Churn Prediction in Telecommunications Sector using Machine Learning

Andreea-Maria COPĂCEANU
The Bucharest University of Economic Studies, Romania
andreea.copaceanu@csie.ase.ro

In these days, due to the increasing competition, churn prediction has gathered greater interest in business, especially in the telecom industry, since gaining new customers is more expensive than retaining the existing ones. The primary objective in telecom churn analysis is to accurately estimate the churn behavior by identifying the customers who are at risk of churning. Another objective is to identify the main reasons for customer churn. This paper focuses on various machine learning algorithms for predicting customer churn, though which we build classification models such as Logistic Regression, Random Forest, Decision Tree, and Support Vector Machine. Prediction performance of the classifiers is evaluated and compared through measures such as Area Under the Curve (AUC), accuracy, and recall rate. Such predictive models have the potential to be used in the telecom industry for making better decisions in customer management.

Key words: *Churn Prediction, Machine Learning, Retention, Telecommunication, Decision Tree*

1 Introduction

In a highly competitive environment, companies must constantly innovate and focus on improving the quality of their services. Customer loyalty represents more than just keeping the right customers, but it represents the key to maximizing profits. Acquiring a new customer is from 5 to 25 times more expensive than retaining an existing one. According to Frederick Reichheld of Bain & Company, inventor of the net promoter score (NPS), increasing customer retention rates by 5% increases profits by 25% to 95%.[1] Considering this, customer retention is one of the main challenges of companies, especially in the telecommunication industry, where customers have multiple options in terms of better and less expensive services. The most often cause of customers churn is due to non-satisfaction in the service offered by a provider or due to more enhanced affordable service by another service provider. [2] In an almost saturated market, companies need a method to identify the customers who are most likely to churn, so that they can build proactive retention campaigns. [3] Thus, an appropriate churn prediction model is essential to predict the customer churn. The proposed model should have the capability to accurately identify customers at risk to churn and then find the reasons behind churning, so as to avoid loss of customers and also propose measures to retain them. The effectiveness of a churn prediction model depends on the learning achieved from the data set provided. An appropriately preprocessed data set gives high performance to the classifiers. Therefore, proper preprocessing is required to remove any redundant or useless features that do not have any relation to the target feature. [4] In general, machine learning techniques are greatly introduced as churn prediction methods. [2] These techniques can help building prediction models in order to discover behaviors and future trends and allow companies to make smart decisions, based on the knowledge extracted from the data. The objective of this study is to investigate existing techniques in machine learning, to evaluate the classifier models for customer churn predictions, and to identify the churn key factors, using data from a telecom public data set. From the experiments, we observed that Random Forest produced better accuracy compared to other machine learning algorithms, while Decision Tree produced a better recall rate compared to other algorithms. We identified the factors behind the customers churning by using the Feature Selection technique. The rest of the paper is structured as follows. In Section 2 we present related work. In Section 3 we present the telecom churn case study. In Section 4 we expose the results. Finally, in Section 4 we conclude the paper.

2. Related work

Churn prediction in telecom companies has been addressed in the literature using various techniques, including machine learning or data mining. These techniques are aimed to assist companies to identify, predict, and retain churning customers.

In [3], the authors presented an advanced data mining methodology which predicts customer churn using a call record data set for 3333 customers with 21 features. The author applied the principal component analysis (PCA) technique to reduce the data dimensionality. Three machine learning algorithms were used for classification: Neural Networks, Support Vector Machine and Bayes Network. The overall accuracy values were 99.10%, 99.55%, and 99.70% for Bayes Networks, Neural networks, and Support Vector Machine, respectively.

Different machine learning algorithms were proposed in [5], through which different models were employed, such as Logistic Regression, Support Vector Machine, Random Forest, Gradient Boosting Trees, which are used to predict churn customers. The authors used AUC to measure the performance of the models. According to the AUC values, the method that gave the most accurate model was Gradient Boosting with AUC value of 84.57%.

The author presented in [6] a comparison study using three classifiers K-NN, Random Forest, and XG boost, respectively. The XG boost classifier performed the best, compared to the KNN and RF classifiers, in terms of accuracy score and F score.

A churn prediction model was proposed in [7], as well as clustering techniques to identify the churn customers. The proposed model used classification algorithms, in which the Random Forest algorithm performed the best with 88.63% correctly classified instances. Furthermore, the study also provided factors behind customer churn using the Attribute Selected Classifier algorithm.

