NOVA IMS
Information Management School

MGI

Mestrado em Gestão de Informação
Master Program in Information Management

Churn Prediction Modeling Comparison in the Retail Energy Market

Thiago Sampaio Nogueira

Project Work presented as partial requirement for obtaining the Master's degree in Information Management

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

CHURN PREDICTION MODELING COMPARISON IN THE RETAIL ENERGY MARKET

by

Thiago Sampaio Nogueira

Project Work report presented as partial requirement for obtaining the Master's degree in Information Management, with a specialization in Knowledge Management and Business Intelligence

Advisor: Prof. Roberto Henriques, PhD

February 2021

ABSTRACT

Machine Learning algorithms are used in diverse business cases and different markets. This project has the goal of applying different training models with the purpose of predicting customer churn in a retail energy provider. Following CRISP-DM methodology, the dataset was analyzed, prepared and results were evaluated in order to achieve the best method of forecasting the likelihood of churning in an existent customer base. That information is essential in company’s business planning to maintain and increase its portfolio.

KEYWORDS

Data Mining; Machine Learning; Churn Prediction; Supervised Learning; Retail Energy

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

1. INTRODUCTION

This work project was elaborated with the purpose of predicting customer churn for one of the largest independent retail energy provider in the United States. Churn, also known as attrition, it's defined by Lazarov and Capota (2007) as the discontinuation of a contract. The same authors also explain that this word is an amalgam of change and turn.

1.1. PROBLEM STATEMENT

Due to the high churn existing in this market, modeling customer attrition is essential for independent providers that compete directly with distributors, detaining the largest pool of customers.

The same liberties that allow customer choice regarding their utilities also quickly permit customer to change to the competition. That increases market's competitivity in an already reduced customer base available. Customer inertial is another known process, and acquisition is one of the most expensive operational costs within the company as it's usual in the market to be higher than maintaining the current clients (Iranmanesh, Hamid, Bastan, Hamed Shakouri, & Nasiri, 2019).

It's standard in the industry to elaborate fixed-term contracts with early termination fees charged to clients that churn before the end of a contract, which recover partially the customer acquisition cost, but a large chunk of them can be in variable portfolio, without months long contracts, that has a turnaround of just one billing cycle. These customers are highly subject to market fluctuations due to crises and environmental changes. These can go from the current pandemic to taxes to support green generation and to big consumption spikes due to heat or cold waves.

Sales margins are highly competitive among other independent retail providers due to the smaller pool of customers that chose to move out of their default distribution company in pursue of better rates.

Those challenges require that the attrition modelling current in use, with outdated techniques and tools, to be revamped into a more comprehensive and accurate forecast that can be applied to quarterly sales margin executive reports and retention processes.

1.2. OBJECTIVE AND SUMMARY

The objective of this work is to use Data Mining techiniques while pre-processing the data available and to apply a Machine Learning algorithm with company's own training dataset in order to predict churn based on patterns and correlations identified in customer's behaviour and demographics. With that goal, it was evaluated different models that can define the best framework available to this particular company and market.

In the second chapter, to situate the reader, it contains an overview of the retail energy market with history, challenges and current state. Literature review was performed to review the most recent past works regarding churning prediction modeling with similar problems and its solutions. It's also used as a benchmark for this work evaluation.

Third chapter starts with the basis for Data Mining and Machine Learning with definitions, algorithms, data types and model evaluation metrics. Finally, it dives deeper into predictive modeling theory and three models used by this work: Logistic Regression, Decision Tree and Artificial Neural Networks.

Following, it describes the CRISP-DM methodology that will be used to analyze the dataset and evaluate the models. It's composed of the following sections:

- Business Understanding: a general overview of the current business process.
- Data Understanding: the dataset overview and study in order to apply the correct techniques.
- Data Preparation: the most relevant section of data pre-processing to achieve good results in next two steps.
- Modeling: goes over which models were tested and its parameters.
- Evaluation: section to review results, compare all models and present the best model.

The fourth chapter contains the results of applying the described techniques to company dataset and discusses the best framework available from the ones tested.

Finally, conclusions are summarized in the fifth chapter and the discussion over limitations and recommendations for future works is the sixth and last chapter.

1.3. CONTRIBUTION TO THE COMPANY

Customer churn prediction will be used to forecast customer attrition and future company's sales margin with more accuracy. Also, retention workers will specifically target customers that have a higher chance of churning during the next quarter. Insights of data understanding and results during the evaluation phase will grant managers tools to direct business in order to have a larger market share.

2. LITERATURE REVIEW

2.1. Retail Energy Market Overview

In order to avoid monopolistic measures, United States gradually started to deregulate several markets in late 1970s. Public Utility Regulatory Policies Act (PURPA) from 1978 modernizes regulation for airlines, railroad, trucking, bus service and also started the market change for natural gas and electricity markets, requiring utilities to purchase power from other generators at cost set by the state (O'Connor, 2017).

That's in part a response to the 1970s oil crisis increased energy costs. Market conditions gradually lead to a substantial body of opinion for customer level competition, which started to be enabled by FERC Order 888 of 1996, that allowed wholesale competition through transmission access at reasonable rates to other companies.

The final decision on opening the market to different retail energy suppliers on the consumer level is upon to states due to distribution jurisdiction. As of 2020, 17 states allow retail markets, most initializing by end 1990s and beginning of 2000s (Choice, 2020). Several more states were willing to liberalize its markets, although the Enron crisis of 2001 and consequential California's market failure due to bad practices caused a rollback (O'Connor, 2017).

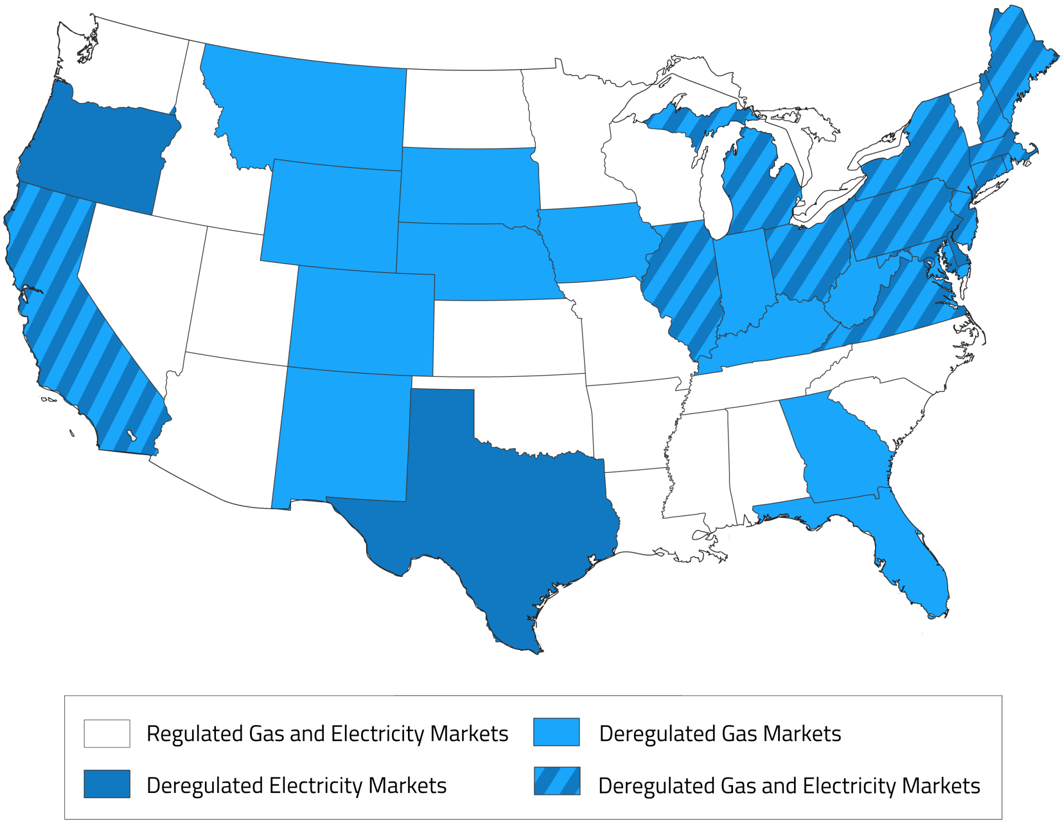

Regulated Gas and Electricity Markets
Deregulated Gas Markets
Deregulated Electricity Markets
Deregulated Gas and Electricity Markets

Figure 2.1 - United States Deregulated Map (Choice, 2020)

Consumer price didn't seem to change between regulated and deregulated markets as expected. Nakajima and Hamori (2010) compared the residential prices in all continental states and concluded that price elasticity kept the same between both markets.

This conclusion can be corroborated by Chen (2019) that studied the last 20 years of rates in Delaware and concluded that after the price cap was lifted, customers got a higher bill in average, with the regulated utility providing the lower and more stable rates. Only 10% of customers switched their electricity supplier, which denotes the highly competitive market of independent retail providers. Still, the author remarks that a flexible market may provide other benefits like integration of new technologies.

However, not always the market liberalization didn't lead to consumer benefits. Carlson and Loomis (2008) analyzed that prices in 10 years after deregulation trended down when in comparison to border states, comparing nominal and real rates paid by different customer classes.

2.2. CUSTOMER CHURN

Customer churn can be defined as the act of a customer closing an active account with a current service provider. Its analysis has a major importance in CRM as a gauge to customer retention and leads to a substantial profit increase (Van Den Poel & Larivière, 2004). In other note, attracting new customers has a cost that is significantly higher than maintaining the existing ones, therefore many companies use customer churn analysis as a marketing strategy to increase revenue and reduce costs to grow its position (Iranmanesh, Hamid, Bastan, Hamed Shakouri, & Nasiri, 2019).

Hejazinia and Kazemi (2014) determine different factors that leads to customer attrition:

- Service price: lower price on competition may drive customer to transition to a different service provider.
- Switching cost: several costs can compose a switch. Besides economical costs like early termination fees and down payments, they can also be physical, emotional and time.
- Competitors with superior technology: more advanced technology and products are drivers for highly competitive markets.
- Quality: service companies rely on the gap between customer expectations and perception of the services provided, with the goal being to reduce this difference.
- Satisfaction: similar to quality, satisfaction measures the difference among the perceived value and customers' expectations. That establishes customer attitude towards a service or product.
- Security concern: it relates to customer preoccupation with how his data is handled in enterprise environment.
- Advertising: the presentation of ideas regards goods or services can attract loyal customers.

Three different types of churn can happen. Involuntary churn is when the company terminates service due to a default or fraudulent usage. Unavoidable churn happens when customers die, moves or is permanently out of the market. Finally, voluntary churn is the customer terminating service due to one or more reasons above, this is the most valuable churn prediction (Yang & Chiu, 2006).

CRM tools have been used by service companies in order to establish their market position via customer acquisition and retention. Those tools gather information regarding customers during regular leads or contract lifetime that is useful for marketing purposes, being also essential to determine the likelihood of a customer canceling his contract (Lazarov & Capota, 2007).

Van Den Poel and Larivière (2004) define four sets of data that can be used to customer retention purposes: customer behavior, customer perceptions, customer demographics and macro-environment variables. All those can all be used in the customer churn analysis and they usually result in big datasets. Still, each one has a different importance to this prediction, with customer demographics being one of the most used in the literature (Lazarov & Capota, 2007).

A review of past works was conducted for customer churn prediction on the services market, focusing on utilities.

Iranmanesh et al. (2019) predicted customer churn on a retail banking dataset using Artificial Neural Networks with specific detail to the variable selection and areas that had a higher customer churn probability by customer type. Also is described techniques to act upon this data in order to bring more awareness of selected customers to try retaining them.

Hung, Yen and Wang (2006) used K-means for customer segmentation and then applied Decision Tree on the segmented clusters for a Taiwanese telecom company to determine churn. Beyond, it used Decision Tree and Neural Network using back propagation in the non-segmented entire set. All tentative were mostly successfully for the first six months, with significant degradation in one of the months. In the same first six months, Neural Networks outperformed both Decision Tree models.

