



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

Prof. Anagnostopoulos Aristidis

External Advisor

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

© 2020 Vigèr Durand Azimedem Tsafack. All rights reserved

This thesis has been typeset by L^AT_EX and the Sapthesis class.

Version: October 18, 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

| | | |
|----------|--|-----------|
| 1 | Introduction | 1 |
| 1.1 | Motivation and background | 1 |
| 1.2 | Theoretical framework and focus of the study | 2 |
| 1.3 | Research questions and objectives | 2 |
| 1.4 | Methodology | 3 |
| 1.5 | Structure of the thesis | 3 |
| 2 | Machine learning: Some theoretical concepts | 5 |
| 2.1 | Data collection and preprocessing | 5 |
| 2.1.1 | ETL | 5 |
| 2.1.2 | Apache spark | 5 |
| 2.1.3 | Dealing with missing data | 6 |
| 2.1.4 | Dealing with imbalance data | 6 |
| 2.1.5 | One-hot encoding | 7 |
| 2.1.6 | Ordinal encoding | 7 |
| 2.1.7 | Categorical embeddings | 7 |
| 2.1.8 | Feature selection | 8 |
| 2.2 | Machine learning models | 8 |
| 2.2.1 | Logistic regression | 8 |
| 2.2.2 | Decision tree classifier | 9 |
| 2.2.3 | Random forest classifier | 10 |
| 2.2.4 | Boosting: Gradient-boosted tree, XGBoost | 11 |
| 2.3 | Evaluation metrics | 12 |
| 2.3.1 | Confusion matrix | 12 |
| 2.3.2 | Precision, Recall, F-Measure | 12 |
| 2.3.3 | Area under the curve (ROC AUC, PR AUC) | 13 |
| 3 | Related work | 15 |
| 3.1 | Review: Techniques and Data | 15 |
| 3.2 | Summary | 16 |
| 4 | Energy provider case study: churn prediction machine learning model | 17 |
| 4.1 | Tools and libraries | 17 |
| 4.1.1 | Python | 17 |
| 4.1.2 | Apache Spark | 17 |

| | | |
|----------|--|-----------|
| 4.2 | Data description and understanding | 17 |
| 4.3 | Data preprocessing and feature selection | 17 |
| 4.3.1 | Handling missing data | 18 |
| 4.3.2 | Dealing with categorical features | 18 |
| 4.3.3 | Imbalanced data | 18 |
| 4.3.4 | Data Normalization | 18 |
| 4.3.5 | Feature selection | 18 |
| 5 | Models and results | 19 |
| 5.1 | Logistic regression | 19 |
| 5.2 | Random forest classifier | 19 |
| 5.3 | Gradient-boosted tree classifier | 19 |
| 5.4 | Decision tree classifier | 19 |
| 5.5 | Support vector machine | 19 |
| 5.6 | Multilayer perceptron | 19 |
| 5.7 | Summary and analysis of the results | 20 |
| 6 | Conclusions | 21 |
| 6.1 | Conclusion | 21 |
| 6.2 | Suggestions for future research | 21 |
| | Bibliography | 23 |

Glossary

ETL Extraction-Transformation-Loading. [5](#)

GBT Gradient Boosted Tree. [iii](#)

MapReduce MapReduce is a programming model and an associated implementation for processing and generating big data sets with a parallel, distributed algorithm on a cluster. A MapReduce program is composed of a map procedure, which performs filtering and sorting (such as sorting students by first name into queues, one queue for each name), and a reduce method, which performs a summary operation (such as counting the number of students in each queue, yielding name frequencies).. [6](#)

ML Machine Learning. [3](#), [5](#)

MLP Multilayer Perceptron. [iii](#)

NDCG@k Normalized Discounted Cumulative Gain at k. [iii](#)

SVM Support Vector Machine. [iii](#)

List of Figures

| | | |
|-----|--|----|
| 1.1 | Thesis research area | 2 |
| 2.1 | One-hot encoding method | 7 |
| 2.2 | A decision tree example based on binary feature Y [18] | 10 |
| 2.3 | A confusion matrix example | 13 |
| 2.4 | ROC and PR curves interpretation | 14 |
| 3.1 | Related works summary | 16 |

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 customer 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 customer acquisition is far greater than that of customer retention, companies try now to focus their attention mostly on retaining existing clients rather than 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 customers 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 more and more popular in almost every industry. Most machine learning applications work as follows: 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.

