PREDICTION OF TELECOM CUSTOMER CHURN.

Group members

1. Jamleck Mathenge
2. Sophia Mbataru
3. Collins Kiptoo
4. Bonface Mutua
5. Getrude Obwoge

Supervisors

Lucille Kaleha

William Okomba

Nikita Njoroge

Moringa School

Data Science Core

DSF-FT2

BUSINESS UNDERSTANDING

Business Overview

Telecom, a telecommunications company, has compiled data on their customers and whether or not they have stopped doing business, or churned, with Telecom. In the competitive world of telecommunication carriers, customer retention is key. Over 95% of Americans already own a cell phone. Moreover, acquiring new customers can cost up to 25 times more than retaining existing customers. Telecommunication carriers look to big data analytics around demographics, usage, customer accounts, connectivity, network performance and reliability, customer support and service issues, and more, to reduce customer churn rates. Predicting customer churn is critical for telecommunication companies to be able to retain customers effectively. For this reason, large telecommunications corporations are seeking to develop models to predict which customers are more likely to change and take action accordingly.

Problem Statement

In the telecom industry, customers are able to choose from multiple service providers and actively switch from one operator to another. In this highly competitive market, the telecommunications industry experiences an average of 15-25% annual churn rate. Given the fact that it costs about 25 times more to acquire a new customer than to retain an existing one, customer retention has now become even more important than customer acquisition. To reduce customer churn, telecom companies need to predict which customers are at high risk of churn.

Justification of the Study

In business, it's much easier to keep an existing customer than to get a new one. Keeping an existing customer is however not easy. Companies that don't invest in maintaining strong customer relationships risk high churn rates, and high churn rates can jeopardize your company's future. When a company loses a customer, there is an obvious loss of their business, which includes any immediate loss to a company's revenue, the loss of what could've been that customer's recurring business, and the potential loss of a referral. This leads to a loss of opportunity because new customers are not nearly as likely to try new products or spend as much. In contrast, an existing customer is 50% more likely to try a new product and spend over 30% more than a new one.

General Objective.

To obtain a data-driven solution that will allow us to reduce churn rates.

Specific Objectives

- To build machine learning models that will predict how likely a customer will churn by analyzing its features: demographic information, account information, and services information.
- To find the best machine learning model for the correct classification of churn/non-churn customers.
- To determine which features affect the customer churn rate thereby giving necessary recommendations.

Research Questions

- What features of the dataset are primary determinants of customer churn and to what extent?
- What are the ways that these findings can be interpreted and how can Mobi-Tel implement cost-effective solutions?
- Will these solutions be feasible in reducing the customer churn rate by 50%?

Metric of Success

Build a machine learning model with an accuracy score of above 75%.

Project Plan

Our project will consist of:

- Cross-Industry Standard Process For Data mining(CRISP-DM) will be used for conducting this research.

- A GitHub repository

- Presentation slides for the project

DATA UNDERSTANDING

The dataset provided information on the following features for each customer:

Exploring Data

This dataset contains 2 tables, in CSV format:

- The Customer Churn table contains information on all 7,043 customers from a Telecommunications company in California in Q2 2022
- Each record represents one customer and contains details about their demographics, location, tenure, subscription services, status for the quarter (joined, stayed, or churned), and more!
- The Zip Code Population table contains complementary information on the estimated populations for the California zip codes in the Customer Churn table
- The following is the breakdown of the dataset features into their specific categories:

(1) Demographic Information

- gender: Whether the client is a female or a male (Female, Male).
- SeniorCitizen: Whether the client is a senior citizen or not (0, 1).
- Partner: Whether the client has a partner or not (Yes, No).
- Dependents: Whether the client has dependents or not (Yes, No).

(2) Customer Account Information

- tenure: Number of months the customer has stayed with the company (Multiple different numeric values).

- Contract: Indicates the customer’s current contract type (Month-to-Month, One year, Two year).
- PaperlessBilling: Whether the client has paperless billing or not (Yes, No).
- PaymentMethod: The customer’s payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit Card (automatic)).
- MontlyCharges: The amount charged to the customer monthly (Multiple different numeric values).
- TotalCharges: The total amount charged to the customer (Multiple different numeric values).

(3) Services Information

- PhoneService: Whether the client has a phone service or not (Yes, No).
- MultipleLines: Whether the client has multiple lines or not (No phone service, No, Yes).
- InternetServices: Whether the client is subscribed to Internet service with the company (DSL, Fiber optic, No)
- OnlineSecurity: Whether the client has online security or not (No internet service, No, Yes).
- OnlineBackup: Whether the client has online backup or not (No internet service, No, Yes).
- DeviceProtection: Whether the client has device protection or not (No internet service, No, Yes).

- TechSupport: Whether the client has tech support or not (No internet service, No, Yes).
- StreamingTV: Whether the client has streaming TV or not (No internet service, No, Yes).
- StreamingMovies: Whether the client has streaming movies or not (No internet service, No, Yes).

Verifying Data Quality

The data set needs to be preprocessed for it to be ready for analysis

DATA PREPARATION

These are the steps followed in preparing the data:

We split the data into numeric and categorical variables so as to be able to value the observed differences according to the type of variable.

This was done to ensure the Validity, Accuracy, Completeness, Consistency, and Uniformity of the Data.

TASK1: The first task was to check if the column names were uniform and readable.

TASK2: The second task was to check for duplicated rows.

TASK3: The next task was to check if there were missing values.

