Customer Lifetime Value

Jeremiah Onyema
Supervisor: Prof. Fabian Transchel
Hochschule Harz
Faculty of Automation and
Computer Science
U35970@hs-harz.de

ABSTRACT

Accurate churn prediction is a growing problem in many business sectors. This is because of the huge amount of data generated per day through various mediums. This has often contributed to making it difficult for organization to draw accurate business intelligence from structured data. The growing application of machine learning to churn prediction modelling in structured data and its promising performance more than the traditional methods, has proven that structured data possess reliable business intelligence yet untapped.

In this work, we aim to build churn prediction model by using bank's customers behavior from structured data. We conducted experiments on existing datasets from Kaggle on ten thousand bank customers. In this paper, our experimental result using random forest classifier which clocked an Area Under Curve (AUC) score of 0.82 shows that structured data impacts the accuracy of churn analysis significantly.

Keywords

Churn Prediction; Unstructured Data; Customer Data Analytics; Machine Learning.

1.0 Introduction

Many organizations are beginning to rely on data analytics as a tool to drive competitive edge in their businesses (Uthayasankar S. etal. 2017).

The recent development in churn prediction using machine learning, has provided more approaches to gaining more business intelligence from structured data (Farid S. 2018). Churn prediction over the decades has proven to be a good indicator for evaluating the probability of a customer terminating his business relationship with an organization. Within the scope of this paper, we will utilize the dataset details of a bank's customers. The target variable in this dataset is a binary variable which reflects whether a customer left the bank (closed his account) or the customer remains with the bank. This will help the bank in establishing marketing and customer retention strategies with more focus on customers with high probability of closing their account.

The main contribution of our paper is:

- Establishing the impact of structured data on churn prediction accuracy through machine learning approach, using a bank's customers data obtained from Kaggle, a machine learning and data science community.

After the introduction in Section 1, we look at related works with major focus on the application of machine learning to structured data for churn prediction in Section 2. Section 3 provides the description of our methodologies and datasets. The machine learning model was developed and applied to the bank's customers datasets to perform churn prediction on the target label in Section 4. We drew our conclusion in Section 5.

2.0 Related Literature

For some decades, there have been vast traditional and machine learning methods applied to obtain business intelligence that could help in customer segmentation, marketing planning and customer retention. Telecommunication industry has received more attention on churn prediction as a result of its large amount of dataset (Gordon & Michael, 2011).

Financial services industry has already witnessed some but limited churn prediction models, more especially in the area of private financial institutions (Ozden and Umut, 2014). Support vector machine has been the predominant methodology for building models for financial industry as it has proven to achieve good results in general (Mohammed et al, 2014). As noted by (Ngai, Xiu, & Chau, 2009), many researches on churn prediction model have been geared towards structured data and the use of decision trees, logistic regression and others. Also in recent years, the intention of replacing the approach of single prediction model with hybrid model approach in order to increase churn prediction accuracy has been at the forefront of prediction model building (Huang & Kechadi, 2013). Most of the works in machine learning methodologies whenever attempts are made to analyze structured data are geared towards the telecommunication industry because of the large amount of dataset generated in the sector. In this paper, we will apply machine learning approach to build a churn prediction model from structured data in the financial industry.

3.0 Methodology

This section describes conceptually the different stages employed in this project, ranging from obtaining the customers data to the final churn prediction as could be seen here.

Figure 1 below shows the algorithm from the sources of the structured data, to its processing and finally the churn prediction modelling. The first stage usually is to collect data from databases, followed by different pre-processing tasks in order to structure the data into a more usable format. The last step is to estimate the churn probabilities for each customer by applying a suitable statistical model.

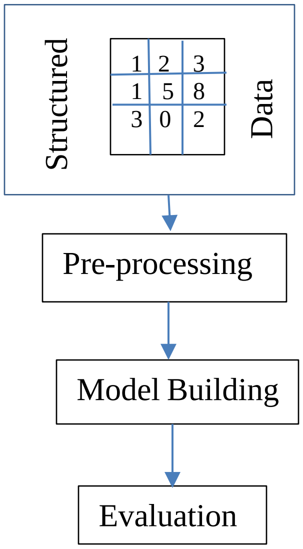


Figure 1. Processes of Churn Prediction

3.1 Structured Data

Structured data are the organized, coherent type of data and usually depend on a data model. These types of data include excel files and relational databases (Enterprise, 2014).

