New Trends and Issues Proceedings on Advances in Pure and Applied Sciences

No 12 (2020) 57-67

www.propaas.eu

User behaviour analysis and churn prediction in ISP

Ahmet Turkmen*, Aalborg University, Fredrik Bajers Vej 5, 9220Aalborg, Denmark
Cenk Anil Bahcevan, Turknet, Sisli, 34394Istanbul, Turkey
Youssef Alkhanafseh, Turknet, Sisli, 34394Istanbul, Turkey
Esra Karabiyik, Turknet, Sisli, 34394Istanbul, Turkey

Suggested Citation:
Turkmen, A., Bahcevan, C. A., Alkhanafseh, Y. & Karabiyik, E. (2020). User behaviour analysis and churn prediction in ISP. *New Trends and Issues Proceedings on Advances in Pure and Applied Sciences.* (12), 57–67.

Received date December 12, 2019; revised date;February 15, 2020 accepted date April 2, 2020.
Selection and peer review under responsibility of Prof. Dr. Dogan Ibrahim, Near East University, Cyprus.

Abstract

There is no doubt that customer retention is vital for the service sector as companies' revenue is significantly based on their customers' financial returns. The prediction of customers who are at the risk of leaving a company's services is not possible without using their connection details, support tickets and network traffic usage data. This paper demonstrates the importance of data mining and its outcome in the telecommunication area. The data in this paper are collected from different sources like Net Flow logs, call records and DNS query logs. These different types of data are aggregated together to decrease the missing information. Finally, machine learning algorithms are evaluated based on the customer dataset. The results of this study indicate that the gradient boosting algorithm performs better than other machine learning algorithms for this dataset.

Keywords: Data analysis, customer satisfaction, subscriber churn, machine learning, telecommunication.

* ADDRESS FOR CORRESPONDENCE: **Ahmet Turkmen,** Aalborg University, Fredrik Bajers Vej 5, 9220Aalborg, Denmark
E-mail address: atu@es.aau.dk

1. Introduction

Companies from various sectors, such as banking, retailing, insurance and internet service providers, are having difficulties with customer retention because they have a limited amount of knowledge about their customers. Having sufficient information about customers will increase revenue and decrease complaints about the company. The required knowledge can be mined from their anonymised data which includes a variety of different information about them. In this research, we aim to analyse and understand customers' intentions as much as possible by applying machine learning algorithms with detailed visualisation and using various data sources offered by the internet service providers (TurkNet). It is no secret that customer retention is more significant and tougher than gaining a new customer. Hence, having a system, where it can automatically propose potential customer who is at risk of leaving the company, is crucial and doable with data mining. Even more significant than having a system is having clean data from which we can get results. However, it is nearly impossible to have clean data before hand because every company has different data points where customer data are stored in various systems within the company. Furthermore, there are some companies that do not rely on recent technologies and cause difficulties when they have to be integrated into the new system.

Despite other articles in the community, our research investigates customers in two different ways, one is trying to predict customers who are willing to churn and the other is clarifying the services (Netflix, Spotify, etc.) that are mostly used by Turk Net's customers in order to make a special agreement with those services.

2. Related work

The community is suffering from the lack of data used in churn prediction and behaviour analysis studies in many sectors. It is a significant advantage to have data from the company itself. Researchers in this area have conducted studies which generally tackle the problem of old and non concrete data. Although there is a lack of data in the community, there is plenty of training in customer churn prediction in many sectors like banking, insurance, telecommunication and internet service providers.

[1] reached the highest accuracy of 93% by using gradient boosting implementations of XG Boost [2]. In their study, they also used the social network analysis of users as a feature, which is rarely seen in the previous churn prediction studies. Among the features extracted from social network analysis, they preferred to add additional features to enrich feature sets, such as SMS, MMS and number of calls. Large data technologies can aid in handling an enormous amount of data when the volume of data does not fit the existing systems. This happened in their study since they dealt with more than 70TB of unstructured data and they used new large data tools, such as Hadoop. [3] did not use all features as input; they selected the combinations of the most important features using the k-means clustering technique. [3] tested decision tree, logistic regression and neural network on the dataset. They achieved an accuracy of 89.08% in churn prediction rate by using feed-forward neural networks. [4] proposed the usage of feed-forward neural network architecture with particle swarm optimisation. Particle swarm optimisation provides weight tuning for the proposed neural network Marikannan, 2019), and this model achieved a better accuracy of 92.90% on their local dataset. Their [4] local dataset belonged to a telecommunication company in Jordan. [5] proposed a genetic algorithm methodology which used a rough set theory. He combined the success of rough set theory with the feed-forward neural network and the genetic algorithm outperformed the other algorithms with an accuracy of 98%.

All studies examined were concerned with feature engineering because of the main nature of churn prediction. The main features are frequency of usage, the volume of the network, short message usage and phone numbers. Although plenty of studies are carried out in the area of predicting churners, the lack of data problem is raised due to privacy concerns.

