BIG DATA PLATFORM FOR TELECOMMUNICATION CUSTOMER CHURN

Student: Francesco Palermo
Student-id: 45539669
Session: Session 1- 2020
Subject: COMP8230 Mining Unstructured Data
Assessment: Assessment Task 1
Submission Date: 12 April 2020
University: Macquarie University

Table of Contents

Executive Summary

Customer churn rate is a metric that measures the percentage of customers who end their relationship with a company within a particular period [1].

The research paper's goal is to create and develop an automatic churn prediction system by using big data technology. This particular work is focused on the telecommunications industry; however, this study might be replicated in other similar situations, such as employee turnover, where the primary goal is to try to retain employees.

Big data platforms will be an essential tool for telecommunication companies in the coming years in order to compete in a very competitive environment. The central focus of this work is to build a churn prediction model which will support telecom businesses to predict current clients who are most likely to churn.

To achieve that, a data lake architecture will be developed to solve issues such as data variety, volume and velocity that were not compatible in a classic data warehouse. Different types of data will be ingested in the system. Customers and handset information will be extracted from tables in relational databases, while customer usage will come from smartphone sensors. Users' social network behavior will be considered as well by extracting sentiment analysis in regards to the satisfaction of the product from social networks such as Twitter and Facebook. Finally, text and audio analytics will be performed to extract useful insights from complaints originating from emails, smartphone applications and phone calls.

The machine learning algorithms used are all related to Decision Trees. Although this might sound restrictive, Decision Trees are very strong algorithms, capable of fitting very large and complex datasets. Furthermore, they are very intuitive and their decisions rules are easy to interpret. These models were used for classification in this churn predictive model.

In conclusion, the result section will explain and present theoretical results of this prototype. It is important to note that some of the graphs are coming from my personal portfolio on this topic, while other illustrations are taken from outside sources to explain possible scenarios.

Industry Context – Domain

Introduction to the industry context

The telecommunications sector consists of various businesses that deliver and offer the infrastructure which enables data transmission and communications between users from all around the world. This sector is comprised of both internet and cellphone service providers.

Investopedia defines it as [2]: "*Think of telecommunications as the world's biggest machine. Strung together by complex networks, telephones, mobile phones and internet-linked PCs, the global system touches nearly all of us. It allows us to speak, share thoughts and do business with nearly anyone, regardless of where in the world they might be. Telecom operating companies make all this happen*".

In the recent years, many discussions have been had in regards to how to increase companies' revenue. The following three main approaches seem reasonable recommendations for increasing profits:

1) Up-sell to existing customers
2) Acquire new clients
3) Expand the retention period of customers, in turn reducing the churn rate

Out of these three methods, the last item will be the focus of this research paper. In particular, this approach is exceptionally important in telecommunications as the cost of obtaining new customers is estimated to be between 5 to 25 times higher than the cost of retaining customers.

Different data islands

Owning different data islands might increase the probability of obtaining more insights from data. Hence, a summary of the important types of data required will be discussed below (see Approach Section for more details)

Data Name	Feature examples	Data type	Data speed
Customer/Handset Information	Sex, income, location, handset price, etc....	Relation database (RDMS)	Batch
Customer Usage	Minutes of voice usage, number of call	Telephone sensors	Real Time
Customer Social media behaviors	Degree of popularity,Num_ReTweet	Social Media Data	Real Time
Customer Complaints	Avg_minutes_cc, Keyword_count	Audio, Text, Application	Almost Real Time

Motivating Scenario

Research question and project goals

The research question for this paper is this: how can we add business value for telecommunications firms? In other terms, can we increase the annual profit by analysing different types of data from different data sources?

In order to answer the research question, we need to set some goals:

1) Identify the key factors of churn. Data exploration and machine learning algorithms are able to recognize which predictors played an important role in the prediction model. Hence, some data visualization is essential to find out the importance of various predictors.

2) Building machine learning classifiers to accurately predict high risk customers who will churn in immediate future. The evaluation metrics available for classification task are various and depend on specific applications. Since we are dealing with a binary classification (0=no churn, 1=churn) confusion matrix, class accuracy and AUC from ROC curve are the preferred metrics.

It is important to note that addressing those objectives in a proper manner will lead to answers to our research question. In fact, building a churn prediction classifier enables us to accurately discover the most likely churners for a proper retention campaigns.

