Predicting Customers Intending To Cancel Credit Card Subscriptions Using Machine Learning Algorithms: A Case Study

Fehim Altınışık, Hasan Hüseyin Yılmaz

Directorate of R&D and Digital Transformation
Türkiye Vakıflar Bankası T.A.O. EBIS
Istanbul, Turkey
Fehim.altinisik@vakifbank.com.tr, hasanhuseyin.yilmaz@vakifbank.com.tr

Abstract

Following paper introduces analysis of machine learning algorithms implemented in order to predict customers of commercial bank who may be in risk of cancelling credit card subscriptions by following three months after a year or less activity. An analysis of various data preprocessing, sampling and structuring procedures using a feature set made up of 106 variables -describing customers' transaction activity, demographics, overall contentment and relative information to consumer experience- also shared. Study also includes performance comparison of Deep Neural Networks against other generic machine learning algorithms on two different cases. Deep Neural Networks were the point of interest of this study and it turns out, them to perform better than generic machine learning algorithms.

1. Introduction

In this paper, conclusion of a Credit Card Churn Prediction Study conducted in a national commercial bank offering credit card services in Turkey are discussed. Since rivalry is intensifying with increasing number of new financial institutions modelling both deposit and Islamic banking applications and increasing costs –both time and fund–, as [1] stated rather than find new customers subscribing for credit cards or convince customers already cancelled their subscription, it is becoming significantly more accurate to execute defensive policy aims retention of existing credit card subscriptions by providing qualified service. It is believed predicting customers' future intentions –rather continue to consume services of institution or not– accurately is the building block to achieve this policy's goals. This study also introduces the first design phase of complex real-time application that aims delivering relevant subset of promotions accurately to relevant subset of customers.

Term "churner" will be used to define a customer who executed cancellation process of credit card service via assist of a call center agent due to low quality service, "survivor" to customers spending via credit cards similar to their usual pattern and "convinced" to customers who -changed their minds by agents contacted via call center or promotions offered- terminated subscription cancellation process and continued using service.

The purpose of this study is making accurate predictions about future status of customer subscriptions and providing necessary information to relevant departments that takes action rather than giving insights or intuitions to business units about root cause of cancellation instantaneously. Thus usage of deep neural networks considered as "black boxes" lacks giving insights, becomes a solid choice contrary to generic regression or tree models. Prediction results of a deep neural network model that classifies between "convinced" and "churners" will also be introduced.

This paper organized as reader will find: definition of problem in Section 2, processes of data exploration, structuring, quality analysis and transformation in Section 3, sampling of training examples, architectural aspects of model in Section 4 and conclusion in Section 5.

2. Definition of Problem

Same terminology chosen as [2] stated for performance and observation periods.

Two significant issues negative correlated with each other detected after analysis of previously conducted credit card churn prediction studies, leaves business units unpleasant by the output. First, models are failing to alert mechanisms responsible that preventing customers' churn, on time, due to long observation periods almost a year-long or more and latter models are not inclusive enough to notice customers' seasonal shifting expense behaviors, due to observation periods' duration falls short.

Goal of techniques used in this study is minimizing tradeoff cost between long observation periods and seasonal effects. Since point of interest is a classification problem– rather than time series analysis or clustering– following a proper model input restructuring procedure will lead models to sustain recall and precision rates high enough, remain unaffected by customers' seasonal shifting expense behaviors, detailed in Section 3

.

3. Data

A feature set used in this study describes customers' credit card transactions, credit card status, demographic data and customers' banking experience assisted by a client such as a customer representative in a branch office or an agent in a call center. Other information summarizes customers' activity independent of credit card product, such as cash withdrawals, scheduled payment directives from a mobile device or subscription status of other services of the bank.

Raw data transformed and experimented with, consists of more than 4 million records. Each record documents an aggregated summary of customers' credit cards' transaction activity and status information.

3.1. Data Exploration

Each Record contains 106 non-linear fields including, continuous floating point, categorical, date-time and binary data. More than %45 of data was in valid/readable –suitable to use as input for model– values. Reasons of why data was skewed and procedures of precaution will be discussed in part 3.3 and 3.4.

