# PREDICTION OF TELECOM SERVICES CONSUMERS CHURN BY USING MACHINE LEARNING ALGORITHMS

Edin Osmanbegović*, Anel Džinić**, Mirza Suljić***
DOI: 10.51558/2303-680X.2022.20.2.53

**Abstract**

*Machine learning, or as it is also called automated learning, is a special subfield of scientific information technologies. The name "machine learning" refers to the automated detection of meaningful patterns in large data sets. Machine learning is gaining importance in many different areas of the economy. One of those areas is the prediction and prevention of consumer churn. There are two basic types of consumer churn, complete churn and partial churn. Machine learning is used to determine the most significant characteristics that play a role in the churn/retention of consumers, and with the help of machine learning it is possible to establish the probability of churn for each individual consumer. Some of the most commonly used machine learning algorithms for this issue are Logistic Regression, Gaussian Naive Bayes, Bernoulli Naive Bayes, Decision Tree, and Random Forest.*

**Keywords:** machine learning, customer churn, customer retention

**JEL:** L86

## 1. Introduction

In recent decades, we have witnessed an explosion in the amount of data created. It is a widely known that companies have more and more data at their disposal, so an increasing number of companies are trying to make the most of the data in order to achieve a competitive advantage. The process of extracting information and knowledge from corporate data is done by using machine learning techniques. The name "machine learning" refers to the automated detection of meaningful patterns in a large set of data, and represents a special branch that deals with the discovery and development of algorithms that can learn and make assumptions based on input data (Banjanović-Mehmedović, 2011). *Machine learning is gaining importance in many different areas of the economy.* One of those areas for service companies in the telecommunications industry face the challenge of losing valuable customers to competitors, which is known as customer churn (Huang *et al.*, 2012). Buyer and consumer, although they can be the same person, most often are not and do not have to be the same person. A customer is an individual who buys a certain product, which means that they directly perform a market transaction that is a market exchange, regardless of the fact what the subject of the exchange is. A consumer is the carrier of the need for a specific product, who at the same time physically or in some other way consumes or uses the product. The retaining of consumers is profitable for a company because of the following (Verbeke *et al.*, 2011):

- Attracting a new customer costs five to six times more than retaining the existing ones.
- Long-term consumers earn more income and become cheaper to provide services. Also, they can attract new customers through positive feedback, while disgruntled users could spread negative rumors.
- Loss of customers leads to opportunity costs due to reduced sales.

There are two basic types of consumer churn (Buckinx and Van den Poel, 2005):

- Full exit - where a consumer completely stops buying from a particular company,
- Partial departure - where the departure takes place gradually, and a certain criterion needs to be defined to define the departure.

* University of Tuzla, Faculty of Economics, Bosnia and Herzegovina, edin.osmanbegovic@untz.ba
** CaDa Solucije doo, Bosnia and Herzegovina, aneldzinic@live.com
*** University of Tuzla, Bosnia and Herzegovina, mirza.suljic@untz.ba

Predicting customer churn in the telecommunications sector has received serious stakeholder attention in order to increase customer loyalty and improve customer relationship management standards (Plazibat and Šušak, 2016). Telecom operators understand the importance of retaining the existing customers.

The costs, caused by adding a new customer, are far greater than the costs of retaining a customer whose needs are not adequately met. There are many reasons for consumers to leave (Dingli *et al*., 2017). The consumer churn also affects the overall reputation of the company, and it results in the weakening of the brand. It is too late to consider a policy change when the number of telecom company customers falls below a certain level. This results in a large loss of revenue (Ullah *et al*., 2019). Telecom companies need to act proactively and prevent customers from leaving, with appropriate incentives targeting those customers with an increased probability of leaving (Idris and Khan, 2012). It does not take a lot of effort to identify the existing users who intend to leave, and thus the community benefits greatly.

In general, the churn rate in the US telecommunications sector is about 2% per year, making for a total annual loss of approximately 15%. Predicting customer with an increased likelihood of leaving is 16 times cheaper than acquiring a new customer and the cost of attracting a new customer is five to six times higher than the cost of retaining an existing customer. By reducing customer churn by 5%, profits increase by 25% to 85% (Ullah *et al*., 2019).

Customer churn prediction models aim to identify consumers with a high likelihood of churn. Accurate segmentation of consumers allows the company to target those consumers who have the highest probability of leaving, in order to increase the efficiency of the resources used for the marketing campaign of consumer retention (Verbeke *et al*., 2011). A small improvement in customer retention, therefore, can lead to a significant increase in profits. Therefore, precise and comprehensible models for predicting consumer churn are needed for identification, to identify consumers who are likely to leave, as well as the reasons why, in order to prevent churn and retain customers (Lalwani *et al*., 2022, Jiao and Xu, 2021).

