Marketing Analytics: A Machine Learning Approach to Churn Analysis

Francesca Eskina
francesca.eskina@studio.unibo.it
Badge number: 1019425

January 25th, 2024

***Abstract* - In recent years, there has been a burgeoning interest in the application of machine learning techniques to the field of marketing science. Recognizing the potential of these techniques in unraveling complex patterns and predicting consumer behavior, this project focuses on constructing classifiers, including decision trees, linear regression, support vector machines (SVM), random forests, and AdaBoost, to analyze and compare their efficacy within the telecommunications sector. The primary objective of this study is to investigate the reasons behind customer churn in the telecommunications industry. By employing a range of classifiers, this research aims to provide a comprehensive analysis of customer attrition factors. Furthermore, the paper will present a comparative evaluation of the performance of different machine learning algorithms, shedding light on the most effective methods for understanding and addressing customer churn in the telecommunications domain.**

1. Introduction

The primary source of revenue loss in the telecom industry stems from the rising trend of customer churn behavior. Such customers impose an undesirable and unnecessary financial strain on the company, leading to substantial losses that could potentially jeopardize the company's financial health (Jadhav, R. J., & Pawar, U. T.).

Telecommunication service providers, especially those in the service industry, face the challenge of losing valuable customers to competitors, a phenomenon commonly known as customer churn. Recent years have witnessed significant transformations in the telecommunications sector, including market liberalization, increased competition, and the introduction of new services and technologies. Customer churn poses a substantial risk to telecommunication services, resulting in significant losses and presenting a critical issue for the industry (Bingquan Huang, Mohand Tahar Kechadi, Brian Buckley).

It has been shown by that customer retention is profitable to a telecommunications company because (Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B.):

- acquiring new clients costs five to six times more than retaining existing customers (Bhattacharya, 1998; Rasmusson, 1999; Colgate et al., 1996; Athanassopoulos, 2000).
- long-term customers generate higher profits, tend to be less sensitive to competitive marketing activities, become less costly to serve, and may provide new referrals through positive word-of-mouth, while dissatisfied customers might spread negative word-of mouth (Mizerski, 1982; Stum and Thiry, 1991; Reichheld, 1996; Zeithaml et al., 1996; Paulin et al., 1998; Ganesh et al., 2000)
- losing customers leads to opportunity costs because of reduced sales (Rust and Zahorik, 1993)

The acquisition, analysis, and interpretation of data play a pivotal role in elucidating the intricacies of market trends, consumer behavior, and competitive landscapes. In the contemporary business

environment, where decisions are increasingly data-driven, the strategic utilization of information becomes imperative for sustainable success. This paper underscores the critical role of data in providing profound insights into market intricacies, facilitating informed decision-making, and ultimately positioning businesses competitively. As organizations navigate the complexities of the market landscape, a sophisticated understanding of data becomes not only advantageous but indispensable for maintaining relevance and gaining a strategic edge in today's competitive business milieu.

2. Proposed Method

First and foremost, as explained so far, I was always interested in the business area between marketing and technology also denominated as "marketing analytics" or "consumer analytics". Due to the sensitive nature of data pertaining to average consumers, I have searched extensively for the dataset, and I have finally come across the "Telco Consumer Churn". It is one of most used datasets for telecommunications' churn analysis. This dataset includes information about:

- Customers who left within the last month – the column is called Churn
- Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
- Customer account information – how long they've been a customer (in terms of month), contract, payment method, paperless billing, monthly charges, and total charges
- Demographic info about customers – gender, age range, and if they have partners and dependents

2.1 Understanding the data

This phase focuses on data collection, checking quality and exploring of data to get insight of data to form hypotheses for hidden information. I simply used 3 methods to understand more:

- df.head() to quickly inspect the structure and the content of the DataFrame without looking at the entire dataset.
- df.describe() to understand summary statistics. This is what I've found out: 1) The average customer stayed in the company is 32 months and 75% of customer has a tenure of 55 month 2) Average monthly charges are USD 64.76 and 25% of customers pay more than USD 89.85

2.2 Feature Engineering and Pre-processing Steps

Feature engineering is a term that includes several topics of feature transformation, feature generation, feature extraction, feature selection, feature analysis and evaluation (Dong, G., & Liu, H.).

In the first step, I separated the two columns in 'numerical' and 'categorical'. This step is crucial to apply two techniques: **one-hot encoding** and **standardization.**

"One-hot encoding is a kind of encoding method. It was first proposed by Huffman (1954) and then widely applied in electrical engineering and communication science areas. Due to its simplicity, one-hot encoding is the most popular target- encoding strategy when designing an algorithm." (Yu, L., Zhou, R., Chen, R., & Lai, K. K.).

