PHASE 3 PROJECT

Student's name: Daniel Muturi Karue

Course: Data Science.

Date: December 2024.

BUSINESS UNDERSTANDING.

Telecommunication is the sending and receiving of messages by computer, telephone, radio and television, or the business of doing this. SyriaTel is one of the three mobile network providers in Syria. It operates a network of GSM 900/1800, 3G 2100 and 4G 1800 cellular networks. it was founded in January 2000 with its headquarters on Sehnaya Road in Damascus. When the company was founded, it was owned by the Egyptian Telecommunication company ORASCOM (25%) and Rami Makhlouf (75%) cousin of Syrian president Bashar Al-Assad. In 2003 ORASCOM sold its ownership in SyriaTel. The company has approximately 3500 employees and 8 million customers as of 2016.

Project statement: The purpose of this project is to provide valuable insights on ways of reducing the loss of money because of customers who don't stick around very long, as well as a predictive model that can be used to make recommendations to the company.

Objectives:

1. To analyze the effect of various factors like total day minutes, total day calls, total eve charge,etc on churn(which is a measurement of the percentage of accounts that cancel or choose not to renew their subscriptions).

2. Identify key predictors and provide interpretable insights to assist SyriaTel a telecommunication company in reducing the loss of money associated with churn.

3. Identify interventions that can reduce churn.

DATA UNDERSTANDING.

Our data contains 3333 rows and 21 columns.

The 21 columns are labelled: "State", "Account Length", "Area Code", "Phone Number", "International Plan", "Voicemail Plan", "Churns" etc.

The data types are four in total: Boolean (1), floats64(8), int64(8), and object(4).

The memory usage is 524.2+ KB.

DATA PREPARATION

First get a summary of the dataframe you have generated using the .info() method. The dataframe contains four data types, namely: bool(1 column), float(8 columns), integers(8 columns), and objects(4 columns).check for the missing values and their count in each column, for our case there were no missing values. Separate the features variables (X) and the target variable (y). I used the churn column as my (y) and all the remaining columns as my (X) apart from the phone number column which I dropped.

I created a train and test split with a test size of 0.2. The test size refers to the training set which is 80% and the testing set which is 20%. I dealt with the categorical data and first I dropped the column “phone number” from the categorical features function and excluded the datatypes integers and floats. I initiated the one hot encoder which converts categorical variables into a numerical format that machine learning algorithms can work with.

I initialized the minmaxscaler to normalize the numeric features.it transforms the numeric features to a specific range usually between 0 and 1 or any other given range

I combined the two dataframes the one hot encoded one and the normalized one. Next, we do the modeling.

MODELING.

First I begin by fitting a logistic regression to the preprocessed training set. Then I checked on the performance on the training data and our classifier was 87% correct on our training data. I then checked on the performance on testing data and this classifier was 87% accurate on our test data.

I draw the ROC (receiver operating characteristic) curve and AUC (area under curve). The ROC graph allows us to determine optimal precision-recall tradeoff balances specific to the problem we want to solve. Then I plotted the visualization of the ROC curve.

For the Decision Tree I created the classifier and fitted it to the training data. I then plotted the decision tree and from the evaluation of the predictive performance the accuracy is 87%

From the confusion matrix I was able to calculate and display:

1. Accuracy = 96%
2. Precision = 1.0
3. Recall = 0.5902
4. F1-score = 0.7422.

This metrics from the confusion matrix was generated from the classification report.

EVALUATION.

The goal of this project was to provide valuable insights on ways of reducing the loss of money because of customers who do not stick around very long (churn).

I used the logistic Regression and Decision Tree Classifier for modeling. I split the data into a training set which was 80% and testing set which was 20%.

The performance Metrics that I calculated to assess the model's effectiveness include AUC ROC- Curve, confusion matrix, Accuracy, precision, recall and F1-score. For the decision Tree Classifier, the accuracy was 87%. For the AUC ROC-Curve the AUC was 0.738, which is good for our model. From the confusion matrix, I generated an accuracy = 96%, precision = 1, recall = 0.5902 and F1-score = 0.7422.

Perfect classifiers have an AUC score of 1.0 while a score of 0.5 is deemed trivial or worthless.

The model performed well and it can be improved by including more data for example what are the charges rate of other telecommunication companies in the market, how is the network coverage in every state or area etc.

CONCLUSION

I identified the key factors influencing churn, to be total day charge, total evening charge and total night charge customer service calls. The predictive model had an accuracy of 87%.

The model was evaluated using a confusion matrix. It performed well on both the training and validation datasets, showing a recall of 59%, meaning it successfully identified more than half of the churn.

RECCOMENDATION

Based on the evaluation, the company can reduce churn by having personalized offers to the high-risk customers for example having a discount on the charges on the day calls or the night calls.

improve on customer service calls so as to get feedback from the high-risk customers on what challenges they are facing and, on the areas to improve on.

DEPLOYMENT

Continue tuning the logistic regression model and update the model with new data. incorporate additional data sources. This project successfully provided insights into churn