3. Problem

Customer satisfaction and revenue of the company could increase by mining raw database and log data. However, there are several issues regarding the source of data and giving meaning to them. First of all, the data, which are collected from real scenarios, are generally not sufficient enough to have direct processing for machine learning algorithms. Furthermore, the data sources are not completely compatible with new data processing methods, such as creating clusters and indexing data on top of it. Hence, some data are sampled with all features from big data collections and they are examined according to one's needs.

The data used in this study contain missing information, misspelling features that point to the same thing and are no streamlined to query newly generated data. Since the behaviour of users might change according to the trend on the internet, having continuous data streaming would make our analysis process much smoother. To overcome the missing information which existed in the sample dataset, different kinds of methods have been implied according to the type of data. Although applied methods (mean, median and mode) are not sophisticated, the approach helps to recover information, and its effects will be discussed. Another issue regarding to sampled dataset is that the units of features were not normalised at all. Applied normalisation approaches could be explained and formulated as the following.

Min-max normalisation basically calculates the minimum and maximum values of a feature set, and in order to standardise the value, it extracts the existing value with a minimum one and divides it by the result of maximum minus minimum value as follows:

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Figure 1 shows the min–max normalisation diagram.

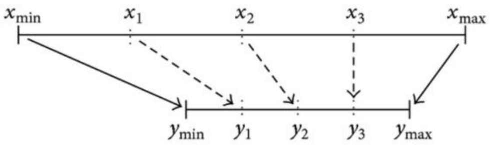


Figure 1. Representation of min–max []

Z-score is another normalisation algorithm that basically focuses on avoiding outlier issues in the dataset with the following formula. It standardises the data by extracting the value from the mean and dividing it by the standard deviation of the feature set as follows:

$$\left(z = \frac{v_i - u}{\sigma} \right)$$

The Standard Scaler was applied to the data for a more general normalisation process that was implemented on the scikit-learn [6] machine learning and feature engineering library. The data became clearer when the discussed procedures used on the data enabled us to make further investigations and observations.

Another problem was having different data sources that were somehow not connected to each other, and which created a lack of communication between departments in the company. In order to deal with it, the following mechanism was created to match the required information from the sources and feed the cluster which is mainly created for handling data. Figure 2shows the simplified case where data are collected from different sources and fed into the cluster which consists of several servers and runs Elastic search on them (It is an open source tool to index data in efficient way.). The main problems were overcome by implementing the described procedures, and while implementing these approaches, it became clear that having a pipeline procedure would be much nicer. The following sections discuss about the pipeline, models, dataset, features on the dataset and behaviour analysis which is driven from DNS queries.

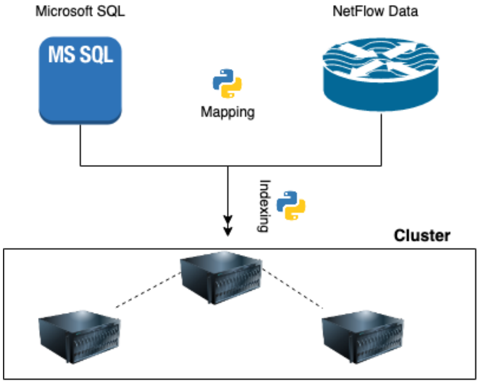


Figure 2. Simplified data combining process

4. Modelling

The modelling section consists of several main steps which analyse, pre-process and process data. Raw data may be unstructured, structured or semi-structured. In this section, we focus mainly on unstructured and semi-structured data. The data include information regarding internet usage, DNS queries and number of complaints to the call centre. However, they are independent of each other and there is no way to use them in an important matter. In the pre-processing step, missing information is clarified, spelling mistakes are fixed and all features within the data are renamed. There is a dilemma when there is a large amount of data and an increase in the likelihood of having wrong observation of the data; yet, correct observation can be gained by looking at a small amount of data. However, in this case, the observation might not be enough to predict the primary goal. The steps explained in Figure 3shows how pre-processing and processing data took place in most of the case.

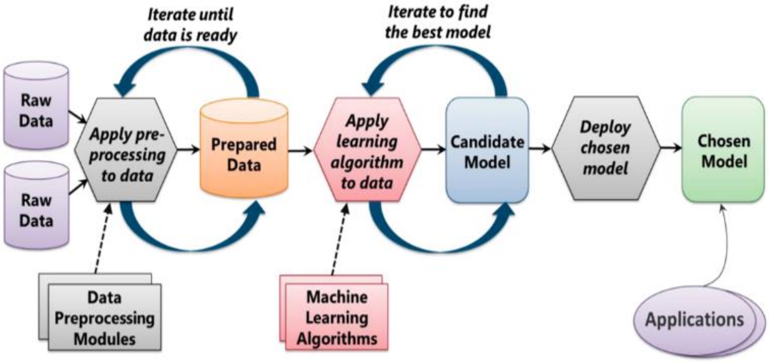

Iterate until data is ready
Iterate to find the best model
Raw Data
Raw Data
Apply pre-processing to data
Prepared Data
Apply learning algorithm to data
Candidate Model
Deploy chosen model
Chosen Model
Data Preprocessing Modules
Machine Learning Algorithms
Applications

