SAPIENZA
UNIVERSITÀ DI ROMA

Customer Churn Prediction - Energy Provider Case Study

Faculty of Information Engineering, Computer Science and Statistics

Master's degree in Data Science

Candidate

Vigèr Durand Azimedem Tsafack
ID number 1792126

Thesis Advisor	External Advisor
Prof. Anagnostopoulos Aristidis	Andrea Ianni, PhD

Academic Year 2019/2020

Thesis defended on October 30, 2020
in front of a Board of Examiners composed by:

Prof. Anagnostopoulos Aristidis (chairman)
Prof. Name Surname
Dr. Name Surname

Customer Churn Prediction - Energy Provider Case Study
Master's thesis. Sapienza – University of Rome

This thesis has been typeset by LaTeX and the Sapthesis class.

Version: September 10, 2020

Author's email: vigerdurand@yahoo.fr

Abstract

The cost of customer acquisition is far greater than cost of customer retention. This is a known fact across all the industry sectors, making retention a crucial business prototype. Customer churn analysis is one of the most important and common drivers laying behind customer retention. In fact, knowing in advance if a client is about to churn can be a quite valuable information; Particularly in the energy field which is going to be the focus of this work.

Energy supply is one of the most competitive industries where large amount of data is usually produced. Therefore, churn prediction in this type of industries is a key tool for customer retention. The present work aims to predict customer churn in energy industry through several data science techniques and methods. The experiment has been held in an Italian energy provider company which provided us with a huge amount of data. we start by explaining some relevant concepts from machine learning and continues to a literature review on the field of customer churn prediction. Then, an empirical study is performed by applying findings from the literature to the data provided by the aforementioned energy provider company. This study can be summarized in two main phases: Data and Modeling. The Data step includes collection, exploration and transformation of data.The Modeling phase refers to the creation and selection of the best machine learning model.

Regarding the results, GBT Classifier outperformed all the other five models that we tried (Logistic Regression, Random Forest, Decision Tree, SVM and MLP) with 73% of accuracy. The model evaluation was done by using the three following metrics: confusion matrix, accuracy and NDCG@k. The study also confirmed that machine learning is a viable tool for predicting customer churn in energy provider companies.

Contents

Glossary

GBT Gradient Boosted Tree. iii

ML Machine Learning. 3

MLP Multilayer Perceptron. iii

NDCG@k Normalized Discounted Cumulative Gain at k. iii

SVM Support Vector Machine. iii

List of Figures

List of Tables

Chapter 1

Introduction

After the industrial revolution and the advent of technical progress, almost all the industrial sectors have become highly competitive in developed countries. The energy sector is certainly not left out since today's costumer won't hesitate to change their energy provider if they do not find what they are looking for or if they get a better offer elsewhere. Knowing that the cost of costumer acquisition is far grater than that of costumer retention, companies try now to focus their attention mostly on retaining existing clients rather searching for new ones.

Communications technologies came with great advantages, making our every day live incredibly easy. however, they also represent a big disadvantage for companies since they have empowered the costumers who are no longer stuck with the decisions of a single company. Given that competitors are only one click away, companies must find interesting ways and techniques to examine their clients, understand their behavior and being able to predict if they are possibly going to leave in a close future. One of the tools that is commonly used in customer churn prediction is machine learning.

The quantity of companies data is continuously increasing, making the usage of machine learning for customer churn prediction a more and more popular in almost every industry. Most machine learning applications work as follow: the dataset is split into a test and training data. The training data is then used to train a model that learns from the data. The model is afterwards used to predict the results on yet unseen test data which are then compared to real values. Last but not least, metrics are used to calculate how good the model is doing using real and predicted values.[1]

The aim of this study was to develop a machine learning application namely an efficient and accurate churn prediction model for an energy provider company. In order to settle the context and make you familiar with the research's realm, we start the report by explaining some machine learning theoretical concepts and afterwards we describe the steps that we took in the development process of our churn prediction machine learning model.