Optus, one of the larger telecom company in Australia, recorded around 10 million customers in 2018[3]. Decreasing by only 1% the churn rate, will equal to a substantial profit rise for Optus. Knowing which aspects were determining factors for churners will also allow the businesses to re-adjust their service offerings by applying discounts or advantageous deals for this specific class of clientele.

Business value through unstructured data mining

Once the research question and project goals have been set up, a suitable architecture needs to implemented. A data lake is the perfect choice for our specific task. A Data Lake is a storage repository that contains a large amount of raw data in its native format. Data can be unstructured, semi-structured and structured.

The first step of a data lake is to ingest all the available data inside this container, after which some components will perform cleaning and transformation of the data that then becomes input for data analytics and machine learning algorithms to provide insights from the initial data. The complete design with each individual component will be presented later (*Approach* section).

The main benefits of a Data lake can be summarized in the following points:

- The emerging growth in data volume, data variety and speed of data transmission allows the quality of analysis to be more accurate

- Data lake consents to install only the required components for a particular application. For example, in our application the data quality component is required in order to find valuable insights (otherwise we might fall into the garbage in – garbage out issue), while as we do not need to have particularly fast queries, the indexing component will not be required.

- Storage capabilities like Hadoop store different data in a convenient way. It is not required to model data in advance as it is store data into an enterprise-wide schema.

Problem Statement

Identifying and explaining the problem

The major challenge in telecommunications companies is customer churn. Customer churn happens when subscribers or clients finish doing business with a particular company or service. It is usually measured as a rate and it plays an important role in profit growth as it is much cheaper to keep current customers than it is to acquire new ones. In this industry, customer churn prevents growth, hence corporations need to develop the means to predict potential customers at risk of churn.

What are the problems that the industry is facing?

As I mentioned in the previous section, customer churn is the biggest issue and one of the most significant concerns for big businesses. Some of the common issues telecom companies are facing are: **poor customer service**, **lack of product quality** and **better competitor price.**

Poor customer service is perhaps the primary cause of churn. In today's era, people expect excellent customer service and experiences and if they do not receive it they are quick to abandon their subscription. Another bad effect that comes from poor customer service is that often customers share bad experiences on social networks and this may damage the image of the company and cause difficulties in acquiring new customers.

Therefore, we have to create variables in the dataset that can be indicators of customer service performance (for example: the average time a customer waits to be assisted from a phone call) and keep track of the "sentiment" of the customer on social media.

Lack of product quality creates a lot of frustration in customers. It may be generated by many factors in telecommunications such as:

- Poor internet connection
- Call drop-outs
- Noise during a phone call
- Expansion in battery usage

Therefore, it is important for our application to get data from phone sensors as well as examine complaints through text mining (email), and audio mining (phone calls between users and technical operators).

Better competitive pricing is the result of an incredibly completive market. Although this is a good thing from the client's point of view, some telecommunications companies struggle to keep track of, and to beat other competitors' prices for similar services.

Identifying the gap between the current and the desired goal

Recently, several scientists attempted to come up with a workable solution to tackle customer churn. Previous efforts used Data Warehouse systems, which encountered many issues in terms of **data variety** and **data volume**.

Data variety refers to the type and nature of the data. Only some types of data can be given as input in a Data Warehouse environment and they need to have a fixed schema. So, data from relational databases such as customer demographic information or billing data were suitable.

Data volume indicates the quantity of generated and stored data. Data Warehouse systems struggle with the size of data and therefore can be incredibly slow.

A solution for that has been applied by sampling the entire dataset, but this would exclude many observations. Therefore, data sources that are too big in size were ruled out due to the complications in dealing with them.

