University of Tartu

Faculty of Science and Technology

Institute of Computer Science

Martin Käärik

A Network-Based Model for Television Services Churn Prediction

Master's Thesis (20 ECTS)

ITM

Supervisors:

Supervisor: Shakshi Sharma

Co-supervisor: Rajesh Sharma

Tartu 2021

Abstract:

A Network-Based Model for Television Services Churn Prediction

Predicting churn helps us understand which customers are likely to replace the company's services with competitors. As the cost of acquiring users is much higher than retaining existing ones, churn prediction has emerged for numerous telecommunication companies as a critical tool to retain an existing customer base.

Usually, churn is predicted by modeling individual customers' behaviour and relatively static features such as demographic data, contractual data, and product information. Recent work has shown that analysing customers' social network improves the accuracy of churn prediction.

Although the network analysis is widely researched for telecommunication customers, little to no research was found for TV service users. This thesis attempts to fill this gap by analysing customers behaviour prior to churning as well as their call logs. Models with and without the network analysis features were trained with XGBoost, Adaboost, Random forest, Logistic regression, and Gradient Boost Classifier. Differences in the prediction results, whether the additional features were added, were presented in this paper. Results indicate that adding information from call logs improves the minority class prediction results.

CERCS: P170 Statistics, network analysis, programming

Keywords: churn prediction, telecommunication company, exploratory analysis, predictive analysis

Võrgustikupõhine mudel teleteenustest lahkujate ennustamiseks

Teenusest loobumise ennustamine aitab paremini aru saada, millised kliendid hakkavad lähiajal tõenäolisemalt kasutama konkurentide teenuseid. Teades, et olemasoleva kliendi hoidmine on säästlikum kui uue leidmine, on paljud telekommunikatsioonіettevõtted hakanud laialdasemalt uurima, kuidas ennetada ja vähendada klientide lahkumist.

Reeglina on uuritud klientide lahkumist läbi staatiliste muutujate, kasutades näiteks demograafilisi andmeid, lepingulisi andmeid, toote informatsiooni jm. Hilisemad uuringud antud valdkonnas on vihjanud, et uurides ka klientide omavahelist suhtlust, on võimalik ennustada, millised kliendid tõenäolisemalt loobuvad teenuse kasutamisest.

Kuigi palju on uuritud seda, et milline on seos kõneteenuste, klientide omavahelise suhtluse ja klientide lahkumise vahel, siis tv-teenuste valdkonnas on uuringuid olnud vähe. Antud töö aitab selle tühimiku täita, uurides tv-teenuste klientide teenuse kasutust 3 kuud enne lahkumist kui ka nende kõneeristust kuu aega enne teenusest loobumist. Loodud mudeleid treenitakse, kasutades järgnevaid masinõppemudeleid: XGboost, Adaboost, Random Forest classifier, Logistic Regression, Gradient Boost classifier. Loodi kaks erinevat gruppi mille tulemusi võrreldi. Esimeses grupis kasutati andmeid, kuhu ei olnud lisatud võrgustikuanalüüsi muutujaid ning teise gruppi lisati ka antud muutujad. Analüüsi tulemusel selgus, et kõneeristuse analüüsist genereeritud muutujad suurendasid mudelite täpsust lahkuvate klientide ennustamiseks.

CERCS: P170 Statistika,programmeerimine, võrgustiku analüüs

Keywords: lahkumise ennetamine, telekommunikatsioonіettevõte, empiiriline analüüs, ennustav analüüs

Contents

1 Introduction

The growth in the use of the internet enables telecommunications companies to offer wider variety of services, and thus earn more revenue per existing customer. For example, telecommunication companies can bundle TV services with broadband internet services to earn more revenue per customer as bundling services make every client even more valuable for the telecommunications company [39]. Bundling TV and telecommunication services is also convenient as they have similar payment structures– postpaid service with usually a fixed monthly fee.

Churn occours when a customer stops using a telecommunications provider's services. Retaining existing customers is considered cheaper than acquiring new clients, therefore, reducing customer churn is crucial for companies that benefit from long-time customers [2]. Customer churn within telecommunication companies is a widely researched topic [14], as there is high competition among the telecommunication service providers [1, 5]. Researchers analysing this problem benefit from the fact that telecommunication companies have access to a wide array of customers' data (demographic information (gender, age), contractual details, customers' location during calling activity, service usage, call logs, and numerous feedback forms), which can help in understanding possible churn reasons.

The decision to leave any service may be affected by customers' friends or acquaintances who recently left the same service. Network effects on service usage can be studied by using social network analysis (SNA). It combines graph theory and social communication to understand interrelations between actors [7]. It is believed that when a customer leaves the company for a competitor, he is also more likely to affect people close to him. Therefore it may result in a cascading effect of churners from the service [15]. However, results are inconclusive whether SNA is an effective way to analyse churn within telecommunication companies. Kusuma et al. [27] claim that network features add no predictive power, whereas Zhang et al. [45] claim that incorporating network attributes can significantly improve churn prediction accuracy.

In this study, churners within TV services were investigated. In particular, a dataset of 50,000 customers were analysed, which was provided by one of the largest Nordic telecommunication service providers [1]. Two different types of customer attributes were used to analyse the data:

1. **Customer attributes:** features characterizing individual customers during the observed month and features that characterized customers up to 90 days before the observed month. Features included information from TV services as well as telecommunication services.

2. **Social Network attributes:** network centrality attributes were generated from call logs, which were extracted by using the information provided by the same service provider.

Both types of features were used to train machine learning models (Logistic regression, Random Forest, XGBoost, Adaboost, and Gradient Boost Classifier). The best performing model

[1] Due to the existing privacy contract, the name of the company is not disclosed.

was XGBoost with the synthetic minority oversampling technique. It had the best performance with and without network attributes. Model without network attributes achieved minority class prediction F1 score of 0.4, model AUC of 85.4%, and model accuracy of 96.8%, whereas the model with network attributes had a minority class F1 score of 0.82, model AUC of 97.0% and model accuracy of 99.0%. Results by the best-performing model were promising. After adding network analysis features, minority class prediction F1 score improved by 0.42, AUC by 11.6%, and model accuracy by 2.2%.

Although there have been multiple studies covering attrition in the telecommunication sector, few of them cover value-adding/TV services offered by telecommunication companies. Cross-linking information between services could help us to understand whether phone usage affects churn within value-adding services. Furthermore, to the best of the authors knowledge this is the first work, which has investigated customer churn prediction for TV services, where prediction models not only rely on customer attributes but also the network formed using the mobile call logs, which has enhanced the prediction results of machine learning models.

1.1 Outline

This thesis is organized into chapters as follows: Chapter 2 covers churn prediction models' background within the telecommunication sector, where value-added services and social network analysis is discussed. Chapter 3 is a review of the literature on customer attrition in the telecommunication sector, where methods and results of previous studies in this area are summarized. In Chapter 4, the features used to perform analysis are visualized and summarized. In Chapter 5, the methodology to perform the predictive analysis is discussed. In chapter, 6 evaluation and results from the analysis are reported. In Chapter, 7 findings and limitations of the research are discussed.

2 Background

In the following chapter, we introduce the overview of the necessary information to understand the background of churn prediction within the telecommunication sector. Firstly, value-added services for telecommunication companies are introduced, followed by a background of social network analysis, as it is used to understand connections between actors in the system.[38] Finally, a brief overview of centrality measures is presented

2.1 Value-added services for Telecommunication companies

Within telecommunication sector, value-added services (VAS) are the services that support the primary business of the telecommunication company. For example, 44% of Telia company 2019 revenues derived from mobile services, whereas TV services accounted for 6% of revenue and other services accounted for 9% of revenue [13]. Therefore it can be argued, that the primary source of revenue (mobile services) could be viewed as primary services.

VAS for mobile services may include services that are offered through strategic alliances with content providers. Services may include parking, bus tickets, coupons, apps etc. According to Anckar & D'incau [3], value-adding services bring mobile value in five different areas: mobility-related needs, spontaneous needs and decisions, time-critical needs and arrangement, entertainment needs and efficiency needs and ambitions.

Entertainment services in the telecommunications industry are seen as good value-adding services as they help to "fill time" instead of "kill time" [3, 26]. TV services could be viewed as a medium to consume entertainment. In addition, bundling TV services with mobile services helps to reduce churn, thus increasing average revenue per customer [39]. Within this thesis, TV services are considered as value-adding service.

2.2 Social network analysis

All of us are part of some social network. It may consist of family, friends, co-workers, or people with whom we are participating in the same activity. Typically, we are part of multiple social networks. For example, friends from high school and friends from college can form two different social networks. There may be overlapping of people in networks. For example, if my high school friend is also my co-worker, he would be part of both networks.

The structure of the network is characterized by vertices(nodes) and edges. Nodes are usually defined as actors within the network, whereas edges represent direct connections or links between the nodes. [38] In the context of this thesis, customers are represented as nodes and calls between the customers are defined as edges. The Author is aware that customers may use other forms of communications besides calling to interact with each-other(e.g via online messaging apps). However that kind of information was not available for use.

Nodes and edges of a social network can be visualized through graphs. When describing a social network through a graph, terms, techniques, and concepts from graph theory are used. Although graph theory and social network analysis are not the same, they are based on common concepts. [6, 38]

2.2.1 Homophily

People who are connected usually have something in common because interactions between people who have similar interests are more likely to happen.[29] This idea is supported by the phenomenon called homophily. Literature about homophily is consistent that it helps to characterize network systems[4]. Many different characteristics can affect homophily. For example, age, gender and education strongly affect one's relation with others. In addition, network position, behaviour, and values also show homophily, although to a limited extent [29].

The principle of homophily can also describe characteristics for the majority of customers when limited information is given for analysis. For example, it is demonstrated that when only 20% of the population had fractions of their social media information available, the remaining population attributes could be described with an accuracy of 80% [30].

There are also multiple studies covering homophily by analysing mobile network data [4, 18]. Although there has been a decrease in calling activity in the UK (Table. 1), it is shown that cellphone data can still be used to link population [18].

Table 1: Summary of call volumes (millions of minutes) In the UK between 2007 and 2018. Source: [36]

Year	Summary of call volume in millions of minutes
2007	164,341
2008	145,288
2009	130,037
2010	124,207
2011	111,753
2012	102,604
2013	92,683
2014	81,709
2015	73,333
2016	64,260
2017	52,977
2018	47,019

As shown in Table 1, call volumes have decreased over 60% during 2007-2018 in the The

United Kingdom.

2.2.2 Centrality measures

There are multiple approaches to understand which nodes within the graph are most influential [43]. For example, we may have acquaintances on Facebook that have added many people to their friends' list. However, it does not automatically mean that they are more influential than local politicians, who are more considerate on adding friends. Similar logic may apply to other networks. Therefore, multiple centrality measures have been developed so we could have better understanding of important actors within the network [43]. Some of the commonly used centrality metrics to analyse customer networks are following:

1. **Degree centrality:** Degree centrality shows how many connections does a single node has. It does not differentiate between the quantity and quality of the connection. [22] For example, somebody may be well connected in the network, but it does not mean that he will be an influential figure compared to some other player.

2. **Betweenness centrality:** Betweenness centrality helps us understand which nodes in the graph have the largest number of shortest paths passing through them. A given node may have high betweenness centrality, although it may be connected with a small number of other vertices. These nodes may act as bridges between two groups of nodes. [43]

3. **Closeness centrality:** Closeness centrality shows the node's importance by measuring how close the node is to all other nodes in the network. Nodes with higher closeness centrality can reach every other node with fewer steps than nodes with lower closeness centrality.[43] The Graph needs to be connected to calculate closeness centrality. If there are multiple unconnected groups, harmonic centrality must be calculated instead [28].

4. **Eigenvector centrality:** Eigenvector centrality helps us understand the node's importance within the network. If a single node is linked to the nodes that are also relatively important, then the node itself becomes more impactful. Therefore, it has a high eigenvector centrality score. [43] For example, if somebody has 1000 friends on Facebook but the friends are not influential, then the eigenvector centrality score would be lower than somebody with 100 influential friends. Pagerank centrality is also affected by the eigenvector centrality scores [9].