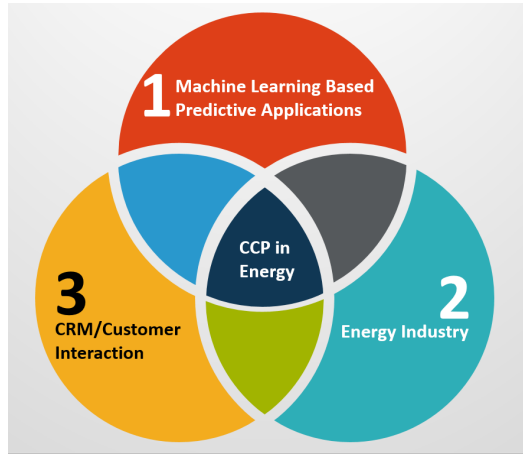


Figure 1.1. Thesis research area

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 and the best performing one is selected to be used in production. Based on these objectives the following research questions are formulated:

1. What is the current state of customer churn prediction in the literature?
2. What is the current state of customer churn prediction on the energy supply field?
3. Which models can be used to accurately predict customer churn given customer feature data in energy supply field?
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 customer data from January 2019 to March 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

[ML](#) is a branch of artificial intelligence that systematically applies algorithms to synthesize the underlying relationships among data and information.[\[4\]](#) Lately, it has become a common solution to several problems that companies face daily. This chapter presents a high level explanation of some machine learning concepts which the comprehension is necessary to fully understand the experiment performed within the framework of this thesis.

2.1 Data collection and preprocessing

Without data, no machine learning project could be made possible. It is therefore crucial to find interesting ways to collect and process the data to make it ready for any machine learning algorithm. Several methods are often used for this purpose. The most relevant for the scope of this thesis will be presented in this section.

2.1.1 ETL

The acronym [ETL](#) stands for Extract, Transform, and Load. ETL tools are pieces of software responsible for the extraction of data from several sources, their cleansing, customization and insertion into a data warehouse.[\[5\]](#) These tools are very often used as helper in the process of data collection and preparation for a machine learning application.

More explicitly, an ETL tool is three database functions combined into one entity to pull data out of one or multiple database(s) and then place it into another database. The process can be summarized as follows: data is taken (extracted) from a source system, converted(transformed) into a desired format, and finally stored into a data warehouse or other systems.

2.1.2 Apache spark

The quantity of digital data generated in today's companies is continuously increasing thus making their analysis more difficult. To overcome this problem several solutions have been implemented through the years. Google's MapReduce revolutionized

large-scale analysis, enabling the processing of massive datasets on commodity hardware and cloud resources, providing transparent scalability and fault tolerance at the software level.[6] Open source implementations of MapReduce include Apache Hadoop.

Industries initially adopted Hadoop because it is a framework based on a simple programming model (MapReduce), it provides a computing solution that is scalable, flexible, fault-tolerant and also cost effective. Apache spark comes into play when the concern is to maintain speed while processing large datasets. It was introduced by Apache Software Foundation aiming to speed up the Hadoop computational process.

Apache Spark is designed to accelerate analytics on Hadoop while providing a complete suite of complementary tools that include a fully-featured machine learning library (MLlib), a graph processing engine (GraphX) and stream processing. Spark is natively designed to run in-memory, enabling it to support iterative analysis and more rapid, less expensive data crunching. Spark runs programs in memory up to 100 times faster than Hadoop MapReduce and up to 10 times faster on disk. Spark's speed and efficiency are some of the the key reasons behind it's popularity.

2.1.3 Dealing with missing data

Data is the hub of every machine learning project. As a matter of fact without a good and clean dataset, useful results cannot be obtained from the data science process. Missing data is a major problem that statisticians and data scientists face quite often.

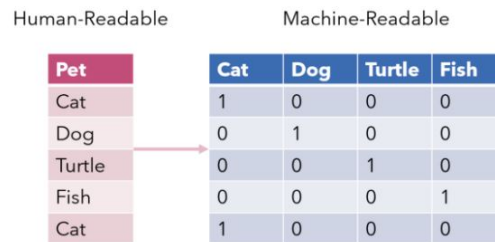
Dealing with missing data is absolutely necessary because most statistical models operate only on complete observations of predictor and target variables. Different methods can be adopted to deal with missing data. Incomplete observations can be deleted or missing values can be replaced by an estimated value based on other information available. This process is called missing data imputation.[7] To handle missing data, three main steps are generally followed: (i) finding the reasons for missing data; (ii) analyzing the proportions of missing data by feature and finally; (iii) choosing the best imputation method. Analyzing the cause of missing data is very important since it plays a major role in the choice of the imputation technique to be used.

