Customer churn prediction with popular machine learning algorithms

Ahmet Çalış

Kocaeli University, Faculty of Engineering, Turkey

e-mail: ieahmetcalis@gmail.com

Justyna Kozłowska

Politechnika Białostocka, Wydział Inżynierii Zarządzania

e-mail: j.kozlowska@pb.edu.pl

Abstract

In today's increasingly competitive environment, it is necessary to follow the needs, demands, and expectations of customers closely for the enterprises and to respond in the most appropriate and fastest way. It aims to gain customer loyalty by developing mutual relations with customers and thus to provide long-term benefit to the enterprise. Today, the cost of earning new customers is much more than the cost of keeping existing customers. Providing promotions, discounts, gifts, or benefits to the customers who are anticipated to churn may hinder the churn customer and thus make more profit in the long term. However, if the wrong prediction is made, this causes unnecessary promotions or gifts to the customer. So for the company, this means unnecessary costs. Therefore, it is important for companies to correctly estimate the churn of customers. With the help of technology, enterprises can analyze the data they collect from different sources by using various data mining methods and obtain more valid information about the customers, and thus develop more effective communication with customers and ensure their continuity. The aim of this study is to analyze the results of customer churn prediction using various data mining techniques and classification algorithms of machine learning. The data analyzed were obtained from a telecommunication company. In the data set, there were 7166 customer records including the data about customer churn. This study also aims to estimate customers' churn with the highest rate. With the train test split, the data set was divided into 70% - 30% training and test data set. Scale and log transformations are performed on data. The performance of the models obtained by classification algorithms was examined.

Key words

churn prediction, classification algorithms, machine learning, telecommunication sector

Introduction

The continuity of existing customers is very important, in addition to acquiring new customers in every business. In the telecommunication sector, the continuity of existing customers rather than the acquisition of new customers causes them to be attractive as a research subject both in terms of cost and to protect and increase the market share [CX Network Report, 2016]. Although researches carried out on cancellation analysis have become widespread in the telecommunication industry on mobile and landlines, researches for internet service providers have become widespread in recent years. Considering the sector based on the saturation of the market and increasing the competition in the market, customer continuity activities are important. Companies need to continue existing customers both in terms of prestige and cost. Only in terms of cost, it will be seen that costs of continuity activities are smaller than marketing, sales, and installation costs for new customers [Athanassopoulos, 2000]. Especially in the telecommunication sector, the cost of lost customers is high in the days when the competition is at very high levels [Jain, Khunteta, and Srivastava, 2020].

Companies that want to reduce the cost of lost customers and want to maintain their presence in the sector need a decision support mechanism such as churn analysis within the scope of continuity activities. The problem here is the multiplicity and complexity of customer data in the telecommunications sector. It should be possible to interpret the data required for the churn analysis and to interpret the data according to the solution methodology to be applied. If any data containing the customer's behaviour is properly interpreted churn analysis should be valuable. The aim of this article is to analyze the results of customer churn prediction using various data mining techniques and classification algorithms of machine learning. In the literature, there are studies that address customer churn prediction in telecommunication industry [e.g. Amin, 2019; Höppner at al., 2020; Verbeke at al., 2012] but they usually focus on one method or model. In this study, a few methods and algorithms are used and the results are compared in terms of accuracy.

1. Telecommunications sector

Telecommunication (remote communication) has passed the French word télécommunication. Sensation, writing, painting, icon, or all kinds of information transmitted by wire, radio, optics, and other electromagnetic systems means that they are being published or received. In 1819, the Danish physicist Hans Christian Oersted found that a needle with magnet movement could be a signal tool. Next, in the 1820s the French physicist Andre Marie Ampere invented the moving needle. Behind, the modern era of telecommunications began in the world with the presentation of the first electromagnetic telegraph by American inventor Samuel Morse in the US Congress In 1844. Point-to-point communication started with a telegram and later, in the 1870s, the work of Elisha Gray and Alexandr Graham Bell developed with the invention of the phone [Özçağlayan, 1998]. Both investigated possibilities to send multiple messages using frequency separations on a telegraph circuit. Gray has been able to use this approach to send some simple voice messages but has focused his work on developing a better telegraph. Bell focused on this new issue by seeing the same audio transmission possibilities, and by 1876, he had patented the first phone and prevented Gray. By the 1880s, the phone was on the verge of entering homes, as well as for business use [Keskin, 2001].

At the beginning of the 1900s, the sudden emergence of many operators with non-interconnected local networks did not pose a major problem for telecommunications. But over time, with the integration of the network, the complexity of the issue has been revealed for the service to switch between these networks and offer it to the user smoothly. With the technological level of the 1920s, the only way to overcome this was to satisfy the market in the form of providing end-to-end service. In this process, digital communication is a very important change in putting individuals into a customer position. After the military bureaucracy, civil bureaucracy, multinational companies, medium-sized enterprises, and small-sized enterprises, the people came to the order with the personalization of computers in the 1980s. [Klemperer, 2000]. Developments in the electronics industry have enabled the development of satellite systems in mobile communication services and transmission, as well as the developments in power plants and consumer devices. There have also been technological developments in wire communication. Instead of conventional overhead line transmission systems, the transmission of coaxial cable and fiber optic cable transmission systems has increased speed, quantity, quality, and reliability, and the costs decreased [Keskin, 2001]. While circuit-switched systems were used in the sector from the 1840s until the 1990s, packet-switched technology developed for security purposes caused great changes in the field of telecommunications. The internet, which is the most common example of this, has been opened to civilian use since 1988 and has begun to change the structure of the telecommunication networks that occurred in about 150 years. It is expected that significant changes will be experienced in communication with the use of 3rd generation networks, which are a real combination of digital mobile phones and the internet, which have become widespread since the mid-1990s [Safel, 2001].