1.1 Motivation and background

In terms of the economic model, the electricity industry has evolved in time from a vertically integrated state-owned monopoly company (not subjected to the normal rules of competition) to a liberalized market where generators and consumers have the opportunity to freely negotiate the purchase and sale of energy.[2] Nowadays, it is crucial for an energy provider to offer a quality service and to invent innovative strategies to increase customer satisfaction in order to retain the maximum number of clients and thus, remain competitive on the market. Machine Learning is a great tool that helps in achieving that goal. Indeed, machine learning based applications turn to be a fruitful avenue of research for data-intensive energy industry.

Some of the existing machine learning studies in energy industry include reliability and preventive maintenance, commonly known as failure detection.[3] Equipment failure in the energy industry, especially on coal-fired power plants, potentially cause injuries or even the death of workers. Artificial intelligence is helpful in preventing this problem. AI algorithms analyze equipment data and detect failures before they happen to save money, time, and people's lives. Regarding customer churn predictive analysis which is the goal of this study, after some research, we sadly noticed that it is not extensively studied in energy industry. However, given the actual competitive state of this market, it deserves more attention.

1.2 Theoretical framework and focus of the study

In this study, we mostly focus on exploiting the current state of the literature to empirically build several models for customer churn prediction in energy industry exploiting the provided data. Then suitable metrics are used to evaluate the build models in order to select the best performing one to be used in an Italian energy provider context.

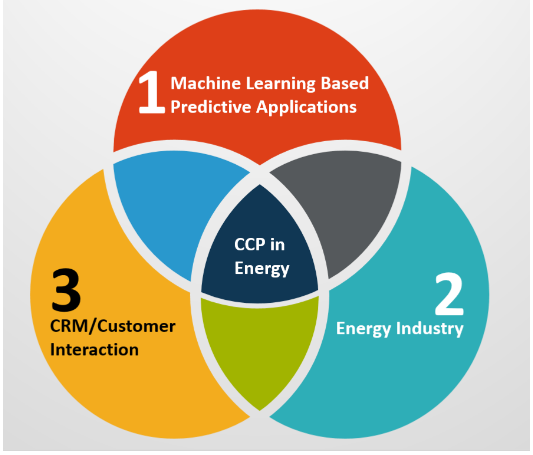
1 Machine Learning Based Predictive Applications
2 Energy Industry
3 CRM/Customer Interaction
CCP in Energy

Figure 1.1. Thesis research area

1.3 Research questions and objectives

The primary goal of this thesis is to accurately predict the future churn or status of costumers (stays/churns) for an Italian energy supply company for the next 2 months. Machine learning is the tool that will be used to achieve our goal. Thereby, a theoretical overview of related machine learning concepts is needed in order to create a good model. The obtained models are compared using some metrics an the best performing one is selected to be used in production. Based on these objectives he following research questions are formulated:

1. What is the current state of costumer churn prediction in the literature?
2. What is the current state of costumer churn prediction on the energy supply field?
3. Which models can be used to accurately predict costumer churn given customer feature data in energy supply filed?
4. How can they be evaluated?
5. How different models compare to one another?

1.4 Methodology

The study made in this thesis consisted in three steps. Foremost, some research are formulated based on the desired outcome and the literature. As a first part, we conducted general overview of the machine learning concepts that are necessary to fully understand this thesis. The second part consisted in searching the literature to find related work that were used as inspiration for the last part. Finally, starting from the literature review, we selected some ML models and some evaluations metrics to build a churn predictive application.

Data was provided by an Italian energy provider and consisted in real costumer data from May 2019 to June 2020. Unfortunately, for privacy reasons the data cannot be disclosed alongside this thesis.

1.5 Structure of the thesis

The second chapter presents a high level overview of critical methodologies and concepts useful to understand the study performed in this thesis. In the third chapter we perform a review of related existing studies in the field of customer churn prediction. Next, in the fourth chapter an empirical study is conducted to prepare the data and build the churn prediction machine learning model. In chapter five we analyze the model results. Finally in chapter six, we discuss the results, eventual limitations, make the conclusions along with the proposals for future studies on the topic.

