Churn Prediction

Le Thi Thanh Tam

9/1/2022

Objective

Case : Which customers are likely to churn ?

- **Customer Churn** occurs when customers leave/stop doing business with the company or service.
- The ability to predict when a customer is at a high risk of churning is valuable for every business with returning customers. Churn is defined as the number of customers cancelling within a time period divided by the number of active customers at the start of that period. In order to apply a modeling technique to predict churn, we need to understand the customer behavior and characteristics which signal the risk of customers churn.
- For this analytics, I will look into a bank customer data to predict whether the customer will leave the credit card services of the bank.
- I will use R for this project. The number of data is not too large so can use R directly. (R is one of the predominant languages in data science ecosystem and makes it simple to efficiently implement statistical techniques and thus it is excellent choice for machine learning tasks).

Model

Classification Problem

First I transformed some categorical variable into numeric (i.e income) and readable code. I also split the data into train and tests sets with a test size of 20%. I tried two different models and evaluated the accuracy based on the test error rate

- **Logistic Regression** the small p-values associated with almost all of the predictors
- However **RandomForest** with the lower test error rate
- Before starting I transformed the type for each column.
  - Change the values of Attrition Flag, Gender, Marital Status as factor, the category Income into Numeric

1. Logistic Regression

```
##
## Call:
## glm(formula = Attrition_Flag ~ ., family = binomial, data = churn2)
##
```

```
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -2.0930  -0.5003  -0.2914  -0.1377   3.7374  
## 
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)    
## (Intercept)               1.851e+00  3.584e-01   5.165 2.41e-07 ***
## Customer_Age             -1.628e-03  4.675e-03  -0.348 0.727730    
## Gender1                  -5.884e-01  1.252e-01  -4.701 2.59e-06 ***
## Dependent_count           6.213e-02  2.948e-02   2.108 0.035050 *  
## Education_Level           4.770e-03  6.009e-03   0.794 0.427306    
## Marital_Status1          -1.705e-01  7.843e-02  -2.174 0.029686 *  
## Marital_Status3          -3.246e-03  1.420e-01  -0.023 0.981767    
## Card_CategoryGold         9.035e-01  3.710e-01   2.435 0.014878 *  
## Card_CategoryPlatinum     1.392e+00  7.133e-01   1.952 0.050981 .  
## Card_CategorySilver       5.331e-01  1.900e-01   2.806 0.005018 ** 
## Total_Relationship_Count -4.325e-01  2.686e-02 -16.105  < 2e-16 ***
## Months_Inactive_12_mon    4.229e-01  3.656e-02  11.568  < 2e-16 ***
## Contacts_Count_12_mon     4.829e-01  3.600e-02  13.413  < 2e-16 ***
## Credit_Limit             -1.636e-05  6.406e-06  -2.554 0.010641 *  
## Total_Revolving_Bal      -7.183e-04  7.214e-05  -9.957  < 2e-16 ***
## Total_Amt_Chng_Q4_Q1     -5.758e-02  1.981e-01  -0.291 0.771301    
## Total_Trans_Amt          -2.024e-04  1.887e-05 -10.728  < 2e-16 ***
## Total_Ct_Chng_Q4_Q1      -4.084e+00  2.249e-01 -18.162  < 2e-16 ***
## Avg_Utilization_Ratio    -5.491e-01  2.431e-01  -2.259 0.023899 *  
## Income                    6.407e-06  1.742e-06   3.678 0.000235 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 6948.7  on 7972  degrees of freedom
## Residual deviance: 4835.9  on 7953  degrees of freedom
## AIC: 4875.9
## 
## Number of Fisher Scoring iterations: 6
```

Some findings from glm : * **Gender1** has small p-value means it is associated with our target. The negative coefficient for this predictor suggests **Male** is less likely to churn. * **Income** has a positive relationship with the churn likelihood. * Person made **The total transaction amount** large tended to not churned . etc

Better assess the accuracy of the logistic regression model

- First split data into training and test sets
- Fit a logistic regression model on train data set
- Predict probabilities of churn customers on test set
- Compute the predictions and compare them to the actual churn customers
- test error rate equal 11%
- We recall that the logistic regression model, the small p-values associted with almost all of the predictors.

- In theory, consider the distribution of the predictors X (EDA part) is approximately normal in each of the classes, the logistic regression model may be unstable.
- Therefore we will consider to obtain better model for this project

RandomForest

- The classification accuracy of the model on the test set is 94.4% , test error rate is 5.6%

After developing the model, the model is applied to all customers such that we obtain the likelihood of churning for each customer. Ranking the results gives you the top X customers who are about to churn.