# Customer Churn Prediction

## Understanding Problem Statement

Decreasing the Customer Churn is a key goal for any business. Predicting Customer Churn (also known as Customer Attrition) represents an additional potential revenue source for any business. Customer Churn impacts the cost to the business. Higher Customer Churn leads to loss in revenue and the additional marketing costs involved with replacing those customers with new ones.

In this challenge, as a data scientist of a bank, you are asked to analyze the past data and predict whether the customer will churn or not in the next 6 months. This would help the bank to have the right engagement with customers at the right time.

## Objective

Our objective is to build a machine learning model to predict whether the customer will churn or not in the next six months.

**Training set**

**train.csv** contains the customer demographics and past activity with the bank. And also the target label represents whether the customer will churn or not.

| Variable | Description |
|---|---|
| ID | Unique Identifier of a row |
| Age | Age of the customer |
| Gender | Gender of the customer (Male and Female) |
| Income | Yearly income of the customer |
| Balance | Average quarterly balance of the customer |
| Vintage | No. of years the customer is associated with bank |
| Transaction_Status | Whether the customer has done any transaction in the past 3 months or not |
| Product_Holdings | No. of product holdings with the bank |
| Credit_Card | Whether the customer has a credit card or not |
| Credit_Category | Category of a customer based on the credit score |
| Is_Churn | Whether the customer will churn in next 6 months or not |

## Observations from EDA:

1. As this the **binary classification problem and the target data are imbalanced**. Imbalanced class distributions influence the performance of a machine learning model negatively so I did upsampling to overcome this issue.
2. There are no null values.
3. **Female customers are more likely to churn than male customers.**

4. **Customers having more vintage are more likely to churn (**The bank should take care of the customers who continued their service till 1 year and 3 to 4 years as they can be retained if they are going to churn. Continuous feedback from customers is a better way to decrease the churn rate).
5. **The customers who have not done any transaction in the past 3 months are more likely to churn** (The bank must reach out to the customers to know the reason. Feedback from customers always helps them to understand and improvements in the services).
6. **Customers who are having poor credit scores are more likely to churn**.
7. **Customers having an income of 10L and more are more likely to churn.**
8. **Customers who have not done any transaction in the past 3 months are more likely to churn.**
9. **Customers who are holding only 1 product are more likely to churn**.

**Techniques of data pre-processing:**

**Vectorization:** Dataset which contains different categories of classes then we have to do one-hot encoding to convert those categorical values into integer representations, which can be handled by our model. Used **one-hot encoding** in this project to convert categorical values to integer values.

Applied normalization min max scaler for 4 columns.

**Model Selection Approach:**

As a base model, I have chosen the Ridge classifier model. Then I have tried Logistic regression, Decision tree, Random Forest, and XG boost prediction but Logistic regression gave me the best f1 score out of all.

I have split the train data into train and test data sets to evaluate the model through confusion matrix and classification report.

**Ridge model**

The model achieved 62% accuracy on the training set and 59.6% accuracy on the test set.

**Logistic Regression model**

The model achieved 62% accuracy on the training set and 59.5% accuracy on the test set.

**Decision Tree model**

The DT model improved accuracy and achieved 76% accuracy on the
training set and 68.5% accuracy on the test set. But not performed well on the public leader board.

**Random forest model:**

The RF model improved accuracy a lot and achieved 88% accuracy on the training set and 77% accuracy on the test set.

**XG Boost model:**

The model achieved 65% accuracy on the training set and 61.5% accuracy on the test set.

But out of all of the models, Logistic regression gave me the best f1 score on the public board. So **Logistic regression is my final model here.** I have tried many different parameters to improve the accuracy.

Thank you for this opportunity, Vidhya Analytics.

Sonya Naidu