Hopkinton Credit Union

Not for profit, not for charity, but for service

Bank Customer Churn Prediction using ML Techniques

SREEJITH MADHAVANKUTTY

07/24/2023

Introduction

“No great marketing decisions have ever been made on qualitative data.” — John Sculley (CEO of Apple Inc.)

Customer Churn prediction means knowing which customers are likely to leave or unsubscribe from a service. For many organizations such as Hopkinton Credit Union (HCU), this is an important prediction. This is because acquiring new customers often costs more than retaining existing ones. Once you’ve identified customers at risk of churn, you need to know exactly what marketing efforts you should make with each customer to maximize their likelihood of staying.

Customers have different behaviors and preferences, and reasons for cancelling their subscriptions. Therefore, it is important to actively communicate with each of them to keep them on your customer list. You need to know which marketing activities are most effective for individual customers and when they are most effective. Impact of customer churn on businesses such as Hopkinton Credit Union (HCU) is huge because they recently they are lost many of their customer base to bigger banks such as Bank of America , JP Morgan etc. A company with a high churn rate loses many subscribers, resulting in lower growth rates and a greater impact on sales and profits. Companies with low churn rates can retain customers.

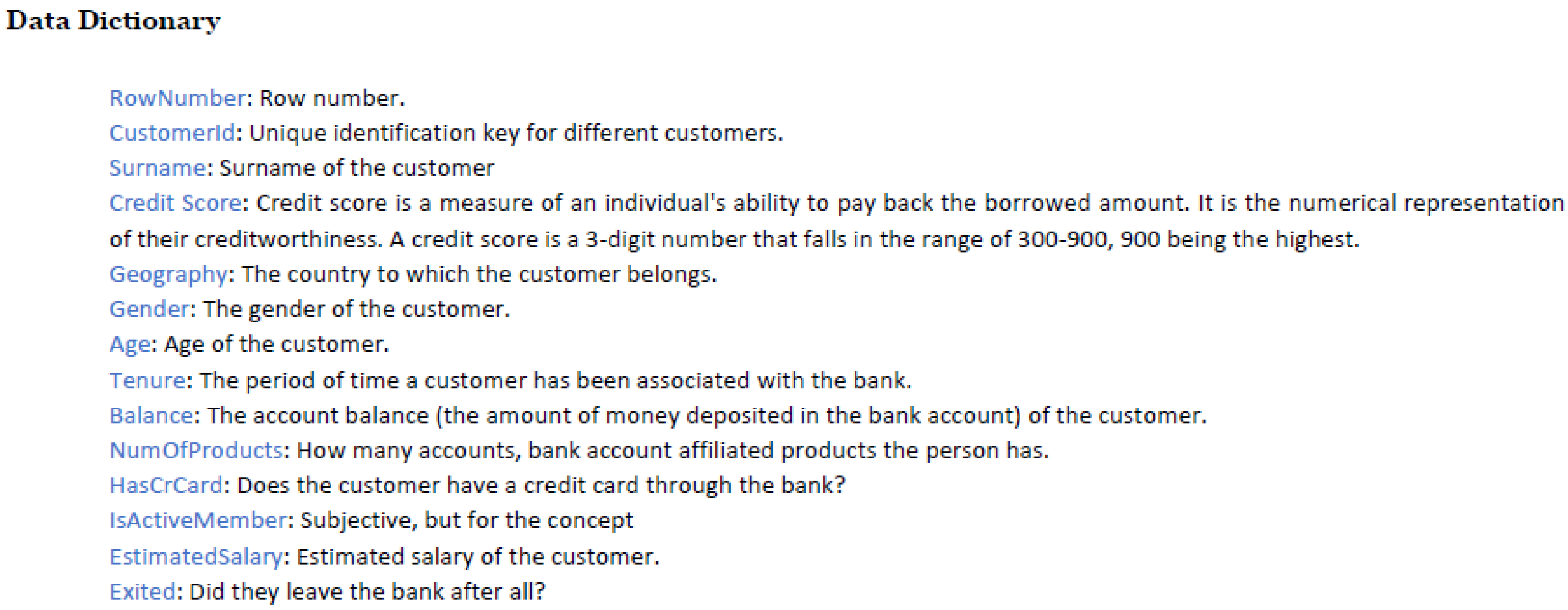
Problem Statement

In this project on bank customer churn prediction using machine learning, I am trying to explain how a local bank - Hopkinton Credit Union can use predictive analytics to help its marketing staff to identify which aspects of the service influence a customer's decision in this regard. Management can concentrate efforts on the improvement of service, keeping in mind these priorities. Using the data collected from existing customers, build a model that will help the marketing team identify potential customers who are relatively more likely to churn.

Given a Bank customer, we will **build a neural network-based classifier** that can determine whether they will leave or not in the next 6 months.