Figure 3. Machine learning process. Source: 'Introduction to azure machine learning' by David Choppel

The step 'apply learning algorithm to data' consists of all machine learning algorithms which are somehow tested against different scenarios or promises to have better accuracy according to the literature review. In this study, we created a pipeline which provides an end-to-end continuous process, from gathering data from the source (the cluster) and implementing all required steps, that is one click away.

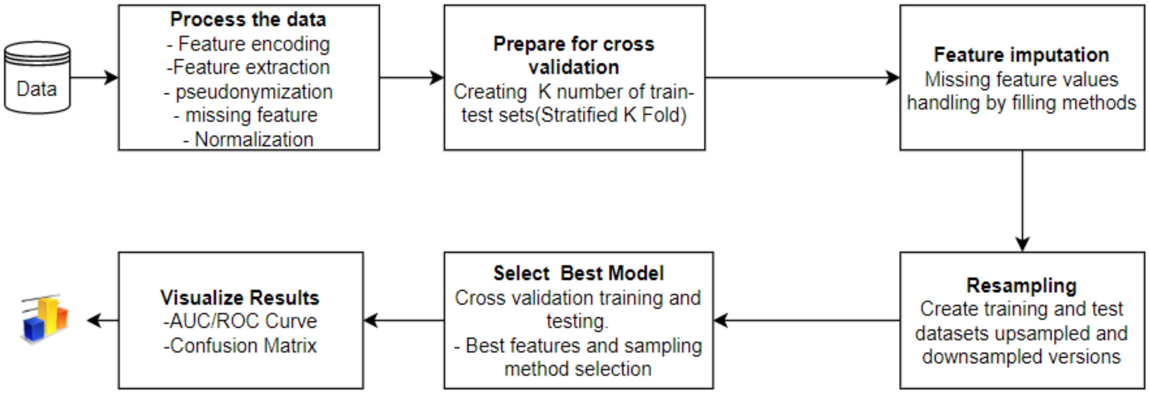

Data
Process the data
- Feature encoding
-Feature extraction
- pseudonymization
- missing feature
- Normalization
Prepare for cross validation
Creating K number of train-test sets(Stratified K Fold)
Feature imputation
Missing feature values handling by filling methods
Resampling
Create training and test datasets upsampled and downsampled versions
Select Best Model
Cross validation training and testing.
- Best features and sampling method selection
Visualize Results
-AUC/ROC Curve
-Confusion Matrix

Figure 4. Pipeline construction of the process

Figure 4 shows the pipeline for data processing in the projects which involve machine learning algorithms; in the first step, pre-processing takes place, such as pseudonymisation; then the encoding of categorical values (such as having an invoice on paper or receiving it as an e-mail (0,1)) takes place in the feature set and in the selection of features. Since the number of churners is greater than non-churners, in the second step, a stratified k-fold approach is implied, which is basically a diving test and training sets for each strata after applying some shuffling in order to decrease the bias on sampled data. For each iteration, a different portion of data is selected as training and test sets, which decreases the likelihood of having wrong observation on data. After that, each individual step in data processing is shown in Figure 4, which is considered independently whether the configuration of run enabled the options or not. Most of the machine learning algorithms are tested with the same testing dataset with a stratified k-fold in order to compare the differences and effects of pre-processing steps in the algorithms.

4.1. Dataset and features

The dataset contains more than 35 features that are aggregated from different sources and are converted to comma-separated value (CSV) types and are written to the SQL Server database of the companies. The features covering a3-month background check is presented in Table 1.

Table 1 Most important features

Feature	Explanation
number_of_service_channel_complaints	Number of complaints about the service
invoice_review_support_count	Number of invoice support
internet_connection_error_complain_count	Number of no internet connection complaints
number_of_call_centers	Number of call centres