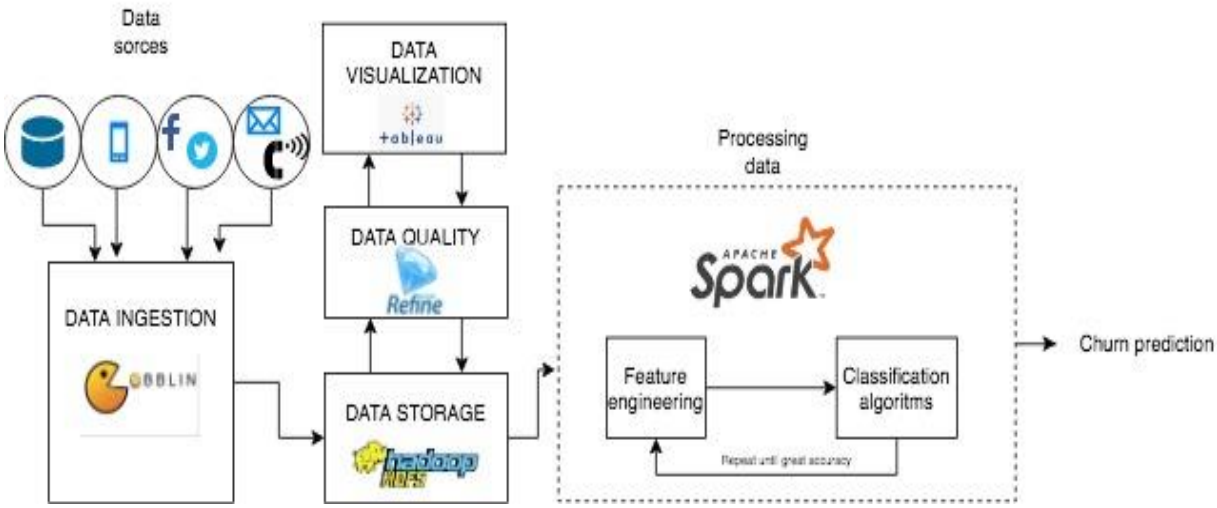
It is easy to think that machine learning algorithms applied on top of the Data Warehouse system may fail to provide accurate results using a limited amount of data. This is where I found the inspiration of building a system that can better deal with those issues. In fact, all these problematic processes with Data Warehouse's can be solved in an easier way by using big data platforms.

In conclusion, other benefits for using Data Lakes in comparison with Data Warehouses (in addition to those that have already been explained above) can be summarized in the following table:

Data Lake	**Data Warehouse**
Any row data. Users decide which parts of data need to be cleaned with urgency and insights are generated faster.	Highly curated data. This is time consuming and we can get data insights too late when the market has already changed.
Can use open source tools. Since they are low cost, smaller size companies can compete with larger companies.	Mostly commercial tools. Mainly large and wealthy telecommunications firm can buy these expensive tools.
Different users with different skills can benefit from it. Data Scientists, Data Analysts, and data visualization experts.	Only for business professionals.

Approach

The following diagram shows the proposed Data Lake structure.



Each individual piece of this architecture will be explained in this section.

Data sources refers to the different types of data we are dealing with in a telecommunications company (see industry context for a summarized version). It is possible to deal with (from left to right in the image):

- Customer information: contains information about demographic aspects such as sex, income, location or marital status. This information is usually retrieved from relational databases in batch mode (data is collected at regularly scheduled intervals).

- Handset information: contains information about handsets. Examples are price of phones, whether it is brand new or second hand or the length of months' customers are using it. This information is usually retrieved from relational databases in batch mode.

- Customer usage: contains statistics in regards to the use of the handset. Average time for a call, total number of calls per months or average data consumed per month are examples of features. Data comes from smartphone sensors and is available in real time.

- Customer social media behaviors: this information is meant to keep track of the behavior of customers on different social media platforms such as Twitter or Facebook. Number of friends, degree of popularity, number of

re-tweets or rate of product satisfaction are only some features that can be extracted. Data is available in real time. Techniques such as sentiment analysis are helpful to create some attributes (e.g. rate of satisfaction).

- Customer complaints: this information can be crucial since it is highly correlated to churn. It can be retrieved from emails, phone calls to technical operators or complaint forms that users submit through smartphone applications. Techniques such text and audio mining are useful to create some of these features.

Data ingestion is the first real big data component utilized in the architecture. It provides tools to retrieve data from different data sources and load into the Data Lake. In this phase, files are simply loaded without being treated. Data can be loaded in batches, near real time or real time.

The suggested data ingestion component is *Apache Gobblin* which is an open source data ingestion tool established by LinkedIn. Gobblin moves data into Hadoop from different sources such as databases, rest APIs, and FTP/SFTP servers. Gobblin can be used in standalone mode or in distributed mode on the cluster.

Data storage is an important component of my Data Lake architecture. It needs to provide low cost storage for my data and it needs to be easily accessible and available.

The suggested data storage component is *HDFS (Hadoop distributed file system*). Hadoop File System was created using distributed file system design. It is run on commodity hardware. It uses a very reliable and low cost hardware and it is able to handle very large amounts of data. In order to store the huge amounts of data received, HDFS spreads it into different machines (Map Reduce).

