Student ID Number: 19075153

Assignment 1
Part A (Literature Review)

Semester 2 2021

Student Name: Bernard O'Leary
Student ID: 19075153

PAPER NAME: Data Mining and Machine Learning

PAPER CODE: COMP809

Due Date: Friday 27 Aug 2021 (midnight)

TOTAL MARKS: 100

INSTRUCTIONS:

1. **The following actions may be deemed to constitute a breach of the General Academic Regulations Part 7: Academic Discipline,**
   - Communicating with or collaborating with another person regarding the Assignment
   - Copying from any other student work for your Assignment
   - Copying from any third-party websites unless it is an open book Assignment
   - Uses any other unfair means
2. **Please email DCT.EXAM@AUT.AC.NZ if you have any technical issues with your Assessment/Assignment/Test submission on Blackboard immediately**
3. **Attach your code for all the datasets in the appendix section**.

Review of Machine Learning Based Churn Prediction in the Telecommunications Industry

Bernard O'Leary
*COMP809_2021_02 Data Mining
and Machine Learning Sem 2, 2021*

***Abstract*—Growth of people and devices connecting with each other online over the internet drives the need for more and better telecommunications services. The commoditization of the telecommunications industry is to the extent that "internet" can be seen increasingly as a utility, more similar to water or electricity, rather than a commercial service or luxury, indeed many nations treat internet and telecommunications access as a human right. As government regulation around internet access increases, so too does the ease of migration between internet and telecommunications service providers. With increased demand for access, increasing options for customers among myriad service providers, competition between telecommunications companies to retain and win customers is continually increasing. Data collected on customer behaviour increasingly enables telecommunications companies to analyse and predict future behaviour, including the propensity of a customer to "churn". Churn refers to the moving of telecommunications services consumers (customers) from one service provider to another. Churn can be driven by several causes and combinations of reasons can trigger a customer to churn. Becaeuse of the complexity of the problem and the volumes of data and customers involved, machine learning techniques lend themselves well to the problem of prediction of customer churn. This paper reviews several approaches that have been suggested by the research community for applying machine learning techniques to the problem of prediction of customer churn. Two main themes will be addressed – data pre-processing approaches, and machine learning model selection techniques. Six peer-reviewed journal articles are reviewed. (*Abstract*)**

I. Introduction

With the introduction of the Internet of Things and unprecedented events such as the Covid-19 pandemic that has driven significant increase in adoptions of online and digital/cloud-based services, the competition for customers by telecommunications service providers continues to increase and these companies need to find better and more innovative ways to not only win new customers, but to retain customers.

The estimated average churn rate for a mobile telecommunications company is 2.2% per month [1], under these conditions telecommunications companies that fail maintain churn prevention technologies that are at least as effective as the marketing and analytics technology being used to draw customers to the competition, market share can be lost rapidly, resulting in revenue loss and poor business performance.

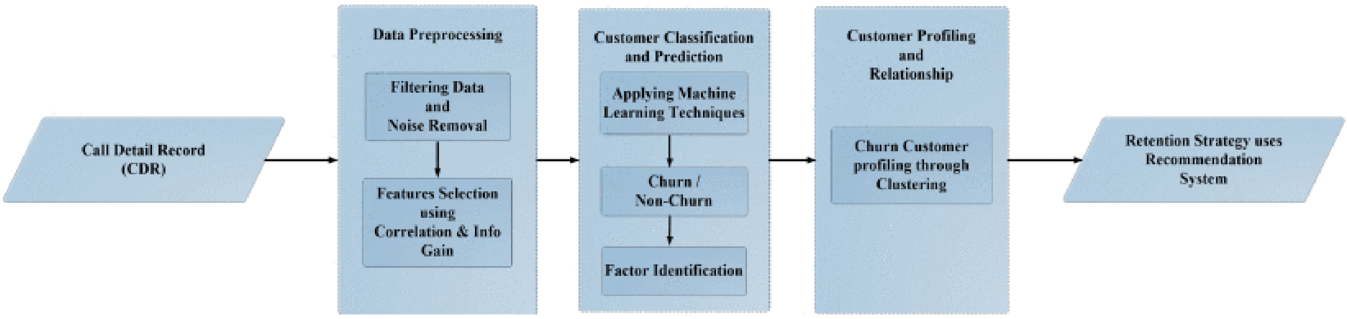
Telecommunications companies collect tremendous amounts of data about customers, including how customers use the network customer billing and payment behaviours, usage preferences, etc. For example, in July of 2015, China Telecom 10.5 trillion data records of user-domain data, requiring hundreds of terabytes (TB) of storage per day [4]. Datasets of this volume and richness make telecommunications companies an excellent case-study for the effectiveness of machine learning techniques being applied for the purpose of churn prevention.