Data Collection

The Neural Network-based Classifier/ML Model work in this project will use the datasets taken from Open-Source Dataset from Kaggle. This is a comprehensive dataset collected by the bank’s marketing team. The dataset contains 10,000 sample points with 14 distinct features such as CustomerId, Credit Score, Geography, Gender, Age, Tenure, Balance, etc.



Data Dictionary

RowNumber: Row number.
CustomerId: Unique identification key for different customers.
Surname: Surname of the customer
Credit Score: Credit score is a measure of an individual's ability to pay back the borrowed amount. It is the numerical representation of their creditworthiness. A credit score is a 3-digit number that falls in the range of 300-900, 900 being the highest.
Geography: The country to which the customer belongs.
Gender: The gender of the customer.
Age: Age of the customer.
Tenure: The period of time a customer has been associated with the bank.
Balance: The account balance (the amount of money deposited in the bank account) of the customer.
NumOfProducts: How many accounts, bank account affiliated products the person has.
HasCrCard: Does the customer have a credit card through the bank?
IsActiveMember: Subjective, but for the concept
EstimatedSalary: Estimated salary of the customer.
Exited: Did they leave the bank after all?

Data Exploration

As part of the EDA, will perform Uni-variate analysis & bivariate analysis with Data Visualization. Generate insights from the dataset to find proportion of customers churned and retained.

Some hypotheses to test are :-

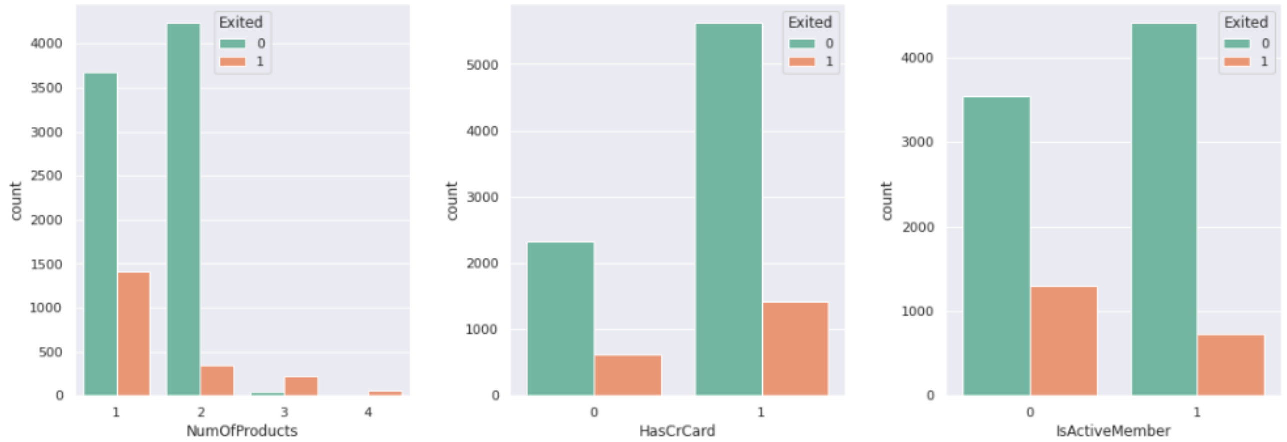
(i) Customers with low credit score tend to churn, reasonable

(ii) On average older customers are the most to churn, but its questionable

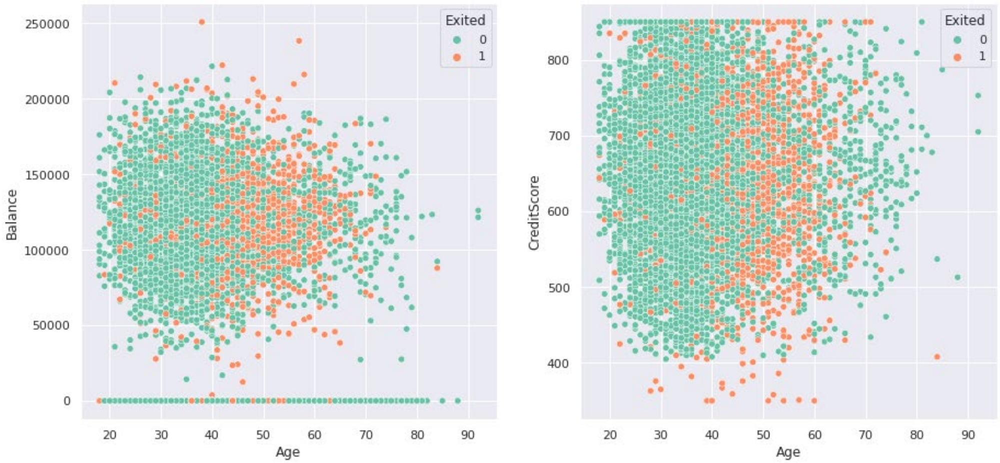
(iii) Customers with high balance are churning probably they are getting attracted by other banks offer to raise the wealth and this corresponds with their estimated salary