5. **Pagerank centrality:** Pagerank centrality measures were developed by Sergey Brin and Lawrence Page in 1998 when they published the paper: "The Anatomy of a Large-Scale Hypertextual Web Search Engine". The paper's main goal was to describe how to build large scale search engine that could exploit the information in the hypertext [9].

   Pagerank centrality is based on the idea that when a person crawls randomly through the internet, then what is the chance that he ends up in a specific webpage. The success of the

website, according to the pagerank centrality score, is characterized by three elements. [9, 16]

(a) The quality of the linkers– what is the pagerank value of the linkers websites

(b) The link propensity of the linkers- how many sites do the linkers link to

(c) The number of links the webpage receives.

By combining multiple centrality metrics, most important actors within the network get spotted.

3 Related work

In this Section of the thesis, literature that has focused on churn within telecommunication companies in different aspects is reviewed. Literature concerning TV services, where network analysis features were implemented was not found. Therefore, papers that cover network analysis results from mobile services churn prediction are presented. Firstly, we discuss papers that have looked into churn models and what the results have been so far. Secondly, methodologies from studies where network analysis is used are summarized.

Retaining existing customers cheaper than finding new customers [2]. That is why Customer retention is a widely analysed topic in industries where business models are built on recurring revenue streams, such as the telecommunications industry [14, 31] as well as in other sectors such as financial services [12], banking [37] and TV services [10, 21].

3.1 Churn prediction based on customer attributes

Client behaviour prior to churning can be analysed by using various methods. For example, if we do not want to capture how customer behaviour has changed during the last few months prior to churning compared to previous periods, a static approach could be used. But on the other hand, if we want to understand whether customer behaviour changed prior to churning, then time series analysis (dynamic approach) could be used [11]. Often various classifying methods are used to develop churn prediction models, such as decision trees [10, 14, 31], Support vector machines [11], Regression models [10, 44, 45], and Neural networks [44, 45]. For churn analysis, decision trees could be the preferred option as their results can be successfully visualized, thus they can be interpreted with ease. In addition, analysis can be performed on categorical as well as numerical features and no prior assumptions are required to perform analysis. [23]

Feature selection for analysis is essential to achieve good predictive results. According to Verbeke et al. [44], 6 to 8 variables may be enough to predict churn with high accuracy, therefore having few features with good quality may be an economically more viable solution than having numerous features. The study also found that service usage attributes got the highest predictive power. This finding is also supported by the study conducted by Mitrovic et al. [31], where different features were used to predict churn characteristics of prepaid and postpaid clients in a telecommunication company.The study also found that the most practical features to predict customer churn in a postpaid segment are:

1. Features that characterize individual customers during the observed month
2. Features that are calculated based on customers 1-st level neighbours
3. Features that characterize individual customers up to (m-3) months before the observed month

Churn within the television services is not as widely researched as in mobile services. For example, Burez & Van den Poel [10] have analysed European pay-TV company churn by gathering contractual data, socio-demographic information, historical subscription data and financial information. Their study had a static approach, where customer behavioural dynamic was not captured. Furthermore, social ties between customers were not considered. Nevertheless, conducted models had higher AUC scores than a random model.

3.2 Churn prediction based on call logs

The following subsection focuses on describing existing approaches used by related studies to describe call graphs.

Customers leaving the service may convince their friends and acquaintances to follow suit, although influence over friends decreases exponentially over time after churn has taken place. Furthermore, increased homophily increases the chance of defection as well as the strength of the ties between the defector and his friends.[35]

Mitrovic et al.[31] described interactional information in the social network analysis through the RFM model, where recency(R) stood for the time between the customers' last call and the end of the observed period, frequency(F) stood for a number of calls a single customer made during the period under analysis and monetary (M) value stood for the duration of calls made by the customer. Interactional data has also been viewed as interactions between the churner and its neighbours [15]. Interactional customer data has been used by multiple related studies that predict customer churn through SNA. [15, 27, 31, 40, 41, 45]

Structural features within the data are described by using centrality measures. Most often, the simplest approach- degree centrality is used in the research [31]. For example, Kusuma et al. used 2nd and 3rd-degree count besides direct neighbours of a node to explain connectivity between nodes and Nanavati et al. used the Pagerank approach. [27, 33] Dynamic data illustrates customer behavioural aspects over a time period rather than describing a static state at a specific time. This approach is said to be indisputably more useful than a purely static approach [17]. A dynamic approach is applied by taking snapshots from the network at different times and merging them later for additional analysis [34].

In Table 2, one can see which feature groups have been used in the SNA analysis in the recent relevant studies.

Table 2: Summary of features used in recent studies within the telecommunication sector

Study	**Sample size**	**TV services**	**Interaction**	**Structural**	**Dynamic**	**RFM**	**SNA**
[33]	2.9m nodes	No	No	Yes	No	No	Yes
[15]	2.1m nodes	No	Yes	Yes	No	Yes	Yes
[41]	28m subscribers	No	Yes	Yes	No	Yes	Yes
[40]	4.8m nodes	No	Yes	Yes	Yes	Yes	Yes
[31]	4.8m nodes	No	Yes	Yes	Yes	Yes	Yes
[45]	>50,000 subscribers	No	Yes	Yes	No	Yes	Yes
[27]	700m call records	No	Yes	Yes	No	Yes	Yes
This study	50,000 subscribers	Yes	Yes	Yes	Yes	No	Yes

We have discussed that there have been numerous studies predicting churn in the telecommunication sector by using social network analysis. Related papers used similar methodologies for the research, but none of the studies combined the calling data and information about TV services to the best of our knowledge. The thesis provides a systematic literature review of the relevant publications and the results of the thesis could be used to further develop churn prediction models in the telecommunications sector.

Furthermore, this paper considers features and methodologies used in the previous studies [10, 27] but also supplementary SNA effect is considered by looking at studies with similar aims. [15, 41]

4 Dataset description and descriptive analytics

This Chapter provides information about the data that was used for the analysis. The Chapter describes from which sources the data was collected and which features were selected. Finally, a descriptive analysis was performed to describe the features within the dataset. Two distinct datasets were used to conduct this analysis. One of the datasets had network features added, whereas the other did not.

4.1 Customer selection

All of the customer based attributes used in the analysis described how the customer uses TV and telecommunication services during the last three months prior to service cancellation or before the agreed end date.

For the analysis, churn is defined when the contract is cancelled, and the customer leaves the company to use alternative services. Therefore, customers who have indicated that they continue to use the service under different terms (another contract for example) are excluded from the churning customer base. In addition, customers who have died during the period under analysis have also been excluded from the dataset as their churning reason is not related to service quality. Approximately 50,000 contracts were analysed with a churning to no-churning ratio of 1 to 33.

All non-churning customers were active between the period from 2020/03 to 2021/02, whereas the period under analysis started in 2020/04 and ended in 2020/12. This approach helped to remove clients from the dataset who may have had service usage patterns of a churning client, but they were considered non-churning in the constructed model. In addition, all customers in the customer base were required to have a contract that has been active for longer than two years to eliminate clients that may have been locked to use the service during the period under analysis. In addition, all of the contracts related to employees were excluded. Every non-churning client was assigned a first date of a random month between 2020/04 and 2020/12 as their end date. For churning clients, the churning date was chosen as the date when the contract was cancelled. After customers were selected, three months of data prior to the agreed end date/ churning date was analysed.

4.2 Features

4.2.1 Customer attributes

Customer attributes included information from the TV/telecommunication service usage as well as information from the call logs.

Some of the features could not be allocated to a specific contract. For example number of active telecommunication contracts for a single client at the churning date, therefore some fea-

tures are grouped by customer level instead of contract level. Please note that a single customer may have multiple contracts. For majority of customers, TV and internet service was bundled (Fig: 1).

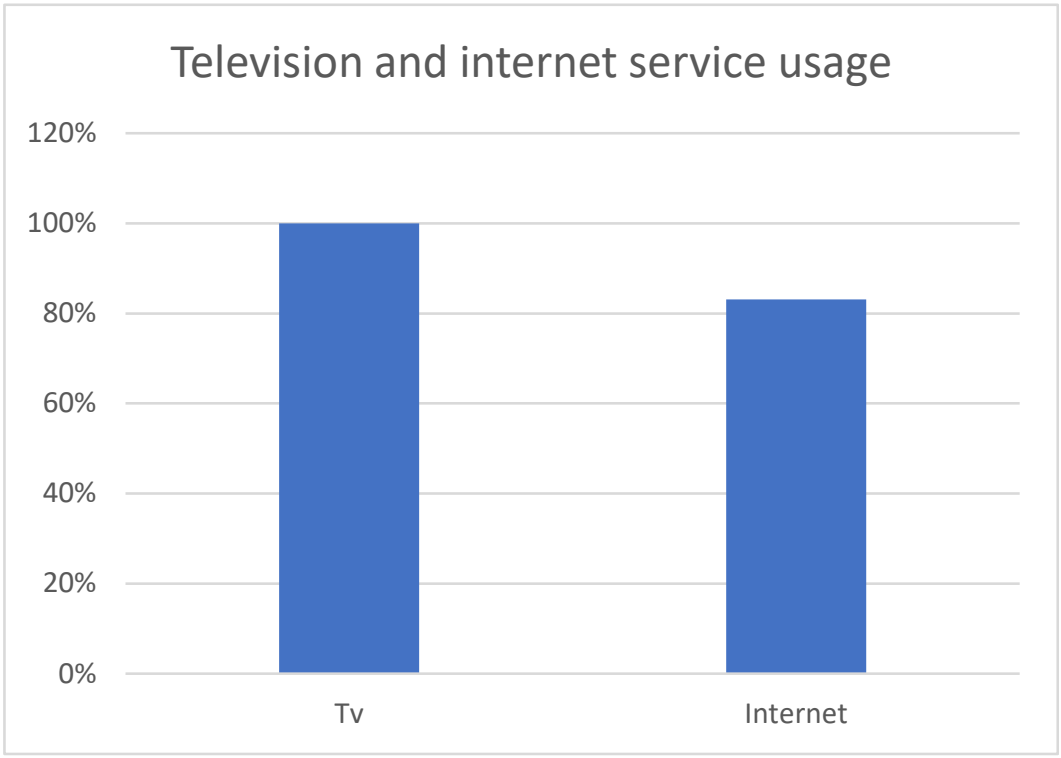


Figure 1: Internet and TV service usage among analysed customers

Prior to churn prediction analysis, multiple changes were made to the dataset (normalization, imputation, creation of dummy variables). For more information, please see Chapter 5 (methodology).

4.2.2 Features related to TV usage

Altogether 13 unique features were extracted from databases to describe TV service usage of a specific contract.

Table 3: TV usage related customer attributes used in the analysis

	Short description of the feature	**Standardized feature**	**Additional information**
1	TV service name and TV service type	No	Type as whether the service requires cable or not
2	Number of additional TV services used	Yes	All of the additional services were combined into a single feature
3	Number of times customer watched TV shows during last three months while using rewind	Yes	
4	Number of unique shows customer watched during last three months while using rewind	Yes	
5	Average TV show watching length while using rewind	Yes	
6	Modem model name	No	Removed rows without values
7	Modem working	No	
8	Modem problematic	No	Two problem codes were merged together as one. A single customer may have had problematic and working modem at the same time.
9	Days since modem update	Yes	
10	TV invoice one month prior to churning	Yes	Same month invoice could not be used as mid-month churners would have unique values
11	TV invoice two months prior to churning	Yes	
	Synthetic features		
12	row 10 + row 11	Yes	Removed customers with empty values
13	Standard deviation of TV invoices	Yes	

Numeric features were standardized with formula explained in the Fig. 7.

4.2.3 Telecommunication and Socio-demographic features

In addition to internet services, some customers had their mobile broadband contracts under the same telecommunication company. Therefore, some additional information could be extracted. Within this section, contractual and demographic information used in the analysis is presented.