Kristof Coussement and Poel (2008) compare applying Support Vector Machines model for churn prediction in a big newspaper subscription dataset with two different parameter-selection techniques for Logistic Regression and Random Forests. It concludes that although SVM performed best than Logistic Regression, Random Forest was still the best model to predict churn for this dataset.

Class imbalance is an issue on most of churn analysis due to the churn rate being reduced. Zhu, Baesens and vanden Broucke (2017) compare several techniques to compare class imbalance using sampling methods and ensembles solutions. The conclusion was that sampling methods not always improves the metrics depending on the classifier, and complex sampling doesn't have any superior performance to random sampling. However, ensemble methods performed the best with random under sampling.

M.A.H. Farquad, Vadlamani Ravi and Bapi Raju (2014) applied a hybrid algorithm using Support Vector Machines and Naive-Bayes Tree on a banking dataset over socio-demographics and behavioral data. Several class imbalances techniques were used, with SMOTE having the best sensitivity of 91.85%.

Olle and Cai (2014) experimented on creating a hybrid model with Logistic Regression and Voted Perceptron to predict churn and also the reason for churn on an Asian telecommunication company. It was found that the hybrid approach performed better than individual ones, however with lower difference due to the dataset available.

Huang and Kechadi (2013) also used a hybrid approach to customer churn prediction, with a model that combines supervised and unsupervised algorithms into a telecom dataset and several benchmark datasets to evaluate performance against single modeling techniques. The hybrid approach outperformed most of classifiers when comparing accuracy.

Pribil and Polejova (2017) used CRISP-DM methodology to apply churn modeling into a Czech energy company using Logistic Regression (three different data pre-processing models) and Decision Tree. The most accurate of all was the Decision Tree model with 88.08% of overall data accuracy. Also, it was verified which variables were more influential:

- Consumption change category between two billing cycles.
- Product Type
- High tariff consumption
- Total consumption
- Estimated consumption
- Age
- Billing cycle month
- Contract length
- Low tariff consumption
- Monthly payments

Cristian (2016) also applied CRISP-DM into a Dutch energy company dataset to compare several predictive models: Logistic Regression, Decision Trees, Random Forests, AdaBoost, Support Vector Machines and Neural Networks. After PCA was applied, both Random Forest and AdaBoost measured an AUC 0.98 for churn prediction. The customer contract and behavior variables were the most useful when modeling, outpacing perception, socio-demographics and macro environmental variables.

Due to the nature of retail energy companies, they possess more structured data related to clients than behavioral data. Moeyersoms and Martens (2015) analyzes how high-cardinality parameters from structured data, such as ZIP codes and surnames can be successfully used in customer churn prediction. Using a large dataset from a Belgium energy company, three transformations techniques were applied, two regular transformations such as dummy encoding and semantic grouping and one proposition from the paper: transformation to continuous attributes. Three methods were used to this last transformation: weight of evidence, supervised ration and Perlich ratio. The study found that including those attributes contributes to the model performance and more data lead to better results, making the transformations to continuous attributes outperform regular transformations.

Author	Title	Techniques	Domain	Best Performance
Hung, S.-Y., Yen, D. C., & Wang, H.-Y.	**Applying data mining to telecom churn management.**	K-Means, Decision Tree and Neural Networks.	Telecom	99% R-Squared in Neural Networks.
Kristof Coussement, & Poel, Dirk V. D.	**Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques.**	Logistic Regression, Random Forest and SVM.	Media	87.21% ROC AUC on Random Forest.
Zhu, B., Baesens, B., & vanden Broucke, S. K. L. M.	**An empirical comparison of techniques for the class imbalance problem in churn prediction.**	Several sampling algorithm over Random Forest and SVM.	Telecom	79.52% ROC AUC on Random Forest without sampling.
M.A.H. Farquad, Vadlamani Ravi, & Bapi Raju, S.	**Churn prediction using comprehensible support vector machine: An analytical CRM application.**	Several sampling algorithm over SVM and Naive-Bayes Tree and hybrid between both.	Banking	91.85% of Sensitivity with SMOTE on Hybrid.
Olle, G. D. O., & Cai, S.	**A hybrid churn prediction model in mobile telecommunication industry.**	Logistic Regression, Voted Perceptron and hybrid between both.	Telecom	72.1% ROC AUC on Hybrid.
Huang, Y., & Kechadi, T.	**An effective hybrid learning system for telecommunication churn prediction.**	Logsitic Regression, Decision Tree, k-NN, SVM, OneR, PART, SePI, k-NN-LR, KM-BoostedC5.0 and hybrid of K-Means and FOIL.	Telecom	90.33% ROC AUC on Hybrid .
Pribil, J., & Polejova, M.	**A churn analysis using data mining techniques: Case of electricity distribution company.**	Logistic Regression and Decision Tree	Energy	88.08% Accuracy on Decision Tree.
Cristian, R.	**Churn Prediction for the Dutch Energy Market.**	Logistic Regression, Decision Trees, Random Forests, AdaBoost, Support Vector Machines and Neural Networks.	Energy	98% ROC AUC on Random Forest and AdaBoost.
Moeyersoms, J., & Martens, D.	**Including high-cardinality attributes in predictive models: A case study in churn prediction in the energy sector.**	Logistic Regression, Decision Tree and SVM with an without high-cardinality variables.	Energy	74.39% ROC AUC on SVM with high cardinality variables encoded in weight of evidence.

Table 2.1 - Literature Review Summary

In summary, it can be concluded after this literature review that the energy market is highly competitive in United States de-regulated areas and that churning prediction in different markets have a good overall performance and can be applied due to metrics evaluation of higher 70% ROC AUC and, in some cases, even higher than the 90% mark of ROC AUC. Decision Trees methods also exceled when compared to other modelling techniques, specially within the energy market and electricity retail operators.

3. METHODOLOGY

3.1. DATA MINING AND MACHINE LEARNING

According to Hand and Adams (2014), Data Mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner. The relationships and summaries derived through a data mining exercise are often referred to as models or patterns. These patterns sometimes are unknown data relationships that are critical to make business decisions.

In order to carry this out, Machine Learning algorithms are used to identify these relationships and make predictions based on the data. That model is then trained with a pre-processed historical database, evaluated its performance and finally applied to a real-time database in order to support decision (Ge, Song, Ding, & Huang, 2017). It's common that business have big databases, increasing the complexity of these relationships over the last decades. At same time, more awareness of this knowledge value and decreasing costs due to technical advancements lead to an increased popularity of these analysis.

Bishop (2006) divides Machine Learning algorithms in three different types: supervised learning, unsupervised learning and reinforcement learning.

In supervised learning, the training data is composed by input vectors with their corresponding target vector. Therefore, the model is trained to identify relationships that leads to the target value. This type of learning can be further divided in categories such as classification problems, in which the target vector is a finite number of discrete categories; and regression problems when the output consists of continuous variables.

Unsupervised learning has no target value in training data, the goal is to find group of similar records within the data, known as clustering, find data distribution within the input space, which is called density estimation, and finally project high-dimensional data in two or three dimensions with the purpose of visualization.

Finally, the author describes the reinforcement learning as the process to find the optimal output based of trial and error and its defined rewards.

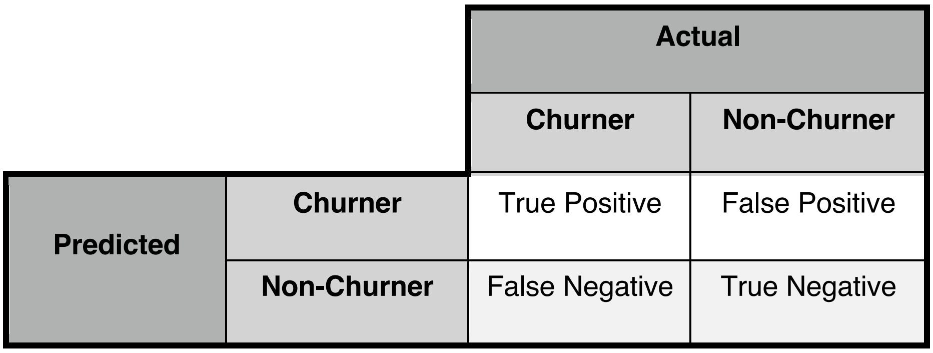
This work is focused in defining churn customers for an existing training dataset of customers that churned previously. Therefore, it's a supervised learning exercise where it will predict future classification of churned or non-churned customers based on an input vector.

According to Sammut and Webb (2017), one of the main Machine Learning problems is the classification of very unbalanced dataset. That means the dataset contains classes in which the number of records is very small when compared with the whole. It's also typical of binary classification models, such as this one, when records belong to positive and negative classes. Usually the positive class, as customers that churned, is a way smaller percentage of the total of records which includes customers that are still doing business with the company.

This balancing needs to be addressed before the model is trained. Several techniques can be used to achieve this effort as per Shelke, Deshmukh and Shandilya (2017): undersampling and oversampling are the main methods used and equivalents, active learning is used when data can be manually labeled to increase its records and the cost-based method provides a solution at the algorithm level. As this work focus on existing labeled dataset and algorithms, it'll focus on sampling methods.

Undersampling algorithms work reducing records on the majority class, usually done randomly of statically (informed undersampling). In oversampling, new records are added to the minority class, they can be added randomly with replication of existing records or synthetically, where artificial records are generated.

The evaluation of a binary classifier can be summarized in the contingency table, also known as confusion matrix, in order to compute the diverse evaluation measures existing in machine learning. This table contains four groups in which it can be divided results whether they belong to a given class c or not and whether they were classified as positive or negative (Sammut & Webb, 2017). In this exercise, the class is the customers that churned and summary can be seen on table below:



		Actual	
		Churner	Non-Churner
Predicted	Churner	True Positive	False Positive
	Non-Churner	False Negative	True Negative

Table 3.1 - Confusion Matrix

Given the values above, several metrics can be computed to measure the output predictions of a given model to reality. Accuracy is defined by Mehdiyev, Enke, Fettke and Loos (2016) as the sum of ratios of all correct classifications to the number of total classifications, giving a general overview of those results.

$$\frac{(TP + TN)}{(TP + FN + FP + TN)}$$

Equation 3.1 - Accuracy

The two metrics below can be used as how well the information retrieved regarding the relevant information requested by the user (Sammut & Webb, 2017). Precision can be defined by the formula below in which describes the total relevant results by the total number of information retrieved to the user in that given scenario. In this specific case, it's the actual churned customers predicted divided by all churned predicted.

$$\frac{TP}{(TP+FP)}$$

Equation 3.2 - Precision

Sensitivity, which can also be mentioned in literature as Recall or True Positive Rate, is the ratio of total relevant results returned by the total number of actual relevant results in the dataset.

$$\frac{TP}{(TP+FN)}$$

Equation 3.3 - Sensitivity

Specificity or True Negative Rate is similar to the metric above but relates to the documents not relevant to the user, meaning the negative results.

$$\frac{TN}{(TN+FP)}$$

Equation 3.4 - Specificity

F-Score, also called F-Rate or F-Measure, can be used instead of separated measures above, grouping Precision and Sensitivity together, as follows.

$$2 * \frac{(Precision * Sensitivity)}{(Precision + Sensitivity)}$$

Equation 3.5 - F-Score

Finally, we can have a very often used metric called ROC Curve, as stands for Receiver Operating Characteristic curve, that is a relationship between Sensitivity and Specificity for a binary classifier. One of the most important metrics of ROC analysis is the area under curve or AUC, which can be between zero and one. As defined by Sammut and Webb (2017), AUC = 1 the model classifier scores every positive higher than negative, and AUC = 0 the opposite. AUC = 1/2 then the model can be performing random classification.

3.2. PREDICTIVE MODELLING

Based on past work research, often Logistic Regression, Decision Trees and Artificial Neural Networks are used in churn prediction, therefore, this paper will analyze the dataset with these three algorithms. Below it follows an overview of each technique.

