SYRIATEL TELECOMMUNICATION COMPANY CUSTOMER CHURN PREDICTION

Introduction

The communication technology industry is one of the most competitive industries nowadays. The main problem that practically all telecommunications industries worldwide are currently facing is customer churn.Churn, in the context of telecommunications, is the process by which customers leave a business and stop using the services it provides either because they are unhappy with those services or because they can find better options from other network providers at more reasonable prices. This could result in a loss of revenue or profit for the business. Additionally, keeping clients has grown to be a challenging task.

In order to provide their clients with the best services possible and keep them satisfied, businesses are working hard to introduce new cutting-edge applications and technology.Since losing them would result in a large loss of revenue for the business, it is imperative to identify those clients who are likely to quit the organization in the near future in advance. Accurately predicting churn can lead to higher customer retention rates, increased market share and improved business performance.This project seeks to accomplish this procedure, which is known as churn prediction.

Problem statement

Business is becoming extremely saturated in this competitive world. SyriaTel telecommunications company wants to take the required actions to retain customers in order to stabilize their market value because the cost of recruiting new customers is significantly higher than the cost of maintaining existing customers and also in order to reduce losses incurred from customer churn.

Objectives

Main objective:

- To predict customer churn using a classification algorithm model

Specific objectives:

- To do exploratory data analysis on the data
- To fit different classification algorithm models to determine which one works best for churn prediction
- To select the best model
- To make predictions using the selected model

- To check the accuracy of the predicted variables

Data Understanding

This data is from syriatel telecommunication company and it was obtained from kaggle. The data has 21 columns and 3333 rows. The target column, churn, is a bool column where True means the customer did churn and False means the customer did not churn, making this a binary classification problem.
The data was quite clean ie it did not have any missing values or any duplicated. The data had 4 categorical columns which were transformed to be numerical for modeling purposes.
The data had outliers but they were not dropped because they might be useful for predicting customer churn. Univariate analysis was performed to understand the distribution of our individual variables, and later compared to our target variable, churn, in bivariate analysis so that we can understand how the target variable is distributed among the predictor variables.

Modeling

We checked how the variables were related to each other and dropped the variables that were highly correlated so as to prevent multicollinearity when modeling. The categorical variables were then transformed to be numerical for modeling purposes.
The data was split into training and testing sets so that we can train the model using the training set and assess it using the testing set. Different classification algorithms were fit to the data and since most of them require data to be scaled, we standardized the predictor variables in the training set and the testing set.
The data had class imbalance problems so we resolved that using SMOTE and used the resampled data to train the model.
The classification models used were; logistic regression, K- nearest neighbors , decision trees and random forest. For all the models, a baseline vanilla model was fit and the score(accuracy) of the model checked, then we used gridsearch to tune the model by checking how different hyperparameters combinations affected the model's accuracy and the best model was fit to the training data and assessed using both the training and testing data.
The gridsearch used 5-fold cross validation. The final model was the one that had the highest accuracy .

Results and discussion

- Logistic regression:
  The baseline model correctly predicted the class of 79.6% of the training set and 76.3% of the testing set. After using gridsearch to tune the model and improve its accuracy,we got the best model.

The best logistic regression model correctly predicts the correct class for 79.7% of the data in the training set and and for unseen data, 77% of the data. That is a good generalizable model and since it used l1 penalty, it means it used lasso regularization which performed feature selection for the model.

- Decision tree:
  The baseline decision tree seemed to be overfitting since it correctly predicted the class of 100% of the training set and 86.2% of the testing set.After using gridsearch to tune the model and improve its accuracy,we got the best model.
  The final decision tree model correctly predicts the class of 90.3% of the data in the training set and 92.6% of the data in the testing set.This is so much better compared to the logistic regression model.

- K- nearest neighbors(KNN):
  The baseline model seemed to be overfitting since it correctly predicted the class of 91.7% of the training set and 75.3% of the testing set.After using gridsearch to tune the model and improve its accuracy,we got the best model.
  The best knn model correctly predicts the class of 93.3% data in the training set and 78.5% data in the testing set which is still lower than the decision tree.

- Random Forest:
  The baseline random forest already performs quite well, it correctly predicts the class of 100% data in the training data and 92% data in the testing data but after tweaking the parameters, the model performed worse than the baseline model, it correctly predicted the class of 90.2% of the training set and 88.6% of the testing set.