2.1.4 Dealing with imbalance data

Imbalanced training dataset means that one class is represented by a large number of observations (majority class) while the other is underrepresented, namely represented by only a few number of observations (minority class). This is a typical issue that is observed with every churn prediction problem. It may produce an importance deterioration of the classification accuracy, leading for example to all the observations being predicted as part of the majority class. Several techniques are usually adopted to handle the imbalance situation. The two most common are the under sampling and over sampling approaches.[8]

2.1.5 One-hot encoding

Most machine learning and deep learning models require all input and output variables to be numeric. This means that if the dataset contains categorical data, that data must be converted into numerical form before fitting it to a machine learning model. One-hot encoding method is one of the most commonly used strategies to convert data from categorical to numerical form. This technique requires very little work. With this method, categorical variables are converted into several binary columns.



| Human-Readable | Machine-Readable | | | |
|----------------|------------------|-----|--------|------|
| Pet | Cat | Dog | Turtle | Fish |
| Cat | 1 | 0 | 0 | 0 |
| Dog | 0 | 1 | 0 | 0 |
| Turtle | 0 | 0 | 1 | 0 |
| Fish | 0 | 0 | 0 | 1 |
| Cat | 1 | 0 | 0 | 0 |

Figure 2.1. One-hot encoding method

One clear disadvantage of this method is the fact that the distance between one-hot encoded vectors does not carry much information. Another major disadvantage is that it aggressively consumes storage resources.[\[9\]](#)

2.1.6 Ordinal encoding

Another import and widely used method to convert features from categorical form to numerical is ordinal encoding. this method is generally used when the variable to be converted is ordinal, namely when the variable comprises a finite set of discrete values with a ranked ordering between values. For instance, if the possible values of the variable are *first*, *second* and *third*, integers *1*, *2* and *3* can be used to encode the variable.

An advantage of this method is that since the integer values have a natural order relationship between each other, machine learning algorithms may be able to understand and use this relationship. A clear disadvantage is that it cannot be used to encode every type of categorical feature since it is adapted only for features presenting a natural ordinal relationship among the possible values.

2.1.7 Categorical embeddings

The standard in natural language processing (NLP) is to encode the input such as words into continuous vector representation which are called embeddings.[\[10\]](#) Embeddings are a solution to dealing with categorical variables while avoiding a lot of the pitfalls of one-hot encoding. Recently, there has been some interest in learning embeddings [\[11\]](#), [\[12\]](#), [\[13\]](#) for general categorical variables instead of using the standard encoding techniques. Formally speaking, an embedding is a mapping of a categorical variable into an n-dimensional vector.

This provides us with 2 advantages. First, we limit the number of columns we need per category. Second, embeddings by nature intrinsically group similar categories together.

2.1.8 Feature selection

Feature selection is the process of manually or automatically selecting the features contributing the most to the prediction of the target variable. As the number of variables and data has increased due to more advanced data gathering, it is essential to include only the most critical and useful variables for the model one is building. Feature selection have three main objectives: (i) Allowing the model to achieve better predictive performance; (ii) getting faster and more efficient predictions; (iii) Allowing to get a more understandable and interpretable model. Adding unnecessary variables to the model also adds unnecessary complexity and can lead to overfitting, while missing essential variables lead to the reduction of the predictive performance. Feature selection methods can be divided into three main category: *Filter* methods, *Wrapper* methods and *embedded* methods.[14]

In *Filter* methods, a relevance criteria is initially decided. Subsequently, the features are ranked based on the previously decided criteria. A threshold is set to select the highest-ranking features. Some commonly used metrics are the following: (i) variance which is used to remove constant features; (ii) chi-square which is a statistical test that is used to verify the dependency of two variables; (iii) correlation coefficients that can be used to remove duplicated variables.

Wrapper method feature selection process is based on a specific machine learning algorithm that we wish to fit on a given dataset. A greedy search approach is used by evaluating all the possible combinations of features against the evaluation criterion. even if this method is effective, it is also computationally expensive.

In *embedded* methods, the feature selection process is completed within the machine learning algorithm itself. In other words, the feature selection process is performed during the model training.

Another method which is quite common is to use principal component analysis (PCA), which is a linear extraction method that transforms the data into a low-dimensional subspace. The idea is to retain most of the information but reduce the features into a smaller vector.[15]

2.2 Machine learning models

Several machine learning algorithms have been used in this thesis. In this section we provide a theoretical explanation for all of them.