Entire table made up of customers whom classified as *"survivors"* and *"convinced"* while data of customers classified as *"churners"* separated using *rule based expert systems*, before this study conducted.

3.2. Data Structuring and Sampling

A set of features unrelated to credit card activity also included in each customers aggregated record. Previous in-house data exploration studies conducted by business units proved correlation between credit card subscriptions' life time and these subset of features non-related to credit card product. Features X6, X7, X12, X13, X14, X15 defined in Table 1 are elements of this subset.

In order to structure and sample data set for model predicting set of customer intending to cancel credit card subscriptions between January and March, procedure below will be followed:

Assume months of repetitive years positioned on an axis and let months February, March and April defined in set $\propto$ as:

$$\propto = \{n, n+1, n+2\}$$

In this case for each element in $\propto$, a performance period p defined as:

$$p = \{n-13, n-1\}$$

A dictionary d initialized from set of $\propto$'s and p's as:

$$d = \{x = 2, y = 2 \mid \propto_0 \text{ to } \propto_x : p_0 \; to \; p_y$$

Dataset of two classes initialized as *"churners"* and *"survivors"* for each pair in d. Roughly 30.000 customers were quantized for each element in d. This led to have balanced datasets of 90.000 customers for each *"survivor"* and *"churner"* classes and a total of 180.000 samples for training sessions.

3.3. Data Quality Analysis and Transformation

An "outlier" is an extremely high or an extremely low data value when compared with the rest of the data values according to [3]**.** Outliers may affect shape and structure of input data after discretization, bucketing and one-hot encoding due to adding of many more columns representing data. Hence "Robust Regression" suggested by [4] applied to eliminate those having fields with outlier values. However it turned out, it may led to a loss function in unstable state.

As [5] pointed out even, if most of features' values may be in corrupted or unreadable format, it still may deliver valuable insight about classes. Hence a set of imputation methods have used to maintain integrity of data. However it is tricky to find suitable imputation method for the data or even methods since different features may require different imputation procedures. Observations showed certain combinations of optimizers and imputation methods may led to a loss function in unstable state in cases of neural networks and multilayer-perceptron independent of model architectures as demonstrated in Fig. 1 and Fig. 2:

