for Logistic Lasso Regression Performance Metrics

selection process identified four significant variables: tenure, monthly charge, monthly contract, and fiber optic internet services. These variables were selected based on maximizing the auc score. The logistic lasso regression used to identify non-churn in high-risk customers did not identify any significant variables.

5.2 Logistic Regression Performance

The logistic regression model fit on the selected features, exhibited strong performance. Accuracy, sensitivity, and specificity were measure using a 10-fold cross-validation. The model achieved an average accuracy of 89.43%, indicating its overall ability to predict churn. However, the sensitivity score of 78.11% indicates that it had a slight bias towards false positives.

Additionally, the logistic regression model provided coefficients for each of the selected variables. Tenure and monthly charge had coefficients of -0.1059 and 0.0546, respectively, indicating that shorter tenure and higher monthly charge are correlated with higher churn rates. The presence of a monthly contract had a coefficient of 3.1177, while the presence of fiber optic internet service had a coefficient of -2.0892, indicating that customers with monthly contracts and without fiber optic internet service are correlated with higher churn rates.

5.3 Logistic Lasso Regression Results

The logistic lasso regression analysis aimed to identify the variables that were correlated with non-churn status in "high-risk" customers. Using a grid search function with 5-fold cross-validation and lasso regularization, The optimal lambda value for the regularization was determined to be 1000.

In this analysis, the logistic lasso regression model provided an average accuracy of 81.26%, however, the sensitivity was notably lower, with an average of 27.68%. This suggests that despite the high accuracy, the model does not fit the data very well. Additionally, a subsequent logistic lasso regression model fit on all of the available data using the optimized lambda value reveals that no variables were reduced to 0. Due to both of these, conclusions regarding which variables are correlated in non-churn in high-risk customers can not be made.

6 Recommended Courses of Action

Based on the results of this analysis, I have identified two possible courses of action could be considered to reduce churn rates. First, it may be beneficial to offer reduced rates to newer customers, given the correlation between shorter tenure, high monthly charges, and churn. Second, offering a bundled package of reduced fiber optic internet services to customers who opt for yearly or two-year contracts instead of monthly contracts could incentivize customers to chose longer-term contracts.

While both of these may offer decreased churn rates, it is important to note that correlation is not causation. If either or both of these are implemented, it is important to continuously monitor churn rates to ensure it is having the desired effect.

7 Suggestions for Further Analysis

While the logistic lasso regression failed to identify variables that may be correlated with non-churn, that does not it is impossible to identify factors that may be contributing to non-churn in high-risk customers. It might be advantageous to implement a targeted customer survey, focusing on those who have been predicted to churn. The feedback collected from thes surveys may help identify and address issues and concerns in high risk customers, leading to better customer satisfaction and potentially lower churn rates.

List of Variables Included in the Analysis

A description of each variable can be found in the provided data dictionary.

- State
- Area
- Children
- Age
- Income
- Marital
- Gender
- Churn
- Outage_sec_perweek
- Email
- Contacts
- Yearly_equip_failure
- Techie
- Contract
- Port_modem
- Tablet
- Internet Service
- Phone
- Multiple
- OnlineSecurity
- OnlineBackup
- DeviceProtection
- TechSupport
- StreamingTV
- StreamingMovies
- PaperlessBilling
- PaymentMethod
- Tenure
- MonthlyCharge
- Bandwidth_GB_Year