2.2.1 Logistic regression

Logistic regression is one of the most common machine learning algorithms. It is generally used as a baseline model when the problem to be solved is a classification problem. This model is particularly appropriate when the dependent variable is dichotomous (binary).[16] The name logistic regression is derived from the function used at the core of the method to transform the linear predictions, the logistic

function. The logistic function which is also known as sigmoid function is a very popular function that machine learning borrowed from the statistical field. It is an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits. The logistic function can be mathematically expressed as shown in equation 2.1.

$$\text{logit}(x) = \frac{1}{1 + e^{-x}} = \frac{e^x}{1 + e^x} \quad (2.1)$$

Every linear methods try to fit a curve between data points. While *linear regression* uses the least-squares method to measure the error namely the distance between the data points and the line, logistic regression in contrast uses maximum likelihood in its fitting process. Maximum likelihood tries to maximize the probability of obtaining the observed dataset using the likelihood function. The chosen maximum likelihood estimators are those maximizing the likelihood function and agreeing the most with the data. Equation 2.2 shows a mathematical representation of logistic regression.

$$\text{logit}(\pi(x)) = \beta_0 + \beta_1 x_1 + \dots + \beta_n x_n \iff \pi(x) = \frac{e^{\beta_0 + \beta_1 x_1 + \dots + \beta_n x_n}}{1 + e^{\beta_0 + \beta_1 x_1 + \dots + \beta_n x_n}} \quad (2.2)$$

where $\pi(x)$ is the probability of the predicted event, β_i is the regression coefficient for each explanatory variable x_i . The probability of belonging to the predicted class is obtained by solving $\pi(x)$ from the equation.[17]

2.2.2 Decision tree classifier

Another commonly used model for solving classification problems in machine learning is decision tree. A decision tree is simply a supervised machine learning algorithm where the data is continuously split according to a certain parameter. It is a very simple model. Given several features, the decision begins with one of these features; if that is not enough, we use another one, and so on.

It is widely known and used in many companies to aid the decision making process and risk analysis. It was widely used from the 1960s to the 1980s for building expert systems. The rules were entered manually, that is why this model lost its popularity after the 1980s. The advent of mathematical methods to build decision trees brought this model back to the battle of automatic matching algorithms. The following is the general algorithm for creating a decision tree:

1. Determine the best feature from the training data set.
2. Divide the training data into subsets containing the possible values of the best feature.
3. Recursively generate new decision trees using the created data subsets.
4. We stop when we can no longer classify the data.

There are different types of decision trees, among which CART, C4.5, CHAID, QUEST, and more...[18] CART which stands for classification and regression trees is the most commonly used by the models that were considered in this study. Figure 2.2, displays a simple binary decision tree.

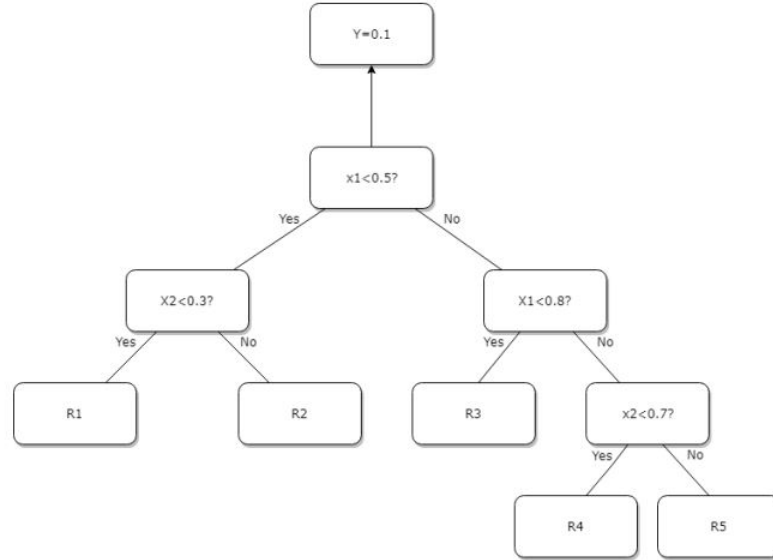


Figure 2.2. A decision tree example based on binary feature Y [18]

among the advantages of decision trees, we can enumerate the following: (i) they are easy to understand and interpret since the tree can be visualized and the obtained results explained easily; (ii) they can work on data with little preparation, for instance they don't need data standardization; (iii) they accept both numeric and nominal data while other learning algorithms specialize in a single type of data.

On the other hand, decision trees present several downsides: (i) they can be complex, they don't generalize well (overfitting), This can be adjusted by tuning the maximum depth of the tree or the minimum number of samples in the leaves; (ii) they may be unstable due to variations in the data; (iii) some concepts can be difficult to learn using decision trees since they are not easy to express, *XOR* is a good example; (iv) they can be biased towards the ruling class, thus the data has to be balanced before training the system. (v) hitting the optimal decision tree is not guaranteed.