Table 4: Telecommunication and socio-demographic attributes used in the analysis

	Short description of the feature	**Standardized feature**	**Additional information**
1	Number of incidents	Yes	Summarized TV and Telco incidents
2	Number of active telco contracts as per churn date	Yes	
3	Duration	Yes	Removed clients with a duration of less than 24 months
4	Gender	No	
5	Language	No	
6	Rounded age	Yes	Rounded feature to anonymize the data
7	Invoice channel	No	Whether invoice information is communicated through SMS, E-mail, self service.
8	County	No	
9	approx. download speed	Yes	
10	Internet service code	No	
11	Telco invoice 1 months prior to churning	Yes	Please note that not all of the customers were using telco services besides TV services
12	Telco invoice 2 months prior to churning	Yes	Please note that not all of the customers were using telco services besides TV services
	Synthetic features		
13	row 11+ row 12	Yes	
14	Standard deviation of telco invoices	Yes	

As seen from Table 4, in total of 7 additional features (rows 2, 9, 10, 11, 12, 13, 14) related to telecommunication service usage were added, resulting in a total of 27 unique features to describe the dataset. Preprocessing also included transforming categorical data into binary columns. In total of 119 features were generated for the machine learning models after the categorical variables were turned into dummy variables. This dataset did not include features extracted from network analysis.

4.2.4 Calling network

Calling logs were collected only from customers that were included in the initial analysis sample (50,000 clients). Every customer contact number was searched within the database for outgoing and incoming calls. Calls that were made in the last 30 days prior to the churning /end date were summarized. Around 85% of customers could be found within the database and in total of 371,440 unique numbers were extracted. Calling logs did not include mobile services

such as mobile parking. In Table 5, generalized information about the call logs could be found. On average, a single node had 2.9 connections. Therefore, the density of the graph, as well as clustering coefficient, are low. Density measures how many edges exist between nodes compared to all possible edges that could exist. The Clustering coefficient for a single node shows how connected are the neighbours of the node between each other and the average clustering coefficient calculates the mean clustering coefficient value within the graph.

Table 5: Summary of network extracted from the call logs

Statistic	**Value**
Number of nodes	371,440
Number of edges	546,765
Average degree	2.944
Density	7.93E-06
Average clustering coefficient	0.000197

The weight of the edges was described by the amount of calls two customers held during the period under analysis. If both customers were presented in a sample with different churning date/ agreed end date, then the strength of their connection would be proportionally stronger. No standardization of edge weight was conducted for such cases due to the following reasons: 1) Less than 2% of clients within the sample would be affected 2) On average, five calls were made between customers and 89% of nodes had weight lower than ten (Fig. 2).

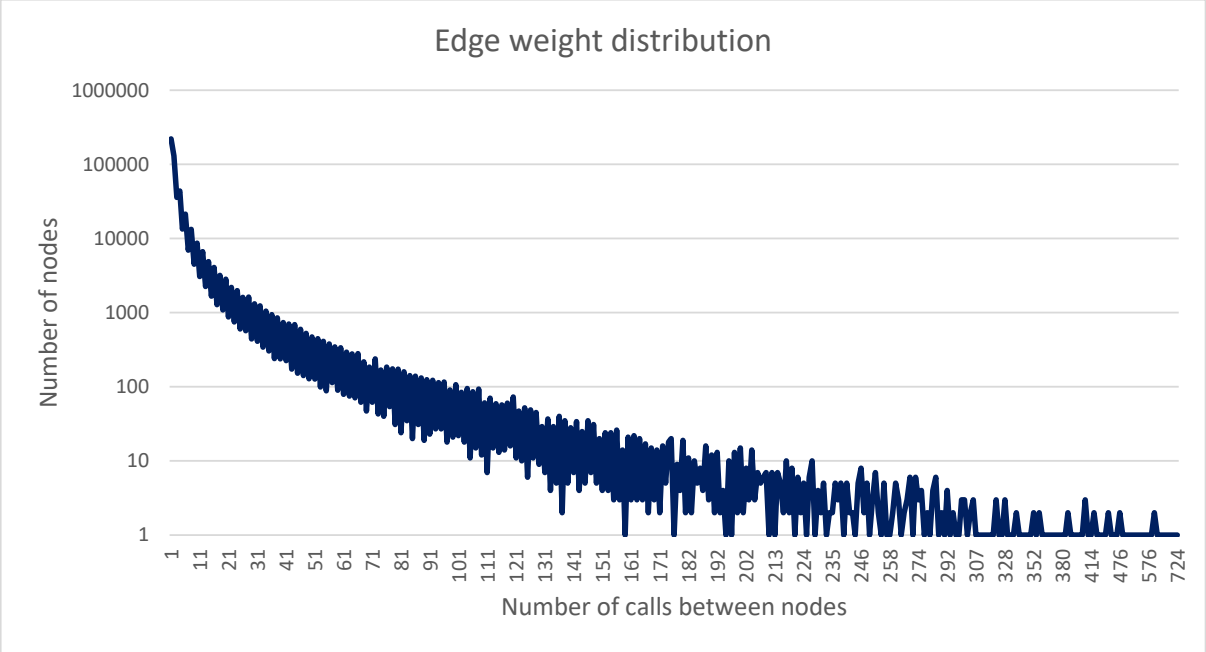


Figure 2: Call volume distribution among the clients within the dataset

As one can see from figure 2, the majority of edges had weight value below 10. The highest

recorded edge weight stood at, 588 meaning that one customer had 588 calls with someone else within 30 days prior to the end date/ churning date.

There were also some limitations regarding the data collection. We had limited information about customers who used an alternative mobile operator because neither their outbound nor inbound calls could not be fully tracked unless the other party uses the mobile network under analysis. Due to a limited amount of available data, undirected network was formed.

Extracted data was visualized, showcasing different communities and betweenness centrality values of nodes (Fig. 3).

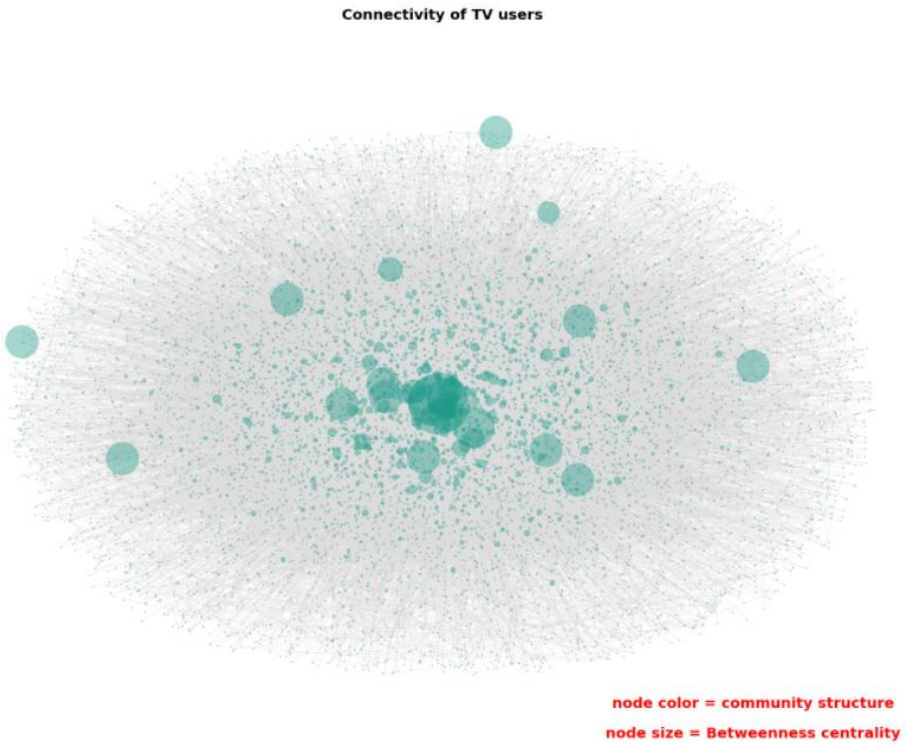


Figure 3: Connectivity of the customers used in the analysis

As seen from Figure 3, all of the customers were labelled into the same community. Another peculiar finding was that some of the nodes with higher betweenness centrality were presented in the edge of the graph, although nodes with high betweenness centrality usually act as a bridge between multiple groups of nodes.

4.2.5 Features extracted from the call logs

The main goal of this paper was to understand whether calling information can improve churn prediction for TV services. In total, 7 features were extracted regarding the calling data. The assumption that more popular customers were more influential regarding the churn was set. Therefore if popular customers churn, then other customers within the sample are also more likely to churn. This assumption has some supporting evidence in the academic work, although it is said that more validation was required [41].

Analysis indicated that only a small fraction of clients had churning friends because, on average, every customer within the base was connected to 13 nodes. Furthermore, clients who had a friend who left the service did not have a higher correlation with the churning label. Therefore centrality metrics had to be calculated for every client. For more information, please refer to Table 6. Features were chosen as follows.

Table 6: Social network features used for the analysis

	Short description of the feature	**Used StandardScaler**	**Additional information**
1	Highest degree centrality value for the churning customers within the last 30 days	Yes	Used mean imputation for empty values
2	Highest eigenvector centrality value for the churning customers within the last 30 days	Yes	Used mean imputation for empty values
3	Highest pagerank centrality value for the churning customers within the last 30 days	Yes	Used mean imputation for empty values
4	Highest harmonic centrality value for the churning customers within the last 30 days	Yes	Used mean imputation for empty values
5	Highest betweenness centrality value for the churning customers within the last 30 days	Yes	Used mean imputation for empty values
6	Number of contacts churned within the last 60 days	Yes	Replaced empty values with 0-s
7	Number of contacts churned within the last 120 days	Yes	Replaced empty values with 0-s

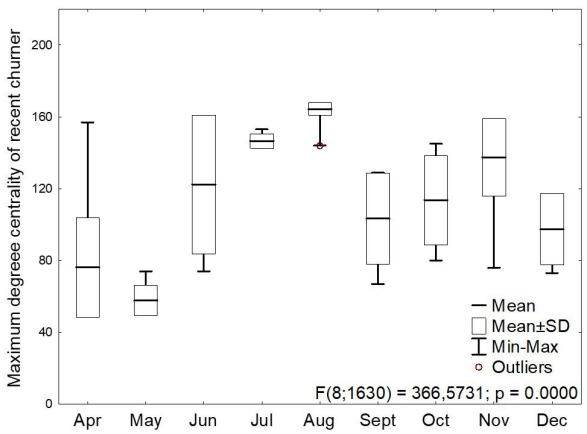
In total of 7 network-related features were added to the analysis. Where possible (eigenvector centrality, betweenness centrality, pagerank centrality), edge weight was considered when calculating the centrality values of nodes. Customer contact numbers that were not found in the call logs database had a mean centrality value imputed.

4.2.6 Limitations in the data extraction

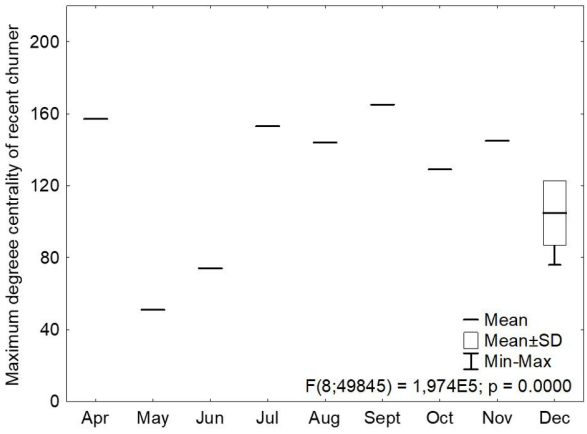
Data extraction had some limitations as some of the features are updated once per month. All of the non-churning customers who had an end date in August had their information retrieved, considering the end date as the first of August. This approach could lead to model overfitting when considering the current network analysis feature extraction technique. To alleviate the problem, the author changed the December date for approximately half of the non-churning

customers to avoid a situation where the centrality values of customers who left in December are not taken into account.

