Universal Plus Transaction Prediction System

Model Report

9 December 2020

Prepared by:

Name	ID
Rohan Kunte	2058395
Stylianos Killas	2084989
Qixin Jia	2058435
Dushant Gohri	2087013
Cheng Cheng	2049865
Mingjian Ma	1959993

Word Count: 1999

Contents

1 Introduction

The following is a report of the details and inspirations behind the development of our transaction prediction model requested by our potential client Universal Plus. Through the CRISP-DM methodology our team turned the business problem into multiple smaller data mining tasks which led the foundation to the model that would be proposed as the solution to the busines problem.

2 Business Understanding

2.1 Business Problem

The business problem is customer churn for our potential client Universal Plus. In the competitive market of banking, retention of valuable customers is essential for the success of the bank. We are asked to create a transaction prediction model able to identify which existing customers of Universal Plus are likely to make a transaction in the future. Our model would allow Universal Plus to take pro-active steps to retain the predicted customers, giving Universal Plus a competitive advantage within the market.

3 Data Understanding

3.1 Strengths and Weaknesses

The data provided by our potential client, demonstrated a highly imbalanced target variable along with very low correlation between the explanatory and target variables. Datasets' strengths were that the data was separated into a full and a small dataset with almost identical target variable proportions and percentage of outliers. This allowed us to work timely and efficiently on the small dataset while maintaining a similar environment to the total population. Furthermore, the dataset provided seemed to be pre-processed since there was no multicollinearity between the variables and the distribution of the variable instances appeared to be to a certain degree normalized. More detail on data characteristics will be developed in 'Data Preparation'.

3.2 Approach to Data Mining

Our approach is to create a clean dataset for training with as little information noise as possible while retaining as much information as possible. Due to the lack of busines information on the variables provided by the potential client, we assumed variables to be

regarded as important and useful by the potential client as to be selected for the particular dataset. Taking also into consideration that the datasets appear to be pre-processed it would be fair to assume that to a certain degree what our team received as data is not raw, unprocessed data.

4 Data Preparation

4.1 Importing

The first step was importing the data and cleaning up any duplicates or corrupted entries likely caused by human error, as to produce a dataset with only relevant and valid information. The 'ID_Code' of each customer in the database was remove as it provides no information for the target variable. Furthermore, the target variable was converted from numeric to factor, with values '1' and '0' representing 'transaction' and 'no transaction. Characteristics of the data mentioned in 'Data Understanding' are shown in Figure 1 and Figure 2.

	Percentage %
Outlier Small	13.73933
Outlier Full	14.19997

Figure 1 - Percentage of Outliers per data set

	Dataset Small %	Dataset Full %
Target = 0	90.324417	90.323761
Target = 1	9.675583	9.676239

Figure 2 - Proportion of "0" and "1" observations in each data set

4.2 Outliers

We proceeded to remove the entries that contained instances outside the respective variables' normal distribution interquartile range or otherwise known as outliers. Our aim was to remove any outliers or extreme values not representative of the normal distribution and are unlikely to re-occur. We produced a dataset focused on the general majority of the population with less noise created by extreme instances, possibly hindering the predictive performance of the model.

Our options were to remove outliers and respective rows or replace outliers with each columns' respective mean. As mentioned in 'Data Approach', we aimed to preserve the rest of the information that resided within the respective rows expecting that information that would otherwise be omitted could improve the accuracy of the model. It resulted in an increase of ~1% in sensitivity at the cost of ~1% decrease in accuracy and specificity respectively. Contradicting our expectations, the results led us to believe that where outliers are identified, the majority of the features lie outside the normal distribution of each feature. For the above reasons our team decided that it would be beneficial to omit outlying records as to reduce noise, since such records provide no useful information.

4.3 Partitioning

We decided that since the dataset significantly unbalanced, we should partition training and test data into 80% and 20% accordingly, to provide as much information as possible for the model to train.

4.4 Balance

Unbalanced datasets are abundant in real life and in academic literature, for creating problems with machine learning and predictive modelling. The method we concluded on was 'Synthetic Minority Over-Sampling Technique' or SMOTE. The necessity for balancing and the decision for using 'SMOTE' was influenced by Nitesh V. Chawala's (2002) paper on 'SMOTE'. The paper emphasizes the imbalance problem in machine learning predictive modelling, while mentioning how often the minority class is of greater value, both attributes present in our model. The paper suggests the necessity for balancing while demonstrating how 'SMOTE' produces more useful and generalized results compared to traditional re-sampling techniques. SMOTE over other methods, achieves a balance between the classes through under sampling and creation of synthetic but unique records, which are created by neighboring data points. This creates a balanced but rather generalized model which is more fitted to predict a greater variety of datasets and would not overfit the existing training set as much as traditional techniques.

