MILESTONE 4

DSC630

Christiansen, Madison

Introduction

A common business problem that is seen throughout many companies is predicting churn rates in a subscription-based business. The churn rate refers to the rate at which customers cancel their subscriptions. Using a predictive model to anticipate why customers are likely to cancel their subscriptions can give the business a better idea of how to market and keep customers. Analyzing customer data such as demographics, subscription type, and more can help give a better understanding as to if the customer is at high risk of leaving.

Models

Having access to accurate and relevant data is needed to build an effective model. The Bank Customer Churn dataset from Kaggle contains information about the individual customers and many features that give insight into their plan usage. This is a large dataset with over 1,000 rows that allows for in-depth analysis and model use. With the many columns of data, the possibilities are wide. For example, looking at the demographics of the customers to understand what type of people are canceling their subscriptions. Or another way to look at it is the services they have and if there is a component that shows why some are canceling their subscriptions. Looking at the different features in the dataset to help predict churn is the top priority for the model.

A logistic regression model is a safe choice for churn predictions. This is a simple and interpretable model which can provide clear coefficients that are easily interpreted. Churn prediction would be a binary classification problem, and the goal is to see whether a customer will churn or not. Logistic regression is good for this problem as it models the relationships between the features and the probability of churn.

Random Forest. A random forest model can capture non-linear relationships between the predictors and target variables. It also allows for a good measure of feature importance as well.

Result Evaluation

When evaluating the results of churn predictions, it is necessary to use the correct evaluation metrics that align with the project. Accuracy is a good initial measurement to see the overall correctness of the predictions. Although a concern is it can be considered misleading if there is an imbalanced dataset where the majority class dominates. Precision is a deeper look at the proportion of correctly predicted churn cases. Recall calculates the proportion of the predicted churn cases out of all the actual churn cases. The F1 score is the mean of precision and recall and gives a balanced revaluation metric. ROC can be used to look at the true positive rate against the false positive rate at the classification thresholds. Along with this, the AUC summarizes the ROC curve performance in a single value.

Learning Goal

The goal of creating an accurate model to understand the customer churn rates is to keep the customers from churning. When a customer cancels their subscription, the business is impacted by losing revenue from the customer and the potential revenue from future subscriptions. High churn rates can lead to negative publicity for the business and further hurt the business's future with new potential customers. These models can allow the company to see how their customers are reacting to certain aspects of their subscriptions and give a better overall understanding of what they need to stay with the company.

Ethical Concerns

One larger concern is if the dataset were to be imbalanced and there is a majority class that dominates. This can lead to a biased model towards the majority class and lead to poor performance in predicting the minority class. If this were to arise, a good way to help the issue is class weighting. The random forest model can offer the option to assign different weights to different classes. If you assign a higher weight to the minority class, the model will be correctly predicting the churn and reduces the impact of the imbalances of class. Another technique is using SMOTE by combining multiple classifiers or employing resampling techniques with the ensemble framework to address the problem.

Contingency Plan

A good contingency plan needs to be flexible and adaptive to address issues that pop up during the evaluation of the project. Having a backup dataset or the ability to add another dataset if the one chosen has poor data quality. Open-minded to exploring alternate modeling approaches. If the initial models do not work out, then try different algorithms and adjust the parameters if needed. Imbalanced dataset, as stated previously, having a plan to deal with this is needed. Understanding the business plan and what is needed for the model to be successful in helping the business. Good documentation throughout the whole project. This allows for the steps, modeling approaches, and results to be understood better.

Data

The dataset that I will be using has many features that will provide a good base when tackling this model. The most important features that will be used include:

- Exited – churn rate.

- Satisfaction Score
- Card Type
- Is an Active Member
- Num Of Products

Although other features will also help and provide a good understanding of the churn rates and the customers that this bank has.

Visualizations

When it comes to explaining a churn dataset visually, there are several effective visualizations that can help convey the information clearly. Churn rate over time. Plotting the churn rate over time can help identify trends and patterns. Churn by important features. Creating a stacked bar chart or a grouped bar chart can show the churn rate for each feature that is important. Churn reasons, a pie chart laying out the reasons for customer churn. Churn prediction model performance. After developing a churn prediction model, visualizations of the performance metrics. A ROC curve or precision-recall curve.

Adjustments/ Expectations

From milestone 2, the expectations and plan set in place are still the ideal plan to go with for this data. With the dataset being as large as it is, I wanted to improve the data preparation and focus on this more to ensure the step is done thoroughly. The data preparation step will be very important as it sets the stage for a well-done model. As for the model being built, both the logistic regression and random forest model will be completed. Whichever model produces the best accuracy and metrics will be taken as the main model. With it being such a large dataset, I

am assuming that the logistic regression model will produce the best results. Along with this, a feature importance will be running to highlight important features of the churn data. Assuming the data prep and model building goes well, all the original expectations are still reasonable, and there will be an outcome that can be shared with the company.

Data Preparation

During data preparation, the main thing to focus on is ensuring the data is ready for model building. To start with the data prep, I went through and looked to see if there were any values within the data set that needed to be added. There were no missing values, meaning all rows and columns were good. With the models being built, there is no need for certain columns in the data set. I was going through and removing the demographic and personal information columns, which are optional for these models. Lastly, one column was needed to turn into a dummy variable.

Model Building

The two models I built were a random forest classifier and a logistic regression. I wanted to make these two models to determine which would have better evaluation metrics. Starting with these models, the data was split by train and test, with the 'Exited" column being the targeted churn column. After this, the test and train data were fit to the two different models, and then the x test was predicted to give the y prediction. After this, both models were then evaluated to *determine* the accuracy, precision, recall, and F1 scores.

Results

When looking at the two models, you can see that the random forest classification shows more accurate metrics. Based on these results, I went and created a feature importance with the random forest model. The top three most important features were "Complain," NumOfProducts," and "Balance."

Random Forest Classifier:

```
Accuracy: 0.9984
Precision: 0.9959839357429718
Recall: 0.9959839357429718
F1 score: 0.9959839357429718
```

Logistic Regression:

```
Accuracy: 0.8008
Precision: 0.0
Recall: 0.0
F1 score: 0.0
```

Conclusion/Recommendation

With the random forest model accuracy at 99%, this is a good indication of the model's ability to correctly predict the churn for 99% of instances in the test set. The precision implies that the model can avoid false positives. The recall of 99% means that the model can identify the positive churn results. From all of these results, the random forest model is the best model overall for this data set to predict churn rates. The random forest classification model can capture the interactions between the various features in the dataset. Based on the feature importance that I ran; you can see that the most important feature was "Complain." This shows us that, in the future, those who submit frequent complaints are more likely to churn.

Reference:

Kollipara, R. (2023). Bank Customer Churn. Kaggle. Retrieved from https://www.kaggle.com/datasets/radheshyamkollipara/bank-customer-churn.