## 2. Research objectives, sample and applied methodology

The subject of the research is the prediction of consumer churn by using machine learning algorithms. Consumer churn occurs when consumers completely stop buying some products or services of a certain company, or when that purchase falls below a certain limit (Gold, 2020).

Two basic types of consumer churn have already been mentioned earlier, complete churn and partial churn. It is completely clear that the occurrence of consumer churn represents something negative for every company, and that is why companies strive to suppress and, if possible, completely eliminate consumer churn. The negative aspects of consumer churn are primarily reflected in the increase in the cost of acquiring new customers, reduced profits, and a decrease in the sales success rate.

Through primary research, hard-to-reach data was collected from individuals using the snowball approach through social media. The snowball approach is used when qualified individuals share the questionnaire with other subjects who are similar to them and meet the qualifications established for the target population (Berg, 2004).

In this research, data was collected from qualified individuals by posting an electronic questionnaire on several Viber groups, Facebook, and Microsoft Teams. The questionnaire was answered by individuals who are users of telecom services.

The electronic questionnaire was created by using Google Forms. A total of 267 responses were collected. There were no incomplete survey questionnaires, as all questions were mandatory. The independent research variable is represented by consumers with their demographic and consumer attributes. The characteristics of the independent variable are expressed through indicators, which, along with the types of attributes, are given in the following table.

Table 2.1 *Characteristics of independent variables, their indicators and types of characteristics*

| Feature | Indicators | Feature type |
|---|---|---|
| Gender | Attitudes about gender | Demographic |
| Marital status | Attitudes about marital status | Demographic |
| Household size | Attitudes about household size | Demographic |
| Employment status | Attitudes about employment status | Demographic |
| Duration in consumer status | Attitudes about duration in consumer status | Consumer |
| Owning a telephone | Attitudes about owning a telephone | Consumer |
| Type of telephone connection | Attitudes about type of telephone connection | Consumer |
| Owning multiple telecommunication lines | Attitudes about owning multiple telecommunication lines | Consumer |
| Owning multiple numbers | Attitudes about owning multiple numbers | Consumer |
| Type of Internet connection | Attitudes about type of Internet connection | Consumer |
| Type of service package | Attitudes about type of service package | Consumer |
| Type of contract | Attitudes about type of contract | Consumer |
| Form of account | Attitudes about form of account | Consumer |
| Method of payment | Attitudes about method of payment | Consumer |
| Average monthly consumption in consumer status | Attitudes about average monthly consumption in consumer status | Consumer |
| Total consumption in consumer status | Attitudes about total consumption in consumer status | Consumer |
| Did the consumer stay with the company | Attitudes about did the consumer stay with the company | A feature related to the consumer churn |

Source: Authors' research

The dependent variable of the research is the coefficient of prediction of consumer churn. The indicator for measuring the dependent variable is the coefficient of prediction of consumer churn expressed as a percentage.

In the data set that was used in this paper, 94% of the respondents said that they do not plan to change their service provider in the next month, while 6% said that they intend to. It follows that the classes of the data set are not balanced. For the purpose of evaluating predicting models, we use Accuracy and F1-score to determine their efficiency (Witten *et al.*, 2011).

The assessment of predictive ability was performed through data mining analysis, using Python software (Albon, 2018), code editor Visual Studio Code, and web code editor Google Collaborate.

Five machine learning algorithms were used (Witten *et al.*, 2011), namely: Logistic Regression, Gaussian Naive Bayes, Bernoulli Naive Bayes, Decision Tree, and Random Forest.

## 3. Research results and discussions

### *3.1 Evaluations of the input attribute importance*

In the phase of data preparation for the research, the value of the input attributes was evaluated in relation to the output attribute. The goal of the attribute evaluation and selection process is to extract irrelevant and redundant attributes from the learning data set (Osmanbegović *et al.*, 2014).

The following table shows the feature importance (Charonyktakis, 2021), determined by calling the "feature importance" function within the Scikit-learn python package.