To visualize the problem, comparative visualizations of degree centrality distribution were conducted (Fig. 4)



(a) Maximum degree centrality of recent churners for clients who churned during the period under analysis



(b) Maximum degree centrality of recent churners for clients who did not churn during the period under analysis

Figure 4: Maximum degree centrality feature values for churners and non-churners from the period of April to December in 2020

As seen from Figure 4, centrality values are fixed based on a month for non-churning clients. Please note that the model does not know the end date for the non-churning customer nor the churning date of the customer who churns, although there may be other non-direct features in the dataset indicating a month that the author is not aware of.

4.2.7 Feature extraction summary

Dataset without network analysis features consisted of 27 unique features. In total, seven features were added during the network analysis resulting in total of 34 unique features.

All of the extracted features were plotted onto a single correlation graph. It indicated a strong correlation in the following areas:

1. Strong correlation was found between features that account for customer TV shows rewinding habits. People who watch TV more often are also more likely to watch more unique TV shows.

2. Customers' invoices 1 month before churning and 2 months before churning are strongly correlated because invoice sizes will not hike unless the customer uses additional services or the company raises prices for customers.

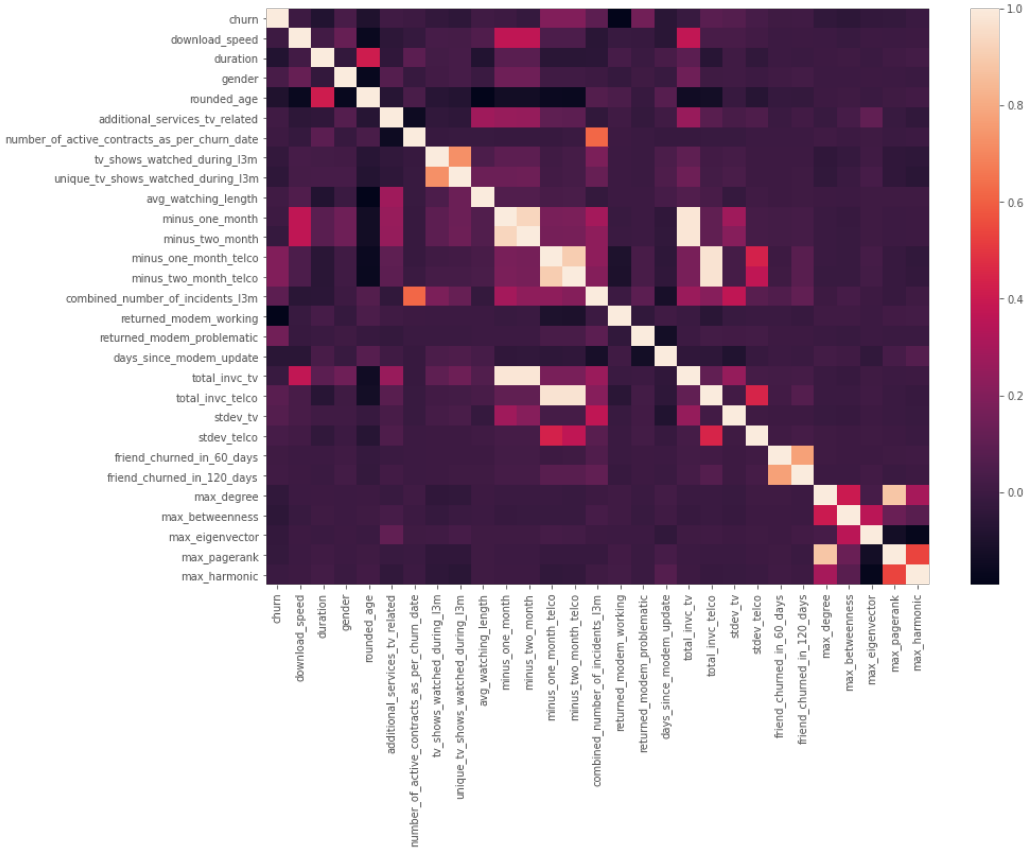


Figure 5: Correlation between variables

No features were strongly correlated to churning behaviour (Fig. 5).

5 Methodology

In this Chapter, we gather ideas from previous chapters to predict the churn of a customer within the sample. First, we give a brief overview of machine learning models that were used to predict customer churn. Secondly, we present the metrics that were used to evaluate the accuracy of the predictions. Finally, prediction results are presented. Summarized visualization of methodology is presented in the Figure 6.

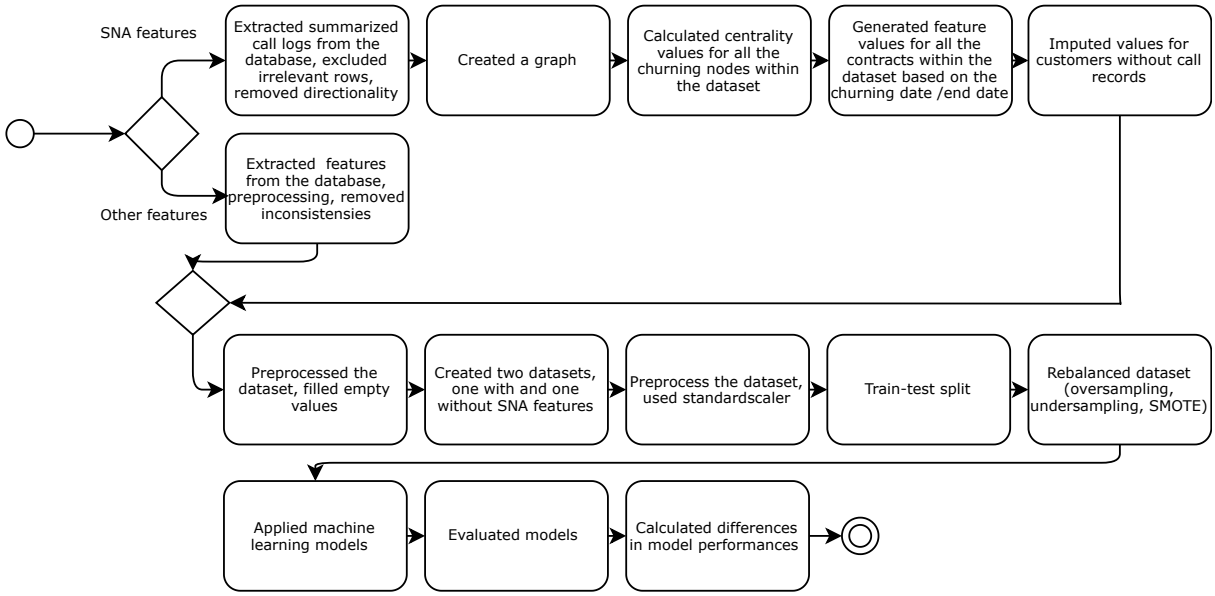


Figure 6: Thesis methodology flowchart

5.1 Dataset preprocessing and balancing

All of the numeric features within the dataset were scaled around the mean value of the feature column (Fig 7).

$$z = (x - u)/s$$

Figure 7: StandardScaler formula

Within the formula, *u* stands for column mean value and *s* stands for the standard deviation of the column. Missing values within the dataset were replaced with 0 unless stated otherwise.

Due to imbalances in the dataset (churn to non-churn ratio of 3 to 100), machine learning models could lead to inaccurate or biased conclusions. Our main goal is to improve ML models so that minority class would be identified with higher accuracy across all ML models. To tackle this issue, the following balancing techniques were used:

1. **Oversampling** - Balance between minority and majority classes is reached by duplicating examples from minority class.

2. **Undersampling**- Balance between minority and majority classes is reached by deleting examples from the majority class.

3. **Synthetic Minority Oversampling Technique (SMOTE)** - Balance between minority and majority class is reached as new minority class values are randomly generated based on existing minority class values.

Results were also calculated without using any balancing techniques for reference purposes.

5.2 Machine learning models

Churn prediction is a binary classification problem, meaning that models have to predict whether a certain data point has a value of 1 or 0 (i.e. churn or no churn) . To predict the churn within the dataset, 5 different machine learning models were used. Hyperparameter tuning for machine learning models was not carried out as it was out of the scope of the thesis. Following models were used:

1. **Logistic regression** is used to predict the probability of a binary class, for example, churn or no-churn. Logistic regression requires no outliers to be within the data, therefore the dataset was rescaled. Besides, Logistic regression does not support multicollinearity (correlation over 0.9) among independent predicting variables. [24]

2. **Random Forest** is an ensemble learning method for classification problems. Random forest constructs multiple decision trees and outputs the average prediction value of the individual trees. Random forest achieves better results than individual decision trees as they tend to overfit the data. [8]

   Within this thesis, Random forest was implemented by using the Scikit-learn library in Python.

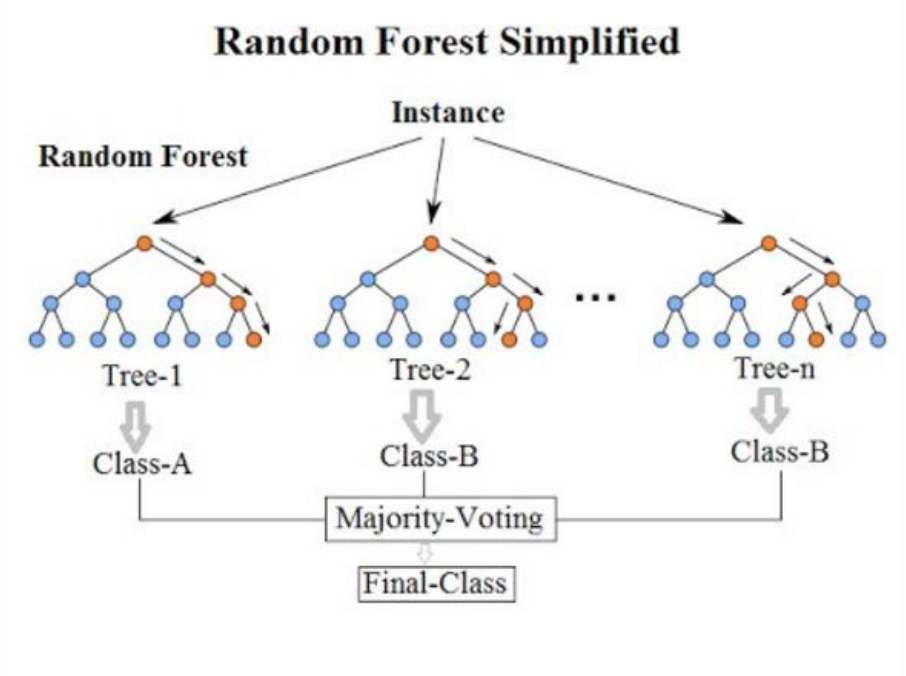


Figure 8: Illustration of Random forest classification [25]

3. **XGBoost** is an optimized version of a gradient boosting algorithm that can accurately solve classification as well as regression problems. [20] Due to the good performance achieved by the parallel tree boosting approach, XGBoost is widely used among the computer science community. Within this thesis, XGBoost was implemented by using the Scikit-learn library in Python.

4. **Adaboost** was one of the first boosting algorithms for binary classification that had success. Boosting creates a strong classifier from weak classifiers based on errors within the subsequent model. This means that every following model has increased weights for cases where classification was more problematic previously. Additional models are added until the training set could be predicted perfectly or a maximum number of models that could be added is reached. Due to the nature of the learning, Adaboost is well suited for out of the box classifications. [19]

5. **Gradient boost classifier** creates a prediction based on multiple decision trees. Each tree is fit onto the modified version of the initial dataset and each model improves based on previous model errors. Within every following model, observations have an unequal probability to be used for model learning. [20]

5.3 Evaluation of the models

To better understand whether a certain approach yields successful results, a good set of metrics needs to be in place. Churning is a classification problem, therefore relevant metrics such as accuracy, precision, recall and F1 score, ROC-AUC curve and confusion matrices should

be used. Each metric used in the model evaluation process, as well as their calculations, are discussed below.

1. **Precision** helps us to understand what proportion of positive predictions were correct. To calculate it, the following formula has to be used.

$$Precision = \frac{TP}{TP + FP}$$

where *TP* stands for True Positive predictions and *FP* stands for False positive predictions within the sample

2. **Recall** helps us to understand what proportion of actual positives were correctly identified. To calculate it, the following formula has to be used.

$$Recall = \frac{TP}{TP + FN}$$

Where *TP* stands for True Positive predictions and *FN* stands for False-negative predictions within the sample. Please note that when a model produces no false negatives then the recall value is 1.