(iv) Tenure, Credit card and being active mean are not explicitly helping in this case to highlight anything big for churn rate

EDA Summary

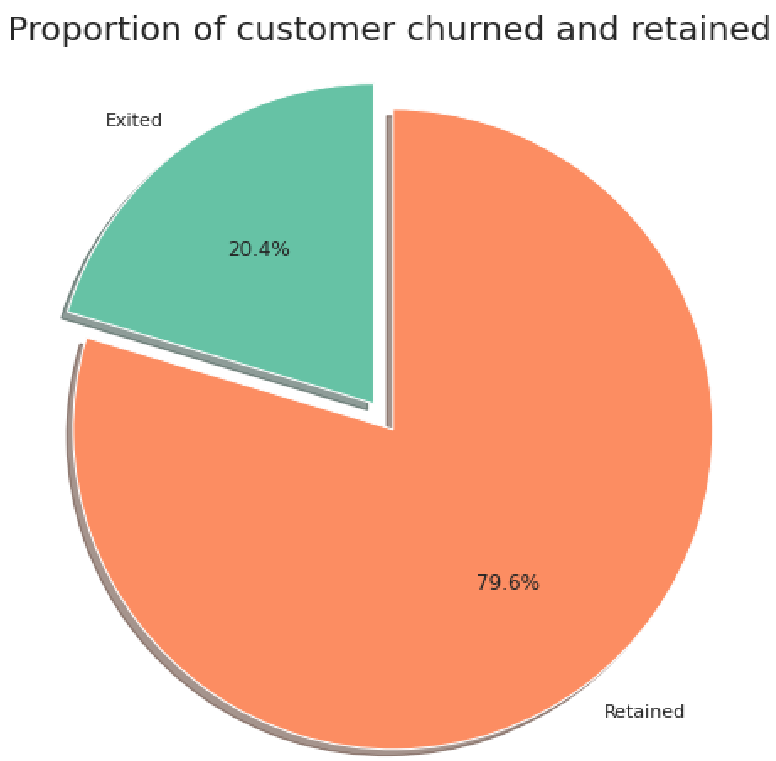


Insights : Customer with 3 or 4 products are higher chances to Churn



Insights :

- 40 to 70 years old customers are higher chances to churn
- Customer with CreditScore less then 400 are higher chances to churn


Proportion of customer churned and retained

Insights :So about 20% of the customers have churned. So the baseline model could be to predict that 20% of the customers will churn. Given 20% is a small number, we need to ensure that the chosen model does predict with great accuracy this 20% as it is of interest to the bank to identify and keep this bunch as opposed to accurately predicting the customers that are retained.

EDA Summary - Visualization

Feature engineering is the process of selecting, manipulating, and transforming raw data into features that can be used in supervised learning. To make machine learning work well on new tasks, it might be necessary to design and train better features. A “feature” is any measurable input that can be used in a predictive model — it could be the color of an object or the sound of someone’s voice. Feature engineering, in simple terms, is the act of converting raw observations into desired features using statistical or machine learning approaches.

The following steps were done as part of feature engineering in this project: -

- Prepare dataset.
- Encoding Categorical data
- Splitting data into training set and test set
- Apply feature scaling.

Feature Engineering

Build and Train ANN using Tensorflow

Used **Google Colab** and Installed **TensorFlow** to build and train ANN

The steps involved are :-

- Initialize ANN
- Adding the input layer and first hidden layer
- Adding the second hidden layer
- Adding the output layer
- Compile ANN
- Train ANN using Training Dataset

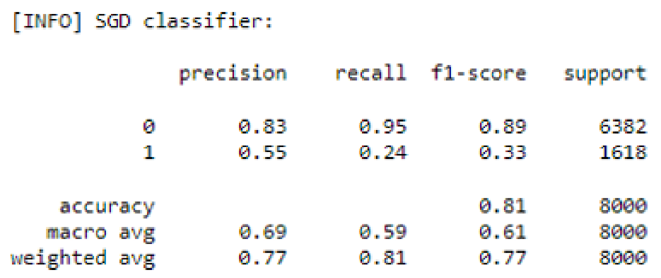
Model Predictions and Evaluating Models using different ML Algorithms

Try to train different machine learning classification models to our data. Once we get the model details for each of the models, we can select the best model from them for our training and testing purposes.