3.2.1. Logistic Regression

Most of supervised learning algorithms are based on estimating a probability distribution p(y | x) (Goodfellow, Bengio, Courville, & Bengio, 2016). Usually this can be determined via a linear regression to retrieve the maximum likelihood estimation using a weighted sum of predictor variables and a random component (Hand & Adams, 2014). In a classification problem, this can define the probabilities of each class, then both probabilities for a given record should sum up to 1.

As a linear regression, this is usually achieved by finding the best parameter vector θ for a parametric family of distribution p(y | x; θ) (Goodfellow et al., 2016):

$$p(y|\mathbf{x};\theta) = \aleph(y;\theta^{\mathcal{T}}\mathbf{x},\mathbf{I})$$

Equation 3.6 - Linear Regression

To specify, given the probability of the ith record yields value 1 is p(i) and even record is independent, then this probability follows a Bernoulli distribution (Hand & Adams, 2014):

$$p(Y(i) = y(i)) = p(i)^{y(i)}(1 - p(i))^{1-y(i)}$$

, where y(i) ∈ {0, 1}.

Equation 3.7 - Linear Regression Probability

In order to get the odds of a given event to take place, in this case to retrieve a 1, it can be used the following formula (Hand & Adams, 2014):

$$\frac{p(y=1|\mathbf{x})}{1-p(y=1|\mathbf{x})}$$

Equation 3.8 - Linear Regression Odds

To achieve a non-linearity of the probability p(y = 1|x), instead of a linear model that can achieve values between 0 and 1, a logistic function can be used to retrieve the log odds:

$$\log\frac{p(y=1|\mathbf{x})}{1-p(y=1|\mathbf{x})}$$

Equation 3.9 - Linear Regression Log Odds

De Caigny, Coussement and De Bock (2018) affirm that logistic regression became the standard on churn prediction due to good predictive performance and robust results. Also, it’s way more comprehensible than other black box methods due to the direct estimation of probabilities as reviewed above and competes with more advanced techniques.

3.2.2. Decision Tree

A Decision Tree is a hierarchical decision process in which usually binary decision trees are used, with two different outcomes in each node (Bonaccorso, 2018). It’s one of most used machine learning algorithms and also one of the oldest. It can be easily understood, as each node is just a simple classification (Sammut & Webb, 2017).

The algorithm is a top-down one, in which it decides the best variable partition on the root of tree and then the corresponding nodes and further variables are branches to this tree (Sammut & Webb, 2017). Given the whole dataset, as defined by (Bonaccorso, 2018):

$$X = \{\overline{x}_1, \overline{x}_2, ..., \overline{x}_M\}\ where\ \overline{x}_i\ \in\ \mathbb{R}^n$$

Equation 3.10 - Decision Tree Input Dataset

Then each level is defined by the tuple below:

$$\sigma = \langle i, t_i \rangle\ where\ i \in (1, n)\ and\ t_i \in (min\ x^{(i)}, max\ x^{(i)})$$

Equation 3.11 - Decision Tree Level

Being the first level, the input variable *i* and the threshold *t* is the value chosen for each range of the input feature. See below the example given by (Bonaccorso, 2018):

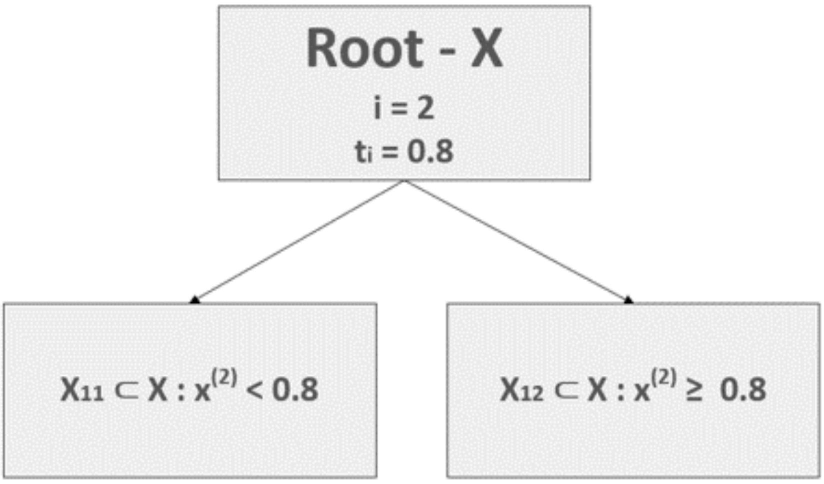


Figure 3.1 - Decision Tree (Bonaccorso, 2018)

The initial set is subdivided into two subsets depending on the input feature *i* = 2 being less or greater than threshold 0.8. Then the algorithm would continue until all nodes are completed for this given record and its input variables, classifying into a given class.

In order to define what is the threshold, the optimal balance should be found. One of the techniques used is to calculate an impurity measure to how heterogeneous is a node. A higher impurity returns several categories, as a 0 one grants a single category. The ideal scenario would be all nodes to have zero impurity, but this is limited by the number of levels (depth) of the tree and its complexity (Bonaccorso, 2018).

Sammut and Webb (2017) define two impurity measures as the most well-known in literature: information-theoretic entropy and the Gini index. Both can be defined as follows:

$$Entropy(S) = -\sum_{i=1}^{c} \frac{|Si|}{|S|} \times \log_2 \left(\frac{|Si|}{|S|}\right),$$

where S is the training set and Si the examples from training that belongs to c.

Equation 3.12 - Decision Tree Entropy

$$Gini(S) = 1 - \sum_{i=1}^{c} \left(\frac{|Si|}{|S|}\right)^2$$

Equation 3.13 - Decision Tree Gini

Although the simplicity and efficiency of Decision Trees are very useful for business cases, the model can be prone to overfitting when the dataset doesn't include records in which the same attributes are from different classes.

3.2.3. Artificial Neural Network

Artificial Neural Networks were designed to mimic human brain learning activities, trying to achieve mathematical representations of biological systems. The original perceptron model and learning curves follows discovery of social sciences on human learning, especially on pattern recognition Sammut and Webb (2017) and Bishop (2006).

Mehlig (2019) establishes that, although the human brain is widely complex when regarding to neural network algorithms, the fundamental is the same: pattern recognition learning based on previous experiences from a training dataset and extrapolating this, generalizing to new records. The basic natural neuron can be seen in the schema below:

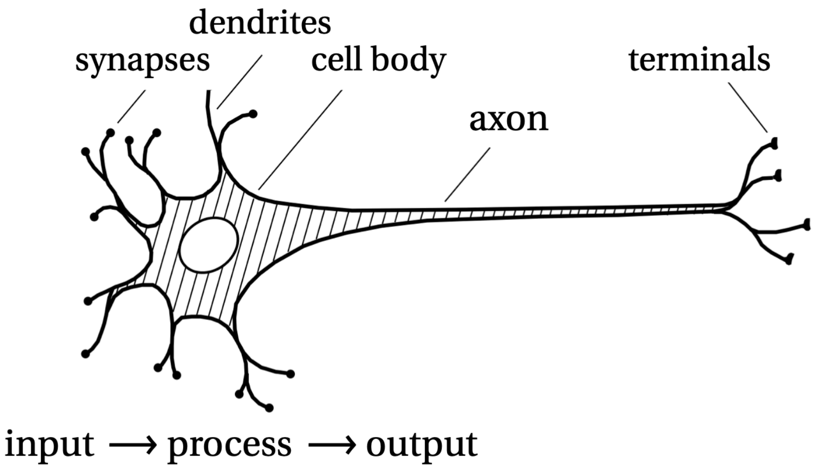


Figure 3.2 - Natural Neuron (Mehlig, 2019)

Dendrites receive electrical inputs via its tips (synapses), the signal is process inside the cell body and finally travels to other cells via the axon and it's transmitted to other neurons synapses using the original neuron terminals (Mehlig, 2019).

Similar to this concept, the basic neuron from a simple artificial network concept involves receiving several variable inputs just as a synapse, applying weights to each one and then processing via an activation function to remap this input dataset into a different dataset and then finally outputting its result (Bonaccorso, 2018). The structure of an artificial neuron can be seen below:

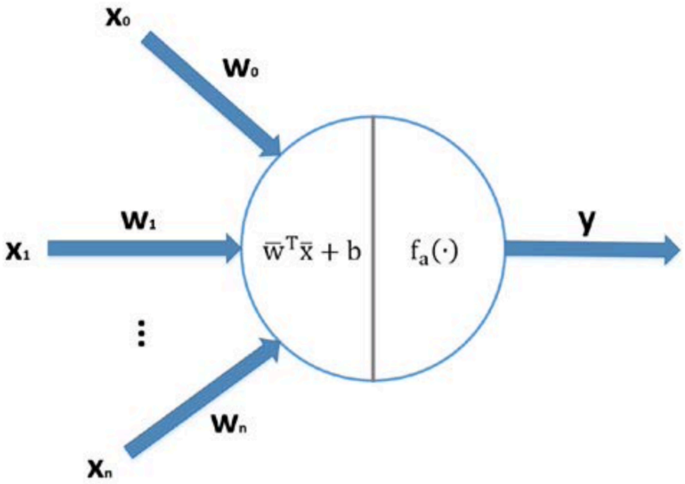


Figure 3.3 - Artificial Neuron (Bonaccorso, 2018)

In mathematical terms, an input vector *x* from the real universe, applying its weight vector *w* with a scalar constant *b* and finally the sum is applied an activation function *f* and generates the expected output, in case of binary classification algorithms the appropriate 1 or 0.

Different activation functions can be used and achieves similar results. As an example, there's a threshold activation function, in which compares the sum to a certain threshold value, if greater or equal is 1, else 0. Piecewise-Linear function can take continuous values and, finally, the sigmoid function transforms the input sum into 0 and 1 or -1 and 1 (Hajian & Styles, 2018).

These simple networks connect neurons in order to process information are called perceptrons. This simplest form returns the state of each neuron, as seen above, of active or inactive via a sign(x) function from a single layer of linear inputs. For other complex scenarios, the activation function can return continuous values (Mehlig, 2019). Following a schematic of a given perceptron by Bonaccorso (2018):

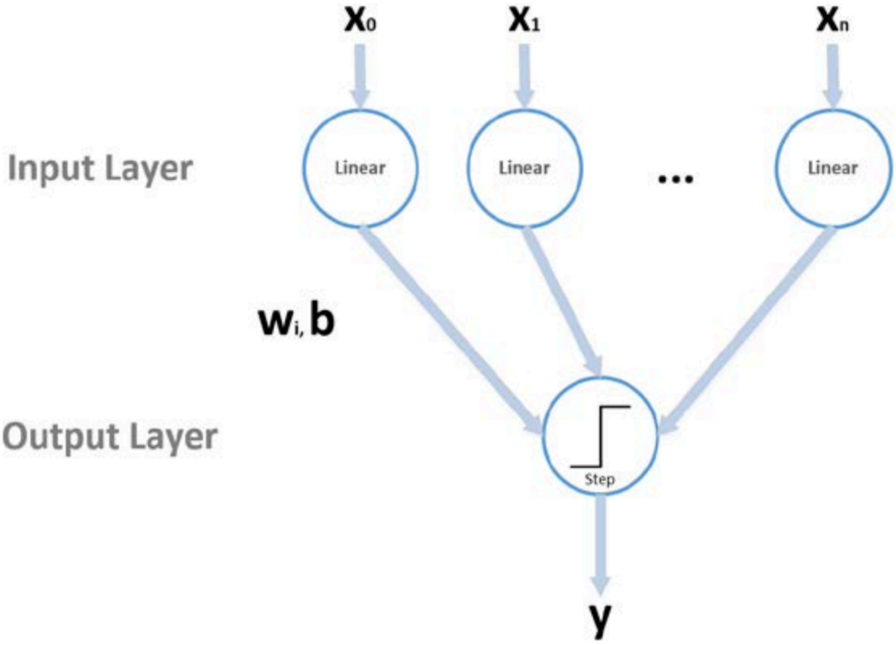


Figure 3.4 - Perceptron (Bonaccorso, 2018)