Table 3.1. *Significance of features*

| Rank | Feature | Type of Feature | Significance (%) |
|---|---|---|---|
| 1. | Household size | Demographic | 11.43 |
| 2. | Form of account | Consumer | 10.47 |
| 3. | Type of contract | Consumer | 10.07 |
| 4. | Type of service package | Consumer | 8.64 |
| 5. | Total consumption in consumer status | Consumer | 8.17 |
| 6. | Marital status | Demographic | 7.44 |
| 7. | Average monthly consumption in consumer status | Consumer | 6.58 |
| 8. | Gender | Demographic | 5.73 |
| 9. | Owning multiple telecommunication lines | Consumer | 5.48 |
| 10. | Owning multiple numbers | Consumer | 5.08 |
| 11. | Type of Internet connection | Consumer | 4.89 |
| 12. | Duration in consumer status | Consumer | 4.88 |
| 13. | Method of payment | Consumer | 4.28 |
| 14. | Type of telephone connection | Consumer | 3.65 |
| 15. | Employment status | Demographic | 3.13 |
| 16. | Owning a telephone | Consumer | 0 |

Source: Authors' research

"Household size" was the most significant feature at 11.43%. This feature was also the most significant demographic feature. It was followed by "Form of account", which was the most significant consumer characteristic, with 10.47%. It was followed by three more consumer characteristics, "Type of contract", "Type of service package" and "Total consumption in consumer status", with 10.07%, 8.64%, and 8.17%, respectively.

After them, in the 6th place, came "Marital status", which was a demographic characteristic and had a significance of 7.44%. In the 7th place was "Average monthly consumption in consumer status", as a consumer characteristic with 6.58% significance. If we assumed that all features were equally important, in that case all features would have a significance of 100%/16 = 6.25%. The first seven features were above the limit, and the next nine features were below that limit.

In the 8th place was "Gender" as a demographic characteristic, with 5.73%. Out of four demographic characteristics, three were in the first half (first eight) of the characteristics. The next six characteristics were all consumer characteristics. "Owning multiple telecommunication lines", "Owning multiple numbers", and "Type of Internet connection" had a significance of 5.48%, 5.08%, and 4.89%, respectively.

"Duration in consumer status", "Method of payment", and "Type of telephone connection" had a significance of 4.88%, 4.28%, and 3.65%, respectively. "Owning multiple telecommunication lines" and "Duration in consumer status" had almost equal significance.

"Employment status" was a demographic characteristic and occupied the penultimate, 15th place, with a significance of 3.13%. The last, 16th place was occupied by "Owning a telephone", which was a consumer characteristic and had a significance of 0%. It was not surprising that this feature had no significance for prediction because all the respondents said that they owned a phone.

There were four demographic characteristics making up 25% of all the characteristics, and the sum of their importance was 27.73%. Consumer characteristics made up 75% of all the characteristics, and the sum of their importance was 72.27%.

*3.2 Analysis of the performance of machine learning algorithms, when only consumer demographic characteristics are used*

Table 3.2. shows the results of individual machine learning algorithms, when self-demographic characteristics are used. Only four demographic characteristics ("Gender", "Marital status", "Household size" and "Employment status") were used in the Python program, and the algorithm analyzed their impact on consumer churn. The total significance of these four features is 27.73%.

Table 3.2 *Performance of machine learning algorithms based on demographic characteristics*

| Algorithm | Accuracy (%) | F1-score (%) |
|---|---|---|
| Logistic Regression | 94.38 | 97.11 |
| Gaussian Naive Bayes | 94.38 | 97.11 |
| Bernoulli Naive Bayes | 94.38 | 97.11 |
| Decision Tree | 94.38 | 97.11 |
| Random Forest | 94.38 | 97.11 |

Source: Authors' research

All the papers mentioned previously hat served to form the standard used a larger number of features and a larger set of data during the research. The standard for accuracy is 81.43%, and for F1-score it was 76.41%. Comparing the individual algorithms with the standards, one can see:

- That the Logistic Regression algorithm achieved an accuracy of 94.38% and an F1-score of 97.11%. Based on accuracy and F1-score, it significantly exceeded the standards. From this, it can be concluded that this algorithm has good abilities to predict consumer churn, i.e., that even a small number of only four features was enough to achieve excellent results. Since the Logistic Regression algorithm outperformed both standards, it can be concluded that it is good enough for predicting consumer churn. During this analysis, some of the advantages of the Logistic Regression algorithm came to the fore, namely that it is effective, the scaling of the features is not necessary, and the tuning of the features is also not needed.
- That the Gaussian Naive Bayes algorithm achieved an accuracy of 94.38% and an F1-score of 97.11%. Based on accuracy and F1-score, it significantly exceeded the standards. From this, it can be concluded that this algorithm has good abilities to predict consumer churn, i.e., that even a small number of only four features was enough to achieve excellent results. Since the Gaussian Naive Bayes algorithm outperformed both benchmarks, it can be concluded that it is good enough for predicting consumer churn. During this analysis, some of the advantages of Naive Bayes algorithms came to the fore, namely that scaling is not necessary and that they achieve good results even on small data sets.
- That the Bernoulli Naive Bayes algorithm achieved an accuracy of 94.38% and an F1-score of 97.11%. Based on accuracy and F1-score, it significantly exceeded the standards. From this, it can be concluded that this algorithm has good abilities to predict consumer churn, i.e., that even a small number of only four features was enough to achieve excellent results. Since Bernoulli's Naive Bayes algorithm outperformed both benchmarks, it can be concluded that it is good enough for predicting consumer churn. During this analysis, some of the advantages of Naive Bayes algorithms came to the fore, namely that scaling is not necessary and that they achieve good results even on small data sets.
- That the Decision Tree algorithm achieved an accuracy of 94.38% and an F1-score of 97.11%. Based on these metrics, it significantly exceeded the standards. It can be concluded that this algorithm has good abilities to predict consumer churn, as even just four characteristics were enough to achieve excellent results. Since the Tree Resolution algorithm outperformed both benchmarks, it can be considered a good predictor of consumer churn. During the analysis, some advantages of the Decision Tree algorithm were revealed, including not requiring normalization or scaling of data and automatically selecting the most significant features.
- That the Random Forest algorithm, just like the other algorithms, achieved an accuracy

of 94.38% and an F1-score of 97.11%. Based on accuracy and F1-score, it significantly exceeded the standards. Based on accuracy and F1-score, it significantly exceeded the standards. It can be concluded from this that this algorithm has good abilities to predict consumer churn, i.e., that even a small number of only four characteristics was enough to achieve excellent results. Since the Random Forest algorithm outperformed both standards, it can be concluded that it is good enough for predicting consumer churn. During this analysis, some of the advantages of the Random Forest algorithm came to the fore, namely error reduction, good performance on unbalanced data sets, reduction of the influence of exceptions, and resistance to overfitting.

### *3.3 Performance analysis of machine learning algorithms, when only consumer characteristics are used*

The following table shows the results of individual machine learning algorithms when consumption features were used. In the Python program, 12 consumer characteristics were used, including "Length of duration in consumer status," "Owning a telephone," "Type of telephone connection," "Owning multiple telecommunication lines," "Owning multiple numbers," "Type of Internet connection," "Type of package service," "Contract type," "Bill form," "Payment method," "Average monthly consumption in consumer status," and "Total consumption in consumer status." The algorithm analyzed their impact on consumer churn.

Table 3.3 *Performance of machine learning algorithms based on consumer characteristics*

| Algorithm | Accuracy (%) | F1-score (%) |
|---|---|---|
| Logistic Regression | 94.38 | 97.11 |
| Gaussian Naive Bayes | 94.38 | 97.11 |
| Bernoulli Naive Bayes | 94.38 | 97.11 |
| Decision Tree | 94.38 | 97.11 |
| Random Forest | 93.25 | 97.11 |

Source: Authors' research

The total importance of these 12 features was 72.27%. All the papers that contributed to the formation of the standard used a larger set of data during the research, and most of them also used a larger number of features. The standard for accuracy was 81.43%, and for F1-score, it was 76.41%. A smaller number of features can lead to an improvement in the performance of algorithms through the reduction of "noise" and overfitting, but it can also lead to deterioration in the performance of algorithms through loss of information. Also, performance may remain the same.

Comparing the individual algorithms with the standards, one can see:

- That the Logistic Regression algorithm achieved an accuracy of 94.38% and F1-score of 97.11%. Based on accuracy and F1-score, it significantly exceeded the standards. From this, it can be concluded that this algorithm has good abilities to predict consumer churn, i.e., that the number of 12 features was enough to achieve excellent results. In terms of accuracy, the algorithm correctly classified 84 out of 89 consumers. There is no intuitive way to interpret F1-score, similar to the interpretation of accuracy, since F1-score is a combined measure. Since the Logistic Regression algorithm outperformed both standards, it can be concluded that it is good enough for predicting consumer churn. During this analysis, some of the advantages of the Logistic Regression algorithm came to the fore, namely that it is effective, the scaling of the features is not necessary, and the tuning of the features is also not needed. A larger number of features had no influence on accuracy and F1-score, i.e., they remained the same, as was the case with the analysis of exclusively demographic characteristics.
- That the Gaussian Naive Bayes algorithm achieved an accuracy of 89.89% and F1-score of 97.11%. Based on accuracy and F1-score, it significantly exceeded the standards. From this, it can be concluded that this algorithm has good abilities to predict consumer churn, i.e., that the number of 12 features was enough to achieve excellent results. In terms of accuracy, out of 89 consumers, the

algorithm correctly classified 80. Since the Gauss naive bayes algorithm surpassed both standards, it can be concluded that it is good enough for predicting consumer churn. During this analysis, some of the advantages of Naive Bayes algorithms came to the fore, namely that scaling is not necessary, and that they achieve good results even on small data sets. A larger number of features had a negative impact on the accuracy of this algorithm, i.e., compared to the analysis with only demographic characteristics, the accuracy worsened by 4.49%. A larger number of features had no influence on F1-score.