All sectors of the technology in the world to make them the perfect customer experience and improve their infrastructure, their ability to compete as a crucial precondition appeared. Technology for the telecommunications industry is the only condition of existence. The technological infrastructure and service quality are the main priority of the telecommunication sector at all times. For example, the network connection was only a problem until yesterday; today, it has significant importance on human life. The process may not be complete yet, but everything we can see in the very near future and will become everyone's interconnectedness.

A quality link infrastructure is the main factor that will ensure the continuity of life without risk. Increasing network traffic and the number of interconnected devices require the telecommunications sector to be ahead of all sectors. Providing the necessary speed and bandwidth, upgrading the network of 5G technologies, maintaining all numbers of customers simultaneously, and not compromising the best customer experience should be the first steps of the telecommunications sector. Investments directly affect customer experience and customer experience directly affects market shares.

Telecommunication companies have to utilize the Internet of Things to continuously monitor national and regional networks that span a wide range of information, to report, to solve problems, to reduce maintenance and repair costs associated with people.

Big data, which is the focus of the whole world, has a different meaning for telecommunications companies. Moreover, a significant portion of the formation of large data provides the telecommunications sector. Aside from the customer's personal data, they have huge data sources generated by the network operation, such as system logs of telecommunication towers and mobile networks, cases, and performance measurements. It is very important that the telecommunication sector can be included in the system by analyzing these data meticulously. Mobile technologies, cloud technology, the Internet of Things, artificial intelligence, digital payments, and all innovative solutions are based on the analysis of big data.

Telecommunication companies are responsible for all personal, commercial, and official connections through their networks. Among the expected features of the telecommunication company's infrastructures is the security of networks in technologies such as e-government applications, banking systems, interconnected medical devices, the technology of digital production systems, creating networks that are resistant to possible leaks and attacks, intervening in the problems and developing the necessary defense methods.

2. The importance of churn analysis in the telecommunications industry

Today, communication needs are often considered not as an extra cost but as a need. The telecommunications sector tends to be mostly saturated in countries with a competitive market for this sector [Verbeke et al., 2012]. In fact, in other words, the growth rate of the market is shifting towards a smaller acceleration day by day. In light of these evaluations, gaining new subscribers is becoming increasingly difficult. This situation leads companies to conduct activities such as churn analysis and carry out activities to keep them in the system. Given that the cancellation rate in the competitive environment is 2.2% per month [Chih-Ping and I-Tang, 2002], it will not be difficult to predict how important these activities are. Considering the following reasons, it is understood why churn analysis is important in the telecommunications sector:

- Churn customers by 2.2% per month indicate a churn of around 25% per annum.
- With churn customers per year for a telecommunications company, the quarterly turnover constitutes a hidden cost.
- Continuity costs cost 5 times less than the cost of new customers [Swift, 2001].

The longer the customers' subscription life, the more profit is considered for the company. Companies tend to prefer long-term contracts, campaigns, or tariffs instead of short-term customer relationships. Due to the value of loyal customers, churn analysis is applied to create loyalty.

3. Data analysis

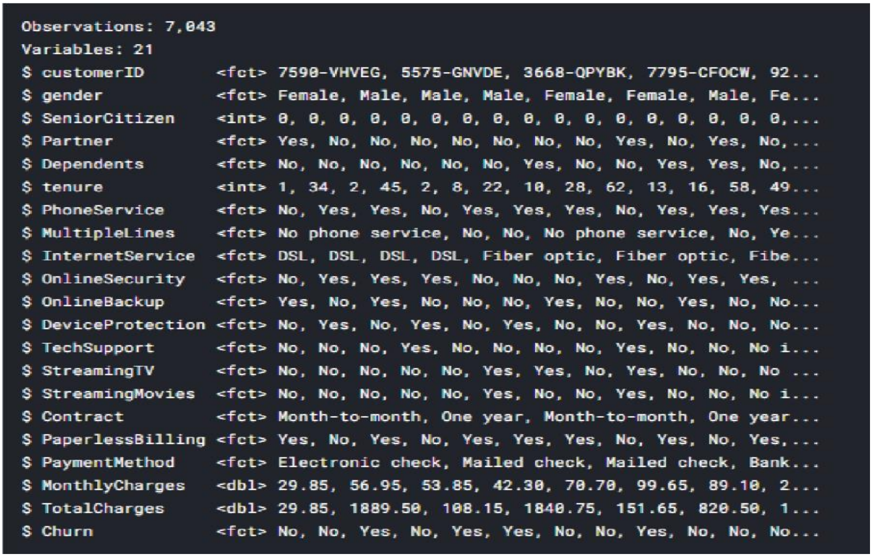
In this study, we used Telco communication company customer data. Telco customer data contains information about home phone and Internet services to 7,043 customers in California in the third quarter that provides which customers leave, stay in use, or sign up for their services (new customer) displays. Data has 21 colons which indicate demographic attributes and whether the customer continues membership or not.