The perceptron above can be incremented and reduced its linearity by extending the number of layers between the input layer and output layer. This is called Multi-Layer Perceptron and it’s one of more used and known techniques to apply neural networks to non-linear problems (Fleck, Tavares, Eyng, Helmann, & Andrade, 2016). Below there’s a simplified representation of one hidden layer for simplicity purposes by Bonaccorso (2018):

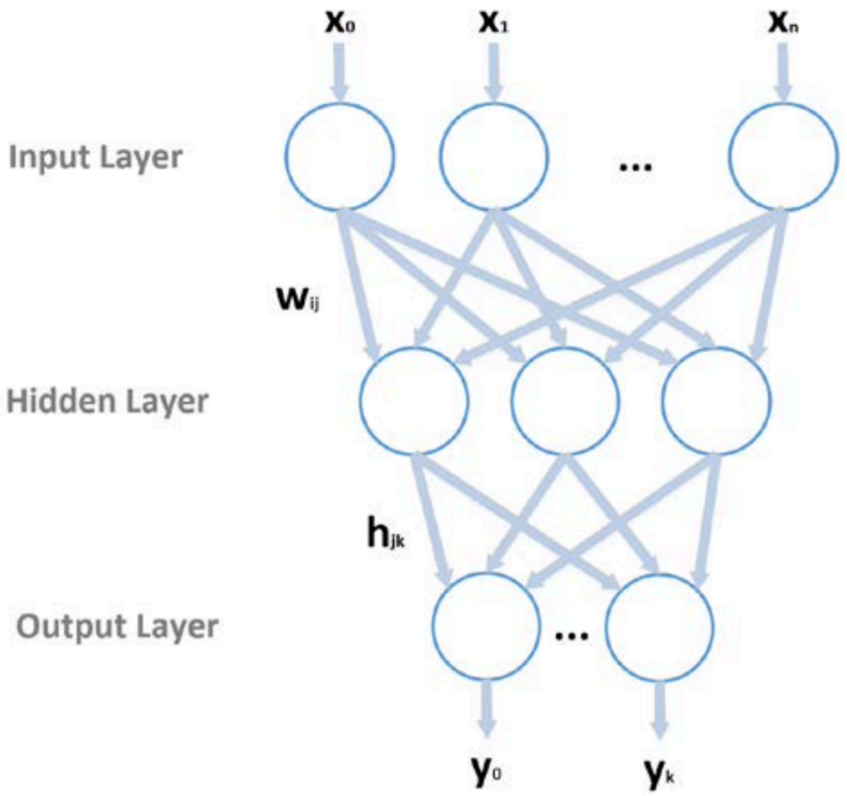


Figure 3.5 - Multi-Layer Perceptron (Bonaccorso, 2018)

It’s important to notice the extra matrix *h* that will weigh the result from input layer and also an activation function should be ideally non-linear in order to increase the results that will be input to the next layer (Bonaccorso, 2018).

In regarding to the learning aspect of a neural network, the weights must be adjusted whenever a training dataset record falls into the wrong classification. Therefore, that’s the main element that must be learned (Bonaccorso, 2018).

One way to adjust weights are by back-propagation of errors for a feed-forward neural network, meaning that during training the algorithm will compute an error factor for each output unit and therefore will back-propagate this error through the network, adjusting each weight correspond to the importance on its calculation (Sammut & Webb, 2017). This can happen during several epochs until the network is adapted to the input dataset, also the learning rate should be applied to determine how strong the error factor will be applied in each run through.

To summarize, (Fleck et al., 2016) compiles several advantages can be taken from ANNs implementation: it’s non-linearity, adaptability, response to evidence, contextual information, fault tolerance and neurobiological analysis. In its disadvantages can be listed a few such as: training can take longer, black-box behavior and demand for a big input dataset.

3.3. CRISP-DM

Even after twenty years since its creation, CRISP-DM (CRoss-Industry Standard Process for Data Mining) is indeed the industry standard methodology being applied in data discovery and machine learning modeling (Martinez-Plumed et al., 2020).

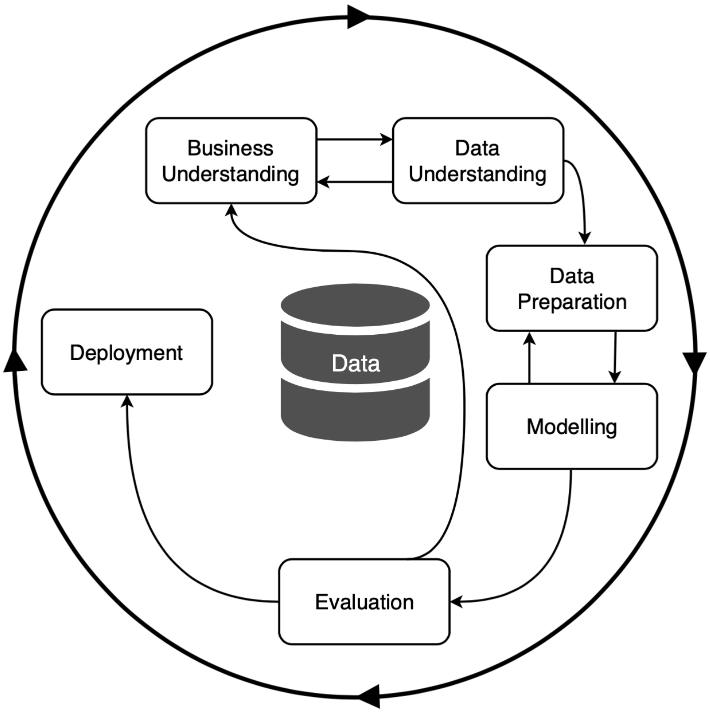


Figure 3.6 - CRISP-DM (Martinez-Plumed et al., 2020)

As shown on graph above, it’s divided in six interconnected phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment. In this chapter we will go over these phases, with exception of Deployment.

The first iteration of this methodology was developed by Chapman et al. (2000) with application and refinement in large industries with successful data mining process established.

As defined by Chapman et al. (2000), Business Understanding goes over the discovery of objectives and requirements from a business perspective and set an initial plan based on the company needs. Data Understanding reviews specifically the data gathered to get familiar with it and identify possible problems. Data Preparation goes over the creation of a final dataset that will be feed as a training data into the model. This task happens multiple times as the project refines model. In the modeling phase, several techniques can be used and different parameters until the best approach is found. In evaluation, key metrics are used to assess the quality of the whole process. Finally, deployment is the presentation to the final customer and applies the process to company’s decision process.

3.3.1. Business Understanding

This project goal is to successfully predict whether a given customer from an independent retail energy provider company located in United States, with a customer base of around two hundred thousand customers, will churn in the quarterly evaluation. As input data, it'll be required three months' worth of customer data acquisition and those customers will be followed in the next year. That's due to market seasonality, in which customer acquisition and churn rate varies during the year and seasons.

The relevancy of this project will be used to determine which accounts can be contacted to reinforce relationship and also which ones have a good prospect of renewing the current contract.

3.3.2. Data Understanding

The training dataset is composed of customer records and its contract information acquired from beginning of January 2019 to end of March 2019. The timeframe was defined due to the seasonality of contract acquisition in which varies each season. Furthermore, these datapoints will be used as a sliding window of forecasting every quarter and applied this forecast on final quartely sales margin report based on training data from previous year and its weather variations. This dataset was exported from the company Data Warehouse and its total number of records is 30998. As several data points of customer contract involves dates, a pre-processing was performed during the extraction in order to transform these dates into interval records, such as number of days of usage and contracted months. *Churned* flag was defined by accounts that have a termination date within its fixed-term contract or, for variable term contracts, anytime during the last year. Accounts which are still active and don't have termination date were set to 0.

Below it's a description of each datapoint gathered and its description:

Variables	Description
AccountId	Record Primary Key
AccountType	Business classification (Ex: Small Business, Residence)
EntityId	Company Id record
MarketCode	Client state
UtilityCode	Client energy distribution utility
Origin	Customer acquisition method (Ex: Online, Tablet)
PorOption	Company insurance for customer

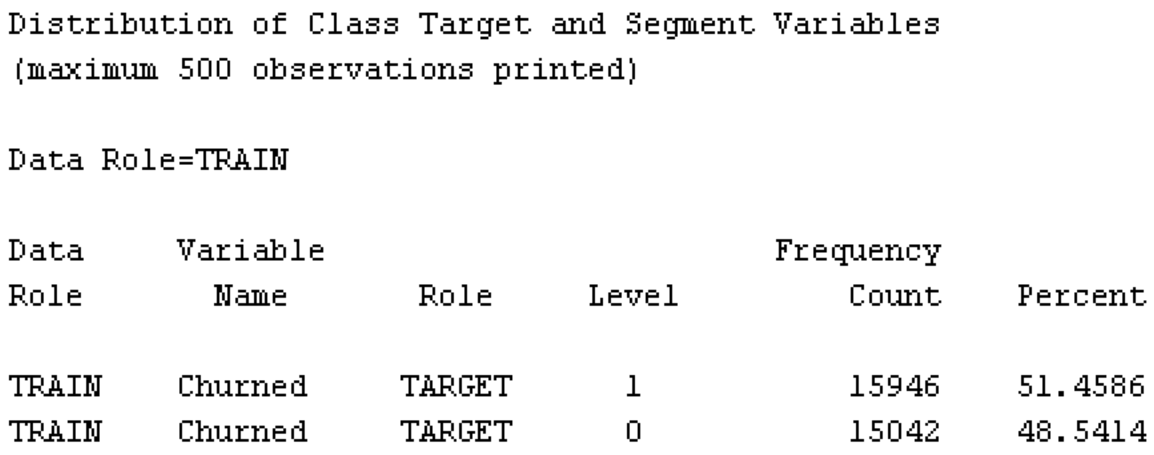
BillingType	Type of bill
AnnualUsage	Annual usage in kWh as of customer acquisition
IsLowIncome	Customer is low income
ContractType	Physical type of contract
ContractDealType	Type of deal (Ex: New or Renewal)
ChannelName	Commissioned sales channel
ContractedTerm	Contract length in month
ContractedProductType	Type of product (Ex: Fixed, Green)
IsVariable	Out of contract customer
Usage1Days	Last customer billing in days
Usage1	Last customer billing in kWh
Usage2Days	Previous customer billing in days
Usage2	Previous customer billing in kWh
Usage3Days	Second before last customer billing in days
Usage3	Second before last customer billing in kWh
Churned	If customer churned

Table 3.2 - List of Input Variables

The dataset above encompasses both socio-demographics and behavioral data, as seen in literature as useful to churn prediction. Variables like AccountType, MarketCode, UtilityCode and IsLowIncome demonstrate social-demographics and Usage variables together with contract ones can be defined as behavioral.

With the objective to start our data understanding, the exported dataset was loaded in the SAS Enterprise Miner application via File Import node, then used Stat Explore, Graph Explore, Variable Clustering and Multiplot nodes.

In relation to data imbalance, this specific dataset is not imbalanced as usually churned data customer is, as this market has a high turn around due to challenges related with energy liberated market. Below is the discrimination regarding the target variable Churned:



```
Distribution of Class Target and Segment Variables
(maximum 500 observations printed)

Data Role=TRAIN

Data      Variable                          Frequency
Role        Name       Role      Level          Count     Percent

TRAIN     Churned     TARGET       1            15946     51.4586
TRAIN     Churned     TARGET       0            15042     48.5414
```

Figure 3.7 - Class Distribution

An example of the analysis performed in this step is the review of each variable, exploring future validations that can be performed, such as the one below that contains some records with empty values. Also, this variable seems to have some correlation to our target variable due to the majority of Green product being churned.