- That the Bernoulli Naive Bayes algorithm achieved an accuracy of 94.38% and F1-score of 97.11%. Based on accuracy and F1-score, it significantly exceeded the standards. From this, it can be concluded that this algorithm has good abilities to predict consumer churn, i.e., that the number of 12 features was enough to achieve excellent results. Since Bernoulli's Naive Bayes algorithm outperformed both benchmarks, it can be concluded that it is good enough for predicting consumer churn. During this analysis, some of the advantages of Naive Bayes algorithms came to the fore, namely that scaling is not necessary, and that they achieve good results even on small data sets. A larger number of features had no influence on accuracy and F1-score, i.e., they remained the same, as was the case with the analysis of exclusively demographic characteristics.
- That the Decision Tree algorithm achieved an accuracy of 94.38% and F1-score of 97.11%. Based on accuracy and F1-score, it significantly exceeded the standards. From this, it can be concluded that this algorithm has good abilities to predict consumer churn, i.e., that the number of 12 features was enough to achieve excellent results. Since the Decision Tree algorithm outperformed both benchmarks, it can be concluded that it is good enough to predict consumer churn. During this analysis, some of the advantages of the Decision Tree algorithm came to the fore, namely data normalization and scaling not being required, and the automatic selection of the most significant features. A larger number of features had no influence on accuracy and F1-score, i.e., they remained the same, as was the case with the analysis of exclusively demographic characteristics.
- That the Random Forest algorithm achieved an accuracy of 93.26% and F1-score of 97.11%. Based on accuracy and F1-score, it significantly exceeded the standards. From this, it can be concluded that this algorithm has good abilities to predict consumer churn, i.e., that the number of 12 features was enough to achieve excellent results. An accuracy of 93.26% means that out of 89 consumers, this algorithm correctly classified 83. During this analysis, some of the advantages of the Random Forest algorithm came to the fore, namely error reduction, good performance on unbalanced data sets, reduction of the influence of exceptions, and resistance to overfitting. A larger number of features had a negative impact on the accuracy of this algorithm, i.e., compared to the analysis with only demographic characteristics, the accuracy worsened by 1.12%. A larger number of features had no influence on F1-score.

### *3.4 Performance analysis of machine learning algorithms, when all consumer characteristics are used*

The following table shows the results of individual machine learning algorithms, when all features are used.

Table 3.4. *Performance of machine learning algorithms based on all features*

| Algorithm | Accuracy (%) | F1-score (%) |
|---|---|---|
| Logistic Regression | 94.38 | 97.11 |
| Gaussian Naive Bayes | 92.13 | 97.11 |
| Bernoulli Naive Bayes | 93.26 | 97.11 |
| Decision Tree | 94.38 | 97.11 |
| Random Forest | 94.38 | 97.11 |

Source: Authors' research

In the Python program, all 16 characteristics were used ("Sex", "Marital status", "Household size", "Employment status", "Length of duration in consumer status", "Owning a telephone", "Type of telephone connection", "Owning multiple telecommunication lines", "Owning multiple numbers", "Internet connection type", "Service package type", "Contract type", "Bill form", "Payment method", "Average monthly consumption in consumer status", and "Total consumption in consumer status "), i.e., both demographic and consumer characteristics were used, and the algorithm analyzed their impact on consumer churn. All the papers mentioned previously, that served to form the standard, used a larger set of data during the research, and most of them also used a larger number of features. The standard for accuracy is 81.43%, and for F1-score it is 76.41%. A smaller number of features can lead to an improvement in the performance of the algorithms through the reduction of "noise" and overfitting, but it can also lead to deterioration in the performance of the algorithms through the loss of information. Also, performance may remain the same.