In [8], the authors studied the problem of customer churn in a big data platform. The goal of the researchers was to prove that big data greatly enhance the process of predicting the churn depending on the volume, variety, and velocity of the data. The Random Forest algorithm was used and evaluated using AUC.

The authors proposed in [9] machine learning algorithms on a big data platform in order to predict the customer churn. The performance of the model was measured by Area Under Curve (AUC), obtaining a value of 93.3%. The model was built and tested through Spark environment, using a large data set provided by SyriaTel telecom company. The model experimented four algorithms: Decision Tree, Random Forest, Gradient Boosted Machine Tree “GBM” and Extreme Gradient Boosting “XGBOOST”. The best results were obtained for the XGBOOST algorithm.

A classification model for churn prediction based on the Rough Set Theory in telecom was proposed in [10]. The proposed Rough Set classification model outperformed the other models like Linear Regression, Decision Tree, and Voted Perception.

The author implemented in [11] a classification model based on Logistic Regression and used statistical methods for feature selection. The experimental results showed that by increasing the threshold values and selecting the right features with different combinations, the model will deliver better results in the churn prediction process.

3. Churn prediction case study using Machine Learning

Based on these concepts, a customer churn prediction case study will be further presented. A telecom data set is going to be used so that to predict churn using machine learning algorithms and detect the main factors that may lead the customer to switch to another telecom provider.

3.1 Data preprocessing

We are going to use a public telecom data set on churn. This data set contains various information about customers including customer care service details, customer value-added services, customer personal details, customer usage pattern and customer bill and payment details.

The data set contains 58 attributes and 51,047 observations, which indicate whether the customer had left the telecom provider or not.

Part of these features that are not relevant for churn prediction, will be removed from the data set. In this regard, the feature CustomerID, which contains numeric identifier of a customer, is not relevant in the analysis, and therefore it is removed.

In the same manner, feature NotNewCellphoneUser is removed, as it contains the opposite values of feature NewCellphoneUser, the latter will be retained in the data set.

Feature ServiceArea is also removed from the data set, as it contains alphanumeric values representing the service area of the telecom company and it is irrelevant for churn prediction. Also, the feature Homeownership is removed from data set.

There are 15 features in the data set containing missing values, as shown in Table 1. The feature HandsetPrice is dropped from the data set, as it contains more than 50% of missing values.

For the rest of the features in the table, missing values are replaced with the mean value of the entire feature column, using the fillna() method.

Table 1. Features having missing or “Unknown” value

No.	Variable	Missing Values
1	MonthlyRevenue	156
2	MonthlyMinutes	156
3	TotalRecurringCharge	156
4	DirectorAssistedCalls	156
5	OverageMinutes	156
6	RoamingCalls	156
7	PercChangeMinutes	367
8	PercChangeRevenues	367
9	ServiceArea	24
10	Handsets	1
11	HandsetModels	1
12	CurrentEquipmentDays	1
13	AgeHH1	909
14	AgeHH2	909
15	HandsetPrice	28982

Categorical variables can hide important information in the data set. Variable values with 2 categories will be converted to numbers using Python map() function, and the other 4 variables with more than 2 categories CreditRating, PrizmCode, Occupation and MaritalStatus, will be transformed using OneHotEncoder function of SciKit package, which maps a column of category indices to a column of binary vectors.

3.2 Data visualization

When a feature is analyzed independently, we are usually mostly interested in the distribution of its values. Churn is the target variable, and it is binary: Yes indicates that that the company lost this customer, and No indicates that the customer was retained. In the data set, 28.6% (14,257) customers are churners, whereas 71.5% (35,519) customers are non-churners, as shown in **Fig. 1.**

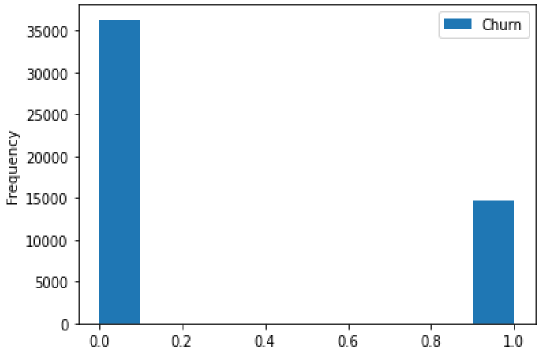


Fig. 1. Churn distribution

Bivariate analysis finds out the relationship between two variables. In this analysis, the distribution of categorical variables according to the Churn variable is plotted, as shown in **Fig. 2**. This shows us a skewed distribution for most part of the variables. Out of 21 variables, 17 variables have two values, and the remaining variables have three or more values.

