

Deep Neural Pipeline for Churn Prediction and Mitigation

Andrei Ionut DAMIAN
Chief Data Scientist
University POLITEHNICA of Bucharest
Bucharest, Romania
damian@cloudifier.net

Nicolae TAPUS
Computer Science & Engineering Department
University POLITEHNICA of Bucharest
Bucharest, Romania
ntapus@cs.pub.ro

Andrei SIMION-CONSTANTINESCU
Computer Science & Engineering Student
University POLITEHNICA of Bucharest
Bucharest, Romania
andrei.simion.c@gmail.com

Bogdan DUMITRESCU
General Manager
Hight-Tech System&Software
Bucharest, Romania
bogdan.dumitrescu@htss.ro

Abstract— Customer Churn is a metric used to quantify the number of customers who left the company. All retail and business to consumer companies carefully analyse customer behaviour to prevent them to cease his or her relationship with the company, in other words to make churn. With the latest development of Artificial Intelligence and the ones particularly related to Deep Learning we have a new set of powerful tools ready to employ within multiple horizontal and vertical domain – such as the horizontal of predictive business analytics domain. One of the main goals of predictive analytics is the research and development of the *almost-perfect* churn detection system. This paper objective is to propose a beyond state-of-the-art churn prediction and mitigation model together based on deep neural models and employing Big Data processing on massive parallel computing using GPU cells.

Keywords— machine learning, predictive analytics, massive parallel computing, GPU, deep learning, customer churn, big data

I. INTRODUCTION

Customer Churn is a crucial metric in evaluating customer satisfaction over periods of time. Predicting customers behaviour has always been an important research area for all retail and business to consumers companies with high impact in the structure of the marketing campaigns. Identifying early who is at risk to leave is the top priority, because it is more expensive and time-consuming to acquire a new customer than retaining old ones.

This paper presents some churn prediction models applied on Pharma Industry taking advantage of big data provided from the largest pharmaceutical group in our country. Churn prediction techniques attempt to understand customer behaviours and attributes which signal the risk and timing of customer churn.

Pharma industry has been used as a target industry for our applied industrial research and we do have actual successful research experiment in our work.

The first step in creating our model was the exploratory analysis, in which you took the raw data for all client, we analyse and process keeping only the attributes that were proved to be relevant to our model. After pre-processing the raw data, we split the customers into 4 clusters, from the “weakest” clients to the best ones using RFM analysis. Then,

we split every macro-segment into other micro clusters, in order to have clients grouped by their behaviour into every large segment.

Another important stage of our project was to determine the definition of churn in our concrete example: A client is flagged as making churn if in the previous year it has a minimum number of one transaction and in the next one has none. We will also analyse the impact of “relaxing” the churn definition in our models.

The next step is choosing the model architecture. In this paper, we will analyse all range of experimented models beside the actual final research results, starting from a simple linear model and all the way down to a complex Directed Acyclic Convolutional-Recurrent Graph (DACOREG). As we will within the paper, our experimentation has been conducted using a pharma industry database of over 150 million observations thus resulting in a real base for benchmarking in conjunction with the plethora of proposed concurrently models.

The proposed and concurrently used models are:

- Logistic Regression
- Simple Decision Trees
- Union of Trees (Forests)
- XGBoost

Our model performance will be quantified using standard metrics, from which we are especially interested in two performance evaluating measures: precision and recall. Recall rate tell us who may churn clients our model identifies from the real ones and precision give us the false positive rate of our model. We tried to balance between a high recall rate and a maximal precision one, trying to get a balance in which we accept only a tiny decrease in recall for a better increase in precision.

Our first models were capable to see year-to-year evolution. We want a model that can see quarterly changes to have a more accurate prediction. Customers behaviour regarding pharma industry is directly influenced by seasons, that is why we need a model able to see couple of months to couple of months. Also, regarding RFM analyses, from our raw data we can collect quarterly information, meaning that we have a total year RFM and a RiFiMi for every quarter. This make the introduction of seasoning in our model more naturally.