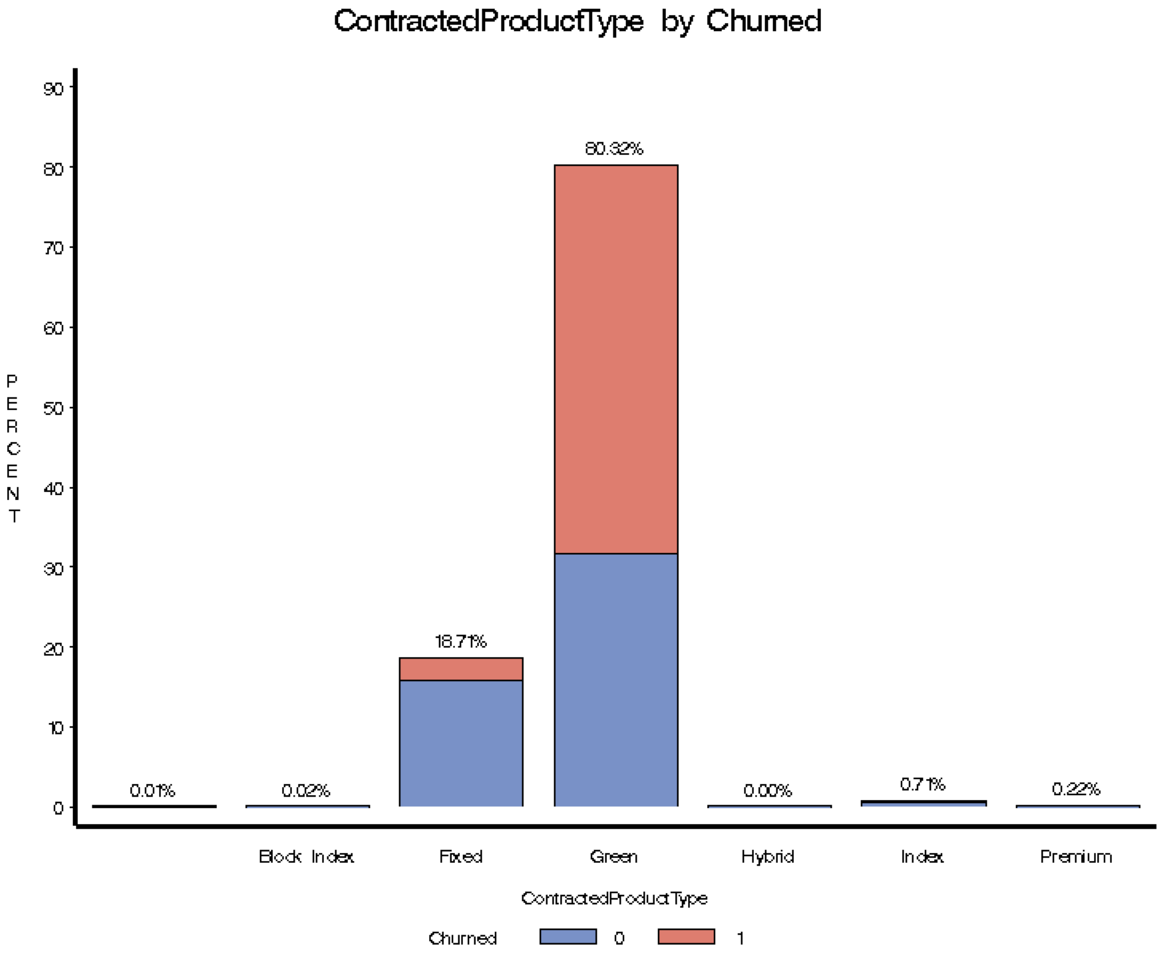


Figure 3.8 - Variable Exploration in Multiplot Node

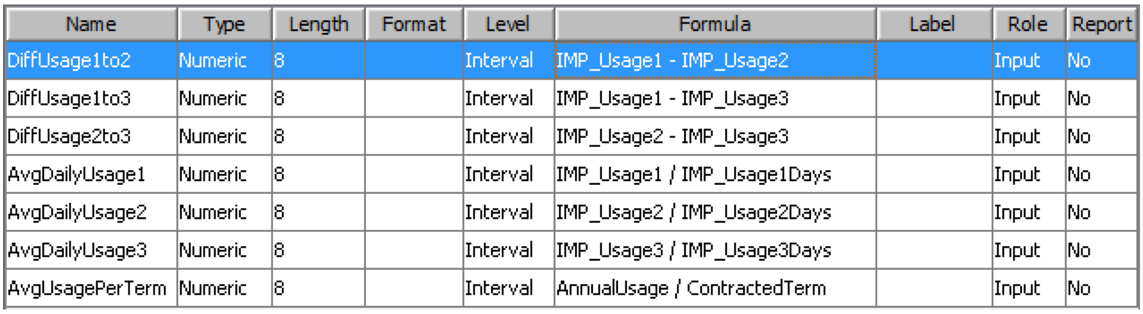
Although reviewed, information such as variable worth and correlation will be performed during data preparation step, as new variables will be generated, outliers removed and imputation of missing data.

3.3.3. Data Preparation

First step in data preparation in this project was to eliminate outliers. After reviewing the data and selecting interval variables to be filtered, the rule of thumb of 3% of observation was used as the interval data selected has skewed distributions. After filtering manually, the variables regarding Usage and Contracted Term, 929 records of 30988 were removed from further analysis, giving a total of 2,99% of outliers.

Next, due to missing values in variables related to Usage and ContractedProductType, a median was used to populate all usage related fields and tree surrogate was the technique defined for the latter class variable. It's important to notice that Usage3 variable had 5444 missing values, which can be understood by customers that rescind their contracts before the three-month mark. On other note, only 2 records didn't have information for ContractedProductType, which may denote data inconsistency on the database.

To get more insight from original variables, transformations were made to generate new variables. The first one was AvgUsagePerTerm which takes the AnnualUsage variable and divides by the ContractedTerm, this will return the average monthly usage in kWh during the total term. Then, three variables were created based on each Usage to divide by the number of days to get daily usage in kWh. After this, all UsageDays variables were filtered out, as this information is basically the amount of days between meter reads and don't compose any business value. Finally, as per literature review, the consumption change between two billing cycles were calculated by subtracting the last billing cycle by the two previous ones into two variables and the Usage2 subtracted by Usage3 in another variable.



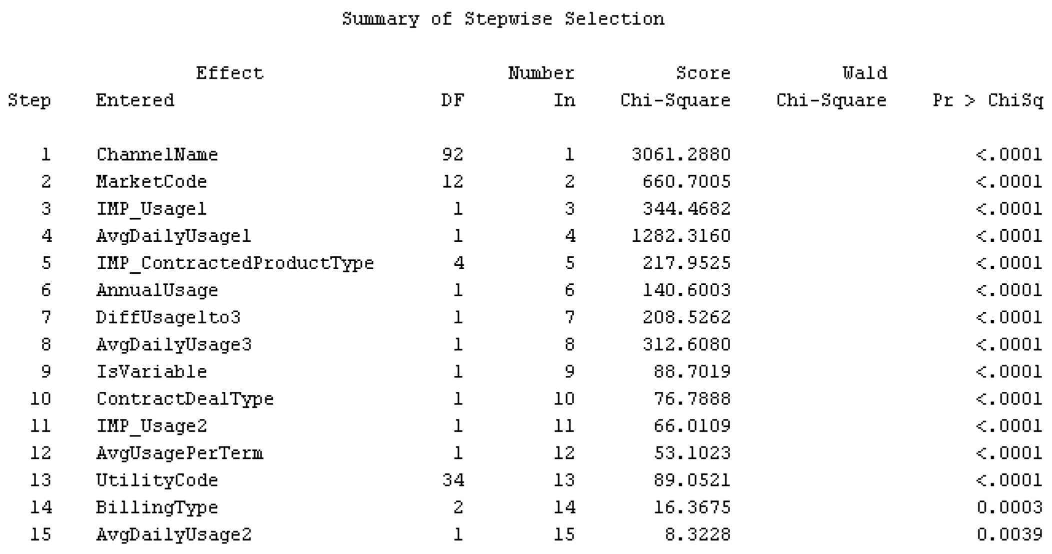
Name	Type	Length	Format	Level	Formula	Label	Role	Report
DiffUsage1to2	Numeric	8		Interval	IMP_Usage1 - IMP_Usage2		Input	No
DiffUsage1to3	Numeric	8		Interval	IMP_Usage1 - IMP_Usage3		Input	No
DiffUsage2to3	Numeric	8		Interval	IMP_Usage2 - IMP_Usage3		Input	No
AvgDailyUsage1	Numeric	8		Interval	IMP_Usage1 / IMP_Usage1Days		Input	No
AvgDailyUsage2	Numeric	8		Interval	IMP_Usage2 / IMP_Usage2Days		Input	No
AvgDailyUsage3	Numeric	8		Interval	IMP_Usage3 / IMP_Usage3Days		Input	No
AvgUsagePerTerm	Numeric	8		Interval	AnnualUsage / ContractedTerm		Input	No

Figure 3.9 - Variable Transformations

After transformation, the Data Partition node was used to divide the dataset into Training and Validation datasets with an allocation of 70% and 30%, partition method used was Stratification of the target variable.

In the final node of Data Preparation, Metadata, it was performed the dimension reduction with the objective to exclude variables that are not relevant for this analysis. A combination of variable relevance (worth) and redundancy (correlation). In order to achieve this, another review of all nodes from Data Understanding step with the prepared dataset was performed with addition of a stepwise logistic regression modeling.

Stepwise logistic regression is a node in which iterations are made while including and excluding variables and assessing the results with each one. Based on iteration results, it suggests which variables should enter. After this step, the following variables were automatically selected: AnnualUsage, AvgDailyUsage1, AvgDailyUsage2, AvgDailyUsage3, AvgUsagePerTerm, BillingType, ChannelName, ContractDealType, DiffUsage1to3, IMP_ContractedProductType, IMP_Usage1, IMP_Usage2, IsVariable, MarketCode, UtilityCode.



```
                         Summary of Stepwise Selection

                 Effect                       Number       Score          Wald
Step    Entered                       DF        In    Chi-Square    Chi-Square    Pr > ChiSq

   1    ChannelName                   92         1     3061.2880                      <.0001
   2    MarketCode                    12         2      660.7005                      <.0001
   3    IMP_Usage1                     1         3      344.4682                      <.0001
   4    AvgDailyUsage1                 1         4     1282.3160                      <.0001
   5    IMP_ContractedProductType      4         5      217.9525                      <.0001
   6    AnnualUsage                    1         6      140.6003                      <.0001
   7    DiffUsage1to3                  1         7      208.5262                      <.0001
   8    AvgDailyUsage3                 1         8      312.6080                      <.0001
   9    IsVariable                     1         9       88.7019                      <.0001
  10    ContractDealType               1        10       76.7888                      <.0001
  11    IMP_Usage2                     1        11       66.0109                      <.0001
  12    AvgUsagePerTerm                1        12       53.1023                      <.0001
  13    UtilityCode                   34        13       89.0521                      <.0001
  14    BillingType                    2        14       16.3675                      0.0003
  15    AvgDailyUsage2                 1        15        8.3228                      0.0039
```

Figure 3.10 - Summary of Stepwise Selection

Along with the step above, it was again analyzed the variable worth in Stat Explore step and some variables which were not selected by stepwise regression were imported due to the high variable worth such as: AccountType, ContractedTerm, EntityId, Origin and PorOption.