Six papers are reviewed in this study that cover a range of different methods for applying machine learning techniques to telecommunications customer churn. Each proposes an approach or a set of approaches that fits somewhere into the typical workflow that a telecommunications organisation will use when applying model for machine learning based churn prediction. The workflow is defined in Figure 1; this illustration is taken from [1]. Although the data source may differ, in that data may not always be call detail records (CDR), which is data generally associated with a telecommunications company's Business Support Systems (BSS) [4] these data are very typical and representative of a telecommunications usage dataset and are used by several of the studies covered, however often customer data such as billing and customer preferences are also used.

II. Background and Motivation

As is mentioned in the introduction, management of the problem of customer churn is of great and increasing importance to telecommunications companies globally. The overarching theme of this report is an examination of the literature that exists covering how machine learning and data mining are used to counter-act the customer churn process by attempting to predict when a customer is likely to churn. Anecdotal evidence suggests that although telecommunications customer churn is an ideal problem for a machine learning and data mining-based approach, many telecommunications companies have been slow to adopt this technology and it is only now starting to be used as the one of the primary means of predicting and preventing churn within many telecommunications organisations.

Motivation for this report stems from this anecdotal understanding and from a personal interest in seeing this powerful technology used to enable telecommunications and similar subscriber-based companies to keep their customers happier. Despite the fact the process of migrating between telecommunications companies has become easier and more user-friendly, it is still not a process that a consumer wants to spend time on if they don't have to. Changing suppliers is a time-consuming process that can result in telecommunication service outages for the consumer, which in an increasingly connected world can be extremely inconvenient. Most consumers are happy to stay with their current service provider and will only churn if they are sufficiently dissatisfied with the experience they have with their current service provider – i.e. if they are essentially driven to churn because they feel they are being charged too much, are not getting good value for their money, have had a poor customer service experience, poor network performance, or a combination of such factors.



Call Detail Record (CDR)
Data Preprocessing
Filtering Data and Noise Removal
Features Selection using Correlation & Info Gain
Customer Classification and Prediction
Applying Machine Learning Techniques
Churn / Non-Churn
Factor Identification
Customer Profiling and Relationship
Churn Customer profiling through Clustering
Retention Strategy uses Recommendation System

FIGURE 1. Proposed model for customer churn prediction. [4]

The themes focussed on in this review are the treatments applied in steps 2 and 3 of the workflow outlined in Figure 1: Data Pre-processing, and the Model Selection aspect of the Customer Classification and Prediction section. The areas are focussed on as they relate to the concern outlined above; that which emphases that if this data mining and machine learning technology are used well – that is, the right data are provided to the modelling process and the right modelling tools are used to develop the predictive model for propensity to churn – consumers will get a better deal and telecommunications as-an industry stands to improve beyond it's current state.

A. *Data Pre-Processing*

Data cleansing is an aspect of the pre-processing block that appears Figure 1, specifically the "Filtering Data and Noise Removal" sub-block. Data cleansing covers several aspects of the machine learning process for real-world telecommunications data. Often the dataset that is intended for use is messy and requires cleansing before it is ready for consumption by a machine learning system. In this situation we are looking for features that contain missing values or incorrect values such as "Null" or attributes that imbalance the dataset [5].