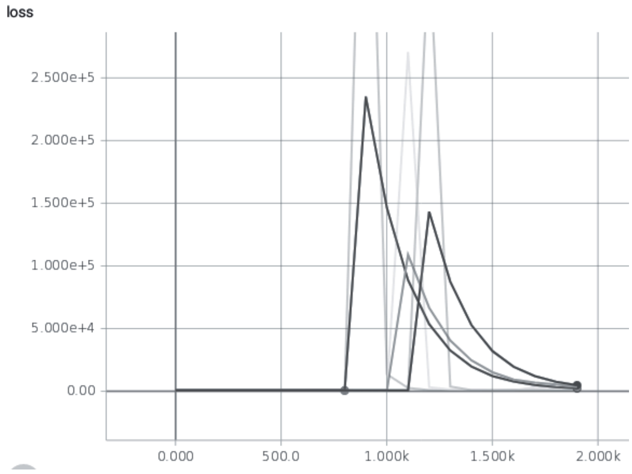

Fig. 1. Unstable loss function samples in use of median imputation

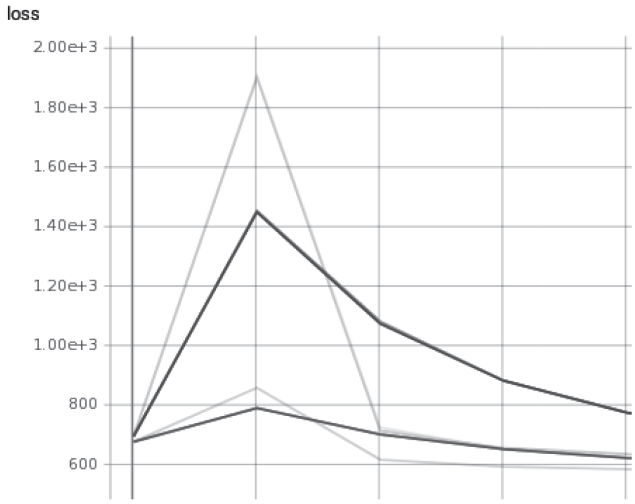

Fig. 2. Decay of stable loss function

Table 1. Description of Features

Feature	Description
X0	Age of customer
X1	Gender customer
X2	Annual salary of customer
X3	Marital status of customer
X4	Ratio amount of pending installment to expense limit.
X5	Amount of interest discount for credit card customer
X6	Number scheduled payment directives on digital channels
X7	Maximum Amount of cash withdrawal within last 3 months
X8	Age of credit card
X9	Flag variable indicates if customer postponed credit card debt within last 3 months or not
X10	Credit card debt balance
X11	Total amount of purchases with credit card within last 3 months
X12	Number of annual call center calls
X13	Average of call center's response time to customer
X14	Minimum duration of any call center call within last 3 months
X15	Number of branch office visits within last 3 months

It turns out filling missing values with some constant provides stable loss function although less accurate models while using median or mean provides higher accuracy although a less stable approach to global minima.

3.4. Data Description

Table 1 includes features from categories described in section 3. Data. Due to The Institutional Data Policy regulations, only a small set of features were described here.

4. Models

A detailed examination on corporate reports reveals, churn predictions made by "Logistic Regression", "Support Vector Machine" and "Multilayer perceptron" models met needs of generic cases as in telecommunication, utilities (electricity, gas) and digital media distribution industries etc. However that is not the case in the commercial banks due to reasons discussed in 2. Definition of Problem. In order to demonstrate the clear difference between models' success rates, other models' results also will be shared, trained, tested and predicting using same sample sets as described in 3. Data, 3.2. Data Structuring and Sampling, for all training, testing and prediction inputs.

A Rectified Linear Unit defined as:

$$f(x) = max(0, x) \quad (1)$$

A Leaky Rectified Linear Unit [6] defined *as:*

$$f(x) = 1(x < 0)(ax) + 1(x \geq 0)(x) \quad (2)$$

[9] Demonstrated how Leaky Rectified Linear Unit increases performance of deep neural network.

Before examining the results let us assume classifications in between *churners* and *survivors* as *case A, convinced* and *unconvinced* as *case B*.

4.1. Logistic Regression

As logistic regression finds a wide range of use cases for problems similar to this study, it is trained, tested and made predictions using different combination of hyper-parameters. Logistic regression performs at its peak with use of various combinations of L2 regularization and Rectified Linear Unit as activation function defined in Equation 1 in *case A*, while it fails to fit an admissible model that classifies customers who may be convinced to continue using their credit cards or not as stated in *case B*.

4.2. Support Vector Machine and Multilayer Perceptron

AI services providers suggest Support Vector Machine and Multilayer Perceptron performs good enough to give insights about customers' intensions in short-term. With use of generic hyper-parameter ranges for Multilayer Perceptron and Support Vector Machine with RBG Kernels, as standard approaches suggests [7][8] it turns out Support Vector Machine and Multilayer Perceptron performs better than Logistic Regression in *case A*, however it also fails to perform sufficient good enough to make a difference in *case B*.

4.3. Deep Neural Network

This study's objective defined as comparing performance measures of deep neural network models against generic machine learning models that dependent on various hyper-parameter combinations discussed below, highest scoring results obtained in *case A* and reasonable enough prediction results in *case B* as shown in Table 2. F1 score based comparison allowed to evaluate results more reasonable way due to high rate of false positive rates encountered in credit card churn prediction cases. F1 score defined as:

$$F1 = 2 * (P * R) / (P + R)$$

Where precision P is:

$$P = \frac{T_p}{T_p + F_p}$$

And recall R is:

$$R = \frac{T_p}{T_p + F_n}$$

Table 2. A detailed comparison of performance measures in case A and B

Feature	Cases	F1 Score
Logistic Regression	Case A	0.67
	Case B	0.49
Support Vector Machine	Case A	0.74
	Case B	0.56
Multilayer Perceptron	Case A	0.72
	Case B	0.54
Deep Neural Network	Case A	0.87
	Case B	0.65

Stable results acquired using a set of hyper-parameters including Leaky Rectified Linear Unit as activation function defined in Equation 2, Adam function as optimizer, L2 regularization and other hyper parameters. A deep neural network having between 12-24 hidden layers as architecture demonstrated in Table 3 where each section of layers consist of 2 to 4 layers in this case. Network provides stable results.