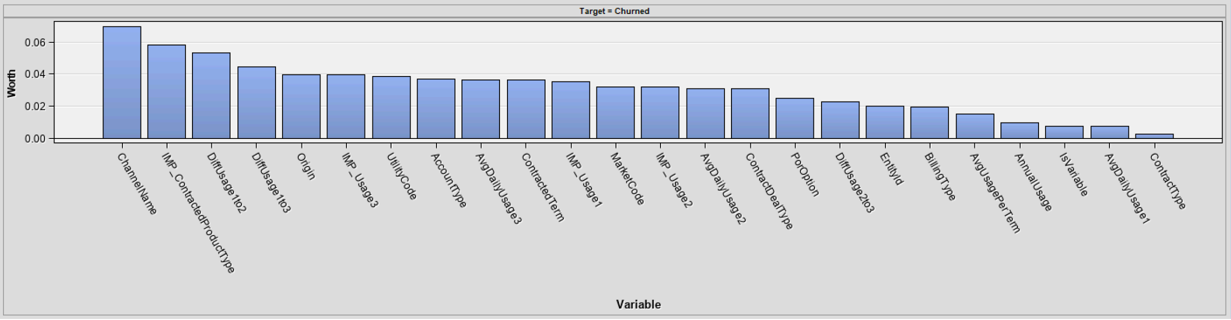


Figure 3.11 - Variable Worth

Finally, it was taken into consideration the correlation between interval variables using the Pearson correlation coefficient matrix. This work used a coefficient higher than 0.8 as correlated variables. Any variable which highly correlates with another selected variable was considered as redundant and removed. For this effort, it was used both Variable Clustering node and a SAS Code that calculates the coefficient and plots into a table.

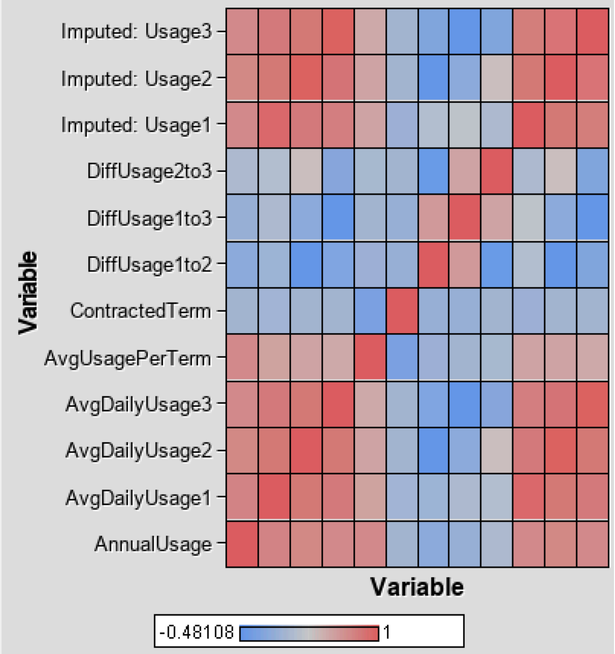
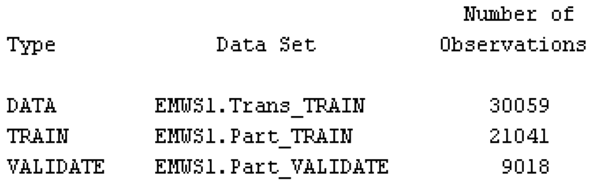


Figure 3.12 - Variable Correlation

As can be noticed above, selected variables AvgDailyUsage1, AvgDailyUsage2, AvgDailyUsage3 are highly correlated, so only the AvgDailyUsage2 was kept.

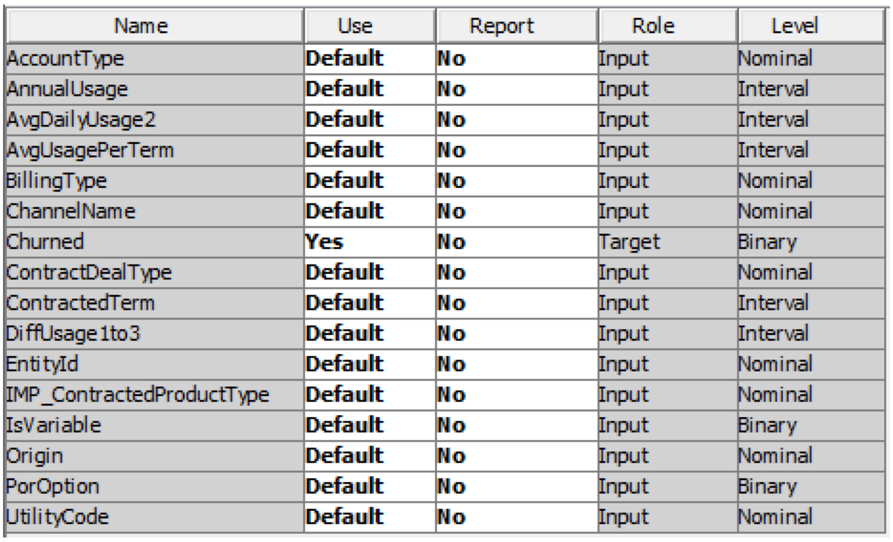
Finally, the total number of records as well with variables selected can be seen below with information that will feed the models selected in next section.



```
                                         Number of
Type            Data Set               Observations

DATA        EMWS1.Trans_TRAIN              30059
TRAIN       EMWS1.Part_TRAIN               21041
VALIDATE    EMWS1.Part_VALIDATE             9018
```

Figure 3.13 - Datasets Sizes



Name	Use	Report	Role	Level
AccountType	Default	No	Input	Nominal
AnnualUsage	Default	No	Input	Interval
AvgDailyUsage2	Default	No	Input	Interval
AvgUsagePerTerm	Default	No	Input	Interval
BillingType	Default	No	Input	Nominal
ChannelName	Default	No	Input	Nominal
Churned	Yes	No	Target	Binary
ContractDealType	Default	No	Input	Nominal
ContractedTerm	Default	No	Input	Interval
DiffUsage1to3	Default	No	Input	Interval
EntityId	Default	No	Input	Nominal
IMP_ContractedProductType	Default	No	Input	Nominal
IsVariable	Default	No	Input	Binary
Origin	Default	No	Input	Nominal
PorOption	Default	No	Input	Binary
UtilityCode	Default	No	Input	Nominal

Figure 3.14 - Final Dataset

3.3.4. Modeling

Proceeding to modeling phase, 70% of final dataset was feed as training input and 30% of final dataset was feed as validation input into different models for comparison purposes. The models used were the following:

- Neural Networks: it was experimented with 1, 3, 5, 7, and 10 hidden layers.
- Decision Tree: it was experimented with the following parameters 2 maximum branches and 3 maximum depth, 3 maximum branches and 4 maximum depth, 4 maximum branches and 5 maximum depth, 2 maximum branches and 3 maximum depth, 5 maximum branches and 6 maximum depth.
- Logistic Regression.

The hidden layers in Neural Networks add non-linearity capabilities to the model in which different options can be tried to achieve the best results.

The maximum branches of Decision Trees defines the number of splits of dataset based on the splitting rule selected, which in this case was the Gini index. For example, 2 maximum branches will split the dataset into binary leaves, as 5 maximum branches can split up to five different clusters. As per maximum depth, it's the number of layers a tree can have, from root node being the layer zero and its first children are depth 1. These numbers were selected to have a bigger variety of outputs as well control in the decision tree extension in order to have better data and model understanding.

3.3.5. Evaluation

During evaluation, first step was to analyze the confusion matrices. The results are shown for the validation dataset in order to assure we don't have any overfitting for the training dataset. Then, the following metrics based on the confusion matrix were compared for all models: accuracy, precision, sensitivity, specificity and f-score.

Beyond those basic metrics, the one that it was took as gold standard for the model comparison was the ROC curve and its area under (ROC AUC) index. This metric can be compared to the exercises appointed in literature review to estimate how well the model performed, as well between each model to grant the one that excelled.

Secondary metrics that were took into consideration were the misclassification rate, which indicates the total number of errors in our results and the cumulative lift which defines the ratio of correct responses by each model by the percentage of dataset deciles.

Finally, a deeper look into the best decision tree model might explain how it behaved within this dataset and can demonstrate possible overall improvements.

4. RESULTS AND DISCUSSION

For brevity purposes, we took the best models to represent below among all parameters chosen to be compared. Logistic Regression was tested once, Decision Tree with 5 maximum branches and 6 maximum depth and Neural Network with 7 hidden layers had the least number of errors. As stated in previous chapter, the total number of records in the validation dataset is of 9018.

		Actual	
		Churner	Non-Churner
Predicted	Churner	TP = 3854	FP = 1903
	Non-Churner	FN = 891	TN = 2370

Table 4.1 - Logistic Regression Confusion Matrix

		Actual	
		Churner	Non-Churner
Predicted	Churner	TP = 3425	FP = 956
	Non-Churner	FN = 1320	TN = 3317

Table 4.2 - Decision Tree (5 B; 6 D) Confusion Matrix

		Actual	
		Churner	Non-Churner
Predicted	Churner	TP = 3593	FP = 1299
	Non-Churner	FN = 1152	TN = 2974

Table 4.3 - Neural Network (7 HL) Confusion Matrix

As seen above, the Logistic Regression although had a great performance predicting churners, it didn't work so well for non-churners. Neural Network improved the non-churner performance but not at same pace the Decision tree did.

Model Description	Accuracy	Precision	Sensitivity	Specificity	F-Score
Neural Network	0,685850521	0,665914613	0,808640674	0,549496841	0,730370229
Neural Network (3 HL)	0,69937902	0,689597315	0,779557429	0,610344021	0,731823128
Neural Network (5 HL)	0,723109337	0,724082935	0,765437302	0,67610578	0,744186047
Neural Network (7 HL)	0,728210246	0,734464432	0,757218124	0,695998128	0,745667739
Neural Network (10 HL)	0,72588157	0,727163702	0,766701791	0,680552305	0,74640952
Decision Tree (2 B; 3 D)	0,682523841	0,729400293	0,630558483	0,740229347	0,676387476
Decision Tree (3 B; 4 D)	0,740740741	0,756555106	0,747945205	0,732740463	0,752225519
Decision Tree (4 B; 5 D)	0,741960523	0,780380334	0,709167545	0,778375848	0,743071657
Decision Tree (5 B; 6 D)	0,747615879	0,781784981	0,721812434	0,7762696	0,750602674
Logistic Regression	0,690175205	0,669445892	0,812223393	0,554645448	0,733955437

Table 4.4 - Model Metrics Comparison

Chart below compares metrics for all models tested and confirms analysis above by having a low rate of Specificity for Logistic Regression and a balanced result overall for the last Decision Tree model with upper values for each metric.

The biggest accuracy is on model Decision Tree (5B; 6D) with value of 0,7476. All other metrics are also above the 0,70 mark and although not all metrics have the best scores when compared with other models, it'll be clear below that is the best model for a deployment implementation.

Besides confusion matrix metrics, we can analyze ROC curve and ROC AUC. This graph gives a nice visual overview of each model and the area it's a more reliable evaluation than accuracy and other metrics alone.