Another aspect of data pre-processing step is feature selection, which is also important for telecommunications data, which is rich in features, many of which are not particularly high-quality. Telecommunications data often consists of very large datasets extracted from Business Support Systems (BSS) and Operational Support Systems (OSS). BSS data are mostly customer behaviour features and include features such as call detail records (CDR), minutes of international calls, and demographic information such as name, date of birth, etc. OSS is mostly data relating to the telecommunications network including packet switch and circuit switch data, and management and configuration data [4]. With the range of features that can be collected on customers and the network, feature selection and feature engineering are an important aspect of pre-processing. The refinement process is important for optimisation of the dataset that is intended to be fed into the machine learning model, to improve the performance of (e.g.) neural networks and decision trees [3].

B. *Model Selection and Optimisation*

Once data pre-processing has been performed, the work of customer classification and prediction can be performed. Much of the work covered in the journal articles reviewed is concerned with selection of an optimal model and comparison of different model's performance. Many of the journal articles reviewed suggest approaches to optimising the model selection process.

Some of the article propose more novel model selection approaches such as the use of so-called "boosting" techniques for taking a relatively weakly performing model and increasing it's performance until it provides performance equivalent to that of a more strongly performing model. Several article propose the combination of models – a "hybrid" approach that combines different combinations of models and seeks to optimise performance this way.

III. Related Work and Literature Review

Hun, Yen and Wang [1] conduct an analysis based on the Taiwan wireless telecommunication market, which opened and issued six operator licenses in 1997. The journal article was published in 2006 – nine years afterwards – a reasonable period of time for the market to exhibit patterns in churn behaviour of customers. The authors use a dataset obtained from a wireless telecommunications company in Taiwan. The data include 160,000 subscriber information records including 140,000 churned customers. The data are from between June 2001 to June 2002. The data are mainly made up of BSS information with some information that could be considered OSS, such as in-net call duration where the percentage of time a customer spends on calls that run within the network (do not traverse telecommunications company networks) is measured. The authors apply Neural Network (NN) and Decision Tree (DT) models to the data and report only that they believe that use of the data mining techniques can assist with accurate churn prediction. This is not unexpected as the paper is relatively old (2006) and perhaps the computing power available at that time was not sufficient to conclusively demonstrate that a machine learning approach is superior to human-based for churn prediction.

Huang and Kechandi [2] describe in their paper a hybrid learning approach which applies different combinations of unsupervised learning and supervised learning to try to optimise the results of the churn prediction model. Their work is focused primarily on this "hybrid" aspect of their approach, and they do not explore what possibilities exist for optimisation of results by data pre-processing.

The algorithms used are k-means, FOIL and Learn-A-Rule. FOIL – First Order Inductive Learning – is a rule-based approach that has the advantage of being easily understood and interpreted. No significant detail is supplied on the Learn-A-Rule algorithm other than a high-level overview, which suggests it is used alongside the FOIL algorithm. The authors state a five-fold cross-validation approach is used to assess the predictive algorithm. In the second part of the experiment the authors introduce Logistic Regression (LR) and the Decision Tree (DT), k-Nearest-Neighbour (KNN) and Support Vector Machine (SVM) algorithms and finally another rules-based algorithm called OneR. The authors then proceed to run a sequence of experiments where these algorithms are applied in combination with one-another, and the results are analysed and reported. ROC and AUC are used to measure the experimental results across 22 well-known datasets. The authors make the observation that greater attention to data pre-processing could yield better performance across their results and also state that the hybrid approach attains maximum possible accuracy across the well-known datasets they are applied against.

Keremati, Jafari-Marandi, Aliannejadi, Ahmadian and Mozzaffari [3] offer an approach that looks at how several existing data mining techniques can be used to improve churn prediction for telecommunications companies. The algorithms that they review include Decision Tree Classifier (DT), Artificial Neural Network (ANN), K-Nearest Neighbours (KNN) and the Support Vector Machine (SVM) approach. The authors outline to mechanism of each approach and then proceed to state their experimental results based on data dataset from a telecommunications company's call-centre's database collected over a 12 month period. They describe the dataset in detail, which is clearly a BSS [4] dataset. Notably, the researchers employ the WEKA (Waikato Environment for Knowledge Analysis) library for several of their data mining experiments. A DT-based methodology is used for feature selection. The experimental results the authors attained suggest ANN performs better than the other three models, but that for optimal results a hybrid approach be applied, where all four models are employed, and model selection is applied dynamically based on the balance and combination of Precision and Recall measures obtained for a dataset.