3.1. Variables in analysis

Each customer record contains the following information, which were used as variables (see Figure 1) in this study:

CustomerID: A unique ID that identifies each customer.,

Gender: The customer's gender: Male, Female
Senior Citizen: Indicates if the customer is 65 or older: (Yes, No)
Dependents: Indicates if the customer lives with any dependents: Yes, No. Dependents could be children, parents, grandparents, etc.
Tenure: Indicates the total amount of months that the customer has been with the company.
Phone Service: Indicates if the customer subscribes to home phone service with the company: Yes, No
Multiple Lines: Indicates if the customer subscribes to multiple telephone lines with the company: Yes, No
Internet Service: Indicates if the customer subscribes to Internet service with the company: No, DSL, Fiber Optic, Cable.
Online Security: Indicates if the customer subscribes to an additional online security service provided by the company: Yes, No
Online Backup: Indicates if the customer subscribes to an additional online backup service provided by the company: Yes, No
Device Protection: Indicates if the customer subscribes to an additional device protection plan for their Internet equipment provided by the company: Yes, No
Tech Support: Indicates if the customer subscribes to an additional technical support plan from the company with reduced wait times: Yes, No
Streaming TV: Indicates if the customer uses their Internet service to stream television programming from a third-party provider: Yes, No. The company does not charge an additional fee for this service.
Streaming Movies: Indicates if the customer uses their Internet service to stream movies from a third-party provider: Yes, No. The company does not charge an additional fee for this service.
Contract: Indicates the customer's current contract type: Month-to-Month, One Year, Two Year.
Paperless Billing: Indicates if the customer has chosen paperless billing: Yes, No
Payment Method: Indicates how the customer pays their bill: Bank Withdrawal, Credit Card, Mailed Check
Monthly Charge: Indicates the customer's current total monthly charge for all their services from the company.
Total Charges: Indicates the customer's total charges, calculated to the end of the quarter specified above.
Churn: Yes = the customer left the company. No = the customer remained with the company.



```
Observations: 7,043
Variables: 21
$ customerID       <fct> 7590-VHVEG, 5575-GNVDE, 3668-QPYBK, 7795-CFOCW, 92...
$ gender           <fct> Female, Male, Male, Male, Female, Female, Male, Fe...
$ SeniorCitizen    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
$ Partner          <fct> Yes, No, No, No, No, No, No, No, Yes, No, Yes, No,...
$ Dependents       <fct> No, No, No, No, No, No, Yes, No, No, Yes, Yes, No,...
$ tenure           <int> 1, 34, 2, 45, 2, 8, 22, 10, 28, 62, 13, 16, 58, 49...
$ PhoneService     <fct> No, Yes, Yes, No, Yes, Yes, Yes, No, Yes, Yes, Yes...
$ MultipleLines    <fct> No phone service, No, No, No phone service, No, Ye...
$ InternetService  <fct> DSL, DSL, DSL, DSL, Fiber optic, Fiber optic, Fibe...
$ OnlineSecurity   <fct> No, Yes, Yes, Yes, No, No, No, Yes, No, Yes, Yes, ...
$ OnlineBackup     <fct> Yes, No, Yes, No, No, No, Yes, No, No, Yes, No, No...
$ DeviceProtection <fct> No, Yes, No, Yes, No, Yes, No, No, Yes, No, No, No...
$ TechSupport      <fct> No, No, No, Yes, No, No, No, No, Yes, No, No, No i...
$ StreamingTV      <fct> No, No, No, No, No, Yes, Yes, No, Yes, No, No, No ...
$ StreamingMovies  <fct> No, No, No, No, No, Yes, No, No, Yes, No, No, No i...
$ Contract         <fct> Month-to-month, One year, Month-to-month, One year...
$ PaperlessBilling <fct> Yes, No, Yes, No, Yes, Yes, Yes, No, Yes, No, Yes,...
$ PaymentMethod    <fct> Electronic check, Mailed check, Mailed check, Bank...
$ MonthlyCharges   <dbl> 29.85, 56.95, 53.85, 42.30, 70.70, 99.65, 89.10, 2...
$ TotalCharges     <dbl> 29.85, 1889.50, 108.15, 1840.75, 151.65, 820.50, 1...
$ Churn            <fct> No, No, Yes, No, Yes, Yes, No, No, Yes, No, No, No...
```

Fig. 1. Variables in churn customer prediction analysis

Source: own elaboration with the use of Python.

4. Exploratory Data Analysis (EDA)

4.1. Target variable analysis

The target variable in this dataset is the “Churn”. The aim of the study is to analyze the data. And, in the next step, to predict this variable.

In figure 2, "count" indicates the number of users on the x-axis. A “churn” variable indicates whether the customer has left the relevant company or not.

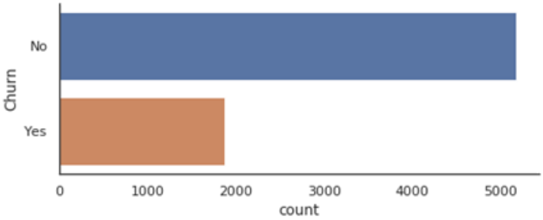


Fig. 2. Count plot for churn

Source: own elaboration with the use of matplotlib package in Python.

The percentage representation of the data in the table is as follows:

- Percentage of Leaving Customers: 27.6
- Percentage of Customers Who Didn't Leave: 72.4

Significant difference between departing customers and continuing customers have been observed. The imbalance between the values may influence the result of the model's training during the training phase.

4.2. Numerical variables' analysis

The graph presented in the figure 3 shows the distribution of the Churn variable according to the duration of use. Customers with low tenure are more likely to leave the company. According to the inference, the 'Tenure' is an important variable that we will use when predicting the Churn variable.