Data quality is a fundamental phase for this architecture. In fact, extracting business value from data is only possible if we perform data cleansing and transformation. The process can be long as we need to deal with missing data, noise data or outliers. However, we only need to treat and clean data we are interested in and not essentially everything that sits in a Data Lake.

The suggested data quality component is *Open Refine*. Samuel Greengard [4] describes it as "OpenRefine, formerly known as Google Refine, is a free open source tool for managing, manipulating and cleansing data, including big data.

The application can accommodate up to a few hundred thousand rows of data. It cleans, reformats and transforms diverse and disparate data"

Data visualization is a phase where telecommunication managers can discover some insights simply by looking at graphs. Here, charts are created on a cleaned version of the data. Clustered box plots/bar charts, where we draw different predictors against the target variable, yield interesting insights. In fact, we can observe which particular predictor category has an effect on churning.

The suggested Data visualization tool is *Tableau*. It is a data visualization software company that is predominantly used due to the capability to create attractive graphs.

Next, the right-hand side dotted-line rectangle is described (see graph on page 11). *Apache Spark* has been used for processing data which includes feature engineering and training and testing models.

Feature engineering techniques were used to extract and chose relevant attributes (features) for model training and prediction purposes. The data needs to be converted from its raw status into features that are understandable for machine learning algorithms.

If this process is done in a proper way, it leads to a significant increase in the predictive power of ML algorithms. The outcome of this phase is to produce a combined wide dataset, where each row in it represents a customer's feature vector. Finally, this object can be fed into classification algorithms.

Classification algorithms is the final link of our chain. Classification is a typical supervised learning task. In supervised learning, the training data you feed to the algorithm includes the desired solutions, called labels[5]. Since the target variable (*Churn*) is a categorical variable with two classes (0: customer do not churn, 1: customer churn) what we need to build is a binary classifier.

The simple idea here is to split the dataset into *training* and *testing* data (usually 80% is reserved to *training* while the rest is reserved for *testing*). In the *training* set we are going to build several ML algorithms and the performance will be evaluated on the test set (which are the set of records the model has not seen before).

Various performance metrics are available for classification tasks such as total accuracy, F1 score or the area under the ROC curve. For our scenario, I would suggest two metrics: **Class accuracy** and AUC.

Special attention needs to be paid for the **Class accuracy** (note that this is different from total accuracy) which specifically measures the percentage of churners that the model correctly classifies. In fact, we are interested in those customers who are more likely to churn rather than those who do not churn (total accuracy contains a percentage of non-churners that the model correctly classifies).

At this stage, one very common issue to consider is **Data imbalance.** This situation involves cases where the class label rate in the dataset is not balanced. Our dataset will encounter this problem since there will be significantly less churners than non-churners. If this is the case, for example 10% churner vs 90% non-churner, ML algorithms do not provide reliable results. Therefore, some methods need to applied in order to resolve this issue. Famous methods are *Up Sampling*, *Down Sampling* and *Weighted Instance*.

The last decision to take is which ML algorithms are more effective for our task. Do we need conventional or deep learning ML algorithms? Conventional machine learning is the best option. In particular, we need to focus our attention on analyzing and evaluating the performance of various tree-based machine learning systems and algorithms for predicting churn in telecommunications businesses.

"Decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules" [6].

The reason why Decision trees are preferred is their interpretability. Our main goal is understanding what causes a customer to churn and decision trees are considered white-box methods. On the other hand, deep learning methods or methods such as SVM or Random Forest are normally considered black box models. Although they can perform well, it is difficult to see what happens behind the scenes.