The main goal of this thesis is to understand which customers churn, therefore recall metric is more closely monitored, as it describes how large percentage of churners were identified out of all churners.

3. **F1 Score** averages results from precision and recall. The highest value the F1 score could reach is 1, which means that perfect precision and recall are reached.

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

4. **AUC - ROC curve** AUC-ROC curve is used to evaluate binary classification problems. It plots the true negative rate against the false positive rate at various threshold values to evaluate how distinguished the two binary classes are. Higher AUC indicates that two classes are well distinguished between each-other

5. **Accuracy score** helps us to understand what is the fraction of predictions the model got right.

$$Accuracy = \frac{CorrectP}{TotalP}$$

Where *CorrectP* stands for the number of correct predictions and *TotalP* stands for the number of total predictions made.

For unbalanced data sets, high accuracy may be common as the majority class could be predicted with more success although the minority class predictions may perform poorly. That is why alternative performance evaluation metrics such as recall and precision should be used.

5.4 Experiment Setup

The aim of the thesis was to find churners among the people who have subscribed to the TV services during the period under analysis. The experiment was set up as follows:

1. **Extracted features from databases** The author prepared two datasets. Both of the datasets included features regarding the TV services and telco services.

2. **Calculated network analysis features** Some of the features were calculated based on the information in the call logs. A graph from call logs information was generated, based on which the centrality values were calculated for each customer contract. Network analysis features were added to the second dataset.

3. **Preprocessing data** Data preprocessing was carried out by removing customers who had certain data points missing, replacing empty cells with 0-s or calculating an average value for column and imputing calculated value to empty cells. After the data was imputed or removed, numeric columns were scaled around the mean value by using Standardscaler from the Scikit library. After transforming categorical variables into binary variables, the first dataset consisted of 119 distinct features and the dataset with network analysis features consisted of 126 features

4. **Train-test split** Data was split into a training set and test set with a ratio of 80%-20%.

5. **Rebalancing datasets** The data was rebalanced and three distinct groups of training sets were created based on the balancing technique that was used. Techniques included under-sampling, oversampling and SMOTE. Please note that the model without balancing was also conducted for benchmarking purposes.

6. **K fold cross validation** For this thesis, the data was split into 5 groups of equal size and K-fold cross-validation was performed. This approach is used to further validate the predictive power of a single model by resampling the training data and by changing the data that is unseen by the machine learning model. [32] K refers to a number of groups the data is split into. For a clearer understanding of K-fold cross-validation, please see Figure 9.

Figure 9: K-Fold cross validation methodology summary [42]

7. **Applied machine learning models** Predictions were performed with 5 models: Random forest, XGBoost, Logistic regression, Adaboost and Gradient Boost classifier. Every model was trained while using different rebalancing techniques. The predictive part was carried out with the first (119 features) and second (119 features + 7 features from call logs) dataset.

8. **Evaluated models** The metrics were chosen as follows. Accuracy and AUC score was calculated at the model level. For the minority (churn) class prediction evaluation, the following metrics were calculated: precision, recall and F1 score.

9. **Calculated differences in model performances** The models and rebalancing techniques were benchmarked against each other and differences in performances are presented.

6 Results

In this Chapter results of applying the prediction models and comparison between them is discussed. In total of 5 different machine learning models as described in Section 5.2 were used with 3 different rebalancing techniques. For benchmarking purposes, models without rebalancing were also tested. Model performance was evaluated by metrics that were described in Section 5.3

6.1 Prediction model results

6.1.1 Logistic regression

Logistic regression model predicted minority class with mean F1 score of 0.2 and mean recall of 0.31 meaning that out of all churning contracts, only 31% were correctly classified. Logistic regression model with SMOTE rebalancing technique had the highest recall score (0.33) amongst the models that did not have social network analysis features added. All the other models had relatively poor performance with F1 scores between 0.14 and 0.18.

Table 7: Logistic regression prediction results, without network analysis features

Logistic regression without network analysis features					
Balancing technique	**Precision**	**Recall**	**F1 score**	**Model AUC**	**Model accuracy**
Oversampling	0.1	**0.74**	0.18	**82.2%**	78.6%
SMOTE	0.3	0.33	**0.32**	80.1%	95.5%
Undersampling	**0.58**	0.08	0.14	81.3%	**97.0%**
No balancing	**0.58**	0.08	0.14	81.3%	**97.0%**
Average	0.39	0.31	0.20	81.2%	92.0%
Logistic regression with network analysis features					
Balancing technique	**Precision**	**Recall**	**F1 score**	**Model AUC**	**Model accuracy**
Oversampling	0.07	**0.71**	**0.13**	**78.0%**	70.0%
SMOTE	**1**	0	0	56.8%	96.9%
Undersampling	0.04	0.42	0.07	55.9%	66.6%
No balancing	0.53	0.03	0.05	73.6%	**97%**
Average	0.41	0.29	0.06	66.1%	82.6%

Unlike other models, adding network analysis features, decreased mean recall and F1 scores by 0.2 and 0.14 points respectively. Model, where SMOTE rebalancing methods were used, could not be calculated and other models also had relatively poor performance. Therefore it is safe to say that logistic regression performed worsened when network analysis features were added.

As one can see from Figure 10, network analysis features rather distracted the classification process. No network analysis features were among the top 15 most important features.

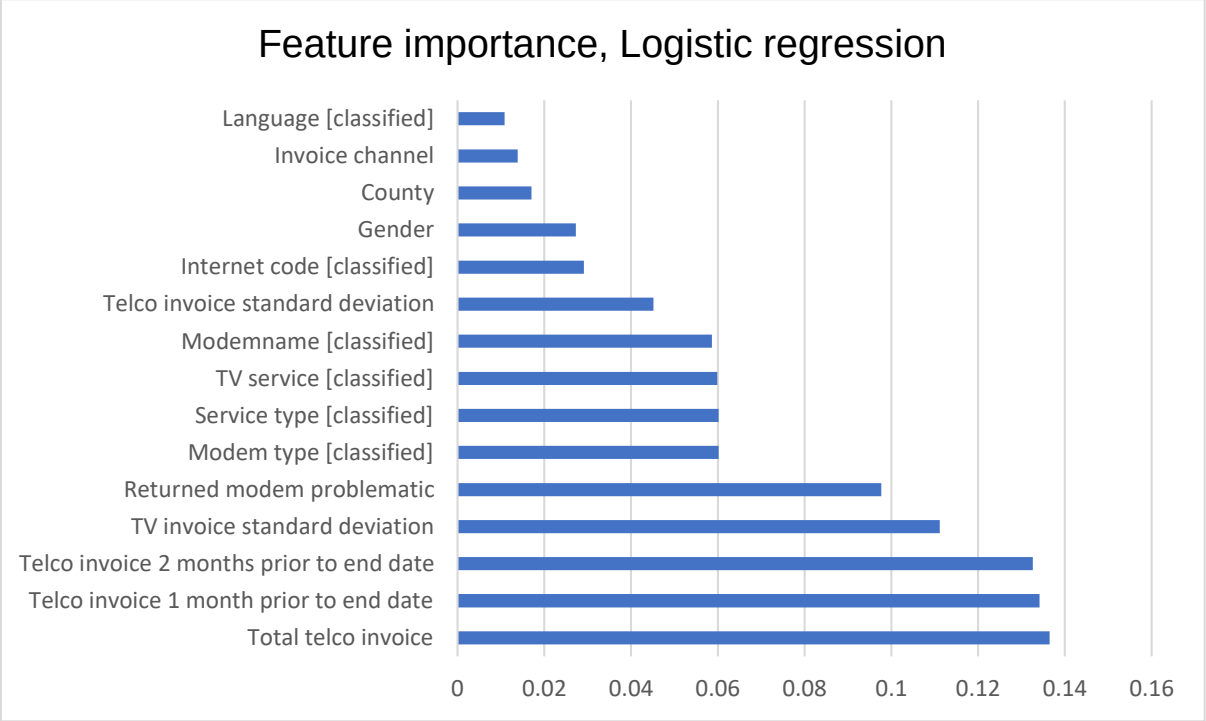


Figure 10: Unbalanced logistic regression feature importance in the conducted model. Network analysis features were added and they are marked as green, whereas customer attributes are marked as blue.

Telcommunication services invoice size had the highest importance for the logistic regression model.

6.1.2 Random forest

Random forest model predicted minority class with mean F1 score of 0.31 and mean recall of 0.22 meaning that out of all churning contracts, only 22% were correctly classified. Random forest model with SMOTE rebalancing technique had the highest recall score (0.27) amongst the models that did not have social network analysis features added. Best performing models according to each metric were highlighted.

Table 8: Random forest minority class prediction results

Random forest without network analysis features					
Balancing technique	**Precision**	**Recall**	**F1 score**	**Model AUC**	**Model accuracy**
Oversampling	0.69	0.18	0.28	**83.1%**	97.2%
SMOTE	0.44	**0.27**	**0.34**	81.8%	96.7%
Undersampling	**0.83**	0.18	0.3	**83.1%**	**97.4%**
No balancing	0.82	0.19	0.31	82.7%	**97.4%**
Average	0.70	0.22	0.31	82.7%	97.2%
Random forest with network analysis features					
Balancing technique	**Precision**	**Recall**	**F1 score**	**Model AUC**	**Model accuracy**
Oversampling	0.89	0.42	**0.57**	**96.0%**	98.0%
SMOTE	0.89	0.42	0.47	95.7%	98.0%
Undersampling	0.87	**0.43**	**0.57**	95.8%	98.0%
No balancing	**0.98**	0.42	0.52	95.6%	**98.2%**
Average	0.91	0.42	0.53	95.8%	98.1%

Adding network analysis features, improved recall and F1 score in average of 0.22 points. Recall scores across the models levelled out, indicating that network analysis features were important. Average F1 score for the minority class stood at 0.53.

As one can see from Figure 11, network analysis features had high importance when predicting whether a contract belongs to a minority or majority class.

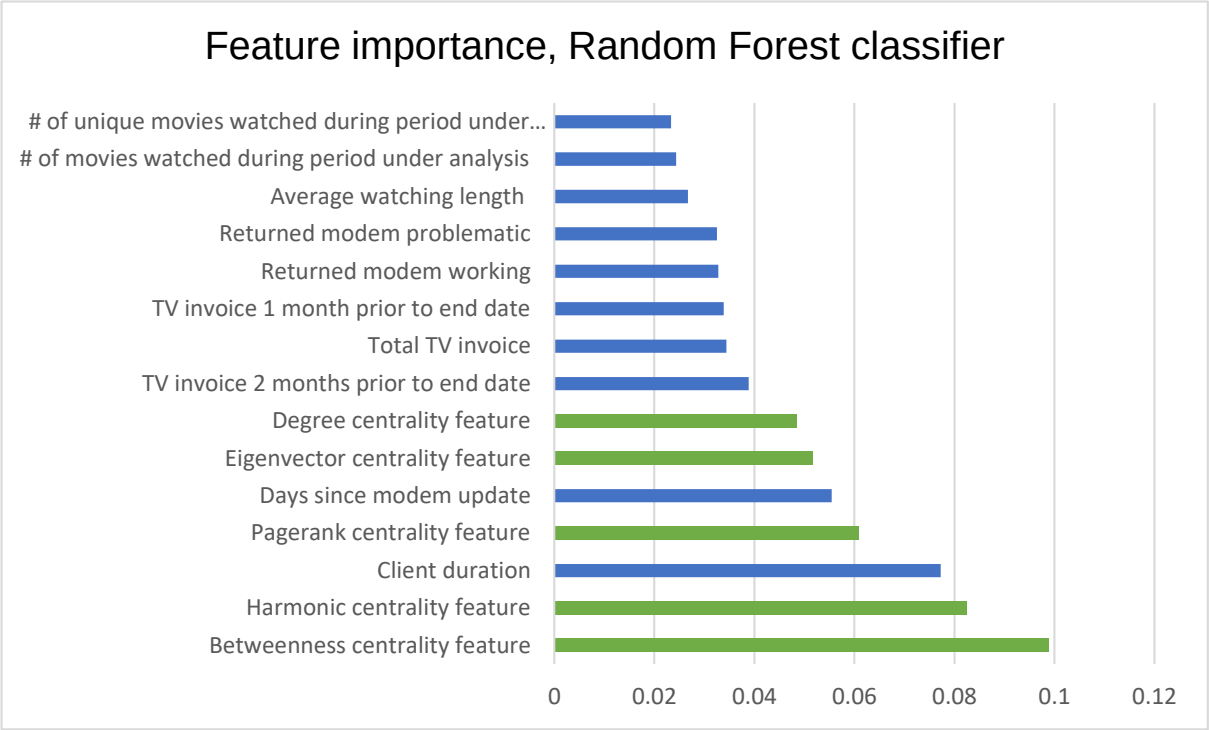


Figure 11: Unbalanced Random forest feature importance in the conducted model. Network analysis features were added and they are marked as green, whereas customer attributes are marked as blue.