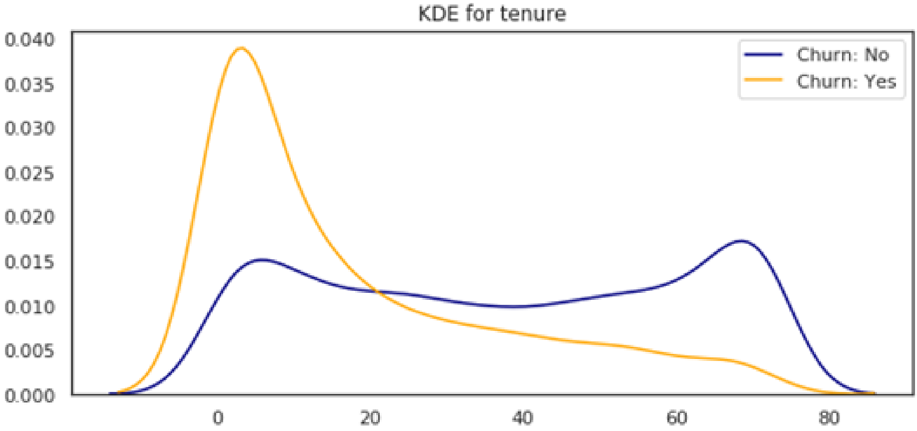


Fig. 3. Churn by tenure

Source: own elaboration with the use of matplotlib package in Python.

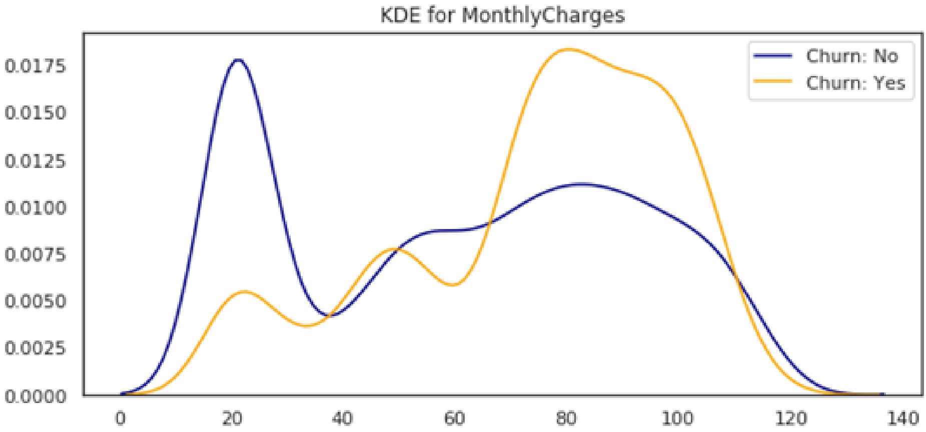


Fig. 4. Churn by monthly charges

Source: own elaboration with the use of matplotlib package in Python.

The graph presented in figure 4 shows the distribution of the Churn variable according to the monthly payment amount. Users with higher monthly payments are more likely to leave the company. According to this inference, it will be one of the important variables that we will use in the prediction.

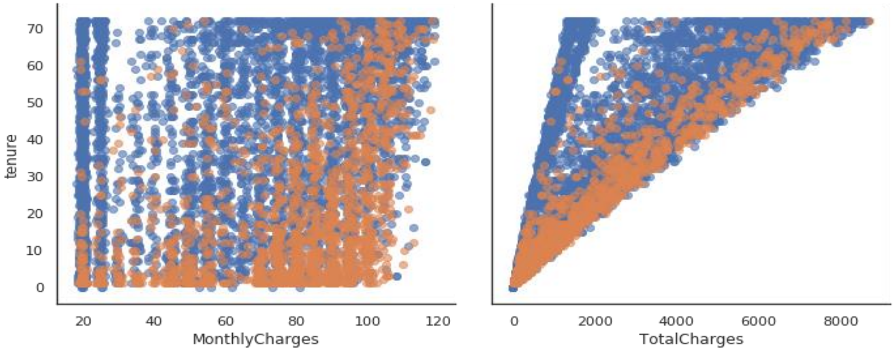


Fig. 5. Churn by charges

Source: own elaboration with the use of matplotlib package in Python.

In figure 5 the orange colour represents customers who churned, the blue color represents customers who are still customers. According to the first graph (on the left), customers with low usage time and high monthly payment amounts seem more likely to churn. In the second graph (on the right), the customer who has low usage and low total charges are more churned from the company.

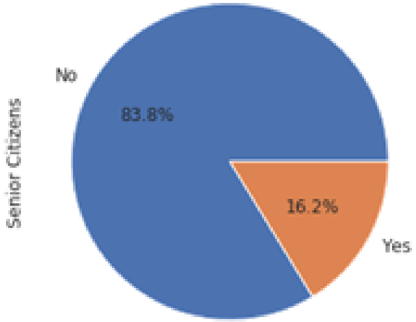


Fig. 6. Pie chart for senior citizens

Source: own elaboration with the use of matplotlib package in Python.

According to the pie chart presented in figure 6, 83.8% of users are under the age of 65, and 16.2% are customers over the age of 65. A value of 0 in the given bar chart presented in figure 7 indicates that the person is under 65 years old. A value of 1 indicates over 65 years of age. The churn rate is higher in users over 65 years of age compared to users under 65 years old.

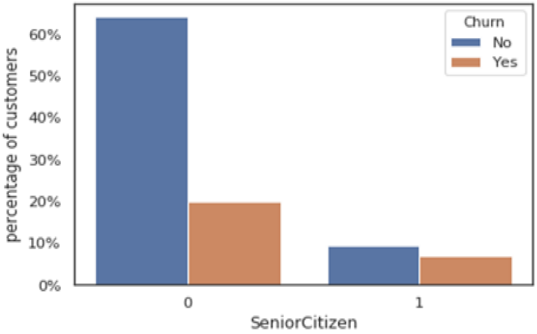


Fig. 7. Churn by senior citizens

Source: own elaboration with the use of matplotlib package in Python.

