***Report on churn rate of Telecom Companies prepared by : (GROUP 6)***

**-HARISH KUMAR UDDANDI, ABHINAV THUPILI, BHARATH CHENNU, and YASH PATEL.**

**Contributions of Each team member**:

1. Harish Kumar uddandi: Deciding the Threshold limits and making the final presentation on behalf of the group.

2. Abhinav Thupili: Understanding the relationship of the variables, deciding on the model to be used, and conclusions to be drawn on the variables as per the churn model.

3. Bharath Chennu: Worked on cleaning the data and missing values in the dataset and created an outline for the presentation.

4. Yash Patel: Understanding the confusion matrix and final report preparation.

**PROJECT GOAL:**

ABC wireless INC is a telecom provider. This project aims to help address their customer churn rate issue. With the help of the company’s historical data, we aim to predict or identify customers who are likely to churn. Churn is the loss of customers to the competitor. This is a severe issue for telecom companies where the competition is cut-throat. Retaining a customer is less expensive than acquiring a new one. The task of our team is to apply analytics and help the management make appropriate decisions to reduce their churn rate and increase client retention.

**Overview of the data:**

The current data set Churn\_train.csv contains the following variables:

* state (categorical),
* account\_length,
* area\_code,
* international\_plan (yes/no),
* voice\_mail\_plan (yes/no),
* number\_vmail\_messages,
* total\_day\_minutes,
* total\_day\_calls,
* total\_day\_charge,
* total\_eve\_minutes,
* total\_eve\_calls,
* total\_eve\_charge,
* total\_night\_minutes,
* total\_night\_calls,
* total\_night\_charge,
* total\_intl\_minutes,
* total\_intl\_calls,
* total\_intl\_charge
* number\_customer\_service\_calls.

**DATA EXPLORATION:**

The current data set contains a large number of NA values. Below is the representation of the missing data.

Graphical user interface, application

Description automatically generated

From above, we can observe that account\_length has significant missing values with a percentage of 15.03%.

Additionally, the dataset also has negative values that could impact the prediction of our model. Here is the representation of the negative values in the dataset.

Graphical user interface, text

Description automatically generated

From the above, we can conclude that account\_length has both significant missing and negative values. This variable is a categorical variable that doesn’t have much impact on predicting the model. Therefore it is best to omit the “Account\_length” variable.

**DATA CLEANING:**

* **Text

  Description automatically generated**The dataset consists of negative values in number\_vmail\_messages. The number of voicemail messages cannot be negative. Since this is a numerical variable is essential to change the value from negative to positive. Below is a representation of how the values have been updated:
* We have used the abs function in R to change the negative values to positive values.
* The dataset still contains NAs and has significant values that could impact the dataset prediction in the model. Any variables with more than 75% NA’s would lack prediction power. The below chunk utilizes any significant variables with NA values of more than 75%.

Graphical user interface, application, website

Description automatically generated

**DATA PREPARATION:**

* **Imputation of missing values:**

After removing significant NAs, the dataset shows that the following variables still have some NAs. Below is the representation of the missing values.

Text

Description automatically generated with low confidenceImputation of Missing values in R refers to replacing the NA with closest with the proximity matrix using the random Forest method. Below snippet shows how the technique works in R:

* **Partition of datasets:**

Once the data has no missing values, we then partition the data using the CARET package in R in training and test sets. The partition index is set at “0.75,” which refers to the training set of 75% of the whole dataset and Text

Description automatically generatedthe test set of 25%.

**MODEL SELECTION:**

* To determine the most accurate model for predicting which customers will churn and which will not:

We discovered a high negative correlation between the number of customer service calls and total day, total evening, and total international charges among those who churned from the below plot:

Chart, scatter chart

Description automatically generated

We can infer several potentially important facts by looking at the correlation. When churn = yes, the more calls to customer support were made, the higher the prices were.

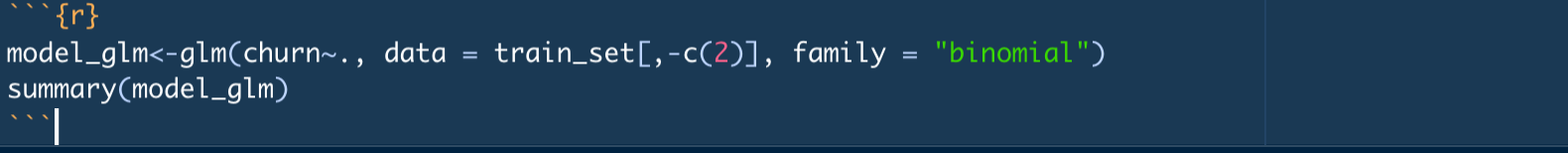
A predictive model based on the regression Model was used to highlight the influence of numerous variables and their importance in forecasting the outcome of the target variable.

Regression can be done in two ways:

\* Linear Regression

\* Logistic Regression

Because the data's target variable is categorical, a logistic regression model is the best option. When predicting a binomial attribute, it's tempting to use linear regression as a model, but the performance likelihood can be negative or more than 1, making it ineffective. A possibility or probability of chances between 0 and 1, as determined by logistic regression, is the desired outcome for this model.