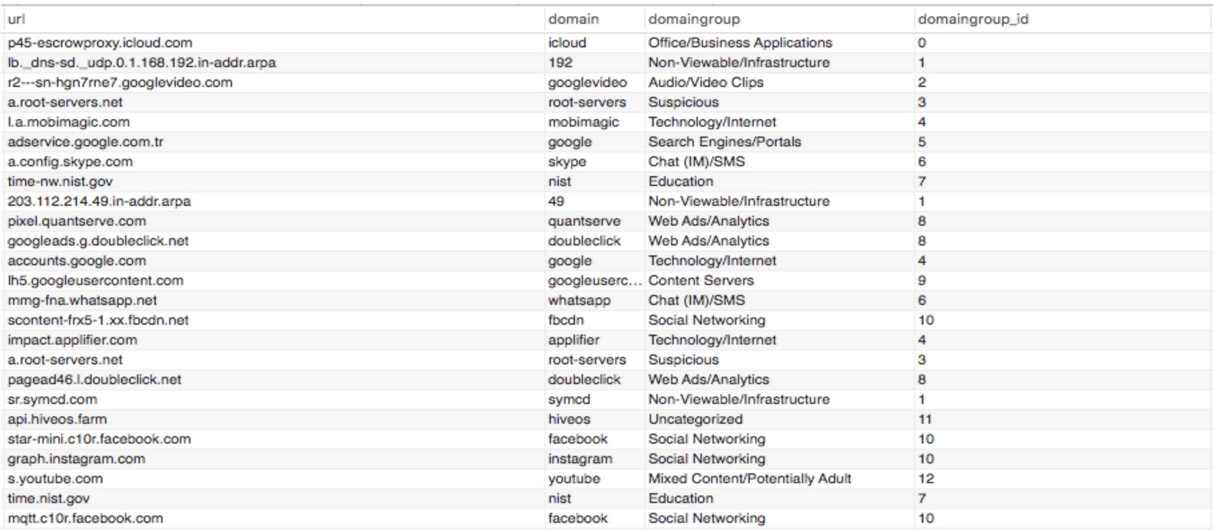
Apart from the features seen in Table 1, there are plenty of features that contain information of the last three months about the customer. Clarifying the most significant features is made by implementing the feature selection part of the random forest algorithm. The training dataset contains 24,935 non-churners and 3,847 churners; despite the training dataset, the test dataset contains 46,138 non-churners and 1,728 churners. Detailed information regarding how those datasets are used and evaluated is given in the evaluation section.

5. Behaviour analysis

Behaviour analysis of the customer is carried out by analysing DNS requests, which are (without any sensitive information of customer) collected and fed into another index in the cluster system. Every day, there are over 4TB of DNS requests from different domains name servers of the company which makes it harder to parse, eliminate and feed into another source from a variety of domain name servers. A DNS request consists of an IP address, a domain, a URL, a timestamp and many other fields; from a DNS query, the URL's category should be matched correctly per one query, for instance, if a DNS query contains 'www.facebook.com' URL, it should be matched with 'social networking'. Therefore, a web service, which updates itself according to changes in the content of the website, is used to categorise URLs. It was costly to make a DNS request to web service; hence, implementing bulk requests with a synchronisation increased the overall performance of the system.

To boost the speed of processing and cleaning data, some preliminary steps have been applied; for instance, when a user, who is surfing on the internet, has some DNS requests which are not made intentionally. To overcome this situation, advertisement URLs, some content delivery network data, extension requests, third-party requests, which do not seem right URLs, are filtered. Meanwhile, having local service which updates itself according to the web service might decrease overhead when dealing with requests to the web. In this sense, there is no need to check a URL which is already categorised; this concept is shown in Figure 5.

Figure 5 is an example of redundant data in DNS requests from the data point of view.



url	domain	domaingroup	domaingroup_id
p45-escrowproxy.icloud.com	icloud	Office/Business Applications	0
lb._dns-sd._udp.0.1.168.192.in-addr.arpa	192	Non-Viewable/Infrastructure	1
r2---sn-hgn7rne7.googlevideo.com	googlevideo	Audio/Video Clips	2
a.root-servers.net	root-servers	Suspicious	3
l.a.mobimagic.com	mobimagic	Technology/Internet	4
adservice.google.com.tr	google	Search Engines/Portals	5
a.config.skype.com	skype	Chat (IM)/SMS	6
time-nw.nist.gov	nist	Education	7
203.112.214.49.in-addr.arpa	49	Non-Viewable/Infrastructure	1
pixel.quantserve.com	quantserve	Web Ads/Analytics	8
googleads.g.doubleclick.net	doubleclick	Web Ads/Analytics	8
accounts.google.com	google	Technology/Internet	4
lh5.googleusercontent.com	googleuserc…	Content Servers	9
mmg-fna.whatsapp.net	whatsapp	Chat (IM)/SMS	6
scontent-frx5-1.xx.fbcdn.net	fbcdn	Social Networking	10
impact.applifier.com	applifier	Technology/Internet	4
a.root-servers.net	root-servers	Suspicious	3
pagead46.l.doubleclick.net	doubleclick	Web Ads/Analytics	8
sr.symcd.com	symcd	Non-Viewable/Infrastructure	1
api.hiveos.farm	hiveos	Uncategorized	11
star-mini.c10r.facebook.com	facebook	Social Networking	10
graph.instagram.com	instagram	Social Networking	10
s.youtube.com	youtube	Mixed Content/Potentially Adult	12
time.nist.gov	nist	Education	7
mqtt.c10r.facebook.com	facebook	Social Networking	10