2.2.3 Random forest classifier

This powerful machine learning algorithm makes it possible to make predictions based on the aggregation of several decision trees. The method uses binary decision trees, in particular CART trees proposed by Breiman et al. (1984). The general idea behind the method is the following: instead of trying to get an optimized method all at once, we generate several predictors before putting together their different predictions.

Random forests are an improvement of bagging for CART decision trees with the aim of making the trees used more independent (less correlated). Some characteristics

of this method are that (i) They give good results especially with large data sets; (ii) They are very easy to implement; (iii) They have few parameters.[19] The general steps for implementing this model are the following:

1. We draw at random from the training set B samples with replacement $z_i, i = 1, \dots, B$ each sample having n data points.
2. For each sample i we build a CART tree $G_i(x)$ according to a slightly modified algorithm: each time a node has to be cut (“split” step) we randomly select a part of the attributes (q among the p attributes) and we choose the best division in this subset.
3. For classification problems which are of interest in this study, aggregation by vote is used: $G(x) = \text{Majorityvote}(G_1(x), \dots, G_B(x))$.

2.2.4 Boosting: Gradient-boosted tree, XGBoost

Boosting is another type of ensemble method just like random forests. The principle of boosting is to combine the outputs of several weak classifiers to obtain a stronger result (strong classifier). The weak classifier must have a basic behavior being a little better than the random one: error rate less than 0.5 for a binary classification. Each weak classifier is weighted by the quality of its classification: the better it classifies, the more important it will be. Misclassified examples will have greater weights (they are said to be boosted) towards the weak learner in the next round so that it addresses the gap.[20]

Gradient boosting is a particular boosting technique which is mainly used with decision trees. The main idea here is again to aggregate several classifiers together but by creating them iteratively. These “mini-classifiers” are generally simple and parameterized functions, most often decision trees in which each parameter is the criterion for splitting the branches. The final super-classifier is a weighting of these mini-classifiers. One approach to build this super-classifier is to:

1. Randomly set the weighting (weights w_i of the mini-classifiers to form the initial super-classifier.
2. Calculate the error induced by this super-classifier, and find the mini-classifier that comes closest to this error.
3. Subtract the mini-classifier from the super-classifier while optimizing its weight with respect to a loss function.
4. Repeat the process iteratively.

Some of the most common boosting ensemble model implementations are *Gradient-boosted tree* and *XGBoost*. These two algorithms are indeed going to be used in this study.

2.3 Evaluation metrics

The choice of the best performing model is a critical task in machine learning since choosing the wrong model makes all the hard work performed useless. In this section, we give a theoretical explanation of the evaluation metrics used in this thesis.

2.3.1 Confusion matrix

A Confusion Matrix or contingency table is a tool for measuring the performance of a machine learning model by checking in particular how often its predictions are accurate compared to reality in classification problems. More specifically, it is a summary of the results of predictions about a classification problem. Correct and incorrect predictions are highlighted and broken down by class. The results are thus compared with the actual values.

This matrix helps to understand how the classification model is confused when making predictions. This not only allows you to know what mistakes were made, but above all the type of mistakes made. Users can analyze them to determine which results indicate how errors are made. To illustrate the idea, we can think of the problem as a binary classification problem where the instance either is classified correctly or is not. In this case, there are four possibilities:

- True Positives (TP): cases where the prediction is positive, and where the real value is indeed positive.
- True Negatives (TN): cases where the prediction is negative, and where the real value is indeed negative.
- False Positives (FP): cases where the prediction is positive, but the true value is negative.
- False Negatives (FN): cases where the prediction is negative, but the actual value is positive.

Several other metrics (*accuracy*, *precision*, *recall*, *f-Score*, ...) can be calculated directly from the confusion matrix. and some of them will be covered subsequently. Figure 2.3 shows a simple illustration of a confusion matrix configuration.

2.3.2 Precision, Recall, F-Measure

Precision, recall and F-Measure are by far the most commonly used evaluation metrics when we want to deal with classification problems. Indeed, in many machine learning problems the performance of algorithms is evaluated using precision and recall measurements.^[22] However, these two measures can have a very different importance depending on the context. That's why data scientist more often prefer an evaluation metric that will combine these two measures equally important one as the other. The most common among those combination metric is called F-Measure (a.k.a f1 score).^[23]

Precision can be defined in machine learning as the conditional probability that a randomly chosen example is correctly classified by the system. This is the ratio

| | | Actual Values | |
|------------------|--------------|---------------|--------------|
| | | Positive (1) | Negative (0) |
| Predicted Values | Positive (1) | TP | FP |
| | Negative (0) | FN | TN |