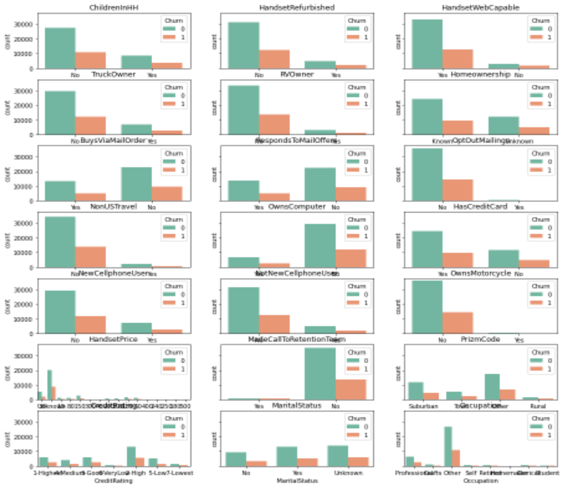


Fig. 2. Categorical variables and Churn Variable

Pearson correlation coefficients of numerical variables are shown in the correlation matrix in **Fig. 3**.

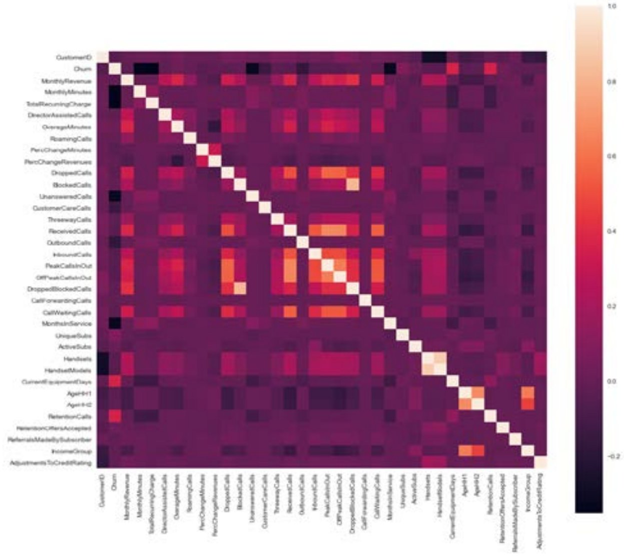


Fig. 3. Correlation matrix

Variables CurrentEquipmentDays, RetentionCalls, MonthlyMinutes, TotalRecurringCharge, UnansweredCalls, OutboundCalls, and MonthsInService are highly correlated with Churn predictor variable. (**Fig. 3**) Using the sklearn feature selection method, the most important variables to predict the churn are obtained and shown in **Table 2**.

Table 2. Feature importance score obtained using Feature Selection

No.	Feature name	F Scores
1	CurrentEquipmentDays	9,936.10
2	TotalRecurringCharge	7,251.70
3	MonthlyMinutes	7,027.36
4	MonthsInService	5,982.62
5	UnansweredCalls	5,768.01
6	RetentionCalls	5,233.20
7	OutboundCalls	1,998.18
8	CustomerCareCalls	882.28
9	MadeCallToRetentionTeam	625.62
10	UniqueSubs	512.07
11	CreditRating_5-Low	164.18
12	PercChangeMinutes	101.38
13	OffPeakCallsInOut	82.09
14	PeakCallsInOut	75.66
15	InboundCalls	72.94

The most highly correlated feature that predicts customer churn based on feature selection (**Table 2**) is CurrentEquipmentDays, which means that the longer a customer has the current equipment for, the more likely he is going to churn. This makes sense when we consider that a customer with an outdated equipment is likely to be looking for a new one. When looking for a new equipment, a customer without brand loyalty will consider several brands for it, thus increasing the chances that he will switch to another provider. Also, the fact that a customer has the same equipment for a long time without upgrading might be a signal that he is not using the service frequently, in which case it would also make sense that the customer churns.

3.3 Software tools

This study will use Python, which is one of the mainstream languages in data science and machine learning, providing libraries for data visualization, preparation, and machine learning tasks. The main packages in Python used for data analysis and machine learning are:

- **Pandas** is one of the main tools used by data analysts, for multiple phases of data science workflow, including data cleaning, visualization, and exploratory data analysis.
- **NumPy** is the fundamental package for numerical computation in Python and provides support for large multidimensional array objects.
- **Scikit-learn** is a machine learning library that implements a range of machine learning, preprocessing, cross-validation, and visualization algorithms.
- **SciPy** provides various numerical tools, such as interpolation, integration, optimization, image processing, statistics, special functions.
- **Matplotlib** is a data visualization library for creating static, animated, and interactive visualizations in Python.