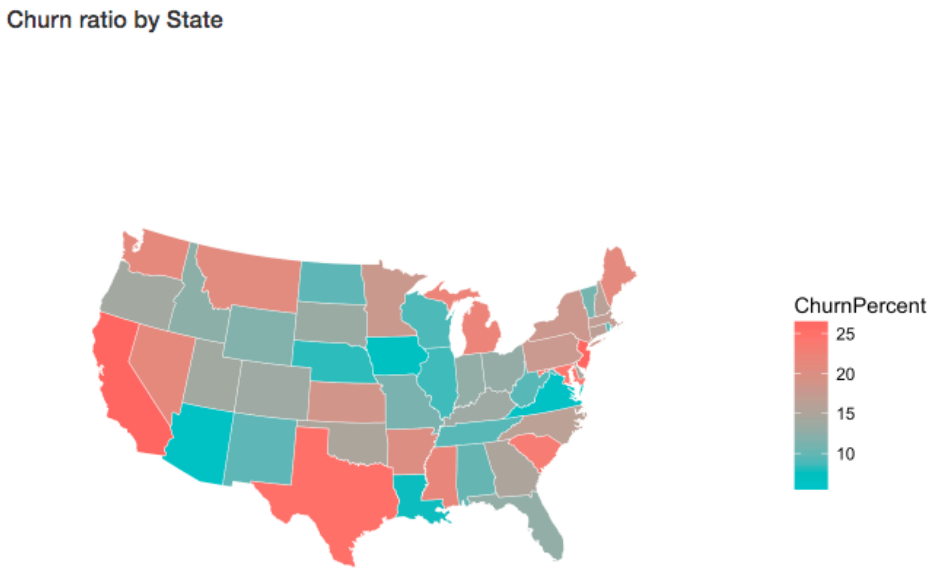
Results

The above approach is a prototype of a big data platform that would help telecommunications companies dealing with the problem of customer churn. As the real results are not actually obtainable at this time of my research, let us comment some of the hypothetical results together with the CEO of the company we are working for.

Explanatory data analysis

The following graphs are obtained from the data visualization department and they are part of what is called EDA (explanatory data analysis).

The first chart shows an heatmap of the United States map [7].



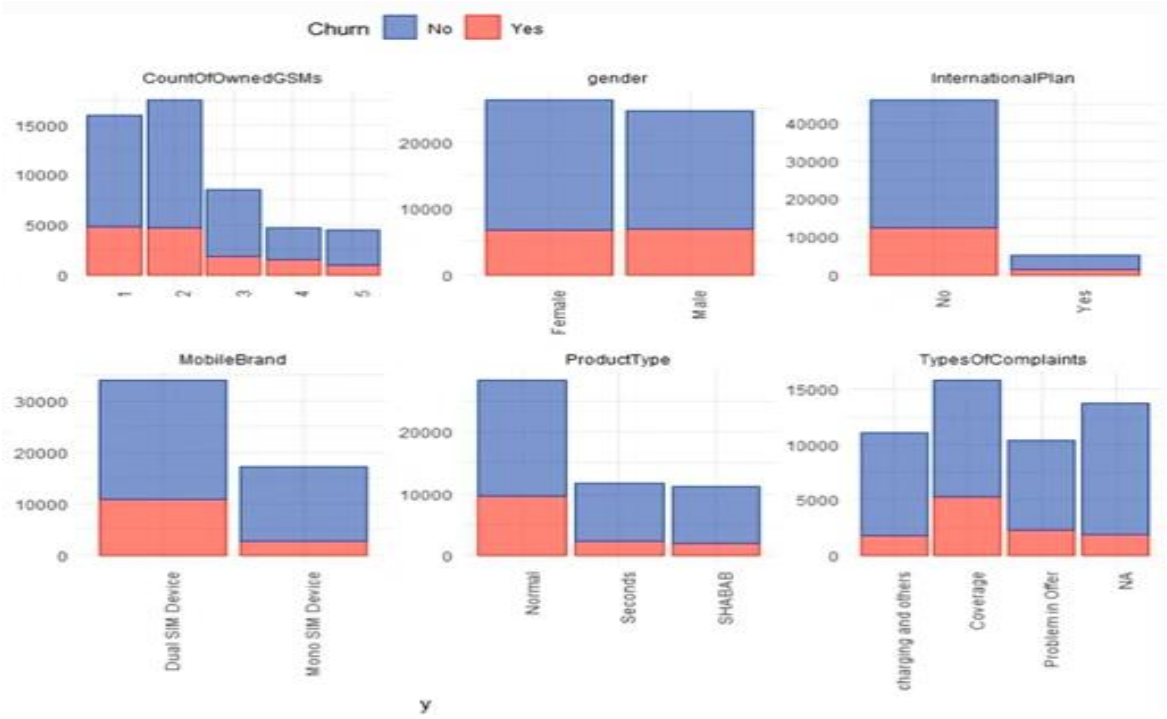
According to the above chart, it is clear that the churn rate for our telecommunications company varies significantly across the states of America.

In particular, states such as Texas, California and New Jersey (the red territories in the map) record a high percentage of churners, while Arizona (blue colour) registers a churn rate below 10%.

This is a great illustration which would enable telecommunications firms to allocate more resources to the states who are showing higher rate of churn.

