SyriaTel - Predicting Customer Churn

Project Overview

SyriaTel wants to be able to identify customers who are about to churn so they can reach out to them in order to stay. Customer acquisition is far more expensive than customer retention so if customers churn SyriaTel will lose out on future revenue and will have to spend a substantial amount of money to acquire new customers. To help them with this problem, I created a gradient boosting classification model that can predict whether a customer will churn or not based on 12 features. I created multiple classification models including logistic regression, decision trees, random forest, XGBoost, and more. In this repository, I have new jupyter notebooks. One for my data cleaning data exploration and the other with my classification modeling. Feel free to look at and run the notebooks, but I will note the GridSearchCV in the data modeling notebook can take a long time to run. Also in this repository is the Data folder which contains two csv files I used for this project.

Business and Data Understanding

Business Problem

For this business problem, the SyriaTel wants to use this model to predict whether a customer will churn or not. Once they identify customers who are likely to churn they are going to reach out to them and provide them with incentives to stay with the company. Customer acquisition costs much higher than customer retention so they do not want customers who are about to churn to go unnoticed. For this reason, the company wants to minimize False Negatives as they do not want to lose out on customers who are about to churn. I am going to focus on models that have high recall scores as they are impacted by False Negatives. I will also provide a few recommendations for SyriaTel that can help them prevent customers from churning.

Data Understanding

The dataset used is from Kaggle.com and it contains 3,333 rows of customer data. Some of the information on the customers includes state they are located, how long they have been with SyriaTel, phone usage, number of customer service calls, amount they were charged for their plan, and whether they churned or not.

From the dataset, SyriaTel has a churn rate of 14%. I wanted to look into this more to see if there are any significant differences in the characteristics of customers who churned and those who did not. I found that customers who churned on average made more customer service calls and used the phone more during the day and night. I ran hypothesis tests to see if these 3 characteristics were statistically significant based on an alpha of 0.05. After running the tests, I rejected the null hypothesis tests that these 3 characteristics are statistically significant.

Modeling and Evaluation

My goal was to create a classification model to predict customer churn and I focused on having a high recall score because I want to minimize the amount of False Negatives. I set a baseline model using a Dummy Classification model that would predict the most frequent value. The baseline model had an accuracy score of 85.5%, recall score of 0%, and precision score of 0% on the test data.

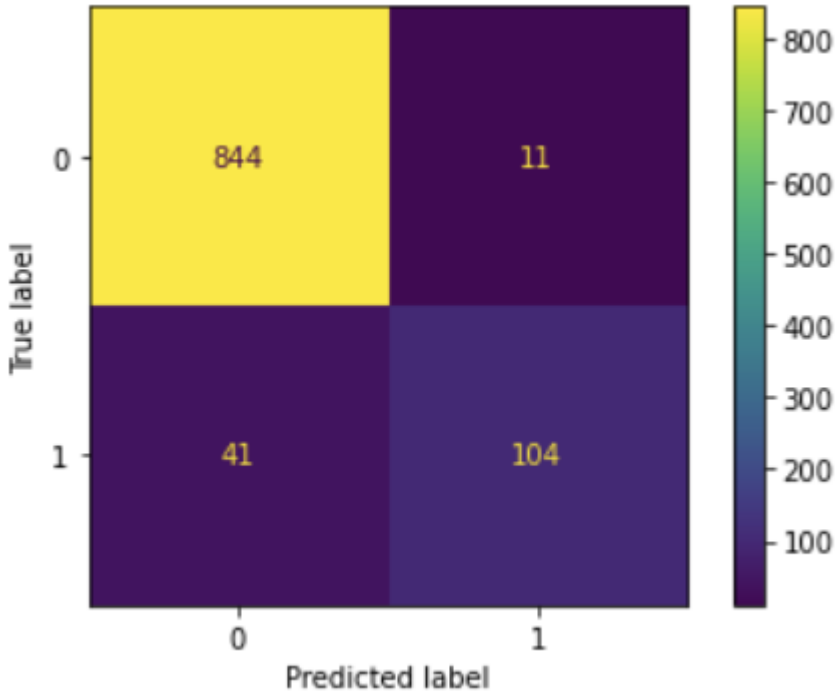
After the baseline model, I decided to create more classification models and then pick the best performing model to run a grid search on. The models I created after the baseline model were logistic regression, decision tree, k nearest neighbors, random forest, gradient boost, adaboost, and xgboost. The gradient boost model performed the best and was not as overfit as some of the other models so I decided to use this to run a gridsearch.

I ran a gridsearch on the gradient boosting model to tune the hyperparameters. The hyperparameters I chose were max_depth, learning_rate, max_features, criterion, and n_estimators. The best model came out with the following parameters: learning_rate=0.2, max_depth=5, max_features='sqrt', and n_estimators=50. This model did perform really well on the training data but it was far too overfit on the test set so I decided not to use it as my final model.

Final Model - Gradient Boost Classifier

My final model was the original gradient boost classifier that I ran. On the test set, the model had the following performance metrics. It was still slightly overfit but it worked well enough on the test set for me to use as my final model.

Accuracy score: 0.948 Precision score: 0.904 Recall score: 0.717 ROC AUC score: 0.901



Conclusion

My final classification model for SyriaTel produced an accuracy score of 94.8% and 71.7% on the test data. This model will work well with what SyriaTel is trying to accomplish. I provided the following 3 recommendations for SyriaTel based off their business problem for predicting customer churn.

1. Use this model to identify customers who are about to churn and provide them with incentives on staying. These incentives will be an upfront cost to retaining these

customers but can save you a lot of money down the road in customer acquisition if these customers do not churn.

2. To prevent customers from churning, I recommend looking into improving your customer service. Customers who churn make many calls to customer service so by improving the customer service call center and overall user experience this can have an impact on people staying with the company.
3. Look into improving international service. Many customers who have churned have international plans and improving this service can help with these customers staying.

Repository Structure

```
├── Data
│   ├── syriatel_clean.csv
│   ├── syriatel_data.csv
├── .gitignore
├── Data_cleaning_and _analysis.ipynb
├── Modeling
├── Presentation.pdf
└── README.md
```

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

- **Jupyter Notebook** 100.0%