SYTRIATEL CUSTOMER CHURN PREDICTION

Business Overview

SyriaTel is a leading provider of telecommunications services in the United States. It offers a wide range of services, including wireless, wireline, and internet services. The company has been in business for over 50 years and has a strong customer base.

Problem Statement

In recent years, the company has been facing an issue with customer churn. Churn is the rate at which customers cancel their service with a company. The company's churn rate has been increasing over the past few years. This is a major concern for the company because it is losing revenue and customers. The company would like to predict customers who are likely to churn. This will help the company to identify customers who are at risk of leaving and take steps to prevent them from leaving.

Objectives

The goal of this project is to:-

- Predict customers who are likely to churn. This will help the company to identify customers who are at risk of leaving and take steps to prevent them from leaving.
- Check for relationship between various variables and churn.
- Find out the features that most predict customer churn.

Data Understanding

The data is SyriaTel customer churn dataset which contains 21 columns and 3333 rows.
I loaded the dataset and checked for missing values, outliers, duplicates and inconsistencies in the data.
I did univariate analysis to understand the distribution of our individual variables, and later compared these to our target variable churn, in bivariate analysis.
I discovered that some of our data was categorical, so I proceeded to change the values into a numerical format.
I also checked for correlations of several variables with each other.

Modelling

- I kickstarted the modelling by using logistic regression and achieved an accuracy of 0.76 which indicates that the model is able to predict the correct outcome in about 76% of the cases, on average. The model had an F1 score of 0.497. The F1 score indicates that the model has an average level of precision and recall when predicting positive cases.
- The second model i chose was a random forest model which improved the accuracy significantly and achieved an accuracy of 0.921. This indicates that the model is able to predict the correct outcome in about 92.1% of the cases, on average. The F1 score is 0.743 indicating that the model achieves a relatively good balance between precision and recall when predicting positive cases. This suggests that the model's positive predictions are accurate, and it can effectively capture a significant portion of the actual positive cases.

- The third model was a hyperparameter tuned random forest model which had a slightly decreased accuracy of 0.919 however it had a slightly improved F1 score of 0.754 indicating that the model achieves a relatively good balance between precision and recall when predicting positive cases. This suggests that the model's positive predictions are accurate, and it can effectively capture a significant portion of the actual positive cases.
- The logistic regression had an AUC of 0.75, the random forest model had an AUC of 0.83 while the tuned random forest model had an AUC of 0.85. The tuned random forest model with the highest AUC of 0.85 indicates the strongest discriminatory power among the three models, making it more effective in classifying true and false cases.

Challenges

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

Conclusion

- The variables that most predict customer churn include; total day minutes, customer service calls, international plan, total international calls and total evening minutes.
- Considering the F1 score and AUC, it appears that the hyperparameter tuned random forest model performs better than the first model. However, it's important to note that these differences are relatively small, the margin of improvement might not be significant enough to warrant choosing one model over the other. Other factors, such as computational resources and model complexity, should also be considered when making a final decision.

Recommendation

The company should do the following things: -

- Offer customers a discount on their monthly bill if they use less minutes during peak hours. This could help to reduce the number of total day minutes that customers use, which could in turn reduce the number of customers who churn.
- Create a customer service portal where customers can easily find answers to their questions. This could help to reduce the number of customer service calls that customers make, which could in turn reduce the number of customers who churn.
- Offer customers an international plan that is more affordable. This could help to reduce the number of total international calls that customers make, which could in turn reduce the number of customers who churn.
- Send customers a text message or email reminder when they are nearing their monthly plan limits. This could help customers to be more mindful of their data usage, which could in turn reduce the number of customers who churn.
- Offer customers a loyalty program that rewards them for staying with the company. This could help to create a sense of loyalty and make customers less likely to churn.

Future work

- Using more advanced machine learning algorithms to see if one predict customer churn better.
- Using more data: The more data that is used to train the model, the more accurate the predictions will be.