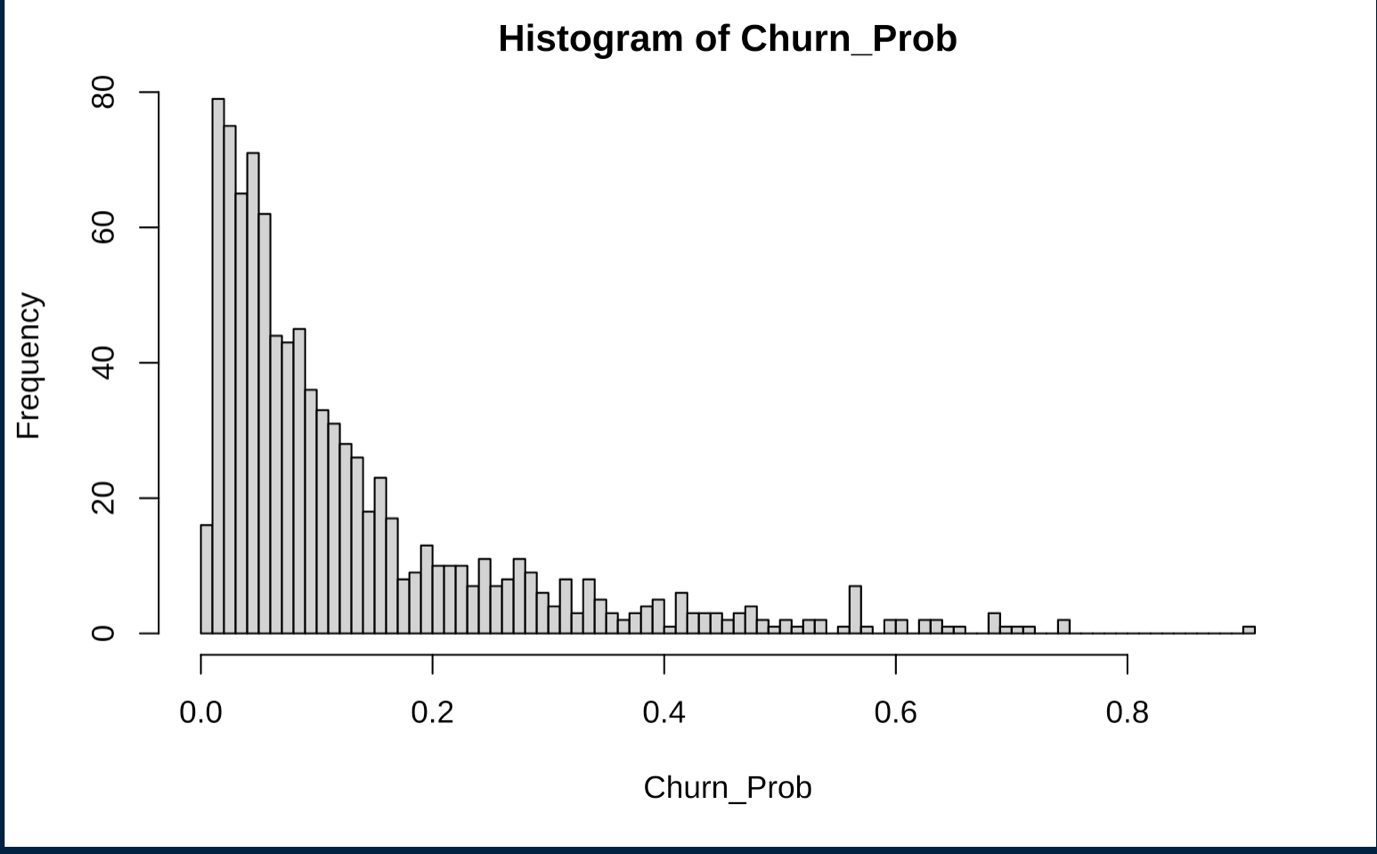
**EVALUATING THE ACCURACY OF THE MODEL:**

* We used the ROC curve to examine the binary characteristic algorithm's effectiveness and determine the best threshold based on our tolerance for false negatives and desire for true positives. Based on its utility as a predictor, we have a curve that displays a relatively excellent result. The x-axis on the graph represents the False Positive, whereas the y-axis represents the True Positive. The area under the curve is a single metric for determining a classifier's usefulness. The area under the ROC curve for a perfect classifier would be 1.
* As a result, the higher the AUC, the more confident we are in our model's predictive ability.
* Below is the representation of the AUC of the Logistic regression model.

Graphical user interface, text, application, email

Description automatically generated

**MODEL PERFORMANCE:**

* The cut-off has been determined using the “Test data set” on a Histogram to understand how the churn rate varies on a frequency. Below is a histogram representation:
* Churn\_Prob contains all the probabilities (from 0 to 1) that a customer from the test set that customers will churn or not. The histogram above reveals the distribution of the potential churn customers we predicted. The histogram tells us that most customers stayed (i.e., did not churn). Since the frequency of a customer not churning was higher between the probabilities of 0.0 to 0.3, with the more significant subset between 0.0 to 0.2.
* We get the "yes" and "no" churn replies for the 'Customers To Predict' data frame by using the 0.2 churning thresholds (cut-off).

Graphical user interface, text

Description automatically generated**PERFORMANCE OF THE MODEL:**

* Below is a representation of the performance of the model after setting a cut-off:

**INSIGHTS:**

* The below plot shows the highest churn rate of states from lowest to highest.
* Top five states which the highest churn rate are:

1. OH
2. GA
3. CA
4. NV
5. ID

**Chart

Description automatically generated**

* The number of customer calls has a direct impact on the churn ratio.
* Graphical user interface, table

  Description automatically generatedIf the customer's calls are too high or too low, the customers tend to churn.
* The telecom international call charges are high, which is significant from the table below.
* Marginal increase in international calls tends to make customer churn, showing they have high costs.

Graphical user interface, application

Description automatically generated

* From below, if the total day calls tend to increase, the churn rate also increases.
* This signifies that the telecom company has high day call rates for a longer duration.

**Graphical user interface, application

Description automatically generated**