3.4 Machine Learning algorithms for Churn analysis

For churn prediction case study, we are going to use Logistic Regression, Random Forest, Support vector machine, and Decision Tree machine learning algorithms. After applying the algorithms, the models will be evaluated in order to decide which model is the best to be used for prediction.

The models have been evaluated using the holdout method, which divided the given data into two independent data sets, 80% for training and 20% for testing. After applying the chosen algorithm to the data set, we will obtain a model that will be evaluated based on performance indicators.

The first algorithm is Logistic Regression and we have a model accuracy of 87,52%, with a precision of 80,82% and a recall of 73,41%. For the second algorithm, Random Forest, we got a classification rate of 95,12%, considered a very good accuracy, with a precision of 92,39% and a recall of 90,25%. The Decision Tree algorithm shows an accuracy of 94,36%, with a precision of 89,55%, and a recall of 90,7%. The last algorithm Support Vector Machine has an accuracy of 87,47%, with a precision of 80,25% and a recall of 74,04%.

The classification report for the four models is summarized in Table 3 and shows the main classification metrics accuracy, precision, and recall for each model. The report is used to measure the quality of predictions for a classification algorithm.

Table 3. Classification report results

Model	Accuracy	Precision	Recall
Logistic Regression	0.8752	0.8082	0.7341
Random Forest	0.9512	0.9239	0.9025
Decision Tree	0.9436	0.8955	0.9070
Support Vector Machine	0.8747	0.8025	0.7404

In order to choose the best model that will predict if the customer is going to churn, we want to maximize the recall, as it represents the ability of the classifier to correctly predict all truly positive cases.

Considering this, the Decision Tree model is the best model to be used in order to predict new values for possible customers that will churn. We can see below the confusion matrix for decision tree, which gives a holistic view of the performance of the model.

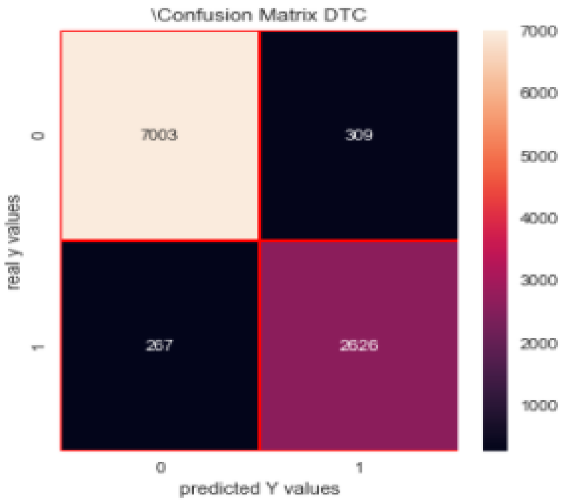
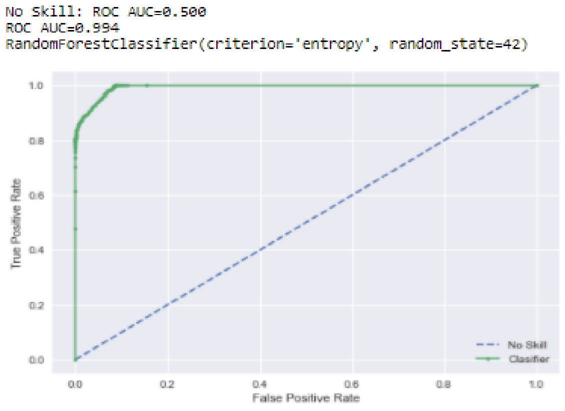


Fig. 4. Confusion Matrix Decision Tree

So, from the total number of customers, we can see that most of them were predicted correctly for churn. 7003 customers have no for churn and were also predicted with no, 2626 have yes for churn and were predicted with yes. Further, 309 were incorrectly classified as churners, when they are not and 267 were predicted as non-churners, when they are going to churn. (**Fig. 4**)