4.5 Feature Selection

We decided to not implement feature selection and keep all the available information for the following reasons. First, the imbalance of the dataset and the low influence of the independent variables to our dependent variable, resulted in poor feature selection processes. Feature selection would only yield acceptable results after balancing, leading to an inconsistent and unstable dataset since each balancing method would produce different variables in feature selection. Second, our team run a correlation matrix to identify the correlation between all the columns with every other column and the target variable. The paper of Thu Pham-Gia (2014), discusses in depth benefits and applications of correlation matrix. Correlation matrix is crucial in multivariate analysis as it allows identification of the relationship between the variables and their components something really useful for our model. The paper is also relevant as the distribution of each variable is observed to be normal, where according to the paper, the concepts of correlation and independence between variables are equivalent. The results showed very low correlation between explanatory variables and the target variable. While this indicates the positive no multicollinearity attribute, it also indicates very little information of the explanatory variables to the target variable.

5 Modelling

5.1 General Linear Model (GLM)

According to Maalouf M. (2011), “Generalized Linear Model” is a good binary predictor for generalized models with the benefit of being able to extrapolate with little or no data. Furthermore GLM performance better when the data is balanced and in conjunction with SMOTE balancing technique our model, developed through ten-fold cross validation, resulted in the best results according to other comparable models. Cross validation was applied to reduce noise and error, as to allow the model to better evaluate predictors and achieve a better result. The probability thresholds for the predictive GLM was changed to 0.4, which according to our calculation of business understanding provided the best results.

5.2 Random Forest

Random Forest was studied from Chao Chen’s (2004) paper. We used Random Forest (RF) as a candidate model because it provides satisfying results while maintaining a generalized error comparable to other models. Furthermore, RF could be adjusted for imbalanced data by either weighting the costs of misclassification or balancing the model. Since our model was balanced through the SMOTE method and since XGBoost which also uses weights to make predictions we seemed fit to use RF balanced through SMOTE. According to Chao Chen’s report, RF balanced with SMOTE achieves improved results over other balancing methods such as SHRINK. Although the RF model was not able to produce satisfactory results about the target variable even though it achieved a high accuracy level. Our dataset is much more complicated and large for RF to produce satisfactory results on target variable.

5.3 XGBOOST

A summary of the information gathered about the attributes and advantages of the model from the paper of Tianqi Chen (2016). XGBoost is a combination of algorithms rather than one single algorithm. The nature of XGBoost is repetition and iteration, multiple runs, each one building on the precedent one, making incremental and continuous improvements to figure out which features matter most. As mentioned in Tianqi Chen’s paper, parallel processing allows for the algorithm to run faster and require less computational power as it cleverly utilizes out-of-core computation by handing disk space to data that does not fit into the main memory, preventing bottle neck in computational power. Able to upscale, analyse and predict results much faster than the other models. The method we’ve chosen, the tree-based method is more suitable for highly imbalanced datasets. XGBoost runs on different records each time, allowing it to deal with possible multicollinearity which may appear in future datasets. Furthermore, XGBoost treats NA values as information through a process called “Sparsity-aware Split Finding”, which makes our approach very flexible and long term,

as this model will be able to perform with less data preparation on future datasets that may have varying levels of NA values. Lastly our approach has the additional benefit of being suitable at dealing with non-linear relationships between the dependent and independent variables, making this a model that would work under a great variety of future datasets.

6 Evaluation

According to the paper of Hossin M. (2015), we decided that we would take into consideration a combination of metrics, with primary metric the AUC value. As accuracy is a is often inconsistent in predictive informativeness of a model it is considered least in our evaluation. Since our business problem refers to the ability to predict the minority class, a combination of AUC value, recall/sensitivity value, along with F-Metric value will be considered. The F-Metric value is important as it represents the summarised predictive ability of the model indicating the balance between the recall and precision metrics. As mentioned in the paper of Hossin M. (2015), AUC is the most important evaluation metric because it can measure accurately the predictive performance of models both in binary and multiclass problems along with being able to handle large volumes of data. This makes our decision model better fitting and sustainable in the long run as it will be expected to perform equally well in multiclass problems as well if requested in the future with minimal remodelling, compared to alternative models mentioned.