The percentage of customers who have more than one phone line or not, their proportions are indicated in figure 8. The proportion of customers who do not use the phone service is lower than costumers who use multiple lines.

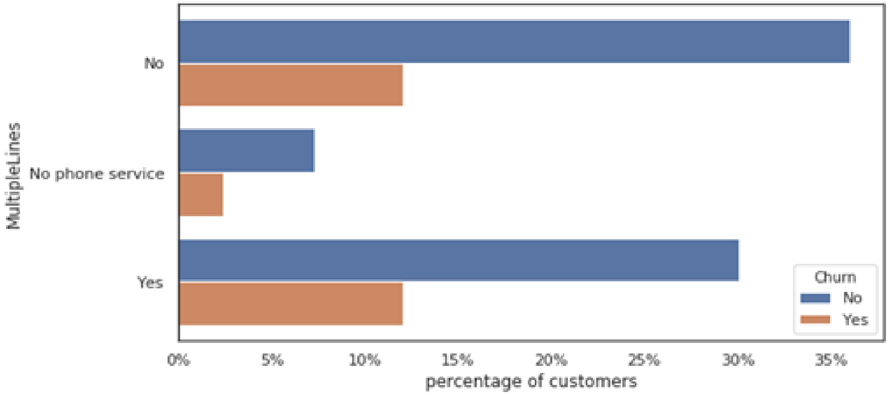


Fig. 8. Churn by multiple lines

Source: own elaboration with the use of matplotlib package in Python.

According to the graph presented in figure 9, customers who do not have internet usage have a lower churn rate. Customer who has fiber optic internet service are more likely to churn than a customer who has DSL.

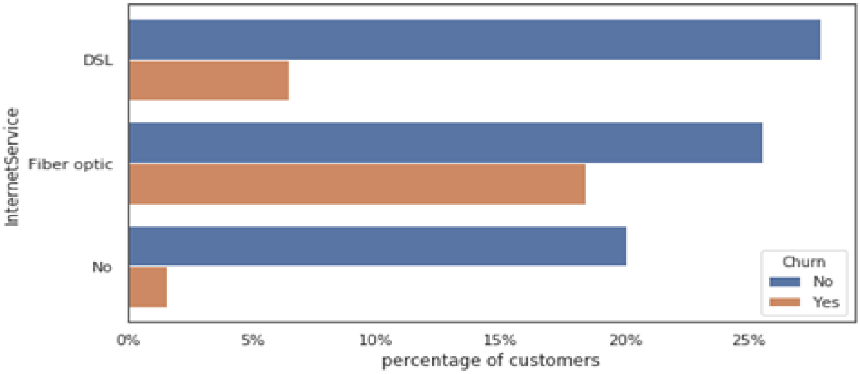


Fig. 9. Churn by internet services

Source: own elaboration with the use of matplotlib package in Python.

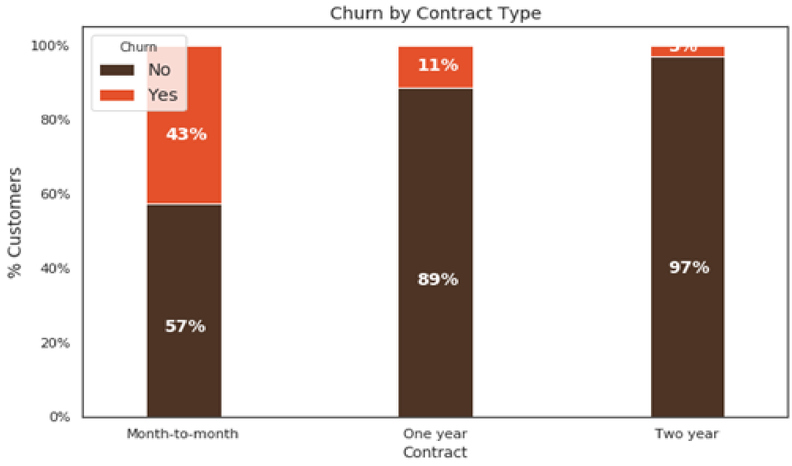


Fig. 10. Churn by contract type

Source: own elaboration with the use of seaborn package in Python.

In figure 10 the churn by contract distribution is presented. The percentage of the month-to-month contract customers is higher than that of customers who have an annual or 2-year contract. Customers with an annual and two-year contract are less likely to churn.

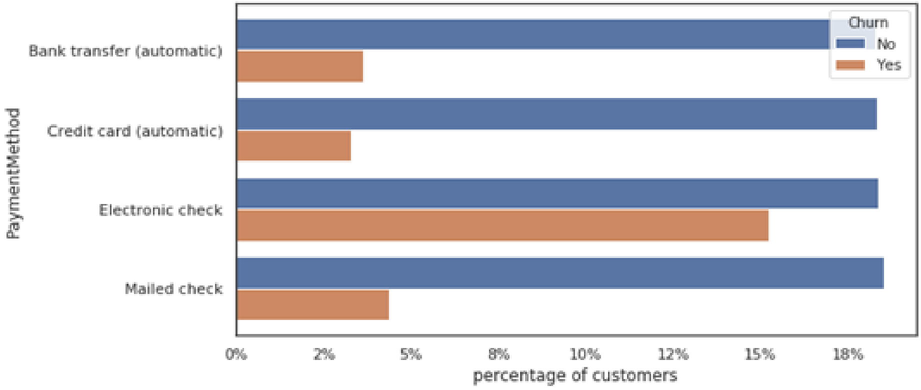


Fig. 11. Churn by payment method

Source: own elaboration with the use of matplotlib package in Python.

