Victoria Hall

6/3/2022

Customer Churn Reduction

Data: <https://www.kaggle.com/datasets/shubh0799/churn-modelling>

Github Repo for code: <https://github.com/vihall95/DSC630_HallVictoria/blob/main/Final%20Project/CustomerChurnReduction_HallVictoria.ipynb>

**Project Statement** - Customer Churn is a challenge that must be addressed. To reduce churn, we will need to be able to predict which customers are more likely to exit the company and intervene ahead of time to prevent customers from exiting the business.

**Understanding the Challenge:**

Customer Churn is commonly understood as the action of a customer exiting a business. It is typically measured by a churn rate. According to Qualtrics.com a churn rate is calculated by dividing the number of customers lost by the total number of customers during a selected time period. The point of a project like this is to reduce the churn rate. It is important to understand and combat customer churn because it can cause a negative view of the brand and impedes on future growth. It also is extremely expensive. It is less costly to keep an existing customer than it is to try and acquire a new one. There are a number of reasons that can cause customers to churn. According to clientsuccess.com there are ten main reasons for customers to churn. Price is often the most common reason for churn however there are number of other reasons including competition, lack of value, poor experience and many more. It is important to note while it is important to try to understand why customers are churning, it is beyond the scope of this project.

The purpose of this project is to utilize different machine learning techniques to create a tool that the company can use to predict whether a customer is likely to churn or not. By providing these predictions any number of business stakeholders like, a marketing or merchant team, can produce strategies to intervene and try to prevent churn. By doing accessing these results, it will the company two save money in two major ways. First, by preventing churn there is not as much of need to replace that customer with a new one, which is a more expensive process. Second, this information allows for targeted intervention. Rather than reaching out to all customers the marketing team can specifically reach out to those likely to leave, saving vital marketing dollars.

**Understanding the Data:**

To predict customer churn, the dataset for building the model had to explored. There were 10,000 rows of data without any missing information. Each row represented a customer. There were demographic features about each customer like which country they belonged to, gender, and age. The were also financial features about the customers like credit score, salary, and whether they owned a credit card. Finally, there were features about their interactions with the company including their current balance, the number of products they owned, and the time of bond with the company (or tenure). A few features were dropped that were not necessary to the model building including customer IDs and their surname. This also helped to protect the anonymity of the customers. Outside of these features, there was a column that identified whether the customer left the business or stayed, which was our target variable for the models. This target feature was a binary identification with one identifying those who left and zero identifying those who did not.

To further explore the data, I plotted the distribution of each feature as well as the target. I also plotted a heatmap to show the correlation between the features. There were a few callouts from these reviews. In the below graphic you can see the distributions of the features. There were more French customers in the dataset than Germans and Spanish, most were between the ages of 30 and 50, and there was a pretty even gender split between male and females. Financially, credit scores were normally distributed, more had credit cards, and there was a wide range of salaries topping out at 200,000 but these were pretty evenly balanced. Most customers had been with the company between one and ten years, either had a zero balance or a balance between 50,000 and 150,000, owned either one or two products. There was an even split for active membership. Our target variable was unevenly distributed. Most of the data was on customers who did not exit the business. This was an additional challenge for model building.

Chart, pie chart

Description automatically generatedGraphical user interface, calendar

Description automatically generated

I also looked at the correlation between the features and the target variable. This showed that age was strongly related to exiting the business. It also showed incidents of multicollinearity going on.

The last step of this data exploration was to prep the data for model building. This was simple. Categorical variables were changed to dummy variables. The data was not highly dimensional, so I opted to keep all the features in the data set. Finally, the target variable of exiting the business was split from the features.

**Model Testing and Evaluation:**

Once the data was properly explored and prepped for modeling, I chose six different models that could support a supervised binary classification task. Each of the six models provided the same result (predicting who would exit the business but they all relied on slightly different approaches.