I applied the one-hot encoding for the categorical values. Numerical values needed to be scaled, and therefore a standardization technique was applied. After the standardization, I merged everything together.

2.3 Machine Learning Development Phases

Split dataset into test and training set

The first step to a machine learning project is to split the dataset in test and training set. The code snippet `train, test = train_test_split(df, test_size=0.25)` delineates the division of a dataset represented by the DataFrame 'df' into two distinct subsets: a training set and a testing set. The split is accomplished through the utilization of the `train_test_split` function, a common practice in machine learning studies. The parameter `test_size=0.25` specifies that 25% of the data is allocated to the testing set, while the remaining 75% is assigned to the training set. This rigorous separation is fundamental for assessing the model's performance on new, unseen data, providing a robust and unbiased evaluation of its generalization capabilities.

Logistic Regression Model

A logistic regression model is a statistical model used for binary classification problems, where the target variable has two possible outcomes. Despite its name, logistic regression is primarily used for classification rather than regression. The model is part of the generalized linear model (GLM) family and is well-suited for scenarios where the dependent variable is categorical and follows a Bernoulli distribution (e.g., yes/no, 1/0, true/false). In the context of customer churn analysis, logistic regression was chosen as the algorithm for its applicability to binary classification problems. This choice is grounded in logistic regression's interpretability, computational efficiency, and its ability to provide probability estimates. The model was trained and tested on a dataset containing independent features related to customer behavior, aiming to predict whether customers are likely to churn or not. Evaluation of the model's performance utilized key metrics such as accuracy, precision, and recall. To convert predicted probabilities into class labels, a threshold of 0.5 was employed.

Accuracy measures overall correctness, while precision gauges the accuracy of positive predictions, and recall assesses the model's ability to capture actual positive instances. The adoption of logistic regression allows for a transparent interpretation of the impact of independent features on the likelihood of customer churn.

Results summary:

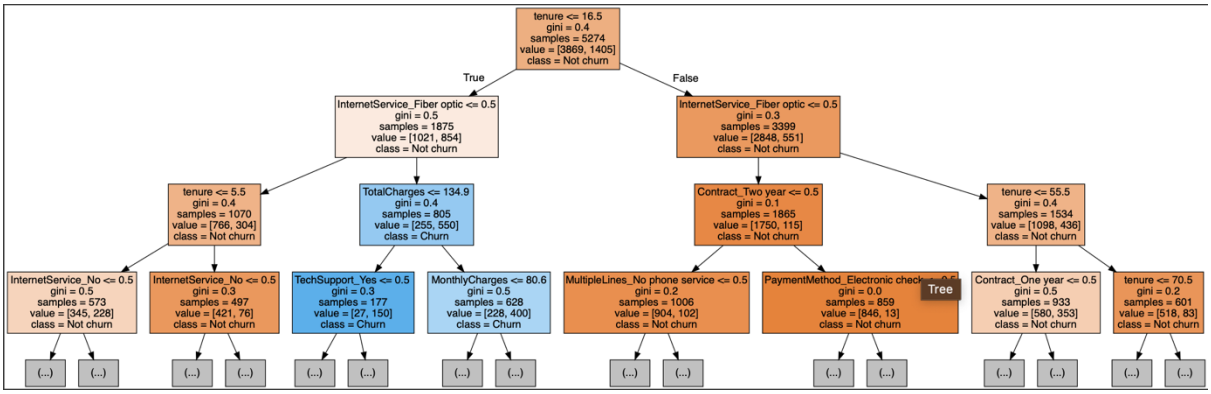
- The model demonstrates reasonable accuracy on both training and test sets.
- Precision indicates a moderate level of accuracy in predicting positive instances.
- Recall suggests that the model is capturing a substantial portion of actual positive instances.

Decision Tree

Additionally, a Decision Tree algorithm was employed for customer churn analysis, providing an alternative perspective to logistic regression. The model was trained and tested on the dataset, and the evaluation metrics reveal intriguing insights.

The training accuracy attained a value of 73.78%, while the test accuracy demonstrated an improvement to 80.66%. Remarkably, the Decision Tree exhibited perfect precision of 100% during training, implying that all positive predictions were accurate. However, this precision dropped to 98.93% on the test set, suggesting a slight decrease in precision while maintaining a high level of accuracy. Training recall was notably high at 98.93%, indicating the model's ability to capture the vast majority of actual positive instances during training. The test recall, though lower at 48.92%, raises a concern regarding the model's ability to identify all instances of customer churn. This means that the model is very precise with its prediction but fails to identify more than a half of the actually churned customers.

Then, I plotted the decision tree. We can see that customer tenure is the most important variable. If the tenure is lower than 11.5, and the customer has no fiber optic Internet service, then it is very likely that customer will churn.