7 Figures

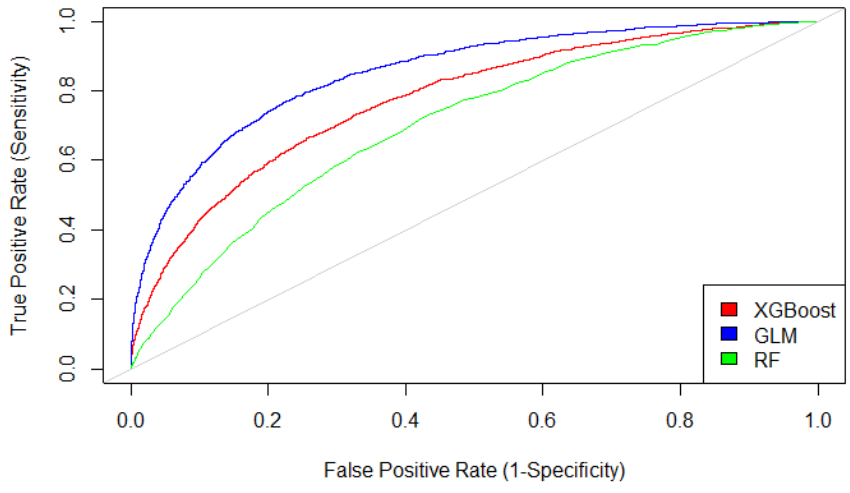

Figure 3 - ROC Chart

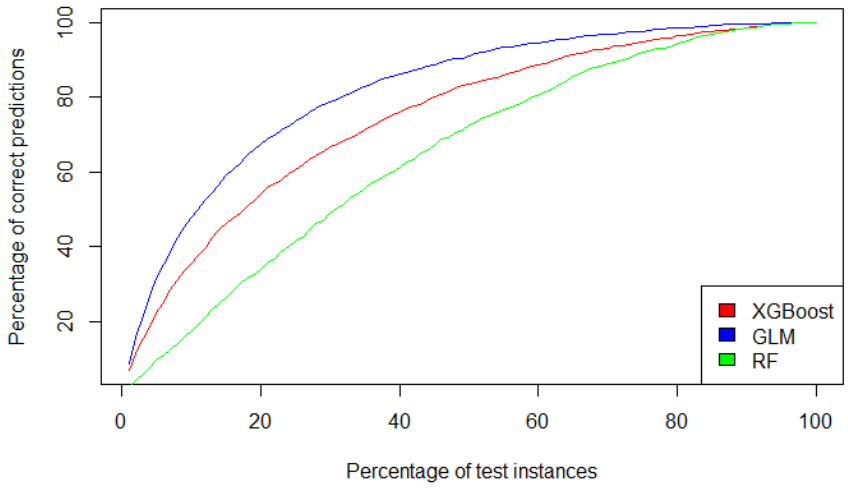

Figure 4 - Gain Chart

	ACCURACY	PRECISION	RECALL	F-METRIC	AUC
GLM	0.6974	0.2214	0.8453	0.3509	0.8519
XGBOOST	0.6529	0.1847	0.7572	0.2969	0.7741
RF	0.8863	0.2621	0.0963	0.1408	0.6999

Figure 5 - Summary Table for Important Metrics

```
Confusion Matrix and Statistics


prediction_XGB     0     1
             0 10141   411
             1  5661  1282

               Accuracy : 0.6529
                 95% CI : (0.6458, 0.66)
    No Information Rate : 0.9032
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1673

 Mcnemar's Test P-Value : <2e-16

              Precision : 0.18465
                 Recall : 0.75724
                     F1 : 0.29690
             Prevalence : 0.09677
         Detection Rate : 0.07328
   Detection Prevalence : 0.39686
      Balanced Accuracy : 0.69949

       'Positive' Class : 1

Confusion Matrix and Statistics
```

Figure 6 - XGBoost Confusion Matrix Test

```
Confusion Matrix and Statistics


prediction_XGB_training     0     1
                      0 29348     1
                      1  4507 40625

               Accuracy : 0.9395
                 95% CI : (0.9377, 0.9412)
    No Information Rate : 0.5455
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8766

 Mcnemar's Test P-Value : < 2.2e-16

              Precision : 0.9001
                 Recall : 1.0000
                     F1 : 0.9474
             Prevalence : 0.5455
         Detection Rate : 0.5454
   Detection Prevalence : 0.6060
      Balanced Accuracy : 0.9334

       'Positive' Class : 1
```