TASK4: The next task was to decide how to deal with the missing values. (Either drop if they are unnecessary or replace the missing values with the best fit.)

TASK5: The last task was to check if the columns had the correct data types.

4. Data Types

The dataset contains categorical and numerical variables

5. Assumptions

The data provided is correct and up to date.

DATA ANALYSIS

In order to better comprehend the patterns in the data and even develop some hypotheses, let's first begin by studying our dataset. In order to find any significant trends, we will first examine the distribution of the various variables.

Univariate analysis.

1. Analysis of categorical variables.

We created a function that uses the dataset's categorical features and creates bar charts from the features.

2. Analysis of numeric variables.

We created a function that uses the dataset's numeric features and creates distplots from the features.

From the univariate analysis we were able to come up with the following inferences:

- About half the customers in our data set are male while the rest are female.
- About 50% of our customers have a partner.
- About 50% of our customers have a partner, but only 30% of our customers have dependents
- As shown in the graph, most customers churn due to a lack of online security.
- Several customers choose the fiber optic service and its only evident that the customers who use fiber optic have a high churn rate, this might suggest dissatisfaction with the type of internet service.
- Customers who opted for Bank withdrawals were more likely to churn than those who chose credit card and mailed checks as payment methods respectively.
- Most of our customers are in the Month to month contract. While there are an equal number of customers in the 1-year and 2-year contracts. Thus are likely to churn with the month-to-month contract.

Bivariate analysis.

We plotted histograms showing the relationship between various features in our dataset and the target feature.
From the observations made from our charts, we came up with the following inferences:

- The figure above shows the relationship between Age and customer behavior. The customer behavior clearly shows that Age has little to no effect on customer behavior but on further analysis, if a customer is considered a senior citizen they are less likely to churn.
- The gender of a customer has no observed effect on the choice of churning where both genders are equally likely to churn.
- The marital status of a customer has little to no effect on customer behavior.
- Customers enrolled in an offer plan are less likely to churn as compared to those out of an offer. However, there is no clear observation that the offer type affects customer behavior.
- The longer a customer relation period the less likely they are to churn. Also, the customer churn rate is highest in the first 10 months of customer interaction.
- Customers who pay via bank are more likely to churn showing this method can be faulty. However, this observation can be a result of the difference in the number of customers using various payment methods.
- Customers on Premium tech support are less likely to churn as compared to those with no tech support. Customers in Premiumtech support account for about 40%.
- Fibre optic internet type is the most popular but it's affected by a high churn rate which is an indication of customer dissatisfaction. Dsl and cable have the same observation respectively.

Visualizations

The visualizations for our univariate and bivariate data analysis are HERE:

Modeling

Data-preprocessing

We drop the churn reason and churn category columns because they had a large amount of missing data.

Train-test split.

The first step in building the model is to split the data into training and test sets.

The training data is used by the machine learning algorithm to build a model while the test set is used to evaluate the model performance on unseen data.

It is important to evaluate a model’s performance to guarantee that our metrics of success are met.

We employ a test size of 0.25 and a random state of 1 to ensure consistency.

Feature Engineering.

It is the process of extracting features from the data and transforming them into a format that is suitable for the machine learning model.

Most machine learning algorithms require numeric values, therefore, all the categorical features should be encoded into numeric labels before training the model. In addition, we need to transform numeric values in the dataset into a common scale which will prevent the big values in the dataset from dominating the learning process.

Training data and Test data pre-processing.

1. One hot encoding - This creates a new binary column for each level of the categorical variable. The new column contains 0s and 1s indicating the presence or absence of the category in the data.
2. Normalization - It is a common practice in machine learning which consists of translating data into a range of (0,1). If the features are scaled differently some may take up more weight than others. Hence, the scale levels the playing field.
3. Class imbalance - Our target feature has a class imbalance hence we use SMOTE to create a synthetic sample class. This is to ensure that the class imbalance does not affect our model performance.
4. Label encoding - It is used to replace categorical values with numeric values. This encoding replaces every category with a numeric variable.

Modeling Techniques.

They are based on the use of algorithms which are a sequence of instructions for solving specific problems.

In the pursuit of the best model, we created a function that takes in model technique parameters and gives an output of the modeling.

We also created a function that finds the best parameters for the modeling techniques.

CONCLUSION

This project entails a business overview, business understanding, data understanding, exploratory data analysis, modeling, and recommendation in order to advise Telecom on the best procedures. The project follows the CRISP-DM methodology showing the various project stages in the proper format. This project is integral to Telecom as it will provide the best model which will be able to predict customers likely to churn and therefore, equip the company with data on customers likely to churn in order for the company to take appropriate action.

In conclusion, when customers leave, they are going to the competition. Understanding churn factors will not only allow Telecom to understand why their customers are leaving but also to what extent the factors that lead to customer dissatisfaction affect the company. Overall, this will lead to the opportunity for Telecom to sharpen its attractiveness in the eyes of its customers by competing in the market well.

RECOMMENDATIONS

1. Telecom should recruit more customers to their offer packages because there is a clear indication that it reduces customer churning.
2. Telecom should improve its relationship with new customers because we found out that the highest churning rates occur in the first 10 months.
3. Telecom should make premium tech support more attractive to customers because customers on the cover have low churn rates.
4. Telecom should make changes to their internet provision services as they all have high churn rates, especially fiber optic internet.
5. Telecom should create products that are attractive to non-senior citizens because non-senior citizens have the highest churn rate.