The following 6 ML Libraries were used in the project:-

- **SGD Classifier**
- **Logistic Regression Classifier**
- **Support Vector Machines (RBF Kernel)**
- **Support Vector Machines (Poly Kernel)**
- **Random Forest Classifier**
- **Extreme Gradient Boost (XGBoost) Classifier**

ML Classifiers - Visualize results (precision, recall, f1-score)

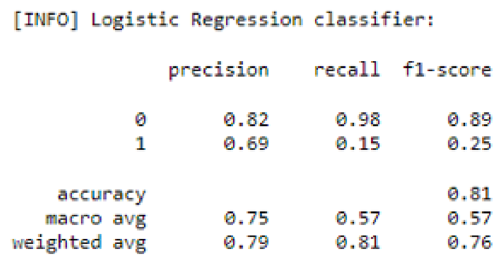


```
[INFO] SGD classifier:

              precision    recall  f1-score   support

           0       0.83      0.95      0.89      6382
           1       0.55      0.24      0.33      1618

    accuracy                           0.81      8000
   macro avg       0.69      0.59      0.61      8000
weighted avg       0.77      0.81      0.77      8000
```

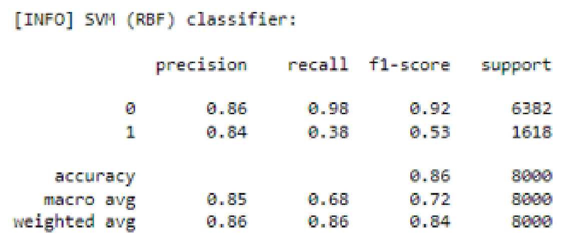


```
[INFO] Logistic Regression classifier:

              precision    recall  f1-score

           0       0.82      0.98      0.89
           1       0.69      0.15      0.25

    accuracy                           0.81
   macro avg       0.75      0.57      0.57
weighted avg       0.79      0.81      0.76
```

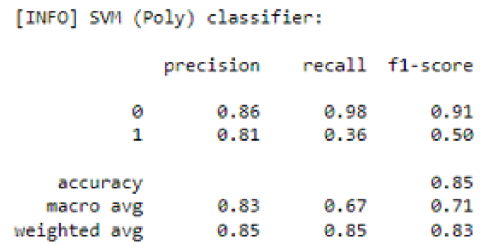


```
[INFO] SVM (RBF) classifier:

              precision    recall  f1-score   support

           0       0.86      0.98      0.92      6382
           1       0.84      0.38      0.53      1618

    accuracy                           0.86      8000
   macro avg       0.85      0.68      0.72      8000
weighted avg       0.86      0.86      0.84      8000
```

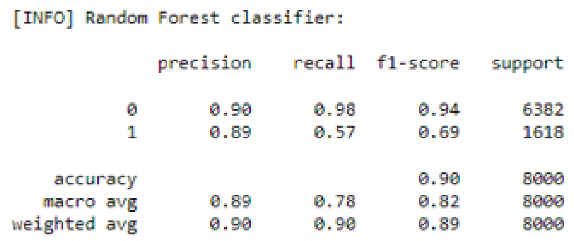


```
[INFO] SVM (Poly) classifier:

              precision    recall  f1-score

           0       0.86      0.98      0.91
           1       0.81      0.36      0.50

    accuracy                           0.85
   macro avg       0.83      0.67      0.71
weighted avg       0.85      0.85      0.83
```

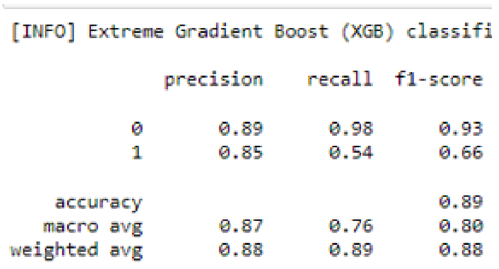


```
[INFO] Random Forest classifier:

              precision    recall  f1-score   support

           0       0.90      0.98      0.94      6382
           1       0.89      0.57      0.69      1618

    accuracy                           0.90      8000
   macro avg       0.89      0.78      0.82      8000
weighted avg       0.90      0.90      0.89      8000
```



```
[INFO] Extreme Gradient Boost (XGB) classifi

              precision    recall  f1-score

           0       0.89      0.98      0.93
           1       0.85      0.54      0.66

    accuracy                           0.89
   macro avg       0.87      0.76      0.80
weighted avg       0.88      0.89      0.88
```

Model Results Summary

Conclusion

The precision of the model on previously unseen test data is slightly higher with regard to predicting 1's i.e., those customers that churn. However, in as much as the model has a high accuracy, it still misses about half of those who end up churning. This could be improved by providing retraining the model with more data over time.

Recommendations for Hopkinton Credit Union Bank

The most important signs to look out for customer churn are the following: -

- client's age
- credit speed
- expected profit.
- account balance
- number of products

So, it’s better to plan the marketing strategy and product offering/service by carefully evaluating these.

To predict customer churn, use a model based on the Random Forest algorithm.

Conclusion & Recommendations