Datasets for prediction model are often grouped into two categories for the purpose of training and testing a model. The training data is often labelled and each customer assigned a class either churn or no churn, based on the observed behavior in the past (Berson et al).

3.2 Pre-Processing

Structured data often require varieties of pre-processing, though often less intensive than unstructured data. Examples include quality data check, feature aggregation and sampling, feature encoding, dimensionality reduction and other pre-processing tasks. For our projects, we applied feature encoding to the categorical features by creating dummy variables. We also checked for data quality by assessing the dataset for missing values and variables, inconsistent values such as wrong label on a row.

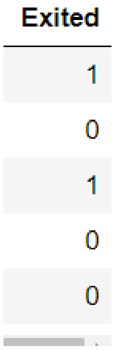
The dataset for this project is pre-processed bank's customers data with demographic features, and target variable which is a binary variable.

3.3 Feature Selection

Feature selection is done to eliminate unnecessary information from data, and ensure that features that could help answer the research questions are selected. The goals of feature selection could be different depending on the stage where it is applied. For example, only words that could significantly impact churn probability are usually considered.

In our paper, the target variable Exited, denoted as churn (1) or no churn (0) was used as the training data as could be seen in Table 1 below.

Table 1: Target Label



Exited
1
0
1
0
0

Also, categorical features in our dataset which do not significantly impact the target label were dropped in order to reduce the dimension of our dataset to a more manageable size.

3.4 Modelling

Determining churning probability is the basic focus of modelling in churn prediction. Support vector machines, Nayes Bayes and other methods have been employed in doing this (Ngai, Xiu, & Chau, 2009). The most widely used method is Random Forest due to its ease of application and robustness in handling complex data. Again, we decided to apply Random Forest in our project because our dataset is made up of more binary variables and also apply holdout method of model validation by dividing our datasets into 10 estimators. In this case, we repeat the hold out method for ten number of iterations.

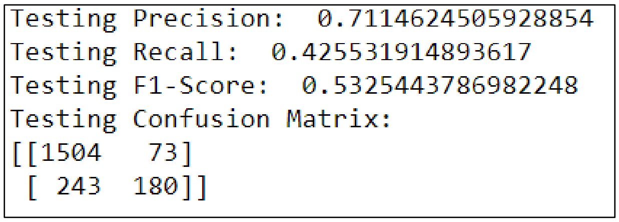
3.5 Evaluation Criteria

In our project, we decided to use Area Under the Curve score to evaluate the performance of our model, and furthermore visualize its output using the Receiver Operating Characteristic Curve (ROC). Our target label (Exited) is based on churn probability between 1 and 0. The Testing Precision, Testing Recall, Testing F1-Score, Testing Confusion Matrix was computed and shown below in Table 2. Our model records a testing precision of 0.71, which shows 71% accuracy whenever it predicts customer's churn. Also, its testing recall of 0.42, proves 42% reliability of the model in identifying customers with a higher tendency of churning.

Confusion matrix as could be seen inside the table 2, is usually categorized into four divisions: True Positive, True Negative, False Positive and False Negative.

Considering the testing confusion matrix below in table 2, 180 customers were correctly predicted by our model as True Positive (churners), 1504 customer correctly predicted as True Negative (non-churners), 243 predicted as false positive (churners) which they are not, and lastly the 73 predicted as False Negative (non-churners), of which is false. The evaluation result of our model shows that structured data significantly impacts churn prediction accuracy. However, it will be important to evaluate further the impact of structured data on churn prediction accuracy using other machine learning algorithms.

Table 2: Evaluation Scores



```
Testing Precision:  0.7114624505928854
Testing Recall:  0.425531914893617
Testing F1-Score:  0.5325443786982248
Testing Confusion Matrix:
[[1504   73]
 [ 243  180]]
```

4.0 Experiment

4.1 *Empirical Dataset*

DataSource:https://www.kaggle.com/shrutimechlearn/churn-modelling

The data used in our project is obtained from Kaggle community. It contains details of a bank customers and the target variable is a binary variable reflecting the fact whether the customer left the bank (closed his account) or he continues to be a customer. The services offered by this private financial institution include credit card issuance, individual account opening and holding. The customers of this financial institution are permitted to exit the bank for whatever reasons they deem fit as they are not legally required to inform the organization aforehand about their decision.