Figure 2.3. A confusion matrix example

between the number of true positive predictions (TP) and the number of positive predictions (TP + FP). In our context, it measures the quality of the predictions, that is, the degree to which customers marked as being churning really are. The **recall** measures the "width of learning" and represents the ratio of the number of true positive predictions to the total number of real positive examples. It measures the percentage of churning clients that was identified by the model, namely the model effectiveness. The break-even point is reached when precision and recall are equal.[24] The expression to calculate both precision and recall can be seen in Equation 2.3.

$$Recall = \frac{TP}{TP + FN} \quad Precision = \frac{TP}{TP + FP} \quad (2.3)$$

The **F-measure**, also called Dice index, can be defined as the harmonic mean of precision and recall. This measurement can be seen as a trade off between precision and recall. Equation 2.4 shows the mathematical formula for this metric.

$$F - measure = \frac{2 * precision * recall}{precision + recall} \quad (2.4)$$

A value close to 1 indicates that the classification is of very good quality.

2.3.3 Area under the curve (ROC AUC, PR AUC)

Other two interesting evaluation metrics in machine learning are the areas under the ROC and PR curves.[25] these two measurements are popular because they are suitable when the data set used to build the model is unbalanced. Indeed, they are going to be perfect for our case, especially area under the PR curve since we wish to have a good recall while keeping the precision reasonably high.

A receiver operating characteristic (ROC) curve is a graph representing the performance of a classification model for all classification thresholds. This curve plots the rate of true positives as a function of the rate of false positives. The true positive rate (TPR) is the equivalent of the recall. It is therefore defined as already shown in equation 2.3, whereas the false positive rate (FPR) can be expressed as shown in equation 2.5. Similarly, A Precision-Recall (PR) curve is

another graph that can be used to evaluate a classification model over multiple classification thresholds. This other curve plots the precision as a function of the recall. In summary, These two curves are quite interesting because they can be used to identify the best threshold (namely the one yielding the best performance) for a particular classifier.

The closer to 1 are the areas under the ROC and the PR curves, the better the classifier is likely to perform. Figures 2.4a and 2.4b help understanding how the areas under the curves should be interpreted.

$$FPR = \frac{FP}{FP + TN} \quad (2.5)$$

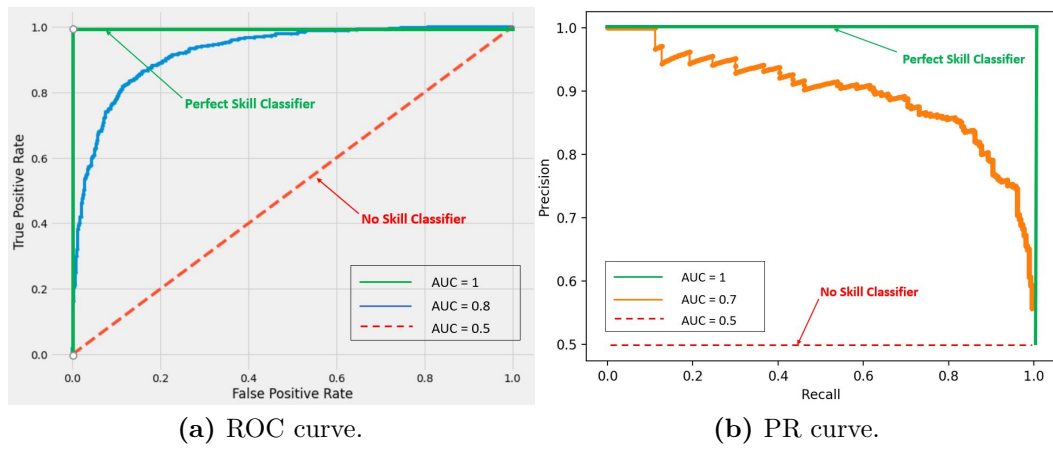


Figure 2.4. ROC and PR curves interpretation

Chapter 3

Related work

The word "churn" - which originated from the contraction between the words *change* and *turn* - describes the phenomenon of losing a customer. It is measured by the rate of churn which is an important indicator for organizations. This churn rate represents the percentage of lost customers over a given period compared to the total number of clients at the beginning of the same period. Reduce this rate is thus one of the most important concerns of every modern company. This study is clearly not the first of his kind. Multiple studies in the field of churn prediction/reduction have been performed, this chapter brings to light some of them reviewing both the techniques and the data sources used.