Table 3. An example of network architecture

Section	Number of Neurons	Dropout
Input Layer	106	-
Layer 1– 4	128	0.4
Layer 5 – 8	128	0.6
Layer 9-12	32	0.6
Layer13-14	32	0.4
Layer15-16	16	0.4
Layer16-18	16	0.2
Output Layer	2	-

5. Conclusion

This paper written with intent of giving an insight to scientists and engineers conducting research on credit card churn prediction, by providing comparison of generic machine learning models against deep learning models. Effect of various preprocessing techniques and introduced a new methodology to structure a balanced training set also shared.

Subroutines defined that sampling balanced datasets out of more than 4 million records and then iterating train, test and prediction processes on models initialized with ranging hyper-parameters. Customers classified as *survivors* and *churners* on top level and then classified as *convinced* and *unconvinced* in between *churners*. Through the process an iterative way introduced to sample data over consecutive months, to include customers' shifting expense behavior's effect in year.

It turns out logistic regression models often performs with moderate accuracy in *case A* (defined in 4. Models) in long-term - as defined duration up to 6-9 months- and fails to observe shifting expense behaviors over year. An alternative approach suggests building mensal or seasonal models trained using datasets of previous years, may provide limited information but intuitive enough. However models built this way, may severely affected by nation's shifting economic indicators. Poor accuracy levels observed in *case B*.

Support Vector Machines and Multilayer Perceptron performs slightly better than logistic regression models both in long and medium term –as defined up to 3-6 months-. Poor accuracy levels also observed in *case B* in preceding models.

Considering results shown in Section 4. Models, Deep Neural Networks seems to be performing better than any other models discussed above, to detect customers intending to cancel their credit card subscriptions, in long and medium term, and moderate performing results in short term as defined less than 3 months. Model also performs slightly better than preceding models in *case B*.

Another aspect authors believe that the reader should consider is, as defined in 3. Data, our feature set only consist of customers' banking activities' aggregated snapshot. As observed in 4. Models, even if this way of structuring dataset and training, saves significant time and processing power, it does not provide enough information about individual itself. An alternative approach to this problem may be, use of Long-Short Term Memory Networks, state of the art deep learning models predicts sequential elements in perfect order in cases, such as natural language processing and time series forecasting. Instead of using customers' aggregated dataset, it is suitable to use customers' credit card transactions data with Long-Short Term Memory networks might be the next successful research's interest in field of customer retention and loyalty.

6. Acknowledgements

Authors offer sincere gratitude Ahmet Genç supported process of writing this paper.

7. References

[1] F. F. &. S. W. E. Reichheld, "Zero defeofions: Quoliiy comes to services.," *Harvard business review,* pp. 68(5), 105-111., 1990.

[2] W. R. L. Z. Y. T. Y. S. Guangli Nie, "Credit card churn forecasting by logistic regression and decision tree," *Expert Systems With Applications,* 2011.

[3] A. Bluman, "Elementary Statistics, brief version,," *New York: McGraw-Hill.,* 2000.

[4] P. J. &. L. A. M. Rousseeuw, "Robust regression and outlier detection," *(Vol. 1). New York: Wiley.,* 1987.

[5] J. W. Graham, "Missing data analysis: Making it work in the real world.," *Annual review of psychology,* pp. 60, 549-576., (2009).

[6] K. Z. X. R. S. &. S. J. He, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification.," *Proceedings of the IEEE international conference on computer vision,* pp. (pp. 1026-1034)., 2015..

[7] B. S. Alex J. Smola, "“A Tutorial on Support Vector Regression”," *Statistics and Computing archive Volume 14 Issue 3,* pp. p. 199-222., August 2004.

[8] V. R. S. B. R. M.A.H. Farquad, "Churn prediction using comprehensible support vector machine:An analytical CRM application," *Applied Soft Computing,* 2014.

[9] B. W. N. C. T. &. L. M. Xu, "Empirical evaluation of rectified activations in convolutional network.," *arXiv preprint arXiv:1505.00853.,* 2015.