Figure 1. URLs with categorisation

When Figure 5 is examined, it is very obvious that some URLs are not visited by the user, for instance, 'a.root-servers.net', 'time.nist.gov' and many other requests do not represent real cases. The data that is not cleaned cannot be used for behaviour analysis, as it may have wrong observations about customers. After the cleaning of data, as shown in Figure 5, most of the URLs have been removed, which either belongs to a background job or an advertisement or something else. It enables us to create clusters based on cleaned data using some clustering machine learning algorithms, such as *k*-means, mean-shift and agglomerative hierarchical clustering algorithms. However, although we have cleaned out redundant information on DNS requests, the volume of data is high. In this case, the proper algorithm to run, which obtains faster results, is the *k*-means clustering algorithm; we have implied clustering algorithms to categorise the DNS data; however, the number of clusters was clarified by elbow analysis which is a heuristic method of interpretation and understanding required by the number of clusters. The result of elbow analysis shows that having four clusters would be the best clustering approach for the data that we have. The algorithm was run and the following results were gathered in different clusters, as shown in Figure 6.

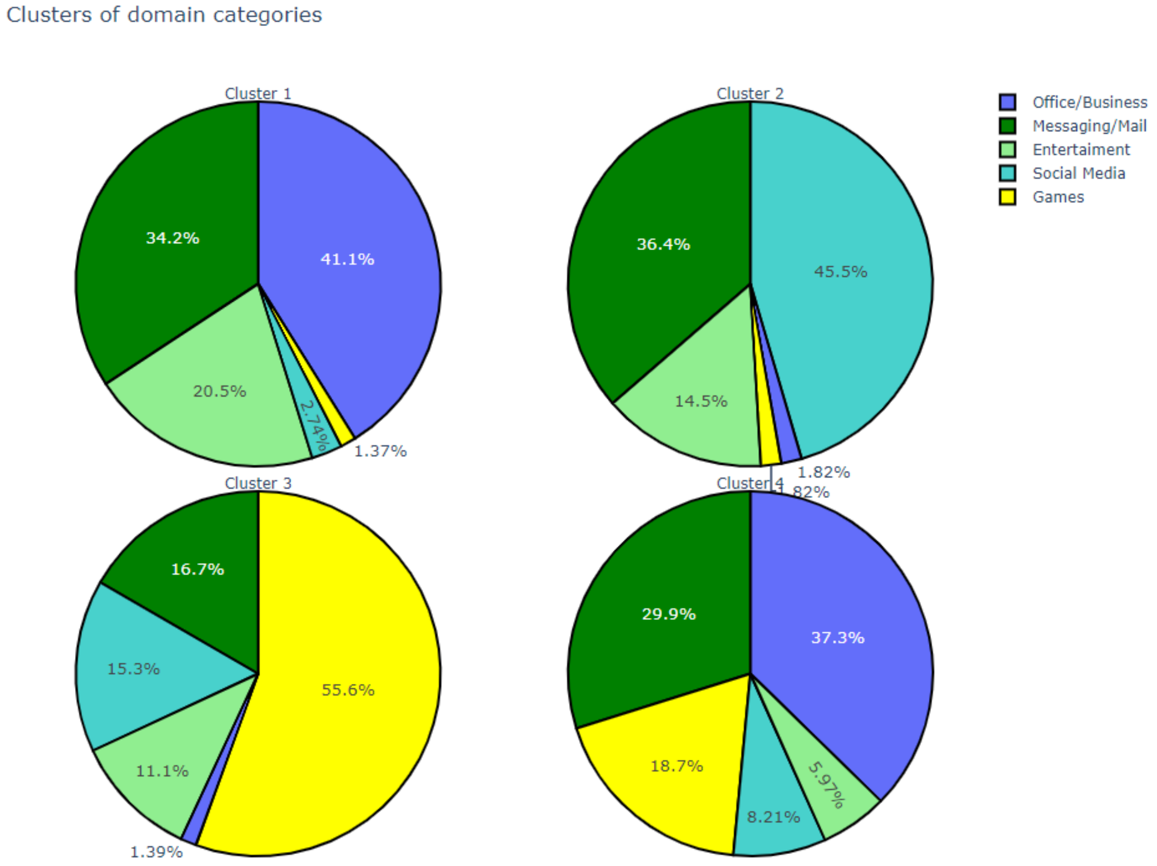


Figure 2. Clusters of domain categories

The four clusters and their detailed information show the percentage of people who are spending their time on which site category. It provided us a quiet vision about our customers' behaviour, according to detailed information on each cluster, our company proposed some kind of campaign to boost the number of customers and existing customers satisfaction as well. For example, if we observe that most of the traffic points have a new kind of application other than a known kind, then the next campaign could be based on that application. The four clusters in Figure 6 shows that customers are grouped from their traffic of game, social media, office and messaging. For instance, an internet service provider would understand the expectations of the network speed from customers; the customers in Cluster 3 expect much higher network speed because of their gaming habits. Apart from DNS request analysis, another approach regarding behaviour analysis that we analysed anonymously was Net Flow traffic and we realised that in some particular time intervals download data is much higher than the rest of the days as shown in Figure 7.