Next, some of the feature distributions will be displayed along with the target variable, in order to capture whether a particular feature's category will be a determining factor for churn or not.

Only categorical features are taken in consideration, but similar results can be illustrated for continuous variables.

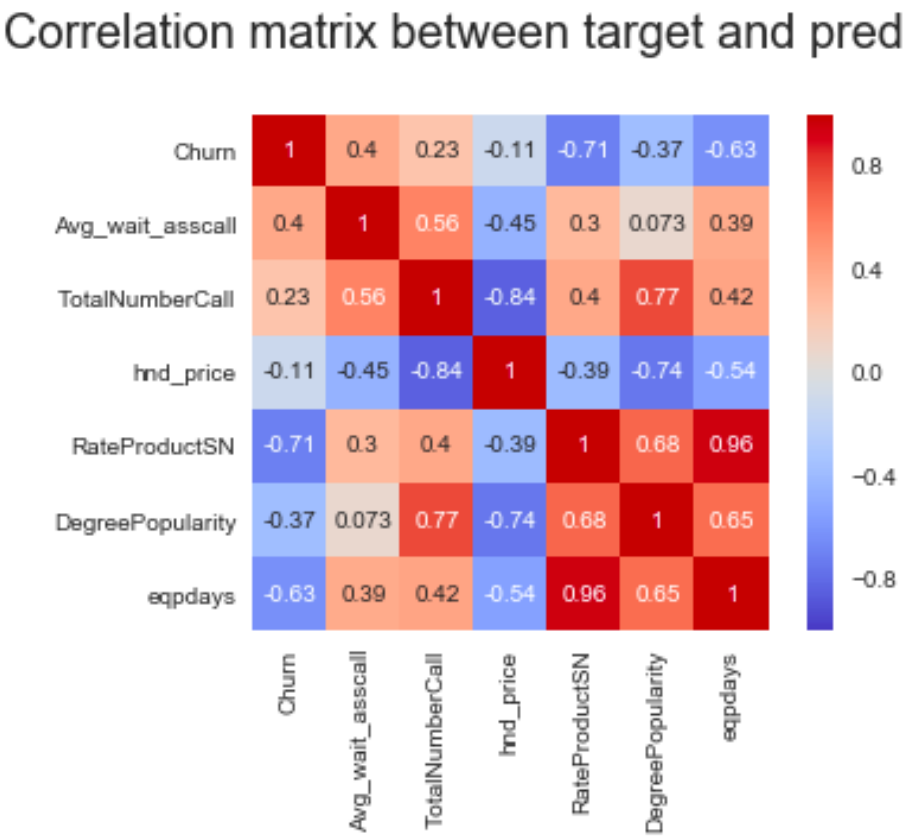


The above clustered bar chart [8] shows some categorical features together with the target variable.

Before commenting on that, it is important to note how these features have been extracted from different data sources. In fact, *Product type* and *MobileBrand* are handset information taken from databases' tables, while *TypesOfComplaints* comes from different data sources such as phone calls, emails or smartphone applications.

Although *gender* distribution does not seem a particularly strong predictor for churn (the number of female customers is quite similar to the number of male customers), *MobileBrand* and *TypesOfComplaints* shows that different categories influence the churn phenomenon. For example, coverage seems to be the *TypesOfComplaints* that leads a customer to churn (the red bar is definitely higher than the other categories).

Another great data representation is the heatmap of the correlation matrix of some of the most significant features together with the predictor variable.



The correlation matrix shows the correlation coefficient between a set of variables. By looking at the first line we can make the following observations:

- There is a moderately positive linear relation (0.4) between *churn* and *Avg_wait_asscall* (average wait time a customer needs to wait for technical support). This means that the longer a customer needs to wait for assistance, the higher is the probability that he/she churns.

- There is a strong negative linear relation (-0.71) between *churn* and *RateProductSN* (rate between 0 and 1 related to how a customer talks about the product on social network platforms). This makes sense because the closer this rate is to 0, which means a customer shares negative thoughts about the product, the higher is the likelihood that he/she churns.

These observations give precious insights to the telecom industry management by helping them to prioritise which aspects of the work-flow to be improved.

Model Evaluation

Next, the data-processing department is showing us the results of the model predictions.

The following diagram shows the result of three tree-based machine learning algorithms and a decision to which model is the best need to be addressed.
The classification metric in used is class specific accuracy (in percentage).