The percentages of customers who use different payment methods are churn or not are indicated according to the payment channel they use (figure 11). Significantly, customers who make electronic payments are more likely to churn than customers who use other methods.

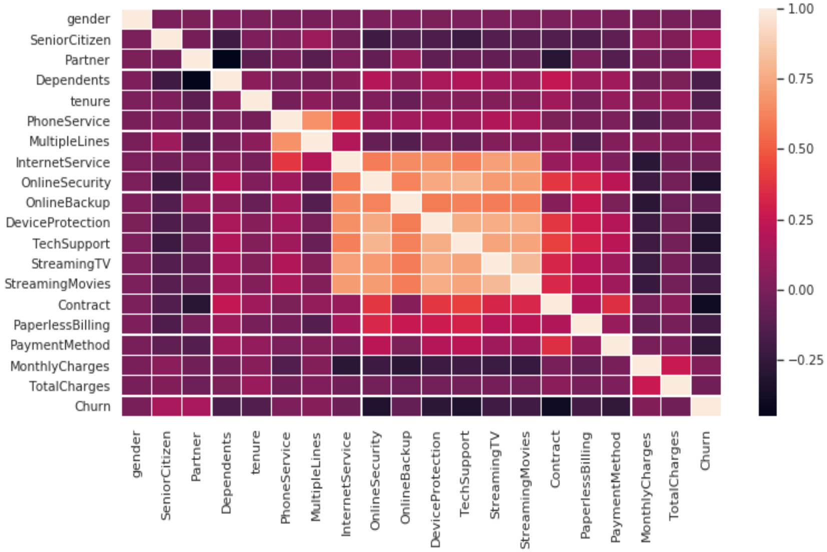


Fig. 12. Correlation matrix

Source: own elaboration with the use of matplotlib package in Python.

At the correlation matrix (figure 12) the relationships between variables are indicated with the colors. If the color is light that means the correlation between those two attributes is strong and positive, on the other hand, if the color is so dark that means there is a strong negative correlation between those two attributes. According to this, as we interpreted from charts, contract and churn have a negative and strong correlation.

5. Prediction results

To use the data for prediction some adjustment had to be conducted. First, a change of the categorical variables to ‘0’ and ‘1’ was needed (dummy variables). When the data was ready for machine learning, the set was split into 4 chunks such as x_train, x_test, y_train, y_test. We used 70% of the data for train, 30% of data for test. The target variable (Y) is the "churn" so that means y_train and y_test just take values from one column which is "churn" and x_train and x_test take the rest of the variables.

As the accuracy of the models are crucial in the analysis, the accuracy score and recall score were used as performance metrics of the model. They can be obtained from the confusion matrix.

5.1. Logistic regression

The accuracy score is obtained as 81% from that model. That means the prediction of the model has correct prediction 81 of 100. It's a good score for a machine learning algorithm but sometimes it's better to use other metrics such as recall score if you have unbalanced data. Because the accuracy score mislead us with unbalanced data. For example, we have 99 non-cancer and 1 cancer people in the data. We are trying to detect a cancer person. So even though the worst algorithm can reach a good score in that situation while saying non-cancer for all people. That means 99% accuracy but in the reality, that means 0% because we missed all cancer people which is in the data. To prevent this, it is useful to use a recall or F1 score. In the study, we used just recall to measure the performance of the model.

Tab. 1. Logistic regression confusion matrix

Confusion Matrix	
1354	130
278	348

Source: own elaboration with the use of Python.

According to the results of the model (table 1), 1354 non-churned person predicted correctly. Even though 130 people didn't churn they predicted as a churned by the model. While 278 people have churned but the model predicted them as not churned and 348 people who churned were detected by the hypothesis.

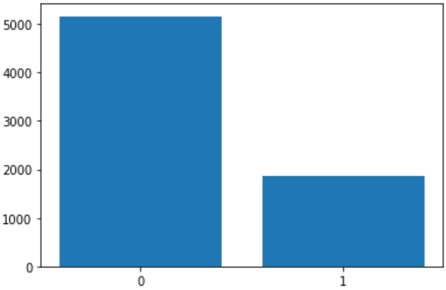
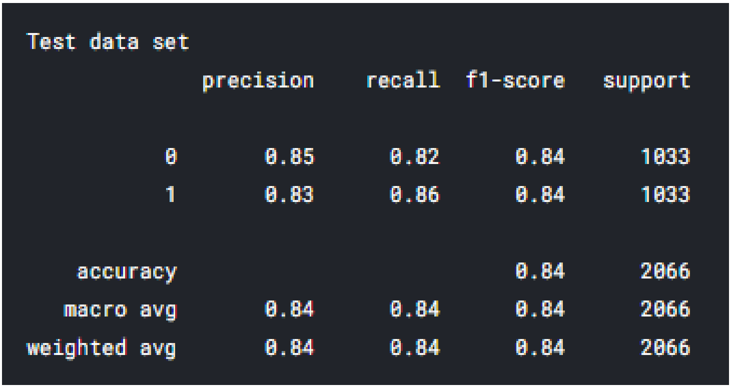


Fig. 13. Countplot for churn

Source: own elaboration with the use of matplotlib package in Python.