Comparing the individual algorithms with the standards, one can see:

- That the Logistic Regression algorithm achieved an accuracy of 94.38% and an F1-score of 97.11%. Based on accuracy and F1-score, it significantly exceeded the standards. From this it can be concluded that this algorithm has good abilities to predict consumer churn. In terms of accuracy, the algorithm correctly classified 84 out of 89 consumers. There is no intuitive way to interpret F1-score, similar to the interpretation of accuracy, since F1-score is a combined measure. Since the Logistic Regression algorithm outperformed both standards, it can be concluded that it is good enough for predicting consumer churn. During this analysis, some of the advantages of the Logistic Regression algorithm came to the fore, namely that it is effective, the scaling of the features is not necessary, and the tuning of the features is also not needed. A larger number of features had no influence on accuracy and F1-score, i.e., they remained the same, as was the case with the analysis of exclusively demographic and consumer characteristics.
- That the Gaussian Naive Bayes algorithm achieved an accuracy of 92.13% and F1-score of 97.11%. Based on accuracy and F1-score, it significantly exceeded the standards. In terms of accuracy, out of 89 consumers, the algorithm correctly classified 82. Since the Gauss Naive Bayes algorithm surpassed both standards, it can be concluded that it is good enough for predicting consumer churn. During this analysis, some of the advantages of Naive Bayes algorithms came to the fore, namely that scaling is not necessary and that they achieve good results even on small data sets. Compared to the analysis with only consumer characteristics, a larger number of characteristics had a positive effect on accuracy, which improved by 2.24%. Compared to the analysis with only demographic characteristics, a larger number of characteristics had a negative impact on the accuracy of this algorithm, i.e., compared to the analysis with only demographic characteristics, the accuracy worsened by 2.25%. A larger number of features had no influence on F1-score.
- That the Bernoulli Naive Bayes algorithm achieved an accuracy of 93.26% and F1-score of 97.11%. Based on accuracy and F1-score, it significantly exceeded the standards. In terms of accuracy, out of 89 consumers, the algorithm correctly classified 83. Since the Bernoulli Naive Bayes algorithm surpassed both standards, it can be concluded that it is good enough for predicting consumer churn. During this analysis, some of the advantages of Naive Bayes algorithms came to the fore, namely that scaling is not necessary and that they achieve good results even on small data sets. A larger number of features had a negative impact on the accuracy of this algorithm, i.e., compared to the analysis with only demographic, i.e., consumer characteristics, the accuracy worsened by 1.12%. A larger number of features had no influence on F1-score.
- That the Decision Tree algorithm achieved an accuracy of 94.38% and F1-score of 97.11%. Based on accuracy and F1-score, it significantly exceeded the standards. From this it can be concluded that this algorithm has good abilities to predict consumer churn. During this analysis, some of the advantages of the Decision Tree algorithm came to the fore, namely data

normalization and scaling not being required, and the automatic selection of the most significant features. A larger number of features had no influence on accuracy and F1-score, i.e., they remained the same, as was the case with the analysis exclusively of demographic and consumer characteristics.

- That the Random Forest algorithm achieved an accuracy of 94.38% and an F1-score of 97.11%. Based on accuracy and F1-score, it significantly exceeded the standards. From this it can be concluded that this algorithm has good abilities to predict consumer churn. During this analysis, some of the advantages of the Random Forest algorithm came to the fore, namely error reduction, good performance on unbalanced data sets, reduction of the influence of exceptions, and resistance to overfitting. A greater number of features had a positive effect on the accuracy of this algorithm, i.e., compared to the analysis with only consumer characteristics, accuracy improved by 1.12%. A greater number of features had no effect on accuracy, when compared to the result of the analysis with only demographic features. A larger number of features had no influence on F1-score.

### *3.5 Discussion of research results*

The analysis of the significance of the characteristics showed that four demographic characteristics ("Sex", "Marital status", "Household size", and "Employment status") have a sum of significance of 27.73%, and make up 25% of all the characteristics.

There are 12 consumer characteristics ("Length of duration in consumer status", "Owning a telephone", "Type of telephone connection", "Owning multiple telecommunication lines", "Owning multiple numbers", "Type of Internet connection", "Type of service package", "Contract type", "Bill form", "Payment method", "Average monthly consumption in consumer status", and "Total consumption in consumer status"), i.e., they make up 75% of all characteristics, and the sum of their importance is 72.27%.

The analysis of the performance of machine learning algorithms, when only consumer demographic characteristics are used, showed that all five analyzed algorithms outperformed both benchmarks.

All five had an accuracy of 94.38% and an F1-score of 97.11%. An accuracy of 94.38% means that all algorithms out of 89 consumers in the test data set correctly classified 84 of them, i.e., they misclassified the five-consumer. There is no intuitive way, similar to the interpretation of accuracy, to interpret the F1-score, since the F1-score is a combined measure. During this analysis, many advantages of the mentioned machine learning algorithms came to the fore. With this, H1 is confirmed, ie., it is possible to predict consumer churn based on demographic characteristics in machine learning algorithms.