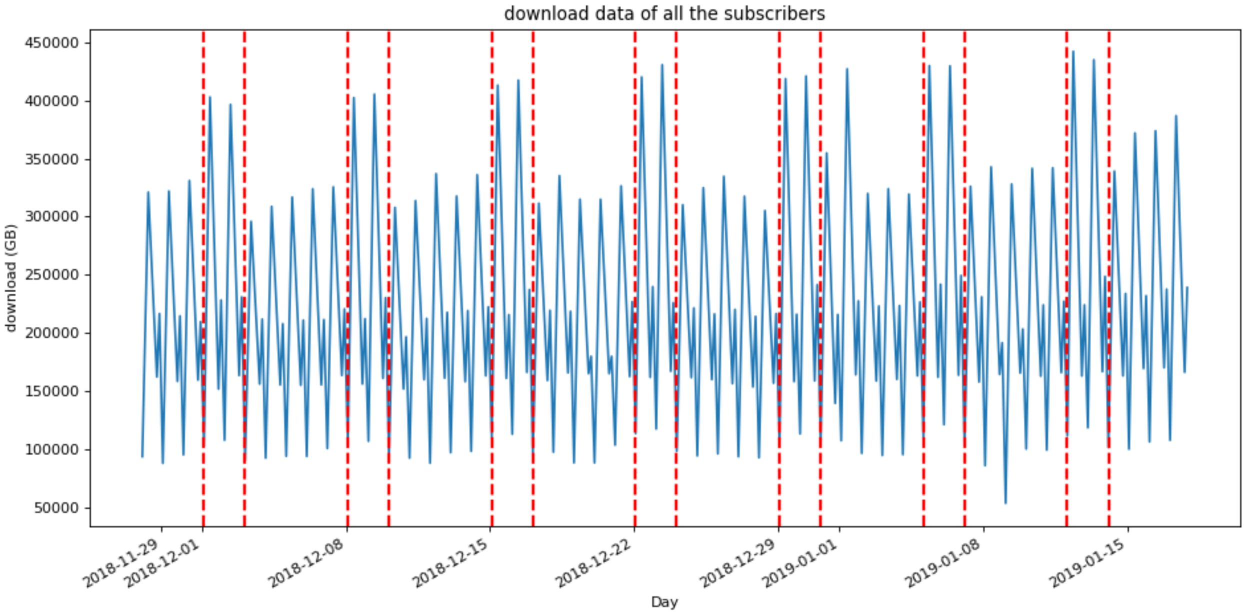


Figure 3. Customers' behaviour on downloading data

The customers could be categorised into different groups according to time periods, which can provide more information about the load on the network infrastructure of the company. When customers are divided into four distinct periods, 65% of our customers are active between 22.00 and 02.00 (10.00 PM – 02:00 AM). It illustrates the overview consumption of our resources and helps us to decide whether there is a need for increasing bandwidth on specified given times for specific channels.

The overall illustration regarding customers and their active periods are shown in Figure 8.

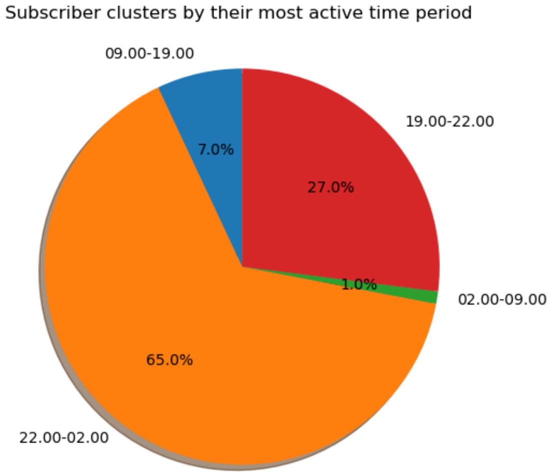


Figure 4. Subscriber clusters and their most active time period

All in all, behaviour analysis is not as simple as you think; however, it is an excellent way to understand your community and customer, and how they are react to your services. Observing new tendencies in the internet environment beforehand provides measurable success.