Betweenness centrality and harmonic centrality features had the highest importance for random forest classifier model.

6.1.3 XGBoost

XGboost model predicted minority class with mean F1 score of 0.34 and mean recall of 0.24 meaning that out of all churning contracts, only 24% were correctly classified. XGBoost model with SMOTE rebalancing technique had the highest recall score (0.34) amongst the models that did not have social network analysis features added.

Table 9: XGBoost prediction results, without network analysis features

XGBoost without network analysis features					
Balancing technique	**Precision**	**Recall**	**F1 score**	**Model AUC**	**Model accuracy**
Oversampling	0.65	0.17	0.27	84.4%	97.2%
SMOTE	0.47	**0.34**	**0.4**	83.3%	96.8%
Undersampling	0.71	0.25	0.37	**85.4%**	**97.4%**
No balancing	**0.86**	0.19	0.3	83.9%	**97.4%**
Average	0.67	0.24	0.34	84.2%	97.2%
XGBoost with network analysis features					
Balancing technique	**Precision**	**Recall**	**F1 score**	**Model AUC**	**Model accuracy**
Oversampling	0.89	0.42	0.57	96.0 %	98.0%
SMOTE	0.94	0.73	**0.82**	**97.0 %**	**99.0%**
Undersampling	0.28	**0.9**	0.43	96.1%	92.7%
No balancing	**0.97**	0.37	0.54	95.4%	98.0%
Average	0.77	0.61	0.59	96.1%	96.9%

Adding network analysis features improved mean recall and F1 scores by 0.37 and 0.25 points respectively. Undersampling had the highest recall score, although the precision was relatively small compared to other models. Best results were obtained by using SMOTE re-balancing technique. It had the minority class classification F1 score of 0.82 points. XGBoost average F1 score for minority class stood at 0.59.

As one can see from Figure 12, network analysis features had high importance when predicting whether a contract belongs to a minority or majority class. Supporting the results from previous model analysis.

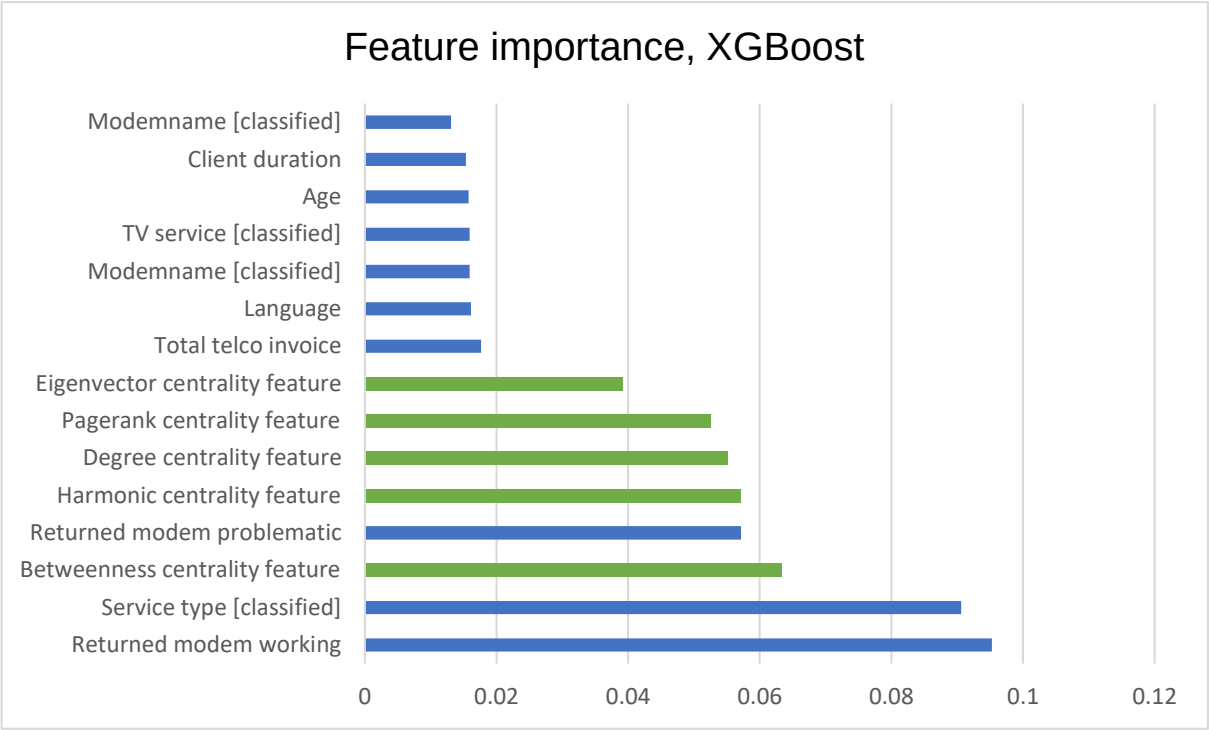


Figure 12: Unbalanced XGBoost feature importance in the conducted model. Network analysis features were added and they are marked as green, whereas customer attributes are marked as blue.

For XGBoost, two most important features to predict customer churn were constructed without network analysis, although as it is seen from the graph 12, network analysis features are relatively important

6.1.4 AdaBoost

Adaboost model predicted minority class with mean F1 score of 0.26 and mean recall of 0.34 meaning that out of all churning contracts, only 34% were correctly classified. Adaboost model with SMOTE rebalancing technique had the highest recall score (0.33) amongst the models that did not have social network analysis features added.

Table 10: Adaboost prediction results, without network analysis features

Adaboost without network analysis features					
Balancing technique	**Precision**	**Recall**	**F1 score**	**Model AUC**	**Model accuracy**
Oversampling	0.11	**0.69**	0.19	82.8%	82.2%
SMOTE	0.34	0.33	**0.33**	82.6%	96.0%
Undersampling	**0.62**	0.17	0.26	**83.0%**	**97.1%**
No balancing	**0.62**	0.17	0.26	**83.0%**	**97.1%**
Average	0.42	0.34	0.26	82.3%	93.1%
Adaboost with network analysis features					
Balancing technique	**Precision**	**Recall**	**F1 score**	**Model AUC**	**Model accuracy**
Oversampling	0.11	0.69	0.19	83.0%	82.0%
SMOTE	0.71	0.58	0.64	94.1%	98.0%
Undersampling	0.22	**0.86**	0.34	94.2%	89.9%
No balancing	**0.96**	0.58	**0.72**	**95.4%**	**98.6%**
Average	0.50	0.68	0.47	91.7%	92.1%

Adding network analysis features, improved mean recall and F1 scores by 0.34 and 0.21 points respectively. Undersampling had the highest recall score, although the precision was relatively small. Oversampling technique also performed poorly with no improvements in F1 score after the network analysis features were added. Best results were obtained when no re-balancing techniques were used. It had the minority class classification F1 score of 0.72 points. Adaboost average F1 score for minority class stood at 0.47.

As one can see from Figure 13, network analysis features had high importance when predicting whether a contract belongs to a minority or majority class. Supporting the results from previous model analysis.

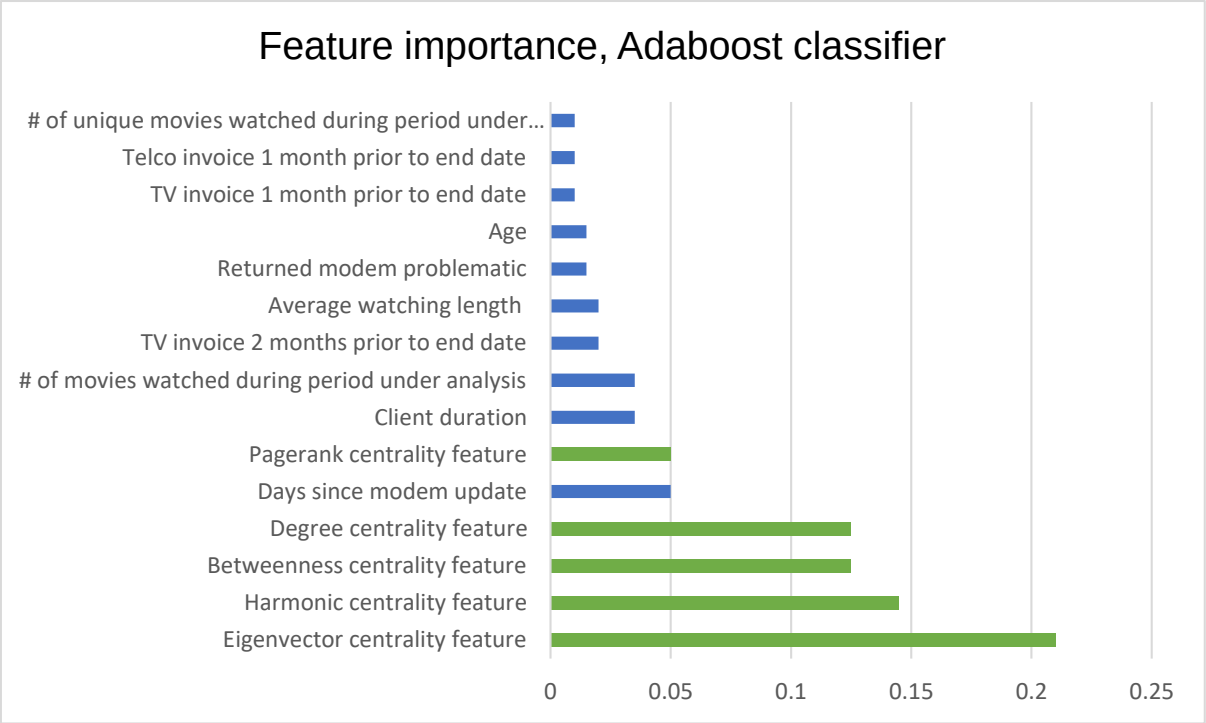


Figure 13: Unbalanced Adaboost classifier feature importance in the conducted model. Network analysis features were added and they are marked as green, whereas customer attributes are marked as blue.

Eigenvector centrality and harmonic centrality features had the highest importance for the Adaboost model.

6.1.5 Gradient boost classifier

Gradient boost classifier model predicted minority class with mean F1 score of 0.30 and mean recall of 0.36 meaning that out of all churning contracts, only 36% were correctly classified. Among the models without network analysis features added, best performance was achieved when no rebalancing techniques were used with recall value of 0.37 and F1 score of 0.36.