Bi, Cai, Liu and Li [4] focus on the utilisation of "Big Data" tools for the inspection and analysis of telecommunications data for the purpose of mitigating the risk of customer churn. Their approach could be considered relatively novel; they place good emphasis on the problem of data quality and the problem of the size of "telco big data", giving the example of - and stating some statistics for - China Telecom. They make mention of a Big Data technology framework called Hadoop MapReduce which they have used for their research. Their approach is to use Axiomatic Fuzzy Sets (AFS) and Subtractive Clustering Method (SCM) to propose a new clustering method, which they call Sematic-Driven SCM – SDSCM. Their paper follows the process of implementing the SDSCM approach. The final step in the pipeline is to use k-means to calculate the clusters, with the centroids being obtained by SDSCM. The paper outlines results attained with the well-known Iris and Wine benchmarking datasets along with a case study on a China Telecom dataset. The research concludes by suggesting the application of the SDSCM approach increases the effectiveness of the SCM and k-means algorithms when applied to telecommunications data and the approach also decreases the risk of imprecise data management.

Ullah, Raza, Malik, Imran, Ul Islam, and Kim [5] present a study that employs the Random Forest algorithm, which is an ensemble method that combines many algorithms, to run a series of experiments against a telecommunications dataset called "churn-bigml" which is a publicly available dataset that contains 3333 records across 16 features, with 14.5% of records being of churned customers. Another "real" dataset is also employed, from a Southeast Asian mobile network provider, which has approximately 64K records across 29 features – the data in this dataset are Call Detail Records (CDR). The authors also use the WEKA (Waikato Environment for Knowledge Analysis) library for their data mining experiments. The authors provide some interesting statistics on the cost of churn to telecommunications companies, including that at an industry-wide churn-rate of approximately 2% the annual cost of churn is approximately $100 billion, this paper was published in 2019 so the figure is presumably still relatively accurate. It is unclear if this is a global or national or regional figure, or the national currency being used (although it is dollar so most likely the USA dollar). Also, the authors state that preventing churn is approximately 16 times less expensive than attracting a new customer. The authors conclude by stating that they have provided some guidelines for decision-makers in telecommunications companies and that they have demonstrated effective results from the Random Forest and J48 algorithms.

Lu, Lin and Lu [6] make similar observations on the cost and rate of customer churn I the telecommunications industry, although do not provide information as right as Ullah et. al. [5]. The premise of the study is that a Boosting technique can be applied successfully to a telecommunications customer churn machine learning model, in order to improve it's effectiveness. Boosting is applied by repeatedly training a relatively weak predictive model until it yields a comparatively strong predictive capability. In this study Logistic Regression is used as the predictive model. The authors use a data set from a mobile telecommunications company that was runs from December 2009 to May 2010 and the data are divided up using an approach that allows the authors to clearly identify churning customers. There are missions of records in the dataset and 70 feature variables, which are reduced to 21 variables after pre-processing. ROC and AUC metrics are used to assess the performance of the approach. The authors report a key differentiator of their result as being the ability to identify in a dataset a "high risk customer group", allowing telecommunications companies to focus their efforts on the group of customers that are most likely to churn.

IV. Opinion

There are clearly myriad approaches that can be applied to machine learning for churn prediction. One of the most significant challenges to getting the right solution in place is data preparation and pre-processing. The authors of the journal articles reviewed uses a range of different data sources to run their experiments, ranging from "pre-canned" data to actual telecommunications company's data. A range of issues with data pre-processing and cleansing were encountered and discussed. Much of the work reviewed however does not place an emphasis on the pre-processing aspect of the machine learning pipeline – for example Huang and Kechadi [2]. This is done consciously and is stated by the authors, however that this is a very important aspect of the commercial data mining process that should not be overlooked. The quality of data and the way data are treated when having machine learning techniques applied and is especially important; if the data supplied is not sufficiently curated, the result can be inaccurate predictions results. As the saying goes, garbage in, garbage out.