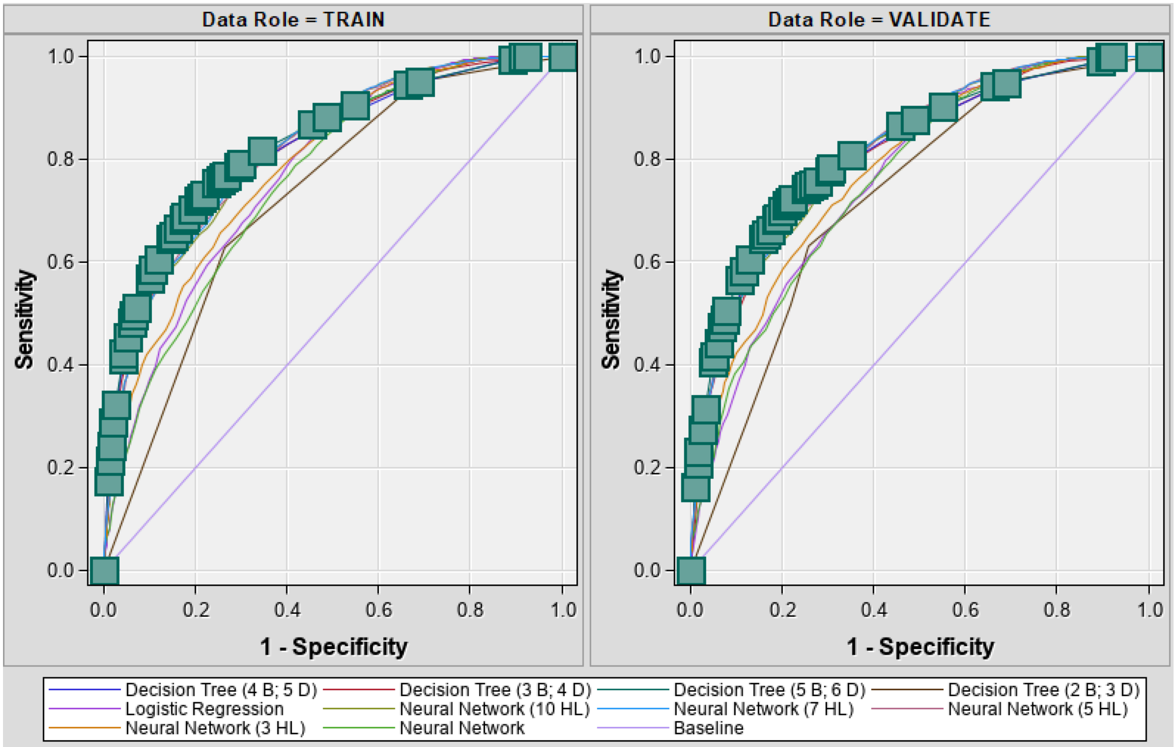


Figure 4.1 - ROC Curve

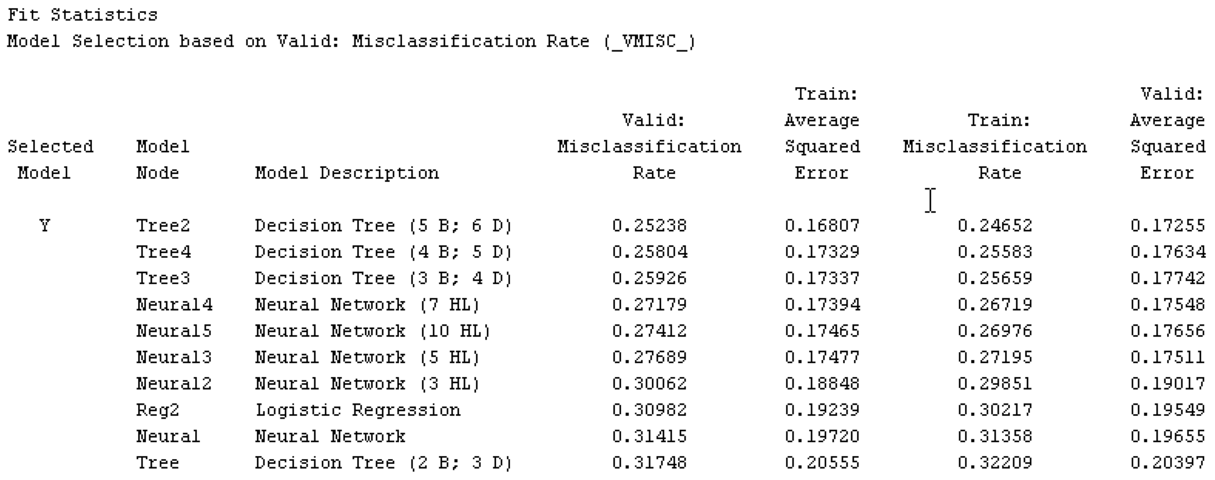
Graph above was taken from Model Comparison node and Decision Tree (5B; 6D) curve was highlighted.

Decision Tree (5 B; 6 D)	Decision Tree (4 B; 5 D)	Decision Tree (3 B; 4 D)	Neural Network (7 HL)	Neural Network (10 HL)	Neural Network (5 HL)	Neural Network (3 HL)	Logistic Regression	Neural Network	Decision Tree (2 B; 3 D)
0.82	0.81	0.81	0.81	0.81	0.81	0.78	0.76	0.76	0.72

Table 4.5 - ROC AUC Index

Most of Neural Networks have 0,81 ROC AUC value, a value similar to Decision Trees models. Although, the model analyzed above still demonstrates a higher capacity to predict churn customers with a value of 0,82.

Coherent with other metrics, misclassification rate is around 25% for both validation dataset and training one. One more time it demonstrates overfitting was not a problem during this exercise. Misclassification rate establishes the total mistakes made by the algorithm.



```
Fit Statistics
Model Selection based on Valid: Misclassification Rate (_VMISC_)

                                                                   Train:                       Valid:
                                                      Valid:       Average        Train:        Average
Selected    Model                                 Misclassification Squared  Misclassification  Squared
 Model      Node        Model Description             Rate          Error         Rate          Error

   Y        Tree2       Decision Tree (5 B; 6 D)      0.25238       0.16807       0.24652       0.17255
            Tree4       Decision Tree (4 B; 5 D)      0.25804       0.17329       0.25583       0.17634
            Tree3       Decision Tree (3 B; 4 D)      0.25926       0.17337       0.25659       0.17742
            Neural4     Neural Network (7 HL)         0.27179       0.17394       0.26719       0.17548
            Neural5     Neural Network (10 HL)        0.27412       0.17465       0.26976       0.17656
            Neural3     Neural Network (5 HL)         0.27689       0.17477       0.27195       0.17511
            Neural2     Neural Network (3 HL)         0.30062       0.18848       0.29851       0.19017
            Reg2        Logistic Regression           0.30982       0.19239       0.30217       0.19549
            Neural      Neural Network                0.31415       0.19720       0.31358       0.19655
            Tree        Decision Tree (2 B; 3 D)      0.31748       0.20555       0.32209       0.20397
```

Figure 4.2 - Misclassification Rate

The final analysis goes through the Cumulative Lift. As before, the Decision Tree (5B; 6D) overperforms all others model and sustain a higher lift through each depth.

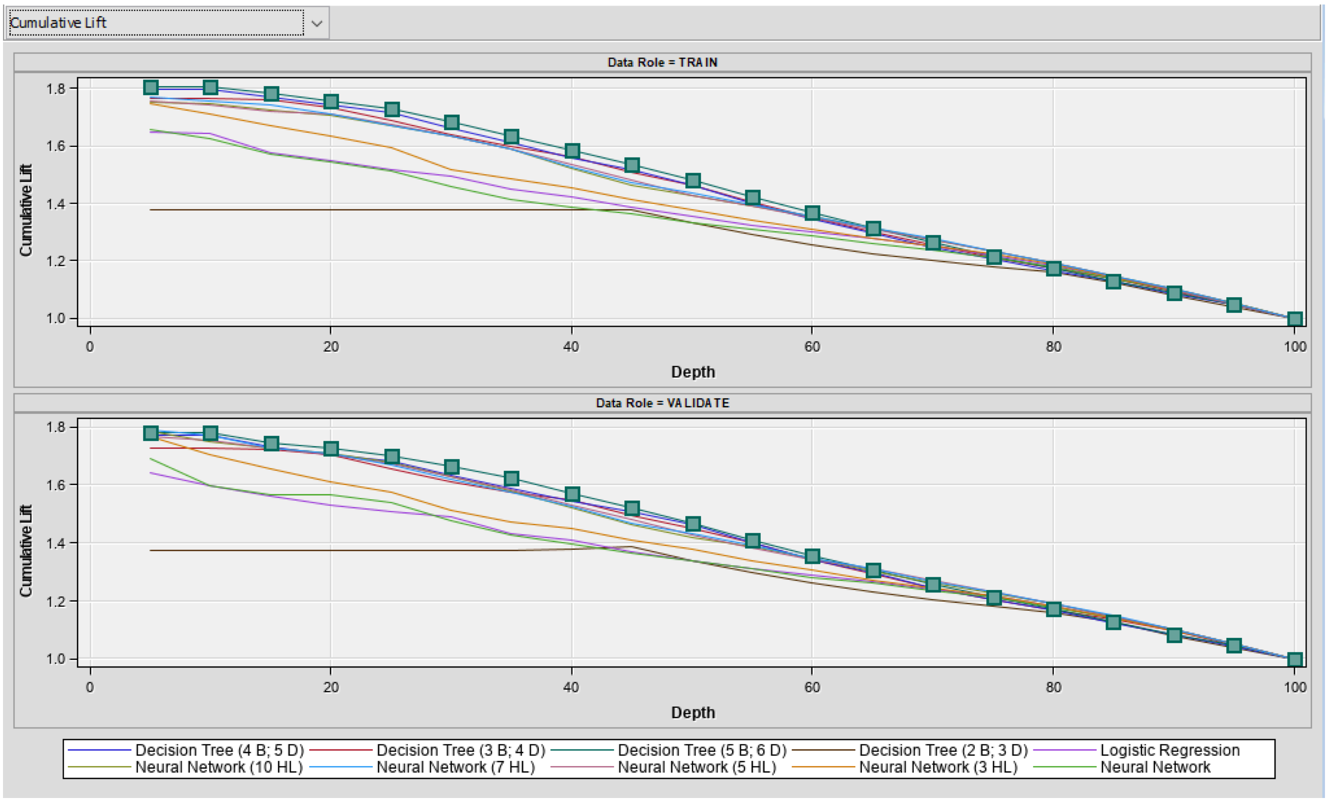
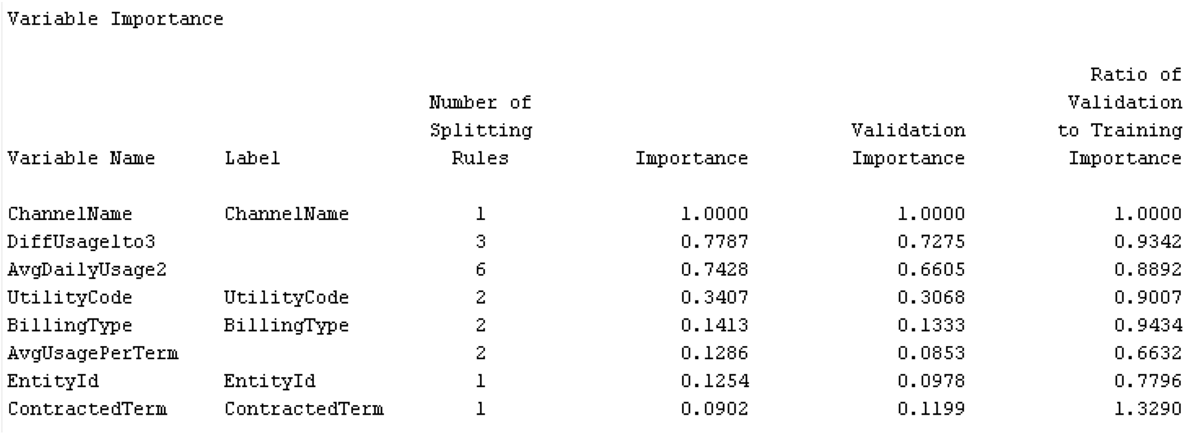


Figure 4.3 - Cumulative Lift

Once the model comparison is completed, we can take a further look at the best model in order to take insights of how this dataset behaved. In the model node, the Variable Importance section outputs how the decision tree was generated, and which variables are more important to the tree, from root to leaves.



```
Variable Importance

                                                                                   Ratio of
                                     Number of                                   Validation
                                     Splitting                      Validation   to Training
Variable Name      Label               Rules       Importance       Importance    Importance

ChannelName        ChannelName           1           1.0000           1.0000        1.0000
DiffUsage1to3                            3           0.7787           0.7275        0.9342
AvgDailyUsage2                           6           0.7428           0.6605        0.8892
UtilityCode        UtilityCode           2           0.3407           0.3068        0.9007
BillingType        BillingType           2           0.1413           0.1333        0.9434
AvgUsagePerTerm                          2           0.1286           0.0853        0.6632
EntityId           EntityId              1           0.1254           0.0978        0.7796
ContractedTerm     ContractedTerm        1           0.0902           0.1199        1.3290
```

Figure 4.4 - Variable Importance in Decision Tree (5B; 6D)

The root node is ChannelName, which have a high cardinality, that infers different sales channels had clustering performance metrics when related to customer churning. This can denote customer relation importance; regulatory rules being followed or better assessment of customer base. The total number of branches set up for this model calculated five different groups of channels. Tree generation was accessible in the node in its completion where this information was taken from.