Figure 7 - XGBoost Confusion Matrix Training

```
Confusion Matrix and Statistics


prediction_GLM     0     1
             0 10770   262
             1  5032  1431

               Accuracy : 0.6974
                 95% CI : (0.6905, 0.7042)
    No Information Rate : 0.9032
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2333

 Mcnemar's Test P-Value : <2e-16

              Precision : 0.22141
                 Recall : 0.84525
                     F1 : 0.35091
             Prevalence : 0.09677
         Detection Rate : 0.08179
   Detection Prevalence : 0.36942
      Balanced Accuracy : 0.76340

       'Positive' Class : 1

Confusion Matrix and Statistics
```

Figure 8 - GLM Confusion Matrix Test

```
Confusion Matrix and Statistics


prediction_GLM_training     0     1
                      0 23021  4348
                      1 10834 36278

               Accuracy : 0.7962
                 95% CI : (0.7933, 0.7991)
    No Information Rate : 0.5455
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5823

 Mcnemar's Test P-Value : < 2.2e-16

              Precision : 0.7700
                 Recall : 0.8930
                     F1 : 0.8270
             Prevalence : 0.5455
         Detection Rate : 0.4871
   Detection Prevalence : 0.6325
      Balanced Accuracy : 0.7865

       'Positive' Class : 1
```

Figure 9 - GLM Confusion Matrix Training

```
Confusion Matrix and Statistics


prediction_RF     0     1
            0 15343  1530
            1   459   163

               Accuracy : 0.8863
                 95% CI : (0.8815, 0.891)
    No Information Rate : 0.9032
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0937

 Mcnemar's Test P-Value : <2e-16

              Precision : 0.262058
                 Recall : 0.096279
                     F1 : 0.140821
             Prevalence : 0.096771
         Detection Rate : 0.009317
   Detection Prevalence : 0.035553
      Balanced Accuracy : 0.533616

       'Positive' Class : 1
```

Figure 10 - Random Forest Confusion Matrix Test

```
  Confusion Matrix and Statistics


  prediction_RF_training     0     1
                       0 33855     0
                       1     0 40626

                 Accuracy : 1
                   95% CI : (1, 1)
      No Information Rate : 0.5455
      P-Value [Acc > NIR] : < 2.2e-16

                    Kappa : 1

    Mcnemar's Test P-Value : NA

                Precision : 1.0000
                   Recall : 1.0000
                       F1 : 1.0000
               Prevalence : 0.5455
           Detection Rate : 0.5455
     Detection Prevalence : 0.5455
        Balanced Accuracy : 1.0000

         'Positive' Class : 1
```

Figure 11 - Random Forest Confusion Matrix Training

8 Conclusion

Our team decided that after inspection of our results as presented in the ‘7 Figures’ part of the report, and more specifically Figure 5 we decided that the best model is GLM. GLM outperforms the other models in all the metrics that we choose to evaluate the models on. GLM according to AUC has the higher informative ability to predict target variables in the specific dataset, while achieving the highest percentage or recall which is our main identifier for target prediction accuracy. GLM also maintains the highest F-Metric from between the models which indicates a healthy balance for the metrics recall and precision. Understanding the business problem and breaking it down into smaller data mining tasks allowed us to conclude on GLM model effectively and efficiently as our predictor. While through data mining we were able to develop a deeper understanding about our dataset and treat it accordingly as to achieve successfully a predictive solution for our potential client Universal Plus.

9 References

Chao Chen, A. L. (2004). *Using Random Forest to Learn Imbalanced Data.* California: UC Berkeley.

Hossin M, S. M. (2015). A Review On Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process Vol. 5, No. 2*.

Maher, M. (2011). Logistic Regression in Data Analysis: An Overview. *International Journal Data Analysis Techniques and Strategies*, 281 - 299.

Nitesh V. Chawla, K. W. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research 16*, 321-357.

Thu Pham-Gia, V. C. (2014). Distribution of the Sample Correlation Matrix and Applications. *Open Journal of Statistics*, 330-344.

Tianqi Chen, C. G. (2016). XGBoost: A scalable Tree Boosting System. *International Concerence on Knowledge Discovery and Data Mining*, 785-794.