A smaller number of features can lead to an improvement in the performance of the algorithms by reducing the "noise" of overfitting, but it can also lead to deterioration in the performance of the algorithms through the loss of information. Also, performance may remain the same.

The analysis of the performance of the machine learning algorithms, when only consumer characteristics are used, showed that all five analyzed algorithms outperformed both benchmarks.

Three algorithms, Logistic Regression, Bernoulli Naive Bayes, and Decision Tree, had an accuracy of 94.38%. An accuracy of 94.38% means that the algorithms correctly classified 84 out of 89 consumers in the test data set, i.e., they misclassified five consumers. Gaussian Naive Bayes had an accuracy of 89.89%, which means% means that the algorithm correctly classified 80 out of 89 consumers in the test data set.

Random Forest had an accuracy of 93.26%. An accuracy of 93.26% means that the algorithm correctly classified 83 out of 89 consumers in the test data set. All five analyzed algorithms had an F1-score of 97.11%. There is no intuitive way to interpret the F1-score, similar to the interpretation of accuracy, since the F1-score is a combined measure. During this analysis, many advantages of the mentioned machine learning algorithms came to the fore.

A higher number of markers had a negative impact on the accuracy of the Gaussian Naive Bayes and Random Noise algorithms, i.e., compared to the analysis with only demographic characteristics, the accuracy worsened by 4.49% and 1.12%, respectively. A larger number of features had no effect on the accuracy of the remaining three algorithms. A number of features had no influence on the F1-score of all five algorithms. A possible explanation for this is the fact that the F1-score is a harmonic mean between precision and recall, and is therefore less sensitive to minor changes in precision and recall. With this, H2 is confirmed, i.e., it is possible to predict consumer departure based on consumer characteristics in machine learning algorithms.

Performance analysis of machine learning algorithms, when all consumer characteristics are used, shows that all five analyzed algorithms outperformed both standards. Three algorithms, Logistic Regression, Decision Tree and Random Forest had an accuracy of 94.38%. An accuracy of 94.38% means that out of 89 consumers in the test data set, the algorithms correctly classified 84, i.e., they misclassified five consumers. Gaussian Naive Bayes had an accuracy of 92.13%. An accuracy of 92.13% means that the algorithm correctly classified 82 out of 89 consumers in the test data set. Bernoulli's Naive Bayes had an accuracy of 93.26%. An accuracy of 93.26% means that the algorithm correctly classified 83 out of 89 consumers in the test data set. All five analyzed algorithms had an F1-score of 97.11%. There is no intuitive way, similar to the interpretation of accuracy, to interpret the F1-score, since the F1-score is a combined measure. In this analysis, the judgment came to express many advantages of the mentioned machine learning algorithms.

A larger number of features had no influence on the accuracy of the Logistic Regression and Decision Tree algorithms, i.e., the accuracy remained the same as in the analysis with exclusively demographic and consumer characteristics.

In the case of the Gaussian Naive Bayes algorithm, the accuracy improved compared to the analysis with only consumer characteristics by 2.24%, but it was still lower than the analysis with only demographic characteristics by 2.25%. In the case of Bernoulli's Naive Bayes algorithm, the accuracy worsened by 1.12% compared to the analysis with only demographic and consumer characteristics.

With the Random Forest algorithm, the accuracy improved by 1.12% compared to the analysis with only consumer characteristics, and remained the same compared to the analysis with demographic characteristics only.

A larger number of features had no effect on the F1-score of all five algorithms. A possible explanation for this is the fact that the F1-score is a harmonic mean between precision and recall, and is therefore less sensitive to minor changes in precision and recall.

This confirms the main research hypothesis, H: It is possible to predict the departure of consumers using predictive machine learning algorithms.

## 4. Conclusion

This paper presented customer churn prediction in the telecommunications, including the Household size, Form of account, Type of contract, Type of service package, Total consumption in consumer status, Marital status, Average monthly consumption in consumer status, Gender, Owning multiple telecommunication lines, Owning multiple numbers, Type of Internet connection, Duration in consumer status, Method of payment, Type of telephone connection, Employment status, and Owning a telephone. Then five machine learning algorithms (LR, GNB, BNB, DT and RF) were used as predictors in this paper. Finally, based on the feature sets and five machine learning algorithms, the comparative experiments were carried out. In the experiments, each subset of the feature was evaluated and analyzed. In addition, the experiments provided:

- the comparative effectiveness of five machine learning algorithms and

- the comparative effectiveness between feature sets.