Chapter 2

Machine learning: Some theoretical concepts

wwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwww

2.1 Data collection and preprocessing

The page is fixed at the dimensions of an A4 paper, therefore you have to print your thesis on A4 paper to obtain the best results. The font dimension

2.1.1 ETL

The page is fixed at the dimensions of an A4 paper, therefore you have to print your thesis on A4 paper to obtain the best results. The font dimension

2.1.2 Apache spark

The page is fixed at the dimensions of an A4 paper, therefore you have to print your thesis on A4 paper to obtain the best results. The font dimension

2.1.3 Dealing with missing data

The page is fixed at the dimensions of an A4 paper, therefore you have to print your thesis on A4 paper to obtain the best results. The font dimension

2.1.4 Dealing with imbalance data

The page is fixed at the dimensions of an A4 paper, therefore you have to print your thesis on A4 paper to obtain the best results. The font dimension

2.1.5 One hot encoding

The page is fixed at the dimensions of an A4 paper, therefore you have to print your thesis on A4 paper to obtain the best results. The font dimension

2.1.6 Ordinal encoding

The page is fixed at the dimensions of an A4 paper, therefore you have to print your thesis on A4 paper to obtain the best results. The font dimension

2.1.7 Word embedding

The page is fixed at the dimensions of an A4 paper, therefore you have to print your thesis on A4 paper to obtain the best results. The font dimension

2.1.8 Data Normalization

The page is fixed at the dimensions of an A4 paper, therefore you have to print your thesis on A4 paper to obtain the best results. The font dimension

2.1.9 Feature selection

The page is fixed at the dimensions of an A4 paper, therefore you have to print your thesis on A4 paper to obtain the best results. The font dimension

2.2 Machine learning models

The captions have a smaller font respect to the text and the label is in boldface. The appearance of the margin notes has been improved.

2.2.1 Logistic regression

The captions have a smaller font respect to the text and the label is in boldface. The appearance of the margin notes has been improved.

2.2.2 Random forest classifier

The captions have a smaller font respect to the text and the label is in boldface. The appearance of the margin notes has been improved.

2.2.3 Gradient-boosted tree classifier

The captions have a smaller font respect to the text and the label is in boldface. The appearance of the margin notes has been improved.

2.2.4 Decision tree classifier

The captions have a smaller font respect to the text and the label is in boldface. The appearance of the margin notes has been improved.

2.2.5 Support vector machine

The captions have a smaller font respect to the text and the label is in boldface. The appearance of the margin notes has been improved.

2.2.6 Multilayer perceptron

The captions have a smaller font respect to the text and the label is in boldface. The appearance of the margin notes has been improved.

2.3 Evaluation metrics

As regards the image formats, please use vector images as much as possible! Use jpg images only for photographs! pdfLaTeX supports the pdf, jpg and png formats.

2.3.1 Confusion matrix

The captions have a smaller font respect to the text and the label is in boldface. The appearance of the margin notes has been improved.

2.3.2 Accuracy

The captions have a smaller font respect to the text and the label is in boldface. The appearance of the margin notes has been improved.

2.3.3 NDCG@K

The captions have a smaller font respect to the text and the label is in boldface. The appearance of the margin notes has been improved.

Chapter 3

Related work

In this chapter I will discuss my stylistic choices of sapthesis. I will show the page layout geometry and I will describe the page style.

3.1 Customer churn prediction

The page is fixed at the dimensions of an A4 paper, therefore you have to print your thesis on A4 paper to obtain the best results. The font dimension

3.2 Customer churn prediction in energy

The page is fixed at the dimensions of an A4 paper, therefore you have to print your thesis on A4 paper to obtain the best results. The font dimension

3.3 Summary

The page is fixed at the dimensions of an A4 paper, therefore you have to print your thesis on A4 paper to obtain the best results. The font dimension

Chapter 4

Energy provider case study: churn prediction machine learning model

The page is fixed at the dimensions of an A4 paper, therefore you have to print your thesis on A4 paper to obtain the best results. The font dimension