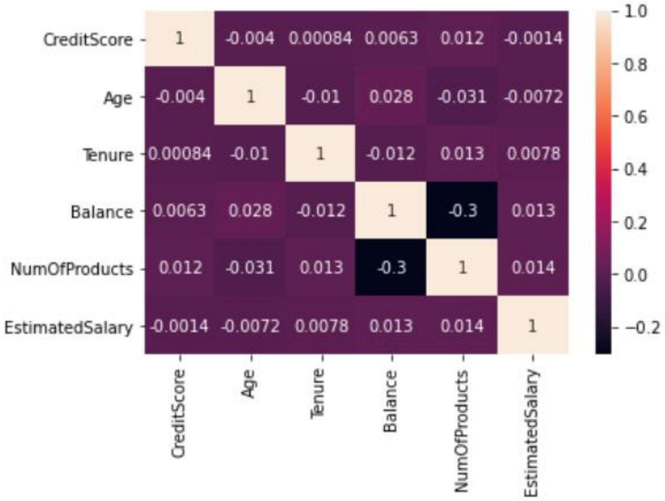
A customer is considered to have churned when the bank receives a request to close the customer's account after exploratory conversations to check the possibility of retaining the customer.

4.2 Data Processing

Our structured data contains some demographic information (age, gender, location), financial information (balance, credit score), account status information (IsActive Member) and others. We first checked if there are missing values in the data. Secondly, we went ahead to carry out feature selection of variables that would significantly impact the churn prediction model, by dropping the features that have no or little impact in the prediction model. We used Filter Feature Selection Method by employing correlation matrix (Heat Map) to determine which features significantly impacts one another in order to accurately carry out feature selection for the model.

Figure 2 visually shows the correlation between the various dependent features.

Figure 2: Heat Map Visualization of Features

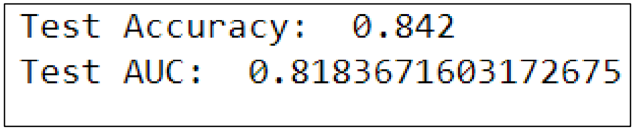


The correlations among variables were visualized using the heat map. It is evident in the figure 2 above, that there is strong negative relationship between Age and Number of Products. While on the positive side, the relationship between balance and age is higher, followed by Estimated Salary and Number of Products. Therefore, these features and others represented in the Heat Map were selected for building the model.

4.3 Results

In our study, our churn prediction mainly focuses on structured data. For building our churn prediction model, we used Random Forest, because of its ease of application and capacity to handle complex or large amount of data. We employed Repeated Holdout Method of validation to evaluate the performance of the model. Table 3 shows the AUC score and Figure 4, the ROC curve for the customers of the bank. The test AUC score of 0.81 proves that in general, our model possesses good quality for accurately predicting churn probability.

Table 3: AUC and Accuracy Score



```
Test Accuracy:  0.842
Test AUC:  0.8183671603172675
```

4.3.1. *Churn Prediction with ROC Curve*

The ROC curve shows that structured data quite holds some substantial information on churning probabilities. The curve shows that structured data can increase the churn prediction accuracy of the bank's customer by 5%, which affirms the theory that structured data possess vital business intelligence for organizations.