Model	Training dataset	Evaluation dataset	Test dataset
ID3	71.0	71.2	71.1
CART	75.6	75.2	68.5
C5.0	72.3	72.1	72.2

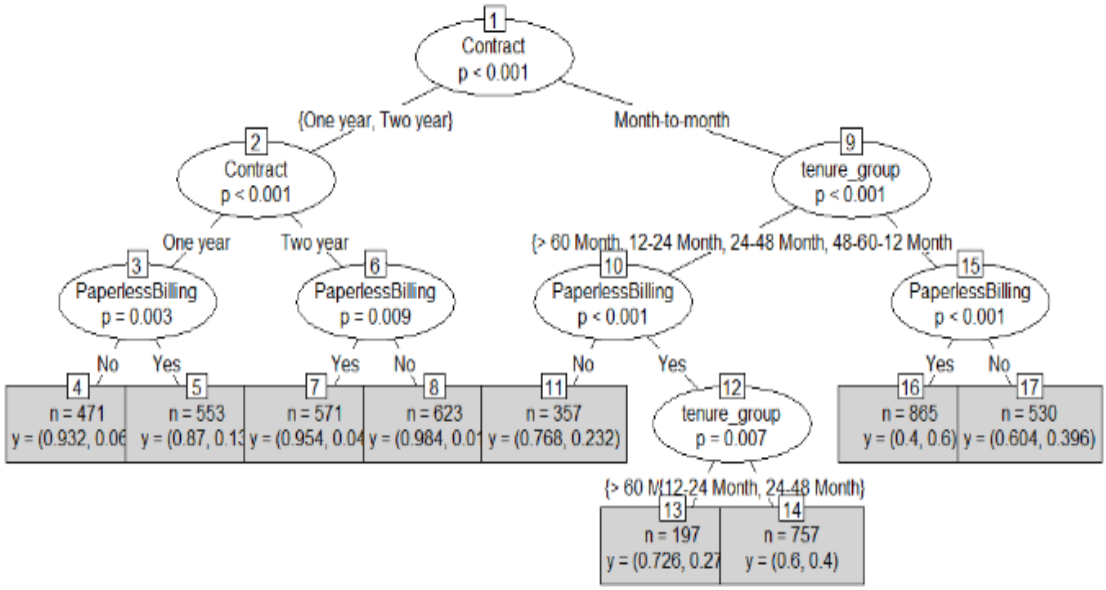
First at all, CART (*classification and regression tree*) clearly overfits.
In fact, while it shows the highest accuracy in training and evaluation datasets, the performance of this model in the test set (never seen data) drops dramatically. This a classic situation of overfitting, which means that this model in not able to generalize well.

The remaining two models, ID3 (*Iterative Dichotomiser 3*) and C5.0 are both adequate as the class specific accuracy over the datasets is pretty much equal. However, C5.0 is preferred because it shows a higher class specific accuracy.

A C5.0 decision tree splits the records based on the predictor that shows the maximum information gain. Each subgroup is then split again, and the process repeats until some stopping criteria is met. Each leaf shows a prediction for a particular subset of the data, and each case belongs to exactly one leaf node.

Once we picked the best model, a tree structured can be produce in order to try to understand the behavior of the model in relation to their predictors. Furthermore, decision rules and predictor importance can be analyzed in details.

The decision tree visualization is displayed below [9].



The most fascinating quality of the decision tree is that decision rules in the form of "*if-then-else"* can be extracted. They are easy to interpret and give great insights. For example:

- Rule n°7: If the customer owns a two-year contract deal and receive the telecom bill via email, there is almost 94% probability that he/she will not churn.