Final model

The best model in terms of accuracy was the decision trees.
A features importance graph was drawn to determine the most influential features in the decision tree model, allowing you to understand which features have the greatest impact on the model's predictions. For our model, the most influential predictors were; customer service calls, total day minutes, International plan, voicemail plan, total eve charge, total intl charge, total intl calls,total night minutes, total eve calls, state LA, state MS, state NJ, state ME,state IL, state HI , state AL and account length(They are listed according to their impact in descending order) .

This shows that customer service calls have the greatest impact in churn prediction thus the company should prioritize having quality customer service calls.

The model was then used to make predictions and evaluated using the testing data and the confusion matrix displayed that the model has 110 true positives, 816 true negatives, 32 false negatives and 42 false positives.

The model :

- Has an accuracy of 92.6% i.e. out of 100 customers, it correctly predicts the class of 93 of them ,churn or not churn.
- Has a precision of 72.3% meaning out of the predicted positives(customers likely to churn), the model correctly predicted 72,3% of them
- Has a recall of 77.5% meaning, out of the actual positives,(the customers that did churn), the model correctly predicted 77.5% of them
- Has an f1 score of 74.8% ie the model has a good balance between precision and recall. It means that the model is effectively identifying the positive class instances while minimizing false positives and false negatives.

Conclusion

The predictor variables;account length, total day minutes,total day calls, total eve minutes, total eve calls, total day charge, total eve charge, total night minutes, total night calls, total night charge , total intl minutes and total intl charge seem to be fairly normally distributed, total intl calls and customer service calls seem to be normally distributed but skewed to the right, area code is a discrete categorical column and number of voicemail messages, most of them were 0 but the rest seem to have a platykurtic distribution.

Wv state seems to be having most of the customers and CA has the least customers, most of the customers don't have an international plan and also most of them don't have a voicemail plan, explaining why the majority of the data in the number of voicemail messages is 0.

The most influential predictors were; customer service calls, total day minutes, International plan, voicemail plan, total eve charge, total intl charge, total intl calls,total night minutes, total eve calls, state LA, state MS, state NJ, state ME,state IL, state HI , state AL and account length(They are listed according to their impact in descending order)

The model has a good balance between precision and recall. It means that the model is effectively identifying the positive class instances while minimizing false positives and false negatives. The model has 92.6% accuracy thus the model is suitable for churn prediction

Recommendations

- The number of customer service calls was identified as one of the most influential predictors of churn. This suggests that providing excellent customer service and addressing customer concerns promptly and effectively can significantly reduce churn. Focus on training customer service representatives to handle customer issues efficiently and provide proactive support to enhance customer satisfaction and loyalty.

- Total day minutes, total eve charge, total night minutes, and total international calls and charges were identified as influential predictors. Analyze these usage patterns further to identify any specific trends or behaviors associated with churn. For example, if

customers who have high evening charges are more likely to churn, consider targeted offers or discounts to encourage retention during evening hours.

- Most customers do not have an international plan or voicemail plan. The company should consider promoting these services to customers who currently don't have them but may benefit from them.

- Since the dataset shows a significant concentration of customers in WV state and fewer customers in CA, the company should consider focusing retention efforts and marketing campaigns on these regions

- The company should Leverage the influential predictors and usage patterns identified by the model to develop personalized retention campaigns. For example, for customers with a high number of customer service calls, consider reaching out to them proactively with special offers or personalized assistance to address their concerns and improve their experience.

Future work

- While the model currently has good accuracy and performance, we should continue monitoring and evaluating its performance on new data. As customer behaviors and preferences change over time, it's important to ensure that the model remains effective and up-to-date.

- Exploring advanced techniques like ensemble methods, gradient boosting methods, XGboost and Adaboost , or deep learning to further improve churn prediction performance.

Challenges

- Due to the presence of high multicollinearity among several predictor variables, we had to remove some of those columns from our analysis.

- The model has some limitations, since decision trees can be biased towards the majority class if the dataset used for modeling has imbalanced class distribution, we had to use SMOTE to solve the class imbalance problems thus any new data fed to the model has to undergo the same preprocessing technique