Table 11: Gradient Boost classifier prediction results, without network analysis features

Gradient boost classifier without network analysis features					
Balancing technique	**Precision**	**Recall**	**F1 score**	**Model AUC**	**Model accuracy**
Oversampling	0.13	**0.69**	0.22	**85.2%**	85.1%
SMOTE	0.71	0.18	0.29	85.0%	97.2%
Undersampling	0.35	0.37	**0.36**	83.8%	96%
No balancing	**0.75**	0.19	0.31	84.8%	**97.3%**
Average	0.49	0.36	0.30	84.7%	93.9%
Gradient boost classifier with network analysis features					
Balancing technique	**Precision**	**Recall**	**F1 score**	**Model AUC**	**Model accuracy**
Oversampling	0.35	**0.87**	0.5	**97.0%**	95.0%
SMOTE	0.88	0.42	0.57	95.8%	98.0%
Undersampling	0.81	0.57	0.67	95.3%	98.3%
No balancing	**0.98**	0.58	**0.73**	96.5%	**98.7%**
Average	0.76	0.61	0.62	96.2%	97.5%

Adding network analysis features, improved mean recall and F1 scores by 0.25 and 0.32 points respectively. Dataset without any balancing yielded the best results with an F1 value of 0.73 and recall value of 0.61 for minority class prediction. Average recall and F1 values across all models stood at 0.61 and 0.62 points respectively.

As one can see from Figure 14, network analysis features had high importance when predicting whether a contract belongs to minority or majority class. Supporting the results from previous model analysis.

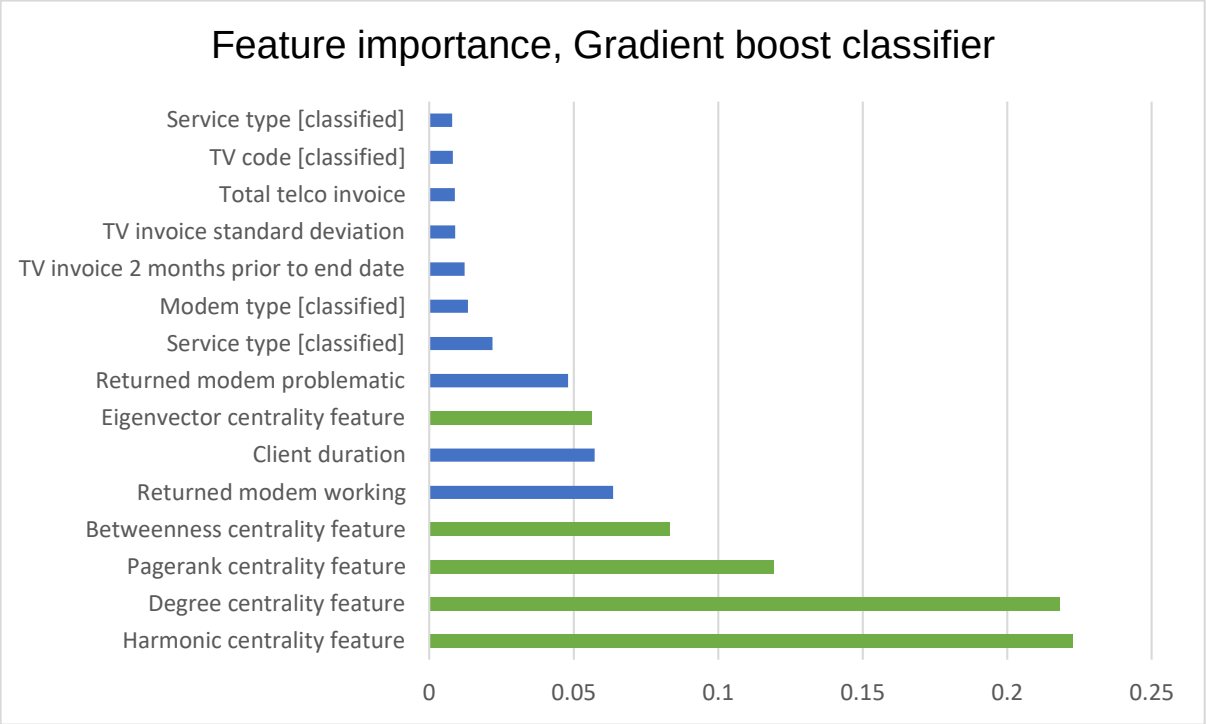


Figure 14: Unbalanced Gradient Boost classifier feature importance in the conducted model. Network analysis features were added and they are marked as green, whereas customer attributes are marked as blue.

Harmonic centrality and degree centrality features had the highest importance for the Gradient boost classifier model.

6.2 Prediction models comparison

To compare all of the models used for the calculations, a comparative graph was conducted. As seen from Figure 15, when no network analysis features were added, various models had somewhat similar performance. Average minority class prediction F1 score stood at 0.28 and XGBoost had the best results with a F1 score of 0.4.

Table 12: Best performing models with and without network analysis features

Best performing models without network analysis features						
Model	**Balancing technique**	**Precision**	**Recall**	**F1 score**	**Model AUC**	**Model accuracy**
Random Forest	SMOTE	0.44	0.27	0.34	81.8%	96.7%
XGBoost	SMOTE	0.47	0.34	0.4	83.3%	96.8%
Adaboost	SMOTE	0.34	0.33	0.33	82.6%	96.0%
logistic regression	SMOTE	0.3	0.33	0.32	80.1%	95.5%
Gradient boost classifier	Undersampling	0.35	0.37	0.36	83.8%	96%
Best performing models with network analysis features						
Model	**Balancing technique**	**Precision**	**Recall**	**F1 score**	**Model AUC**	**Model accuracy**
Random Forest	Undersampling	0.87	0.43	0.57	95.8%	98.0%
XGBoost	SMOTE	0.94	0.73	0.82	97.0 %	99.0%
Adaboost	No balancing	0.96	0.58	0.72	95.4%	98.6%
Gradient boost classifier	No balancing	0.98	0.58	0.73	96.5%	98.7%

As one can see from Table 12, best performance was achieved with the XGBoost model, using SMOTE. Adding features improved predictive power for Random forest, XGBoost, Adaboost and Gradient boost classifier. Highest performing logistic regression model was not added to the table, as performance dropped drastically across all rebalancing techniques after adding network analysis features. For more information please refer to Table 7.

On Figure 16 one can see models performances regarding prediction and recall results of a minority class.

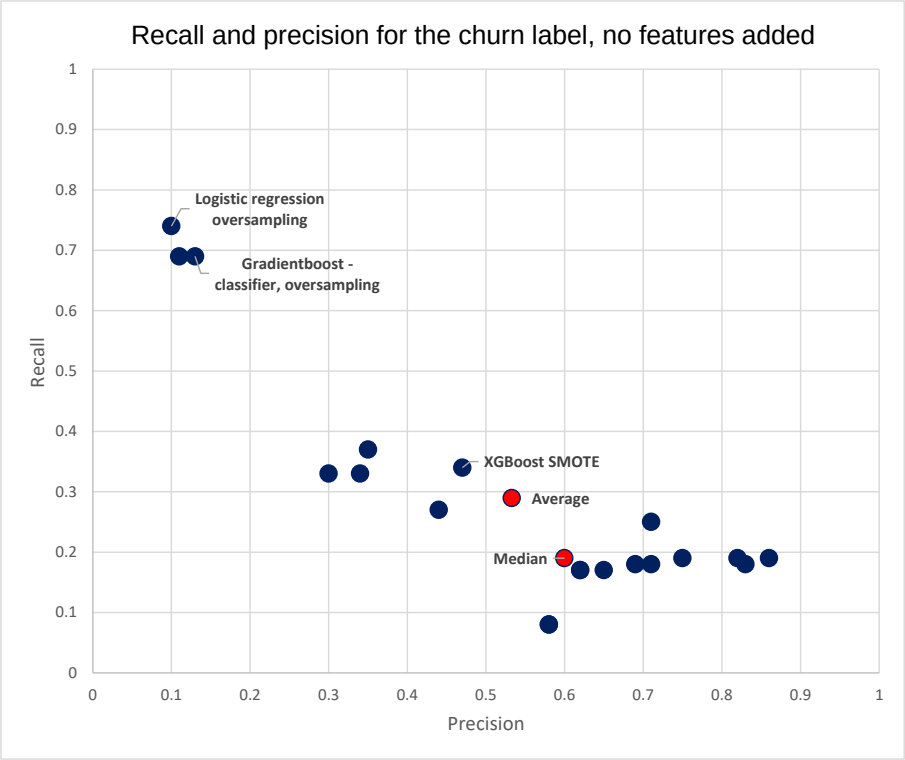


Figure 15: Minority class recall and precision results comparison before network analysis features were considered

When network analysis features were added, best performing models stayed on top, whereas predicting power from logistic regression models deteriorated.

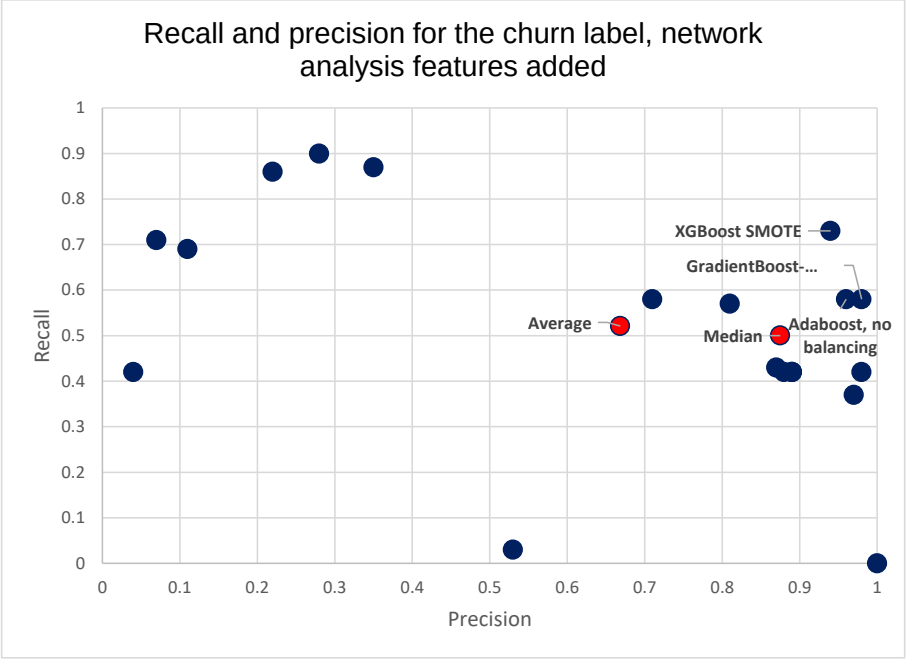


Figure 16: Minority class recall and precision results comparison after network analysis features were considered

To understand the importance of adding network analysis features, graph 17 was conducted. It shows how much has the specific model improved when network analysis features are added.

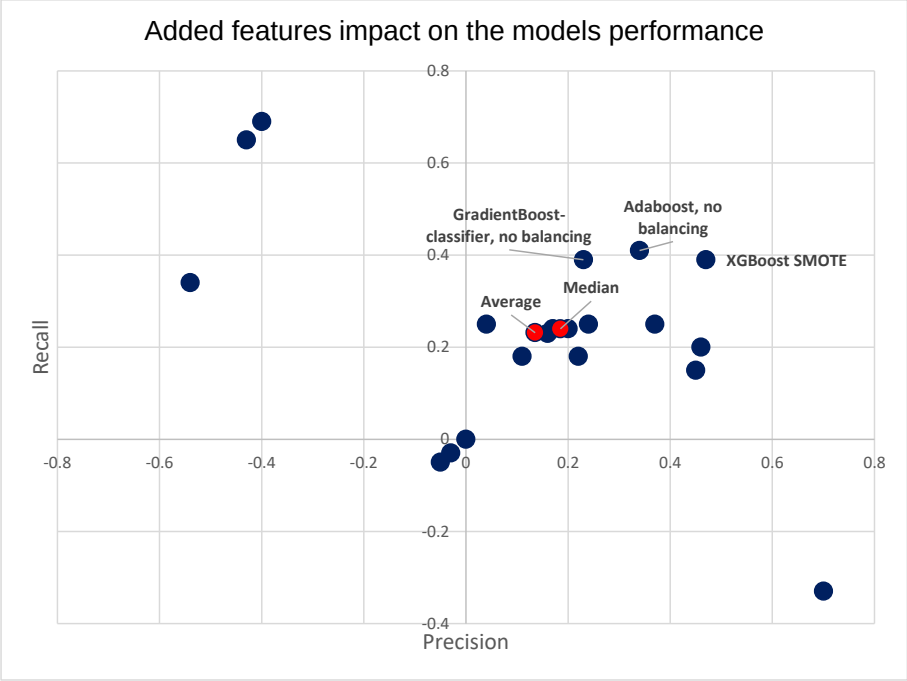


Figure 17: Difference in minority class prediction performance comparison after network features were considered

With added features precision improved in average of 0.14 and recall by 0.23 points. This shows that within this experiment, adding network analysis features improves the predictive power of minority class even for services where the service itself doesn't indicate which customers are linked between each-other. Although further analysis is required, the results are promising.