1. Logistic Regression - Baseline linear classifier for binary data
2. Random Forest - Ensemble Bagging Classifier (Bootstrap Aggregation). Makes predictions by combining classifiers on bootstrapped subsets.
3. K-Nearest Neighbors - Instance Based Classifier. Makes predictions based on the training set only by searching for the most similar instances
4. Support Vector Machine - Maximal Margin Classifier. Finds the hyperplane that splits the data into two groups, and chooses the hyperplane that maximizes the margin (distance between hyperplane and close points)
5. XGBoost - Ensemble Boosting Classifier. Type of gradient boosting. Trains many models in a gradual additive manner and identifies shortcomings with a loss function.
6. Multilayer Perceptron Classifier (MLP Classifier). Relies on an underlying neural network

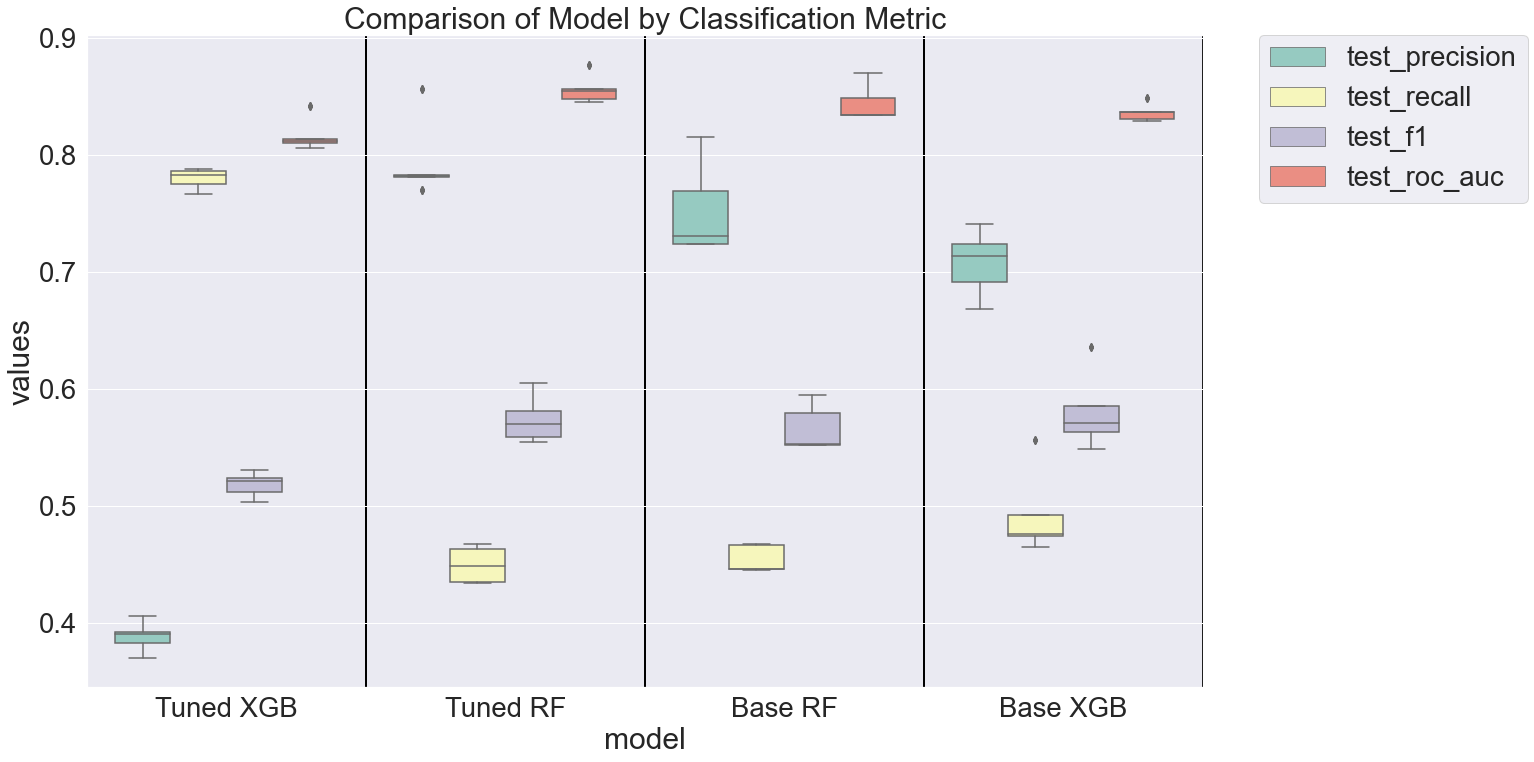
Each of these models was fed into a function that trained and fit the model, then output a classification report and a box plot of the results. There were all baseline models with little interference in the hyperparameters. Each model was evaluated with five split cross-fold evaluations to make sure it would be applicable to data outside of the training data. All the models performed well at identifying which customers would not exit the business. However, the purpose of this churn evaluation is to identify customers who are likely to exit the business so that we can intervene with some sort of marketing or promotions to encourage these customers to stay. I expected this result because the classes were heavily imbalanced. So, for the purposes of this problem, I am really focused on the results of models which correctly identified those who exited the company. For this evaluation I focused on my evaluation on precision, recall, F1 score, and the ROC\_AUC score. I opted to not use accuracy because according to Jason Brownless at machinelearning.com accuracy fails with imbalanced classes. The accuracy of the model will be inflated because it is better able to identify the majority class compared to the minority class.