Further, the DiffUsage1to3 variable was chosen, as an interval variable, it was divided into clusters of three to five branches, depending on the ChannelName parent node. Two of the parent nodes don’t have further ramifications. As expected by literature in (Pribil & Polejova, 2017), consumption change has a big impact during retail electrical market customer churning analysis.

As variable above, AvgDailyUsage2 was another variable generated during Data Preparing section, making with AvgUsagePerTerm the three interval variables with expression during this model training phase. All remaining variables are socio-demographic variables pertinent to each account.

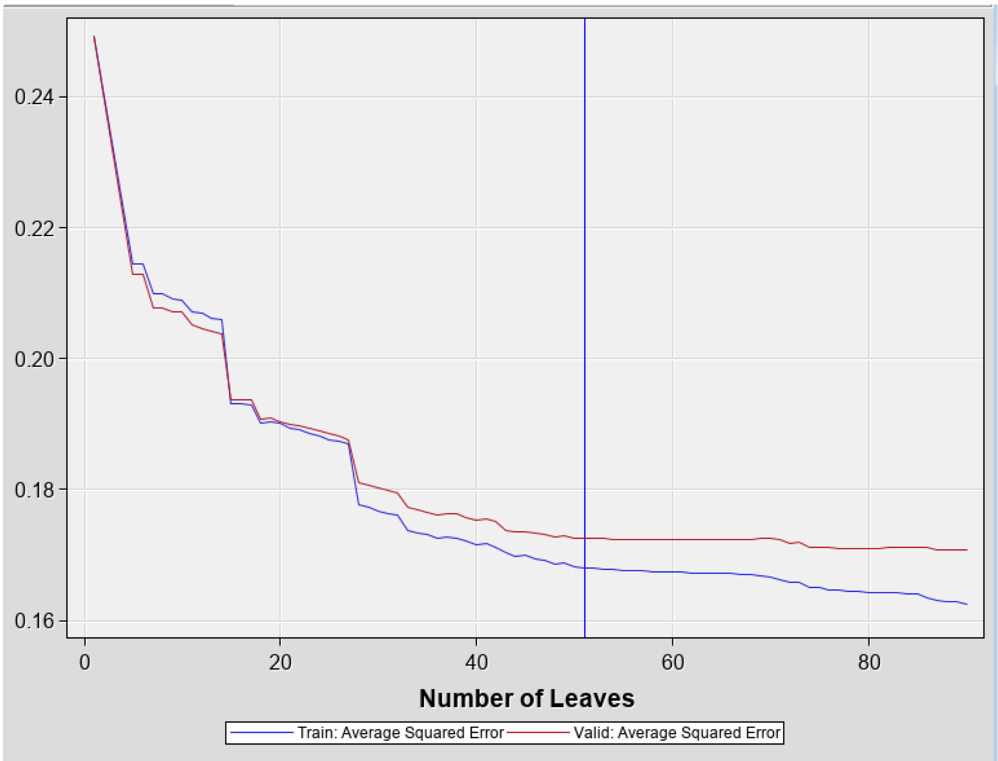


Figure 4.5 - Average Squared Error by Number of Leaves

Another interesting metric is an elbow graphic showing the Average Squared Error metric by Number of Leaves. First it shows the separation of training and validation datasets after around 50 leaves, as well the stagnation that happens on validation after this number of leaves.

5. CONCLUSIONS

This project work applied a data mining workflow with the objective to predict customer churn and train the model every quarter to apply into new customers acquisition and existing portfolio.

With an extensive dataset, a period of three months, from beginning of January 2019 to end of March 2019, was chosen to match with current quarterly financial reporting. The high volatility of the current portfolio allowed to have already a balanced dataset with around half of customers had churned during the timeframe chosen.

After literature review and based on this, CRISP-DM methodology was applied to understand business objectives, then the data understanding process to define variables, visualize input correlations and data size.

Furthermore, in data preparation step outliers were removed, missing data was imputed using different techniques for class and interval variables and variable transformation was performed to have more valuable inputs. Still during this step, data was again analyzed as it's a new dataset. Then, variables were selected by using variables worth and stepwise logistic regression. Finally, data was partitioned in a 70-30 rule for training and validation datasets.

Three different predictive modeling were performed: Logistic Regression, Decision Tree and Artificial Neural Network. As different parameters were used, there was a total of eleven models were trained and validated during this step.

Finally, model comparison was evaluated in order to establish confusion matrices and overall accuracy metrics, including ROC AUC, to define that the best model was Decision Tree with parameters of with 5 maximum branches and 6 maximum depth. The model was further investigated and analyzed variables importance.

This study can be used to drive business processes and marketing strategies to reduce customer attrition, understand customer behavior and increase the company portfolio. It tried to achieve which information is necessary to have a good prediction and from these three models, which one is the best.

6. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORKS

The overall goal of this exercise was to approach a good customer churn modeling. Based on the literature of retail energy market previous efforts, the overall accuracy was lower than expected. This can be related to the data points available for this exercise and the fifteen variables chosen as relevant for the training input. Another possibility would be to use one year timeframe instead of a quarter with the objective to have more input information.

As an academic study, the challenges for oversampling could not be performed during the data preparation step. The dataset was very balanced due to high customer churn and acquisition, giving a big portfolio rotation. Other markets studied had a general churn rate diminished when compared with this dataset.

Although a relevant kickstart, the model was not yet fully applied into the company business process. That avoided the deployment phase to be developed, linking directly SAS Enterprise Miner with company's SQL Server Data Warehouse for input and output of predicted dataset to make it as a repeatable process.

Ideally, the modeling technique should be applied in a recurrent base and adapted as the dataset is changing. Therefore, continuous evaluation of this work could be extended into several quarters with different datasets.

Finally, only three models were tested with this effort. It's possibly beneficial to try the same exercise with other type of models that had success in the literature for customer churn like Random Forests, AdaBoost and Support Vector Machines.

7. BIBLIOGRAPHY

Bishop, C. M. (2006). *Pattern recognition and machine learning*: springer.

Bonaccorso, G. (2018). *Mastering machine learning algorithms: expert techniques to implement popular machine learning algorithms and fine-tune your models*: Packt Publishing Ltd.

Carlson, J. L., & Loomis, D. (2008). An Assessment of the Impact of Deregulation on the Relative Price of Electricity in Illinois. *The Electricity Journal, 21*(6), 60-70. doi:10.1016/j.tej.2008.07.004

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. *SPSS inc, 9*, 13.

Chen, W.-M. (2019). The U.S. electricity market twenty years after restructuring: A review experience in the state of Delaware. *Utilities Policy, 57*, 24-32. doi:10.1016/j.jup.2019.02.002

Choice, E. (2020). Deregulated Energy States & Markets Retrieved from https://www.electricchoice.com/map-deregulated-energy-markets/

Cristian, R. (2016). *Churn Prediction for the Dutch Energy Market.* Vrije Universiteit Amsterdam,

De Caigny, A., Coussement, K., & De Bock, K. W. (2018). A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. *European Journal of Operational Research, 269*(2), 760-772. doi:10.1016/j.ejor.2018.02.009

Fleck, L., Tavares, M. H. F., Eyng, E., Helmann, A., & Andrade, M. (2016). Redes neurais artificiais: princípios básicos. *Revista Eletrônica Científica Inovação e Tecnologia, 1*(13), 47-57.

Ge, Z., Song, Z., Ding, S. X., & Huang, B. (2017). Data Mining and Analytics in the Process Industry: The Role of Machine Learning. *IEEE Access, 5*, 20590-20616. doi:10.1109/access.2017.2756872

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep learning* (Vol. 1): MIT press Cambridge.

Hajian, A., & Styles, P. (2018). Artificial neural networks. In *Application of soft computing and intelligent methods in geophysics* (pp. 3-69): Springer.

Hand, D. J., & Adams, N. M. (2014). Data Mining. *Wiley StatsRef: Statistics Reference Online*, 1-7.

Hejazinia, R., & Kazemi, M. (2014). Prioritizing factors influencing customer churn. *Interdisciplinary Journal of Contemporary Research in Business, 5*(12), 227-236.

Huang, Y., & Kechadi, T. (2013). An effective hybrid learning system for telecommunication churn prediction. *Expert Systems with Applications, 40*(14), 5635-5647. doi:10.1016/j.eswa.2013.04.020

Hung, S.-Y., Yen, D. C., & Wang, H.-Y. (2006). Applying data mining to telecom churn management. *Expert Systems with Applications, 31*(3), 515-524.

Iranmanesh, S. H., Hamid, M., Bastan, M., Hamed Shakouri, G., & Nasiri, M. M. (2019). *Customer churn prediction using artificial neural network: An analytical CRM application.* Paper presented at the Proceedings of the International Conference on Industrial Engineering and Operations Management.

Kristof Coussement, & Poel, Dirk V. D. (2008). Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. *Expert Systems with Applications, 34*(1), 313-327. doi:10.1016/j.eswa.2006.09.038

Lazarov, V., & Capota, M. (2007). Churn prediction. *Bus. Anal. Course. TUM Comput. Sci, 33*, 34.

M.A.H. Farquad, Vadlamani Ravi, & Bapi Raju, S. (2014). Churn prediction using comprehensible support vector machine: An analytical CRM application. *Applied Soft Computing, 19*, 31-40. doi:10.1016/j.asoc.2014.01.031

Martinez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernandez Orallo, J., Kull, M., Lachiche, N., . . . Flach, P. A. (2020). CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. *IEEE Transactions on Knowledge and Data Engineering*, 1-1. doi:10.1109/tkde.2019.2962680

Mehdiyev, N., Enke, D., Fettke, P., & Loos, P. (2016). Evaluating Forecasting Methods by Considering Different Accuracy Measures. *Procedia Computer Science, 95*, 264-271. doi:10.1016/j.procs.2016.09.332

Mehlig, B. (2019). Artificial neural networks. *arXiv preprint arXiv:1901.05639*.

Moeyersoms, J., & Martens, D. (2015). Including high-cardinality attributes in predictive models: A case study in churn prediction in the energy sector. *Decision Support Systems, 72*, 72-81. doi:10.1016/j.dss.2015.02.007

Nakajima, T., & Hamori, S. (2010). Change in consumer sensitivity to electricity prices in response to retail deregulation: A panel empirical analysis of the residential demand for electricity in the United States. *Energy Policy, 38*(5), 2470-2476. doi:10.1016/j.enpol.2009.12.041

O'Connor, P. R. (2017). Restructuring Recharged. The Superior Performance of Competitive Electricity Markets, 2008–2016. Retail Energy Supply Association. In.

Olle, G. D. O., & Cai, S. (2014). A hybrid churn prediction model in mobile telecommunication industry. *International Journal of e-Education, e-Business, e-Management and e-Learning, 4*(1), 55.

Pribil, J., & Polejova, M. (2017). A churn analysis using data mining techniques: Case of electricity distribution company. *Lecture Notes in Engineering and Computer Science, 1*, 355-360.

Sammut, C., & Webb, G. I. (2017). *Encyclopedia of machine learning and data mining*: Springer.

Shelke, M. S., Deshmukh, P. R., & Shandilya, V. K. (2017). A review on imbalanced data handling using undersampling and oversampling technique. *Int J Recent Trends in Eng & Res, 3*, 444-449.

Van Den Poel, D., & Larivière, B. (2004). Customer attrition analysis for financial services using proportional hazard models. *European Journal of Operational Research, 157*(1), 196-217. doi:10.1016/s0377-2217(03)00069-9

Yang, L.-S., & Chiu, C. (2006). *Knowledge discovery on customer churn prediction.* Paper presented at the Proceedings of the 10th WSEAS Interbational Conference on APPLIED MATHEMATICS.

Zhu, B., Baesens, B., & vanden Broucke, S. K. L. M. (2017). An empirical comparison of techniques for the class imbalance problem in churn prediction. *Information Sciences, 408*, 84-99. doi:10.1016/j.ins.2017.04.015