The experimental results showed that: (1) the proposed feature set is large enough, (2) all five algorithms were successful in predicting consumer churn and can be recommended as reliable tools for predicting consumer churn, (3) it is difficult to determine which modeling technique is best for customer churn prediction due to asmall data sample.

However, there are some limitations with our proposed approach. In the future, other information should be added into the feature set in such a way to improve the features. Although sufficient, the sample size should be bigger to allow for the proper use of machine learning algorithms.

A larger sample allows for a larger amount of data to be used for training machine learning algorithms, which should result in better performance (Cios *et al*., 2007). The same is the case with multiple features.

In addition, the imbalance classification problem takes place in this application and we only used the sampling technique to attempt to solve the problem. Therefore, more methods for imbalance classifications also should be appliedin the future.

## References


1. Albon, C. (2018). *Machine learning with python cookbook: Practical solutions from preprocessing to deep learning,* Sebastopol: O'Reilly Media.
2. Banjanović-Mehmedović, L. (2011). *Inteligentni sistemi,* Tuzla: Harfograf Tuzla.
3. Berg, S. 2004. Snowball sampling—I. *Encyclopedia of statistical sciences.* New York: Wiley and Sons. Retrieved from https://doi.org/10.1002/0471667196.ess2478.pub2
4. Buckinx, W. & Van Den Poel, D. (2005) Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. *European Journal of Operational Research.* 164(1)**,** pp. 252-268. doi: 10.1016/j.ejor.2003.12.010.
5. Charonyktakis, P. 2021. *Do We Really Need Feature Selection in a Data Analysis Pipeline?* [Online]. Towards Data Science. Available: https://towardsdatascience.com/do-we-really-need-feature-selection-in-a-data-analysis-pipeline-dc8401621c6c [Accessed 05.12.2022].
6. Cios, K. J., Swiniarski, R. W., Pedrycz, W. & Kurgan, L. A. (2007). The Knowledge Discovery Approach. *Data Mining: A Knowledge Discovery Approach.* New York: Springer. doi: 10.1007/978-0-387-36795-8_2.
7. Dingli, A., Marmara, V. & Fournier, N. S. (2017) Comparison of deep learning algorithms to predict customer churn within a local retail industry. *International Journal of Machine Learning and Computing.* 7(5)**,** pp. 128-132. doi: 10.18178/IJMLC.2017.7.5.634.
8. Gold, C. (2020). *Fighting Churn with Data: The science and strategy of customer retention,* Shelter Island: Manning Publications.
9. Huang, B., Kechadi, M. T. & Buckley, B. (2012) Customer churn prediction in telecommunications. *Expert Systems with Applications.* 39(1)**,** pp. 1414-1425. doi: 10.1016/j.eswa.2011.08.024.
10. Idris, A. & Khan, A. (2012) Customer churn prediction for telecommunication: Employing various various features selection techniques and tree based ensemble classifiers. 2012 15th international multitopic conference (INMIC). Proceedings (pp. 23-27.): IEEE.
11. Jiao, G. & Xu, H. 2021. Analysis and Comparison of Forecasting Algorithms for Telecom Customer Churn. *The 2nd International Conference on Computing and Data Science (CONF-CDS).* Stanford, United States: IOP Publishing.
12. Lalwani, P., Mishra, M. K., Chadha, J. S. & Sethi, P. (2022) Customer churn prediction system: a machine learning approach. *Computing.* 104(2)**,** pp. 271-294. doi: 10.1007/s00607-021-00908-y.
13. Osmanbegović, E., Suljić, M. & Agić, H. (2014) Determining dominant factor for students performance prediction by using data mining classification algorithms. *Tranzicija.* 16(34)**,** pp. 147-158. doi:

14. Plazibat, I. & Šušak, T. (2016) Funkcionalnost programa lojalnosti kupaca u maloprodaji. *Ekonomska misao i praksa.* 25(1), pp. 303-318.
15. Ullah, I., Raza, B., Malik, A. K., Imran, M., Islam, S. U. & Kim, S. W. (2019) A churn prediction model using random forest: analysis of machine learning techniques for churn prediction and factor identification in telecom sector. *IEEE access.* 7(pp. 60134-60149. doi: 10.1109/ACCESS.2019.2914999.
16. Verbeke, W., Martens, D., Mues, C. & Baesens, B. (2011) Building comprehensible customer churn prediction models with advanced rule induction techniques. *Expert systems with applications.* 38(3), pp. 2354-2364. doi: 10.1016/j.eswa.2010.08.023.
17. Witten, I. H., Frank, E. & Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques (Third Edition),* Boston: Morgan Kaufmann.