Of all the models, XGBoost and Random Forest Modeling seemed the best the performing models. The scores are similar, with F1 scores at .57 and .59. Because of this, I chose to explore these models further. Below is box plot of the evaluation metrics for each model.

Chart, box and whisker chart

Description automatically generated

Before tuning the XGBoost and Random Forest Models, I decided to mainly focus my evaluation on a high recall score. A low recall score returns a lot of false negatives, which means, we might miss customers that should have been identified as leaving the business. However, improving this score came a cost. The more I was able to increase the recall score, the more the precision score suffered. But by enhancing the recall the company will be able to more accurately identify customers who look like they will leave the business. There may be higher instances of misidentifying customers that are not leaving the business. But in the end, it is more cost effective to accidentally reach out to some of the wrong people rather than not reaching out to the right people.

The parameter I investigated for XGBoost was the weight applied to the positive class. Applying a heavier weight to our positive classes (who exited the business) helped the model focus on those who left instead of those who did not. Focus on finding the optimal weight to increase recall did mean that precision was sacrificed. For the Random Forest model, I used a grid search to find the best values for the number of estimators, the maximum depth and the minimum sample split and leaves. After finding the optimal values, I fed these tuned models into a similar function the first testing one used. I also included details about the feature importance for each model.

By tuning the XGBoost model in the way I did, I was able to significantly increase the recall score. Precision was sacrificed so the F1 score decreased as well. Tuning the random forest model yielded little change over the baseline model. For both models, age, number of products, and active membership were the most important features.

**Final Results:**

The tuned XGBoost model that optimized the recall score was the best of the models evaluated. The precision score was .38 and the recall score was .81. This model will more accurately identify customers who look like they will leave the business, but we might also have some instances of identifying customers that are not leaving the business. It will be more cost effective to accidentally reach out to some of the wrong people rather than sacrificing customers who are likely to leave.

This model is not perfect. A better model might be derived with more data, specifically more data about those who opted to leave the business. But with the data available this model should be able to identify those who will potentially leave the business so intervention steps can be taken.

Citations:

Brownlee, J. (2021, January 21). *Failure of classification accuracy for imbalanced class distributions*. Machine Learning Mastery. Retrieved June 3, 2022, from https://machinelearningmastery.com/failure-of-accuracy-for-imbalanced-class-distributions/

ClientSuccess. (2021, September 28). *10 common reasons customers churn – and what to do about it*. ClientSuccess. Retrieved June 3, 2022, from https://www.clientsuccess.com/blog/10-common-reasons-customers-churn-and-what-to-do-about-it/#:~:text=Price%3A%20by%20far%2C%20one%20of,wear%20thin%20with%20decision%2Dmakers.

Kumar, S. (2020, June 20). *Churn modelling*. Kaggle. Retrieved June 3, 2022, from https://www.kaggle.com/datasets/shubh0799/churn-modelling

Qualtrics. (2022, April 15). *What is customer churn? learn to measure & prevent it.* Qualtrics. Retrieved June 3, 2022, from https://www.qualtrics.com/experience-management/customer/customer-churn/