Our task was to predict which customers churned, based on the information that was collected 30 or 90 days prior to churning, depending on whether the feature was related to network analysis. Analysis was carried out as described in the Chapter 5

Evaluation metric values presented in this Chapter were calculated by using trained models from the k-folds on the holdout dataset. More information could be found in Figure 9. Please note that only minority class prediction evaluation metrics are presented as it is more insightful, considering that the evaluation dataset was unbalanced (approx. 10,000 non-churners to 300 churners).

7 Conclusions and Future work

7.1 Conclusions:

This thesis analysed the churning characteristics of clients who are using TV services in a Nordic telecommunication company. Churning characteristics were analysed by inspecting customer demographic characteristics, details of the product, contractual information, Invoices and call logs. For the latter, the features included whether customer friends churned 60-120 days before the agreed end date/churning date and how important were the churners according to centrality values.

In the literature review, churn within the telecommunication sector is covered in different aspects. The Chapter also covered the social network analysis methods and results from the previously conducted studies, where network analysis was used. This is the first time TV customer churn has been analysed based on call logs to the best of the authors' knowledge.

Data was gathered during 2020, and the period under analysis started with April of 2020 and ended with December of 2020. Three months of data were analysed for every customer. For churners, the ending date was the churning date, whereas for other customers, the end date was chosen randomly between April and December to mimic real life. The data set consisted of 50,000 clients, of which approximately 1500 churned.

The descriptive analysis reveals a few noteworthy things. Firstly more than 89% of customers had less than ten calls between them, and on average, every customer within the base was connected to 13 nodes. Assuming that most people have more than 13 friends/acquaintances may indicate that customers' communication via phone does not cover all the crucial actors and customers may use other communication methods. This idea is also supported by the fact that the average clustering coefficient was 0.000197, meaning that the analysed graph was sparse.

The churning behaviour was predicted while using five different machine learning models: Random Forest, XGBoost, Adaboost, Logistic regression, and Gradient Boost classifier. The data was unbalanced between majority and minority class. Therefore three rebalancing techniques were used to improve the result: oversampling, undersampling, and synthetic minority oversampling technique (SMOTE).

Every model was trained with two datasets. One of the datasets included network analysis features, whereas the other did not. The Impact of the added features was calculated for models, and the results were following:

Models without network analysis features had an average minority class F1 value of 0.28 and average recall value of 0.29, meaning that out of 100 churning users, 29 were correctly identified. The Best result was obtained by using the XGBoost model with the synthetic minority oversampling technique. XGBoost models in general performed the best with an average F1 score without network analysis features of 0.34 and an average F1 score with network analysis features of 0.59.

Adding network analysis features improved minority class precision on average by 0.23 points and recall by 0.14 points, indicating that adding network analysis features could help to predict churning customers.

In conclusion, the findings support the idea that customer churn could be predicted while using network analysis features. Although the results may be overfitted to this specific dataset, the idea remains promising as the improvement was noteworthy.

7.2 Future directions:

Considering the analysis results, a few aspects should be noted. First of all, a limited amount of research has been carried out on this topic, therefore more research would need to be conducted on the subject, especially with datasets, that include call logs from all of the clients instead of a limited set. In that case, clusters of customers could be formed with higher accuracy and more client-specific features and conclusions could be generated.

Secondly, a better understanding of features which contribute most to churning customers should be analysed further so that a better set of features could be picked at first. Currently used features may have been too static and client attitude towards the firm may not be fully grasped as not enough features are considering client attitude towards the firm during the last three months before churning. Therefore more dynamic features should be considered when additional research is carried out on the same topic.

References

[1] Eesti telekommunikatsiooni turust konkurentsiameti hinnang 2021, 2021. URL https://www.konkurentsiamet.ee/sites/default/files/Dokumentide-failid/konkurentsiameti_hinnang_eesti_telekommunikatsiooni_turust_0.pdf.

[2] D. A. Aaker. The value of brand equity. *Journal of business strategy*, 1992.

[3] B. Anckar and D. D'incau. Value creation in mobile commerce: Findings from a consumer survey. *Journal of Information Technology Theory and Application (JITTA)*, 4(1):8, 2002.

[4] A. Backiel, B. Baesens, and G. Claeskens. Predicting time-to-churn of prepaid mobile telephone customers using social network analysis. *Journal of the Operational Research Society*, 67(9):0, 2016.

[5] C. Bergqvist and J. Townsend. Enforcing margin squeeze ex post across converging telecommunications markets. 2015.

[6] B. Bollobás. *Modern graph theory*, volume 184. Springer Science & Business Media, 2013.

[7] S. P. Borgatti, M. G. Everett, and J. C. Johnson. *Analyzing social networks*. Sage, 2018.

[8] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[9] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. 1998.

[10] J. Burez and D. Van den Poel. Crm at a pay-tv company: Using analytical models to reduce customer attrition by targeted marketing for subscription services. *Expert Systems with Applications*, 32(2):277–288, 2007.

[11] Z.-Y. Chen, Z.-P. Fan, and M. Sun. A hierarchical multiple kernel support vector machine for customer churn prediction using longitudinal behavioral data. *European Journal of operational research*, 223(2):461–472, 2012.

[12] C. Chu, G. Xu, J. Brownlow, and B. Fu. Deployment of churn prediction model in financial services industry. In *2016 International Conference on Behavioral, Economic and Socio-cultural Computing (BESC)*, pages 1–2. IEEE, 2016.

[13] T. Company. Telia company annual report 2019. https://www.teliacompany.com/globalassets/telia-company/documents/reports/2019/telia-company--annual-and-sustainability-report-2019.pdf, 2020.

[14] K. Coussement, S. Lessmann, and G. Verstraeten. A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry. *Decision Support Systems*, 95:27–36, 2017.

[15] K. Dasgupta, R. Singh, B. Viswanathan, D. Chakraborty, S. Mukherjea, A. A. Nanavati, and A. Joshi. Social ties and their relevance to churn in mobile telecom networks. In *Proceedings of the 11th international conference on Extending database technology: Advances in database technology*, pages 668–677, 2008.

[16] Y. Ding, E. Yan, A. Frazho, and J. Caverlee. Pagerank for ranking authors in co-citation networks. *Journal of the American Society for Information Science and Technology*, 60 (11):2229–2243, 2009.

[17] F. Eichinger, D. D. Nauck, and F. Klawonn. Sequence mining for customer behaviour predictions in telecommunications. In *Proceedings of the Workshop on Practical Data Mining at ECML/PKDD*, pages 3–10, 2006.

[18] P. Ferreira, R. Telang, and M. G. De Matos. Effect of friends' churn on consumer behavior in mobile networks. *Journal of Management Information Systems*, 36(2):355–390, 2019.

[19] Y. Freund, R. E. Schapire, et al. Experiments with a new boosting algorithm. In *icml*, volume 96, pages 148–156. Citeseer, 1996.

[20] J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[21] D. L. Garc'ıa, A. Nebot, and V. Alfredo. Visualizing pay-per-view television customers churn using cartograms and flow maps. In *ESANN 2013 Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pages 24–26, 2013.

[22] D. L. Hansen, B. Shneiderman, M. A. Smith, and I. Himelboim. Social network analysis: measuring, mapping, and modeling collections of connections. *Analyzing social media networks with NodeXL: insights from a connected world. Elsevier Inc, Burlington*, pages 31–52, 2011.

[23] M. Hassouna, A. Tarhini, T. Elyas, and M. S. AbouTrab. Customer churn in mobile markets a comparison of techniques. *arXiv preprint arXiv:1607.07792*, 2016.

[24] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

[25] V. Jagannath. Random forest template for tibco spotfire®, Mar 2017. URL `https://community.tibco.com/wiki/random-forest-template-tibco-spotfire`.

[26] R. Kalakota, M. Robinson, and D. Tapscott. *E-business 2.0: Roadmap for Success*. Addison-Wesley Professional, 2001.

[27] P. D. Kusuma, D. Radosavljevik, F. W. Takes, and P. van der Putten. Combining customer attribute and social network mining for prepaid mobile churn prediction. In *Proc. the 23rd Annual Belgian Dutch Conference on Machine Learning (BENELEARN)*, pages 50–58, 2013.

[28] M. Marchiori and V. Latora. Harmony in the small-world. *Physica A: Statistical Mechanics and its Applications*, 285(3-4):539–546, 2000.

[29] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444, 2001.

[30] A. Mislove, B. Viswanath, K. P. Gummadi, and P. Druschel. You are who you know: inferring user profiles in online social networks. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 251–260, 2010.

[31] S. Mitrović, B. Baesens, W. Lemahieu, and J. De Weerdt. tcc2vec: Rfm-informed representation learning on call graphs for churn prediction. *Information Sciences*, 2019.

[32] F. Mosteller and J. W. Tukey. Data analysis, including statistics. *Handbook of social psychology*, 2:80–203, 1968.

[33] A. A. Nanavati, S. Gurumurthy, G. Das, D. Chakraborty, K. Dasgupta, S. Mukherjea, and A. Joshi. On the structural properties of massive telecom call graphs: findings and implications. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 435–444, 2006.

[34] V. Nicosia, J. Tang, C. Mascolo, M. Musolesi, G. Russo, and V. Latora. Graph metrics for temporal networks. In *Temporal networks*, pages 15–40. Springer, 2013.

[35] I. Nitzan and B. Libai. Social effects on customer retention. *Journal of Marketing*, 75(6): 24–38, 2011.

[36] Ofcom. Telecommunications data revenues, volumes and market share update q3 2019. `https://www.ofcom.org.uk/research-and-data/data/statistics/stats20`, 2019.

[37] A. Oyeniyi, A. Adeyemo, A. Oyeniyi, and A. Adeyemo. Customer churn analysis in banking sector using data mining techniques. *Afr J Comput ICT*, 8(3):165–174, 2015.

[38] C. Prell. *Social network analysis: History, theory and methodology*. Sage, 2012.

[39] J. Prince and S. Greenstein. Does service bundling reduce churn? *Journal of Economics & Management Strategy*, 23(4):839–875, 2014.

[40] T. Raeder, O. Lizardo, D. Hachen, and N. V. Chawla. Predictors of short-term decay of cell phone contacts in a large scale communication network. *Social Networks*, 33(4):245–257, 2011.

[41] Y. Richter, E. Yom-Tov, and N. Slonim. Predicting customer churn in mobile networks through analysis of social groups. In *Proceedings of the 2010 SIAM international conference on data mining*, pages 732–741. SIAM, 2010.

[42] Satishgunjal. Tutorial: K fold cross validation, Dec 2020. URL `https://www.kaggle.com/satishgunjal/tutorial-k-fold-cross-validation`.

[43] L. Tang and H. Liu. Community detection and mining in social media. *Synthesis lectures on data mining and knowledge discovery*, 2(1):1–137, 2010.

[44] W. Verbeke, K. Dejaeger, D. Martens, J. Hur, and B. Baesens. New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European journal of operational research*, 218(1):211–229, 2012.

[45] X. Zhang, J. Zhu, S. Xu, and Y. Wan. Predicting customer churn through interpersonal influence. *Knowledge-Based Systems*, 28:97–104, 2012.

I. License

Non-exclusive licence to reproduce thesis and make thesis public

I, Martin Käärik,

(*author's name*)

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to

reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

A Network-Based Model for Television Services Churn Prediction,

(*title of thesis*)

supervised by Shakshi Sharma, Rajesh Sharma.

(*supervisor's name*)

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Martin Käärik

11/05/2021