3.1 Review: Techniques and Data

The goal of the studies on the churn phenomenon is to detect individuals who intend to leave the organization in order to improve decision making and initiate retention actions. It is often analyzed using predictive techniques similar to datamining. Some methods are commonly used for churn prediction, such as logistic regression models ([26], [27], [29], [30]), decision tree models ([27], [29], [29], [30]), support vector machines ([28]), random forests ([29]), neural networks ([30]).

in [28], a data mining method is used to predict the customer churn in mobile telecommunication industry using call detail records dataset that consists of 3333 customers with 21 attributes each and a churn dependent variable with two classes Yes/No. Few attributes include the information about their corresponding inbound/outbound SMS count and voice mail. In this study, a principal component analysis was used in order to reduce the data dimensionality and deal with multicollinearity. the modeling phase was done using three machine learning model, namely support vector machines, neural networks and Bayesian networks. In the evaluation phase of this study, confusion matrix and ROC curve were used as evaluation metrics.

Another interesting churn prediction study is the performed in [27]. in this study, a software called *WEKA* is used to develop a churn prediction model. Each customer was classified as a potential churner or non-churner. The framework discussed was based on Knowledge Discovery Data process. Three different datasets, small, medium and large with varying attributes were considered. The performance of decision trees and logistic regression models are compared by calculating the accuracy and error

rate.

Paper [29] presents a study on subscriber churn analysis and prediction for mobile and wireless service providers. A real and complied dataset by Orange Telecom, 2009 was used. Main emphasis was laid on ensemble methods that encompass single methods to improve the solution to churn prediction problem. These results were compared with that of classic methods, namely logistic regression, decision trees and random forests; the evaluation metric used was ROC score.

Paper [26] proposes a framework of the whole process of churn prediction of credit card holders. The machine learning model used in this study is logistic regression applied on the data of more than 5000 credit card holders provided by a major China commercial bank. Accuracy and ROC curve are used to evaluate the obtained model.

3.2 Summary

Several studies on churn prediction attempt to find the most efficient data mining technique in terms of minimizing the error rate and prediction accuracy. From the researches that have been performed, it turns out that logistic regression random forest and decision trees are the most widely used techniques for churn prediction. In addition, the general remark resulting from this non-exhaustive summary is that there is no standard technique for solving the problem of churn prediction because the quality of the results obtained is closely related to the nature of the data used, their volume and quality, the number and relevance of indicators taken into consideration, the size of the learning sample, and finally the definition of the target variable. Without forgetting the confidential aspect of the data which does not allow the dissemination of the strategic results of the organization, especially in a competitive sector such as energy supply. Figure 3.1 synthesizes the articles that we presented in the previous section.

| Studies | Models | Metrics | Data |
|---------------------|--|-------------------------------|---|
| Toderean2016 | SVM Neural Networks Bayesian networks | Confusion matrix ROC curve | Call detail records |
| Dahiya2015 | Decision trees Logistic regression | Accuracy Error rate | Three different datasets, small, medium and large |
| Yabas2013 | Ensemble methods Logistic regression Decision trees Random forest | ROC curve | Real dataset by Orange Telecom |
| Nie2009 | Logistic regression | Accuracy ROC curve | Real dataset provided by a china commercial bank |

Figure 3.1. Related works summary

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.
- [4] Awad M., Khanna R. (2015). [Machine Learning](#). In: Efficient Learning Machines. Apress, Berkeley, CA.
- [5] Vassiliadis P., Simitsis A., Skiadopoulos S. (2002). [On the Logical Modeling of ETL Processes](#). In: Pidduck A.B., Ozsu M.T., Mylopoulos J., Woo C.C. (eds) Advanced Information Systems Engineering. CAiSE 2002. Lecture Notes in Computer Science, vol 2348. Springer, Berlin, Heidelberg.
- [6] Capuccini, M., Ahmed, L., Schaal, W. et al. [Large-scale virtual screening on public cloud resources with Apache Spark](#). J Cheminform 9, 15 (2017).
- [7] Salgado C.M., Azevedo C., Proença H., Vieira S.M. (2016). [Missing Data](#). In: Secondary Analysis of Electronic Health Records. Springer, Cham.
- [8] Barandela R., Valdovinos R.M., Sánchez J.S., Ferri F.J. (2004). [The Imbalanced Training Sample Problem: Under or over Sampling?](#). In: Fred A., Caelli T.M., Duin R.P.W., Campilho A.C., de Ridder D. (eds) Structural, Syntactic, and Statistical Pattern Recognition. SSPR /SPR 2004. Lecture Notes in Computer Science, vol 3138. Springer, Berlin, Heidelberg.
- [9] Hancock, J.T., Khoshgoftaar, T.M. [Survey on categorical data for neural networks](#). J Big Data 7, 28 (2020).
- [10] F. Almeida and G. Xexeo [Word embeddings: A survey](#). CoRR, vol.abs/1901.09069, 2019. [Online].