6. Evaluation

In this study, classic machine learning algorithms like logistic regression, decision tree, random forest, gradient boosting and k-nearest neighbours (KNN) are compared. While examining the results of the study, metrics like AUC/ROC, Matthew's correlation score and confusion matrix are more important than the accuracy score due to the imbalanced data on churn prediction studies. The high AUC score close to the one is showing high separability of the model. The accuracy score becomes irrelevant in measuring the quality of the model in the imbalanced dataset which has binary output expectations. Matthew's correlation score shows the success of binary classification. The confusion matrix shows the true positive, true negative, false positive, false negative numbers from the model predictions. The number of non-churners is extremely higher than the churners and, in general, the ratio of churners in internet service providers is 3%.

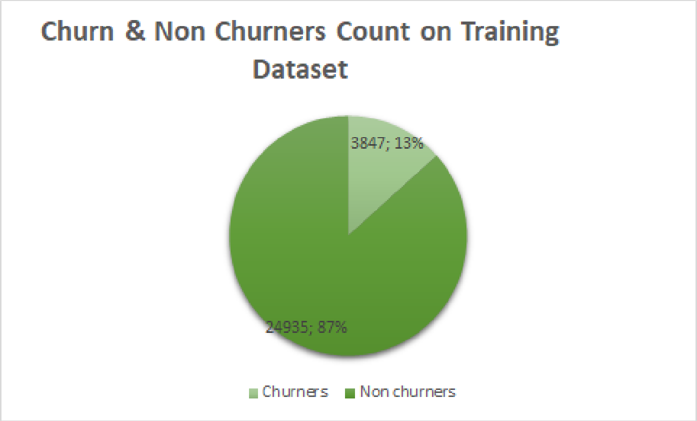
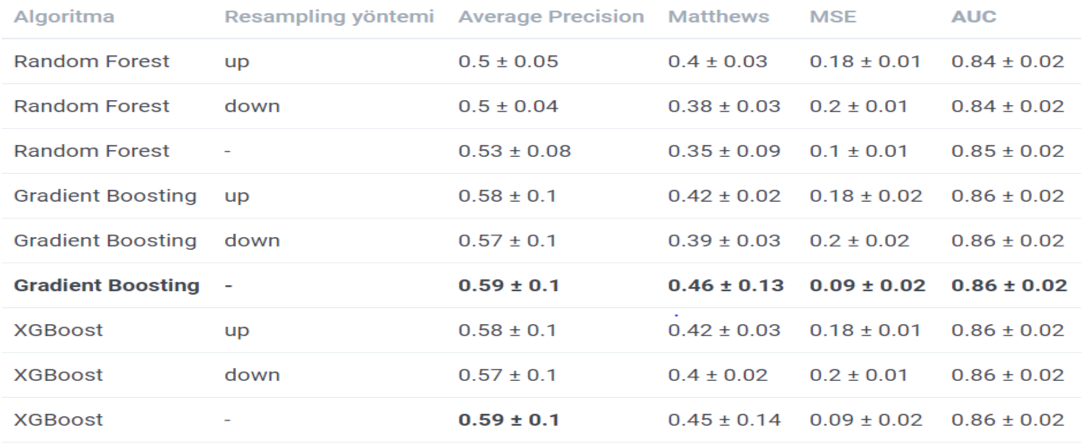


Figure 5. Training dataset distribution

Figure 9 shows the numbers and percentage of churners and non-churners. The training dataset contains 24935 non-churners and 3847 churners. All the models in this study are tested with 46.138 non-churners and 1728 churners. It is a fact that the number of clients who are churning is much higher than non-churners; however, the ratio cannot be the same in the training dataset. The main reason is that the model itself can have a problem with over fitting or under fitting since the percentage of churners is around 3%, and if we assume that everyone will not leave the company, we will have an accuracy of 97%. To prevent well-known machine learning drawbacks, the training dataset is balanced although the test dataset is kept as a real case (with 3% of churners). Sampling methods were used to reduce the level of imbalance of the dataset. The data are resampled for increasing the number of churners and closing the gap between churners and non-churners numbers for training datasets only. Down sampling the number of non-churners in the training dataset provides a more balanced dataset.

The solution applied for the imbalanced dataset problem is two resampling methods, which are up sampling and down sampling. The dataset obtained in the CSV format is anonymised and fetched from the database. It contains 38 features that are categorical and numerical. Categorical variables are converted to numerical values by using one hot encoding.