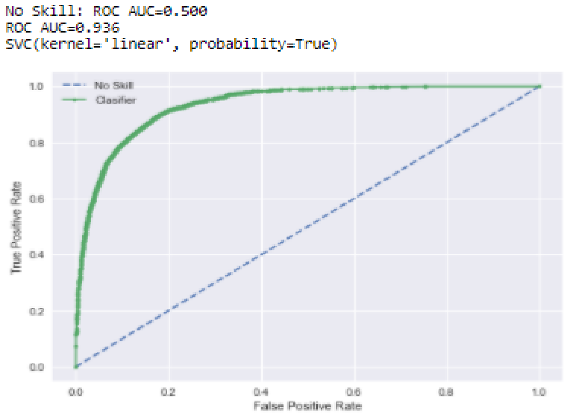
The ROC curve for each model is shown in figures 5-8 according to Table 3, for each model output. The ROC curve shows the trade-off between the true positive rate (TRP) and the false positive rate (FPR). Given a test set and a model, TPR is the proportion of positive (churned) tuples that are correctly labeled by the model. FPR is the proportion of negative (no churn) tuples that are mislabeled as positive.

The AUC values have been plotted for the models as shown in Figures 5-8, where the best result for AUC is 0,994, obtained for Random Forest. (**Fig. 5**)



```
No Skill: ROC AUC=0.500
ROC AUC=0.994
RandomForestClassifier(criterion='entropy', random_state=42)
```

Fig. 5. Random Forest AUC



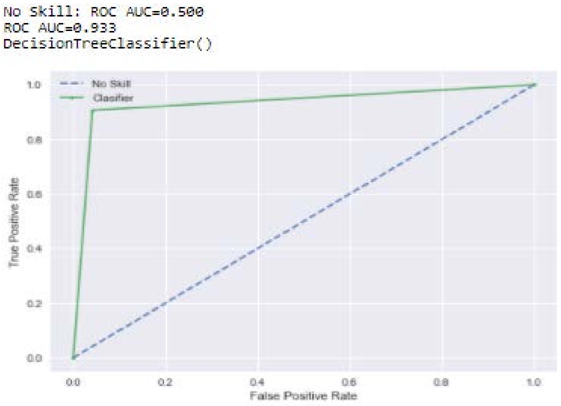
```
No Skill: ROC AUC=0.500
ROC AUC=0.936
SVC(kernel='linear', probability=True)
```

Fig.7. Logistic Regression AUC

4. Results

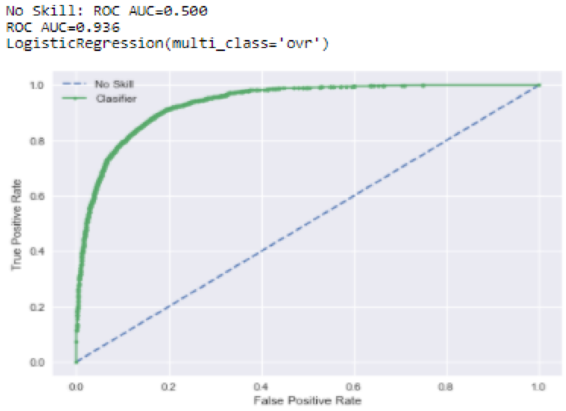
In this case study, we used a public telecom data set, targeting churning customers. The data set contains details about customers including customer care service, customer value added services, customer personal details, and customer usage pattern. For these customers, we observed that those with lower values of total recurring charge, monthly minutes of use, months in service, number of unanswered calls, or number of outbound calls and high values for number of days of the current equipment, are very likely to churn.

In order to avoid this, a classification machine learning algorithm is being used, to predict the customers that are exposed to churn. We have tested four algorithms and have chosen the one that presented the minimal false negative rate, which is Decision Tree. Using this model, the company can predict those customers that are going to churn and take the necessary actions in order to prevent them from churning.



```
No Skill: ROC AUC=0.500
ROC AUC=0.933
DecisionTreeClassifier()
```

Fig. 6. Decision Tree AUC



```
No Skill: ROC AUC=0.500
ROC AUC=0.936
LogisticRegression(multi_class='ovr')
```

Fig. 8. Support Vector Machine AUC

Based on our analysis, a solution might be to include an incentive plan to offer a discounted new equipment to those customers who have the equipment for a long time.

This is based on the high correlation of the number of days of the current equipment with churn, which suggests that churn is more likely when the customer has the same equipment for a long time.

5. Conclusion

The paper aims to find the most accurate model for churn prediction in telecom and, at the same time, to detect the key factors that might lead customers to churn.