- [11] Hannes De Meulemeester, Bart De Moor. [Unsupervised Embeddings for Categorical Variables](#). Fellow, IEEE & SIAM ESAT, STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, KU Leuven Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
- [12] C. Guo and F. Berkhahn, [Entity embeddings of categorical variables](#). CoRR, vol. abs/1604.06737, 2016. [Online].
- [13] Y. Russac, O. Caelen, and L. He-Guelton, “Embeddings of categorical variables for sequential data in fraud context,” in The International Conference on Advanced Machine Learning Technologies and Applications, AMLTA 2018, Cairo, Egypt, February 22-24, 2018, ser. Advances in Intelligent Systems and Computing, vol. 723. Springer, 2018, pp. 542– 552.
- [14] Duboue, P. (2020). [The Art of Feature Engineering: Essentials for Machine Learning](#). Cambridge University Press, isbn 9781108571647
- [15] Jolliffe Ian T. and Cadima Jorge (2016). [Principal component analysis: a review and recent developments](#). Phil. Trans. R. Soc. A.37420150202
- [16] F., Sahar. (2018). Machine-Learning Techniques for Customer Retention: A Comparative Study. International Journal of Advanced Computer Science and Applications. 9. 10.14569/IJACSA.2018.090238.
- [17] Hosmer, W., D. and Lemeshow, S. (2000). [Applied Logistic Regression](#). 2nd edn. John Wiley & Sons, Inc.
- [18] Song, Y. Y. and Lu, Y. (2015). Decision tree methods: applications for classification and prediction, Shanghai Archives of Psychiatry. Editorial Department of the Shanghai Archives of Psychiatry, 27(2), pp. 130–135. doi: 10.11919/j.issn.1002-0829.215044
- [19] Fang, K., Jiang, Y. and Song, M. (2016). Customer profitability forecasting using Big Data analytics: A case study of the insurance industry, Computers & Industrial Engineering, 101, pp. 552–564. doi: 10.1016/j.cie.2016.09.011.
- [20] Freund and Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, Journal of Computer and System Sciences, vol. 55, no 1, 1997, p. 119-139.
- [21] Sokolova, M., N. Japkowicz, et S. Szpakowicz (2006). Beyond accuracy, f-score and roc : A family of discriminant measures for performance evaluation. In Proceedings of the 19th Australian Joint Conference on Artificial Intelligence : Advances in Artificial Intelligence, AI’06, pp. 1015–1021.
- [22] Sokolova, M., N. Japkowicz, and S. Szpakowicz (2006). Beyond accuracy, f-score and roc : A family of discriminant measures for performance evaluation. In Proceedings of the 19th Australian Joint Conference on Artificial Intelligence : Advances in Artificial Intelligence, AI’06, pp. 1015–1021.

- [23] Albatineh, A. N. and M. Niewiadomska-Bugaj (2011). Correcting jaccard and other similarity indices for chance agreement in cluster analysis. *Adv. Data Anal. Classif.* 5(3), 179–200.
- [24] Namburu, S., Tu, H., Luo, J., and Pattipati, K. (2005). Experiments on Supervised Learning Algorithms for Text Categorization. *Aerospace Conference, IEEE*, pp. 1-8.
- [25] Fogarty, J., Baker, R. S., et Hudson, S.E. (2005). Case studies in the use of roc curve analysis for sensor-based estimates in human computer interaction. In *Proceedings of Graphics Interface 2005, GI '05, Canada*, pp. 129-136.
- [26] Nie G., Wang G., Zhang P., Tian Y., Shi Y. (2009) [Finding the Hidden Pattern of Credit Card Holder's Churn: A Case of China](#). In: Allen G., Nabrzyski J., Seidel E., van Albada G.D., Dongarra J., Sloot P.M.A. (eds) *Computational Science – ICCS 2009. ICCS 2009. Lecture Notes in Computer Science*, vol 5545. Springer, Berlin, Heidelberg.
- [27] K. Dahiya and S. Bhatia, Customer churn analysis in telecom industry, in 2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions), 2015, pp. 1–6.
- [28] I. Brândușoiu, G. Todorean, and H. Beleiu, Methods for churn prediction in the prepaid mobile telecommunications industry, in 2016 International Conference on Communications (COMM), 2016, pp. 97–100.
- [29] U. Yabas and H. C. Cankaya, Churn prediction in subscriber management for mobile and wireless communications services, in 2013 IEEE Globecom Workshops (GC Wkshps), 2013, pp. 991–995.
- [30] S. A. Qureshi, A. S. Rehman, A. M. Qamar, A. Kamal, and A. Rehman, Telecommunication subscribers' churn prediction model using machine learning, in 2013 Eighth International Conference on Digital Information Management (ICDIM), 2013, pp. 131–136