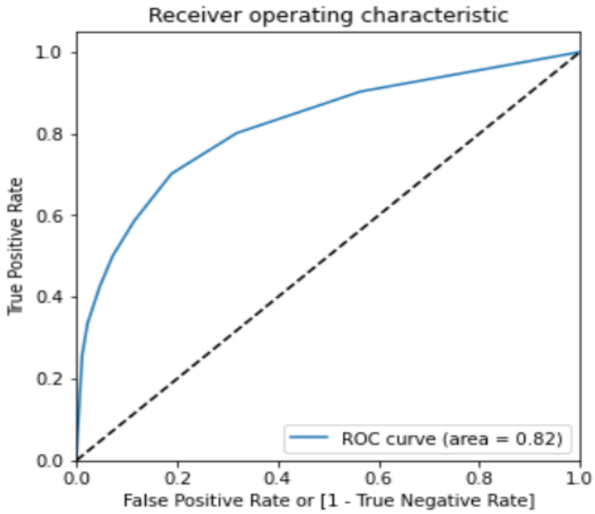


Figure 3: Receiver Operating Characteristic Curve (ROC)

5. 0 Conclusion and Limitations

Many researchers have established the impact of structured data on churn prediction in the telecommunication industry. Therefore, this paper evaluates the contribution of structured data to churn prediction accuracy in the financial sector.

From the results obtained, where the Random Forest Classifier clocked an Area Under the Curve score of 0.82 on the benchmark of 0.5, it is obvious that structured data significantly impacts churn probability accuracy.

Acknowledgment

Special thanks go to my project supervisor Professor Fabian Transchel for his immense contribution and ensuring that formalities of paper writing were observed. I also wish to thank the coordinator of this course Prof. Thomas Leich for the impactful knowledge of research methods and academic writing.

References

[1] Liu S., Brownlow J., Xu G., (2018). Client Churn Prediction with Call Log Analysis. Database Systems for Advanced Applications.

[2] Gordon S., Linoff and Michael JA Berry (2011). Data mining techniques: for marketing, sales, and customer relationship management. John Wiley & Sons.

[3] Ozden G. A. and Umut A. (2014). Dynamic churn prediction framework with more effective use of rare event data: The case of private banking. Expert Systems with Applications, 41(17):7889–7903.

[4} Mohammed A. H. F., Vadlamani R., and Raju S. B (2014). Churn prediction using comprehensible support vector machine: An analytical crm application. Applied Soft Computing, 19:31–40.

[5] Ngai, E., Xiu, L., & Chau, D. (2009). Application of data mining techniques in customer relationship management:

A literature review and classification. Expert Systems with Applications, pp. 2592-2602.

[6] Huang, Y., & Kechadi, T. (2013). An effective hybrid learning system for telecommunication churn prediction. Expert Systems with Applications, 5635-5647.

[7] Enterprise, Q. (2014). webopedia. Retrieved 05.15.2014, from structured data: http://www.webopedia.com/TERM/S/structured_data.html

[8] Berson, A., Smith, S. J., & Thearling, K. (2000). Building Data Mining Applications for CRM. McGraw-Hill Osborne.

[9] Hippner, H., & Rentzmann, R. (2006). Text Mining. Informatik Spektrum, 287-290.

[10] Hastie T., Tibshirani, R., & Friedmann, J. (2001). The Elements of Statistical Learning. New York: Springer New York Inc.

[11] Jaffery T., & Liu, S. X. (2009). Measuring Campaign Performance by Using Cumulative Gain and Lift Chart. SAS Global Forum 2009, S. Paper 196-2009.

[12] Wikipedia (30. April 2014). Von Lift (data mining):http://en.wikipedia.org/wiki/Lift_%28data_mining%29 abgerufen

[13] Uthayasankar S., Muhammad, M. K., Zahir I., Vishanth W.(2017). Critical analysis of Big Data challenges and analytical methods. Brunel University London, Brunel Business School, UB8 3PH, United Kingdom. Journal of Business Research Volume 70, Pages 263-286.

[14] Rita G., Anriba .S. (2012). Evaluations of Conceptual Models for Semi-structured Database System. International Journal of Computer Applications 50(18):5-12 DOI: 10.5120/7869-1145

[15] AnHai D., Jeffrey F.N. (2009). The Case for a Structured Approach to Managing Unstructured Data. Fourth Biennial Conference on Innovative Data Systems Research, Asilomar, CA, USA, January 4-7, 2009, Online Proceedings.

[16] Alice Z., (2015). Evaluating Machine Learning Models. O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472

[17] Farid S. (2018). A Big Data Analytics Model for Customer Churn Prediction in the Retiree Segment. Ryerson University. October 2018 International Journal of Information Mgt 48 DOI:10.1016/j.ijinfomgt.2018.10.005