Algoritma	Resampling yöntemi	Average Precision	Matthews	MSE	AUC
Random Forest	up	0.5 ± 0.05	0.4 ± 0.03	0.18 ± 0.01	0.84 ± 0.02
Random Forest	down	0.5 ± 0.04	0.38 ± 0.03	0.2 ± 0.01	0.84 ± 0.02
Random Forest	-	0.53 ± 0.08	0.35 ± 0.09	0.1 ± 0.01	0.85 ± 0.02
Gradient Boosting	up	0.58 ± 0.1	0.42 ± 0.02	0.18 ± 0.02	0.86 ± 0.02
Gradient Boosting	down	0.57 ± 0.1	0.39 ± 0.03	0.2 ± 0.02	0.86 ± 0.02
Gradient Boosting	-	**0.59 ± 0.1**	**0.46 ± 0.13**	**0.09 ± 0.02**	**0.86 ± 0.02**
XGBoost	up	0.58 ± 0.1	0.42 ± 0.03	0.18 ± 0.01	0.86 ± 0.02
XGBoost	down	0.57 ± 0.1	0.4 ± 0.02	0.2 ± 0.01	0.86 ± 0.02
XGBoost	-	**0.59 ± 0.1**	0.45 ± 0.14	0.09 ± 0.02	0.86 ± 0.02

Figure 6. Comparison of machine learning algorithms

Figure 10 shows the comparison of different machine learning algorithms with a variety of sampling methods; the results are achieved using a clean dataset with stratified k-fold implementation. The gradient boosting algorithm with no resampling method receives the highest accuracy among machine learning algorithms. These results indicate that gradient boosting performs well in imbalanced datasets. Its precision score is 59% and the AUC score is 86%. By looking at the precision value, the AUC score shows that the gradient boosting model is successful at predicting the churners. The results shown in Figure 10 show that gradient boosting with up sampling received better AUC and precision scores than down sampling. Down sampling with all algorithms on the table decreased the metric scores of our model for using this dataset.

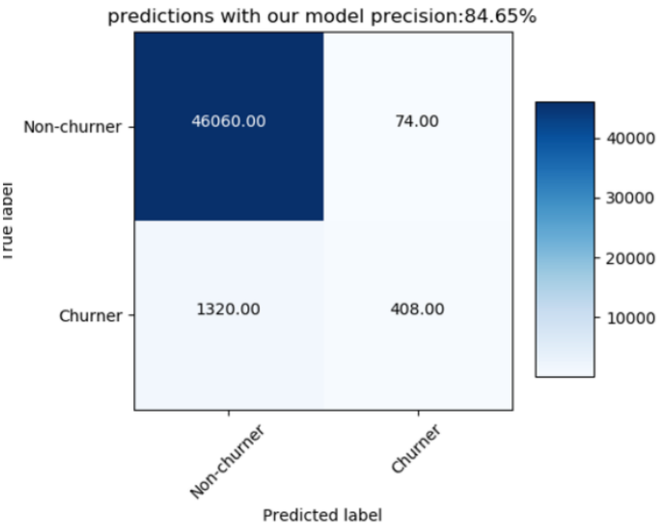


Figure 7. Confusion matrix on the test dataset

The confusion matrix represents the most critical part of the results which is achieved by running machine learning algorithms. Figure 11 shows all the predictions in a given confusion matrix with a low false positive rate. The gradient boosting model predicted 408 churners out of 47866 with a low false positive rate as shown earlier. These results indicate that the model can separate churners from non-churners and the model performs well on the churner prediction.

7. Conclusion

All in all, having strong knowledge about customers can lead to a lot of opportunities, which include gaining new customers, increasing the revenue of the company and creating trust between customers and company. In this research, a variety of approaches have been implemented to understand and analyse customer satisfaction with anonymised datasets. The data comprised different sources, which made the data much more usable in different areas, such as clarifying required network bandwidth, understanding customer intention and having a campaign on a product or not. Although a lot of work has been conducted in this research, there are plenty of works to improve it; creating an environment using large data technologies (such as Hadoop, Hive and Kafka) can be the next step for further research. This will extend the scope of the project and provide more accessibility to recent technologies.

References

[1] Ahmad, A. K., Jafar, A. & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. *Journal of Big Data, 6*(1). doi: 10.1186/s40537-019-0191-6

[2] Chen, T. & Guestrin, C. (2016). *XGBoost*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. doi:10.1145/2939672.2939785

[3] Khan, A. A., Jamwal, S. & Sepehri, M. (2010). Applying data mining to customer churn prediction in an internet service provider. *International Journal of Computer Applications, 9*(7), 8–14. doi:10.5120/1400-1889

[4] Li, K. G. & Marikannan, B. P. (2019). Hybrid particle swarm optimization-extreme learning machine algorithm for customer churn prediction. *Journal of Computational and Theoretical Nanoscience, 16*(8), 3432–3436. doi:10.1166/jctn.2019.8304

[5] Xu, E., Shao, L., Gao, X. & Zhai, B. (2006). *An algorithm for predicting customer churn via BP neural network based on rough set*. 2006 IEEE Asia-Pacific Conference on Services Computing (APSCC06). doi:10.1109/
apscc.2006.23

[6] Buitinck, L., Louppe, G., Blondel, M., Pedregosa, Fabian, Mueller, A., Grisel, O., … Varoquaux, G. (2013). *API design for machine learning software: experiences from the scikit-learnproject* (pp. 108–122). In ECML PKDD Workshop: Languages for Data Mining and Machine Learning.