In the chart presented in figure 13 we can see the number of people who churned and not churned so it's obvious to see that the data is unbalanced. To fix that situation, there are some techniques such as oversampling, undersampling, SMOTE, and ADASYN. A SMOTE was chosen to balance the data because this technique gave better results than others. The results after balancing the data with SMOTE are presented in figure 14.



```
Test data set
              precision    recall  f1-score   support

           0       0.85      0.82      0.84      1033
           1       0.83      0.86      0.84      1033

    accuracy                           0.84      2066
   macro avg       0.84      0.84      0.84      2066
weighted avg       0.84      0.84      0.84      2066
```

Fig. 14. Results of logistic regression after balancing the data

Source: own elaboration with the use of Python.

According to the results, a 3% was gain in accuracy score. Also, the precision and recall scores from the figure 14. Precision indicates how many people predicted by the model as a churn is in fact a churn. On the other hand, recall means, how many of the real churned clients were predicted correctly by themodel. F1 score is the mix of both these metrics precision and recall. Before balancing thedata, the recall score was 56% but now it’s %86. That means we improved 30% recall score by balancing the data.

5.2. K-Nearest neighbor

K- nearest neighbour (KNN) is an algorithm that looks at the nearest data points around the data point that we are trying to predict. After trying lots of options, the optimum number of neighbours as 7. The accuracy score of the model was 76%.

Tab. 2. K-Nearest neighbor confusion matrix

Confusion Matrix	
1333	151
355	271

Source: own elaboration with the use of Python.

As it can be read from the confusion matrix (table 2), even though the accuracy score is 76%, recall is 43% so that means this model is not a good option to make decisions.

5.3. Decision tree

The optimum depth as 7 was found for the decision tree. An accuracy score of 78% was optained in this prediction model (table 3).

Tab. 3. Decision tree confusion matrix

Confusion Matrix	
1287	197
267	359

Source: own elaboration with the use of Python.

Recall obtained as 57%, which means the model is doing correctly about predicting churned people as churned more than half of them.

5.4. Random forest

The random forest contains decision trees. The model detects the most important variables for predicting the target variable with help of decision trees. In this model, 100 decision trees were used. The model gave a 79% accuracy score (table 4).

Tab. 4. Random forest confusion matrix

Confusion Matrix	
1359	125
311	315

Source: own elaboration with the use of Python.

The recall score is 50%. That means the model predicts correctly the half of really churned people in the data.

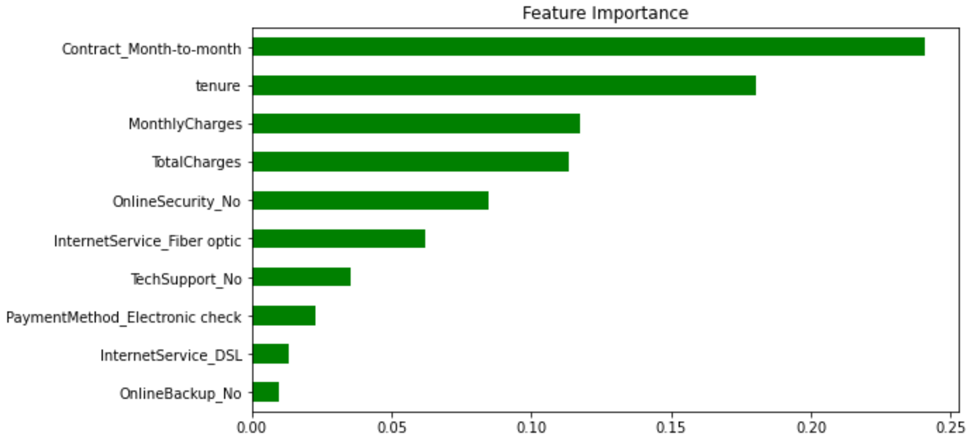


Fig. 15. Feature importance for random forest

Source: own elaboration with the use of Python.

As it can be noticed in the feature importance chart (figure 15), contract month to month is most important one as we thought before according to bar and line graphs.

5.5. SVC (Support Vector Classifier)

The model was created with a linear kernel. Obtained accuracy score is 78%.

Tab. 5. Confusion Matrix SVC

Confusion Matrix	
1386	98
352	274

Source: own elaboration with the use of Python.

Even though the accuracy score is enough (see table 5), this model is not the best option to choose because of the recall score (43%). That means the model is predicting more than half of churned people as non-churned.

Conclusions and recommendations

In this study, the Telecom Company's data were analyzed which contains 7043 person's demographic and categoric attributes. The aim of this study was to predict the customers' churn with machine learning algorithms and compare the results in terms of accuracy. The model created with logistic regression gave an 81% accuracy score. After detection of unbalanced data, the SMOTE technique was used to balance the data. This technique produces synthetic data to balance the target variable. Afterward, a new logistic regression model with balanced data was built. For the accuracy score, a 3% improvement was gained. In addition, a 30% improvement for recall score was achieved which is dramatically better than before.

The best accuracy score was gained with logistic regression (81%), best recall score with a decision tree (57%) for unbalanced data. Although, the accuracy metric is really important for measuring model performance, in this study recall is more crucial to decide about model performance because of the unbalance situation about the target variable. Because of this, if there is no dramatic difference between accuracy scores, it is wiser to choose the model which has a better recall score. According to this, the best model with unbalanced data is the decision tree even though the logistic regression has a better accuracy score, a decision tree is 3% better than logistic regression as a metric of recall. In conclusion, it is recommended to choose a decision tree as a model when using unbalanced data.

Only a logistic regression model was created with balanced data. A new model was better than all of the models that were trained. No attempts were made with balanced data set and other models so it is not possible to comment about the models' performance with balanced data. Though, it is obvious that logistic regression improved the performance dramatically after data balanced.