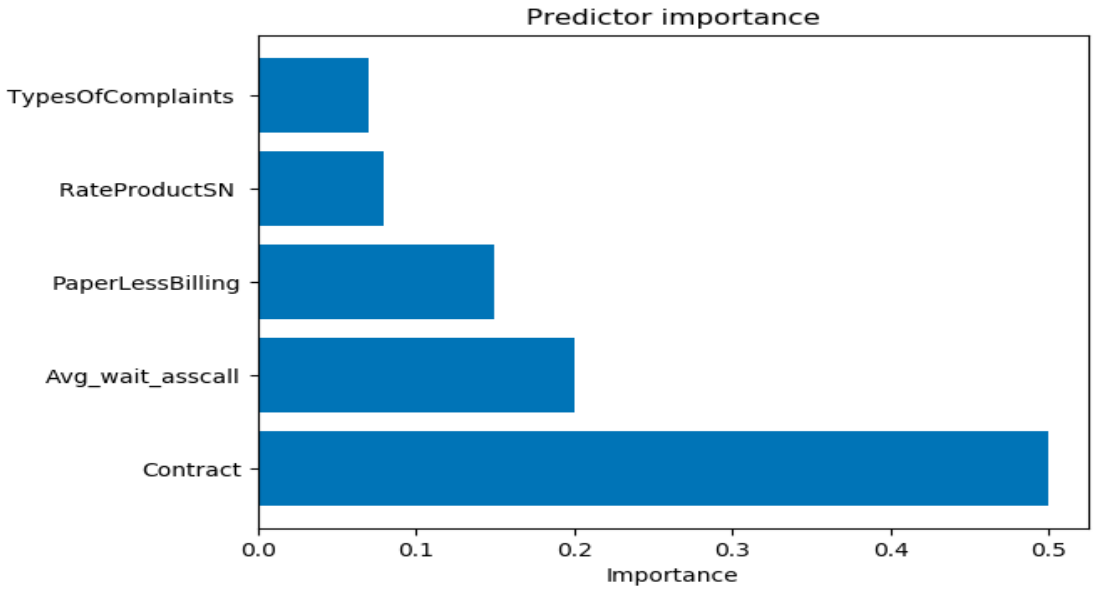
  IF (*Contract* == '2') AND (*PaperBillinig* == True) THEN:
  Churn = 0;

- Rule n°16: If a customer is in a month to month contract, the length of the submission is within 12 months and receive the telecom bill via email, there is 60% chance that he/she will churn.

  IF (*Contract* == 'm/m') AND (*tenure group* < 12) THEN:
  IF (*PaperBillinig* == True) THEN:
  Churn = 1;

The last decision rule clearly defines the group of people that are more vulnerable to churn.

Finally, the predictor importance bar chart assists us by showing the relative importance of the top predictors for a particular model estimation. The sum of the frequencies needs to be one. It is important to note that it does not say anything about the model accuracy, but rather indicates the importance of each predictor in making a prediction.



Features such as *TypesOfComplaints, Contract, PaperlessBilling and RateProductSN* appear to play an important role in customer churn.

Conclusion

The importance of this research paper is to demonstrate the power of using big data technology to assist with decision making in the telecommunications industry.

Understanding the important factors of churn can offer an opportunity for the business to make the right investment choices, by adjusting their offerings and developing appropriate target churn prevention packages to keep its clients and optimize its financial performance.

My final recommendation is to keep monitoring and deploying the system so it always achieves the highest accuracy. In fact, monitoring and deploying are the fundamental elements in a modeling lifecycle. We have to regularly monitor our model in terms of this accuracy strength in order to have sustainable models into the future.

References

1) Gallo, A 2014, 'The Value of Keeping the Right Customers', *Harvard Business Rev*iew, 29 October, viewed on 28 April 2019, https://hbr.org/2014/10/the-value-of-keeping-the-right-customers

2) Brian Beers, Investopedia, 'What Is the Telecommunications Sector?' https://www.investopedia.com/ask/answers/070815/what-telecommunications-sector.asp3)

3) Corinne Recheirt, May 2017, 'Optus doubles NBN customers as mobile subs hit 10 million' https://www.zdnet.com/article/optus-doubles-nbn-customers-as-mobile-subs-hit-10-million/

4) Samuel Greengard, June 2019, 'Learn which data quality management tool is best for your business – and leverage these tools' ability to analyze, manage and scrub data from numerous sources'. https://www.datamation.com/big-data/10-top-data-quality-tools.html

5) Aurélien Géron (2019): "Hands-On Machine Learning with Scikit-Learn, Keras and TensorFlow" (2nd edition; September 2019)

6) Decision Tree, from Wikipedia https://en.wikipedia.org/wiki/Decision_tree#Overview

7) Telecom Churn Analysis, Rstudio http://rstudio-pubs-static.s3.amazonaws.com/277278_427ca6a7ce7c4eb688506efc7a6c2435.html

8) Ahmad, A.K., Jafar, A. & Aljoumaa, K. Customer churn prediction in telecom using machine learning in big data platform. *J Big Data* **6,** 28 (2019). https://doi.org/10.1186/s40537-019-0191-6

9) Susan Li, (November 2017): "Predict Customer Churn with R" https://towardsdatascience.com/predict-customer-churn-with-r-9e62357d47b4