The Boosting approach as is detailed by Lu, Lin and Lu [6] is a novel way to use less modern but more "observable" machine learning techniques and increase the performance to the extent that they are within that of more modern techniques such as deep learning. Boosting might be useful for example if there is a requirement for an "explainable" algorithm, whereby any bias in the algorithm can be explained – something which is not easy using modern deep-learning techniques but is possible with (e.g.) Logistic Regression. This is important as there is increasingly a need for machine learning models to be able to be inspected so that bias inadvertently trained into the model can be identified. This would be useful for example in the case where a telecommunications company need to ensure that it is not inadvertently and unfairly offering customers in more affluent neighbourhoods incentive to stay with the company, but not extending the same offers to customers whose address is associated with lower socio-economic status.

Crowded, mature industries such as telecommunications, are often slow to adopt disruptive technologies as the pressure to maintain one step ahead of the competition is extremely high. Churn prevention for telecommunications is perhaps one of the more obvious applications of commercial machine learning, because there is so much data available and because of the technological nature of the industry. Despite this, commercial applications of machine learning for churn prediction is only just starting to be adopted in mainstream telecommunications companies. Telecommunications companies should look to increasingly adopt machine learning as a means to automate aspects of their business, such as churn prevention.

V. Conclusion and Future Issues

As the field of machine learning advances, new approaches will emerge for application to the problem of churn prediction for telecommunication companies. The ability to retain customers and stay one step ahead of the competition will always be a key concern for the telecommunications industry. Because of this, the industry should be more proactive in adopting machine learning technology.

A useful study that could be conducted would be to survey the level of adoption of machine learning and data mining being used by telecommunications companies for the purpose of managing and preventing churn. It is likely there would be a delta found in most telecommunications companies between their current level of use of this technology and the extent to which they could be using it with relatively little further investment (and relatively high return-on-investment). Assuming this delta exists, it would also be useful to establish the rate of change of the delta – i.e. how quickly telecommunications companies are "closing the gap" between the current level of utilisation of machine learning and data mining technology to manage churn prevention versus the extent to which they could be using it.

There would also be merit in conducting further investigation into how telecommunications companies could increase the quality of the data that is collected on customers. The rationale here would be to establish how companies can collect and store data in a way that optimises it for consumption by a machine learning pipeline. This study would be on the assumption that utilisation of machine learning and data mining for the purpose of reducing telecommunications customer churn is expected to increase over time.

References

[1] Shin-Yuan Hung, David C. Yen, Hsiu-Yu Wang, "Applying data mining to telecom churn management," Expert Systems with Applications, Volume 31, Issue 3, 2006, Pages 515-524,

[2] Ying Huang, Tahar Kechadi, "An effective hybrid learning system for telecommunication churn prediction," Expert Systems with Applications, Volume 40, Issue 14, 2013, Pages 5635-5647

[3] A. Keramati, R. Jafari-Marandi, M. Aliannejadi, I. Ahmadian, M. Mozaffari, U. Abbasi, "Improved churn prediction in telecommunication industry using data mining techniques," Applied Soft Computing, Volume 24, 2014, Pages 994-1012

[4] W. Bi, M. Cai, M. Liu and G. Li, "A Big Data Clustering Algorithm for Mitigating the Risk of Customer Churn," in IEEE Transactions on Industrial Informatics, vol. 12, no. 3, pp. 1270-1281, June 2016

[5] I. Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam and S. W. Kim, "A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector," in IEEE Access, vol. 7, pp. 60134-60149, 2019

[6] N. Lu, H. Lin, J. Lu and G. Zhang, "A Customer Churn Prediction Model in Telecom Industry Using Boosting," in IEEE Transactions on Industrial Informatics, vol. 10, no. 2, pp. 1659-1665, May 2014