In conclusion, the best model that was trained is logistic regression after SMOTE. It is highly recommended to decide with that model for customer churn prediction by Telco C.

ORCID iD

Justyna Kozłowska: https://orcid.org/0000-0001-5164-4023

Literature

1. Amin A., Al-Obeidat F., Shah B., Adnan A., Loo J., Anwar, S. (2018), *Customer churn prediction in telecommunication industry using data certainty*, Journal of Business Research, 1 (1), pp. 1-12.
2. Athanassopoulos A. (2000), *Customer satisfaction cues to support market segmentation and explain switching behavior*, Journal of Business Research, 47 (3), pp. 191-207.
3. Başarslan M.S. (2017), *Customer Churn Analysis in the Telecommunications Industry. Graduate School of Natural and Applied Sciences, Department of Computer Engineering*, Düzce University, Düzce
4. Burez J., Van den Poel D. (2009), *Handling class imbalance in customer churn prediction*, Expert Systems with Applications, 36 (3) part 1, pp. 4626-4636.
5. Coşkun C., Baykal A. (2011), *Comparison of classification algorithms in data mining on a sample (eng. ...)*, Academic Informatics Conference, 2-4 February 2011, İnönü University, Malatya.
6. Coussement K., Lessmann S., ve Verstraeten G. (2017), *A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry*, Decision Support Systems, 95, pp. 27-36.
7. CX Network Report (2016), *Telecoms Focus: How to Reduce Churn and Improve Customer Retention*. Available online at: https://www.cxnetwork.com.
8. Eulor T. (2005), *Churn Predictionin Telecommunications Using Mining Mart*, 5th IEEE International Conference, Data Mining (ICDM), Texas, USA.

9. Günay M., Ensari T. (2018), *New Approach for Predictive Churn Analysis in Telecom*, The 2018 International Conference on Applied Mathematics and Computational Methods in Engineering (AMCME 2018), Venice, Italy.
10. Günay M., Ensari T. (2018), *Customer ChurnAnalysis with Machine Learning (eng….)*, 26th Signal Processing and Communications Applications Conference, May 2-5, 2018, Cesme, Izmir, Turkey.
11. Günay M., Ensari T. (2018), *EEG Signal Analysis of Patients with Epilepsy Disorder Using Machine Learning Techniques*, The Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science in 2018 (EBBT'2018) İstanbul, Turkey.
12. Gürcan M. (1998). *Lojistik regresyon analizi ve bir uygulama*.
13. Höppner S. et al., (2020), *Profit driven decision trees for churn prediction*, European Journal of Operational Research, 284, pp. 920-933.
14. Jain H., Khunteta A., Srivastava S. (2020), *Churn Prediction in Telecommunication using Logistic Regression and Logit Boost*, Procedia Computer Science, 167, pp. 101-112.
15. Vafeiadis T., Diamantaras K.I., Sarigiannidis G., ve Chatzisavvas K.Ch. (2015), *A comparison of machine learning techniques for customer churn prediction*, Simulation Modelling Practice and Theory, 55, pp. 1-9.
16. Verbeke W. et al. (2012), *New insights into churn prediction in the telecommunication sector: A profit driven data mining approach*, European Journal of Operational Research, 218, pp. 211-229.

Prognozowanie ryzyka odejścia klientów za pomocą popularnych algorytmów uczenia maszynowego

Streszczenie

W dzisiejszym, coraz bardziej konkurencyjnym środowisku, konieczne jest dokładne śledzenie potrzeb, wymagań i oczekiwań klientów oraz reagowanie na nie w najbardziej odpowiedni i najszybszy sposób. Ma to na celu zdobycie lojalności klienta poprzez rozwijanie wzajemnych relacji, a tym samym zapewnienie długoterminowych korzyści dla przedsiębiorstwa. Obecnie koszt pozyskania nowych klientów jest znacznie wyższy niż koszt utrzymania dotychczasowych klientów. Zapewnienie promocji, rabatów, prezentów lub innych korzyści dla klienta, który jest skłonny do rezygnacji może go powstrzymać, a tym samym pozwoli osiągnąć więcej zysku w dłuższej perspektywie. Jednakże, błędne przekonanie, że klient jest skłonny do rezygnacji oznacza niepotrzebnie poniesione koszty promocji lub prezentów dla klienta. Dlatego ważne jest dla firm, aby prawidłowo oszacować ryzyko odejścia klientów. Z pomocą technologii, przedsiębiorstwa mogą analizować dane, które zbierają z różnych źródeł przy użyciu różnych metod eksploracji danych i uzyskać bardziej wiarygodne informacje o klientach, a tym samym rozwijać bardziej efektywną komunikację z klientami i zapewnić ich ciągłość. Celem niniejszego opracowania jest analiza wyników predykcji ryzyka rezygnacji klientów z wykorzystaniem różnych technik eksploracji danych oraz algorytmów klasyfikacji uczenia maszynowego. Analizowane dane zostały pozyskane z firmy telekomunikacyjnej. W zbiorze danych znajdowało się 7166 rekordów klientów, w tym dane dotyczące rezygnacji klientów. Celem badania jest również oszacowanie rezygnacji klientów z najwyższym wskaźnikiem. Przy podziale zbioru danych zastosowano proporcje: 70% danych jako zbiór uczący oraz 30% danych jako zbiór testowy. Zbadano wydajność modeli uzyskanych za pomocą algorytmów klasyfikacyjnych.

Słowa kluczowe

przewidywanie ryzyka odejścia, algorytmy klasyfikacyjne, uczenie maszynowe, sektor telekomunikacji