For this purpose, four classification algorithms, Logistic Regression, Random Forest, Support Vector Machine and Decision Tree, have been implemented and analyzed in Python.

The models' performance has been measured by recall value, since the goal in this case is to predict as accurate as possible the customers that are going to churn. The best recall value was 90,7%, obtained for the Decision Tree model. This aspect of the analysis suggest that Decision Tree is the best model to be used in the churn case, in order to predict those customers that are most probable to churn. Another evaluation metric in place, that can be used to measure the models' performance is area under the ROC curve, where the best AUC is 0.994, obtained for the Random Forest model.

The analysis leads to the conclusion that telecom operators can obtain the best predictive models to predict churn, by analyzing historical customer records and further understanding customers behavior. Based on this, decision-making employees can build different marketing approaches to retain churners based on the predictors that have higher importance in scoring the model performance.

References

[1] Gallo, A. (2014, October 29). TheValueof Keeping the Right Customers.
Retrieved from Harvard Business Review: https://hbr.org/2014/10/the-valueofkeeping-the-right-customers;

[2] Ahmed, A., & Linen, D. M. (2017). A review and analysis of churn prediction methods for customer retention in telecomindustries. International Conference onAdvanced Computing and CommunicationSystems (ICACCS). IEEE;

[3] Brandusoiu, I.(2016). Churn Predictionin the Telecommunications Sector UsingSupport Vector Machine. *Annals of theOradea University: Fascicle Managementand Technological Engineering*, pp. 97100;

[4] Mishra, K., & Rani, R. (2017). Churnprediction in telecommunication usingmachine learning. *International Conferenceon Energy, Communication, Data Analyticsand Soft Computing (ICECDS).* IEEE;

[5] Gaur, A., & Dubey, R. (2018). PredictingCustomer Churn Prediction In TelecomSector Using Various Machine LearningTechniques. *International Conference onAdvanced Computation andTelecommunication (ICACAT).* IEEE;

[6] Pamina, J. B. An effective classifier forpredicting churn in telecommunication. *Jourof Adv Research in Dynamical & ControlSystems, 11*;

[7] Ullah, I. R. (2019). A churn predictionmodel using random forest: analysis ofmachine learning techniques for churnprediction and factor identification intelecom sector. (pp. 60134-60149). IEEE;

[8] Huang, Y. Z. (2015). Telco churn prediction with big data. *ACM SIGMOD international conference on management of data*, (pp. 607-618);

[9] Ahmad, A. J. (2019). Customer churnprediction in telecom using machine learningin big data platform. *Journal of Big Data,6(1)*, pp.1-24;

[10] Makhtar, M. N. (2017). Churnclassification model for localtelecommunication company based on roughset theory. *Journal of Fundamental andApplied Sciences, 9(6S)*, pp. 854-868;

[11] Sai, B. a. (2019). Predictive Analysisand Modeling of Customer Churn inTelecom using Machine LearningTechnique. *3rd International Conference onTrends in Electronics and Informatics(ICOEI)* (pp. pp. 6-11). IEEE;

[12] Brandusoiu. (2016). Methods for churnprediction in the prepaid mobiletelecommunications industry. *InternationalConference on Communications (COMM)*(pp. 97-100). IEEE;

[13] Dalvi, P. K. (2016). Analysis ofcustomer churn prediction in telecomindustry using decision trees and logisticregression. *Symposium on Colossal DataAnalysis and Networking (CDAN)* (pp. 14). IEEE;

[14] Reichheld, F. (2011, October).*Prescription for Cutting Costs. Bain &Company.* Retrieved from Bain &Company:https://media.bain.com/Images/B_Prescription_cutting_costs.pdf;

[15] Dalvi, P. K., Khandge, S. K., Deomore, A., Bankar, A. and Kanade, V. A. Analysis of customer churn prediction in telecom industry using decision trees and logistic regression. Symposium on Colossal Data Analysis and Networking (CDAN), 2016, pp. 1-4.

Andreea-Maria COPĂCEANU (b. August 7, 1992) has graduated the Faculty of Cybernetics, Statistics and Economic Informatics at the Bucharest University of Economic Studies, in 2014. She followed a master's degree in Cybernetics and Quantitative Economics, within the same faculty. Currently, she is a PhD student, teaching assistant in the department of the Faculty of Cybernetics, Statistics and Economic Informatics and works in IT as Cloud Engineer in the Support Department. She has high interest in areas such as Data Science, Analytics, Machine Learning, Databases, Big Data, Business Architectures, and Strategic Marketing.