4.1 Tools and libraries

The page is fixed at the dimensions of an A4 paper, therefore you have to print your thesis on A4 paper to obtain the best results. The font dimension

4.1.1 Python

The page is fixed at the dimensions of an A4 paper, therefore you have to print your thesis on A4 paper to obtain the best results. The font dimension

4.1.2 Apache Spark

The page is fixed at the dimensions of an A4 paper, therefore you have to print your thesis on A4 paper to obtain the best results. The font dimension

4.2 Data description and understanding

The page is fixed at the dimensions of an A4 paper, therefore you have to print your thesis on A4 paper to obtain the best results. The font dimension

4.3 Data preprocessing and feature selection

The page is fixed at the dimensions of an A4 paper, therefore you have to print your thesis on A4 paper to obtain the best results. The font dimension

4.3.1 Handling missing data

The page is fixed at the dimensions of an A4 paper, therefore you have to print your thesis on A4 paper to obtain the best results. The font dimension

4.3.2 Dealing with categorical features

The page is fixed at the dimensions of an A4 paper, therefore you have to print your thesis on A4 paper to obtain the best results. The font dimension

4.3.3 Imbalanced data

The page is fixed at the dimensions of an A4 paper, therefore you have to print your thesis on A4 paper to obtain the best results. The font dimension

4.3.4 Data Normalization

The page is fixed at the dimensions of an A4 paper, therefore you have to print your thesis on A4 paper to obtain the best results. The font dimension

4.3.5 Feature selection

The page is fixed at the dimensions of an A4 paper, therefore you have to print your thesis on A4 paper to obtain the best results. The font dimension

Chapter 5

Models and results

The page is fixed at the dimensions of an A4 paper, therefore you have to print your thesis on A4 paper to obtain the best results. The font dimension

5.1 Logistic regression

The page is fixed at the dimensions of an A4 paper, therefore you have to print your thesis on A4 paper to obtain the best results. The font dimension

5.2 Random forest classifier

The page is fixed at the dimensions of an A4 paper, therefore you have to print your thesis on A4 paper to obtain the best results. The font dimension

5.3 Gradient-boosted tree classifier

The page is fixed at the dimensions of an A4 paper, therefore you have to print your thesis on A4 paper to obtain the best results. The font dimension

5.4 Decision tree classifier

The page is fixed at the dimensions of an A4 paper, therefore you have to print your thesis on A4 paper to obtain the best results. The font dimension

5.5 Support vector machine

The page is fixed at the dimensions of an A4 paper, therefore you have to print your thesis on A4 paper to obtain the best results. The font dimension

5.6 Multilayer perceptron

The page is fixed at the dimensions of an A4 paper, therefore you have to print your thesis on A4 paper to obtain the best results. The font dimension

5.7 Summary and analysis of the results

The page is fixed at the dimensions of an A4 paper, therefore you have to print your thesis on A4 paper to obtain the best results. The font dimension

Chapter 6

Conclusions

6.1 Conclusion

The page is fixed at the dimensions of an A4 paper, therefore you have to print your thesis on A4 paper to obtain the best results. The font dimension

6.2 Suggestions for future research

The page is fixed at the dimensions of an A4 paper, therefore you have to print your thesis on A4 paper to obtain the best results. The font dimension

Bibliography

[1] Aurélien Géron (13 March 2017). Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media. ISBN 978-1-4919-6224-4.

[2] Eusébio E., de Sousa J., Ventim Neves M. (2015). Risk Analysis and Behavior of Electricity Portfolio Aggregator. In: Camarinha-Matos L., Baldissera T., Di Orio G., Marques F. (eds) Technological Innovation for Cloud-Based Engineering Systems. DoCEIS 2015. IFIP Advances in Information and Communication Technology, vol 450. Springer, Cham.

[3] Martínez García I.E., Sánchez A.S., Barbati S. (2016). Reliability and Preventive Maintenance. In: Ostachowicz W., McGugan M., Schröder-Hinrichs JU., Luczak M. (eds) MARE-WINT. Springer, Cham.