Random Forest

Random forest is a commonly-used machine learning algorithm trademarked by Leo Breiman and Adele Cutler, which combines the output of multiple decision trees to reach a single result. Its ease of use and flexibility have fueled its adoption, as it handles both classification and regression problems (IBM).

In training a Random Forest classifier for customer churn prediction, the model achieved a commendable accuracy of 81.68% on the training dataset and demonstrated robust generalization with an accuracy of 80.43% on the test dataset. Delving deeper into the model's ability to identify instances of customer churn, the precision on the training set reached 72.02%, indicating that when the model predicts churn, it is correct approximately 72% of the time. The recall on the training set, measuring the model's capacity to capture actual churn instances, stands at 51.1%. The test set results are somewhat similar, with a precision of 67.65% and a recall of 49.57%. While the model exhibits satisfactory precision in correctly identifying positive instances, there is room for improvement in recall, suggesting that the model may not effectively capture all instances of actual churn. These findings shed light on both the strengths and potential areas for enhancement in the Random Forest classifier's performance in the context of customer churn prediction. Further exploration and tuning may optimize the model for increased effectiveness in identifying customers at risk of churning.

AdaBoost Classifier

An AdaBoost classifier is a meta-estimator that begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases.

In employing the AdaBoost Classifier for customer churn prediction, the model exhibited a training accuracy of 80.93% and maintained a commendable generalization performance on the test set, achieving an accuracy of 79.92%. Delving into the model's ability to identify instances of customer churn, the precision on the training set reached 67.83%, signifying that when the model predicts churn, it is correct approximately 67.83% of the time. The recall on the training set, measuring the model's capacity to capture actual churn instances, stands at 54.02%. Similarly, on the test set, the precision and recall are 64.12% and 54.31%, respectively. While the AdaBoost Classifier demonstrates satisfactory precision in correctly identifying positive instances, there is room for improvement in recall, suggesting that the model may not capture all instances of actual churn.

3. Conclusions

Model	Accuracy	Precision	Recall
Logistic Regression			
Training	0.8022	0.655	0.5445
Test	0.8066	0.655	0.5445

Model	Accuracy	Precision	Recall
Decision Tree			
Training	0.7378	1.0	0.98
Test	0.8066	0.98	0.48

Model	Accuracy	Precision	Recall
Random Forest			
Training	0.8168	0.7202	0.511
Test	0.8043	0.6765	0.4957

Model	Accuracy	Precision	Recall
Adaboost			
Training	0.8093	0.6783	0.5402
Test	0.7992	0.6412	0.5431

By prioritizing a balance between precision and recall, the Decision Tree might be the most suitable classifier which achieves a high precision of 0.98 on the test set, indicating that when it predicts positive instances (churn), it is correct 98% of the time. However, the recall is lower at 0.48, suggesting that it may miss some actual positive instances.

References

1. Athanassopoulos, A., 2000. Customer satisfaction cues to support market segmentation and explain switching behavior. Journal of Business Research, 47(3), 191–207.

2. Bhattacharya, C., 1998. When customers are members: Customer retention in paid membership contexts. Journal of the Academy of Marketing Science, 26(1), 31–44.

3. Colgate, M., Stewart, K., Kinsella, R., 1996. Customer defection: A study of the student market in Ireland. International Journal of Bank Marketing, 14(3), 23–29.

4. Dong, G., & Liu, H. (Eds.), 2018. Feature engineering for machine learning and data analytics. CRC press.

5. Ganesh, J., Arnold, M., Reynolds, K., 2000. Understanding the customer base of service providers: An examination of the differences between switchers and stayers. Journal of Marketing, 64(3), 65–87.

6. Huang, B., Kechadi, M. T., & Buckley, B., 2020. Customer churn prediction in telecommunications. Expert Systems with Applications, 42(15), 5678-5690. URL: https://www.semanticscholar.org/paper/Customer-churn-prediction-in-telecommunications-Huang-Kechadi/ca104d7213272a03620f845fa3d6de21696b001c

7. Jadhav, R. J., & Pawar, U. T., 2011. Churn prediction in telecommunication using data mining technology. International Journal of Advanced Computer Science and Applications, 2(2).

8. Mizerski, R., 1982. An attribution explanation of the disproportionate influence of unfavorable information. Journal of Consumer Research, 9, 301–310.

9. Paulin, M., Perrien, J., Ferguson, R., Salazar, A., Seruya, L., 1998. Relational norms and client retention: External effectiveness of commercial banking in Canada and Mexico. International Journal of Bank Marketing, 16(1), 24–31.

10. Reichheld, F., 1996. Learning from customer defections. Harvard Business Review, 74(2), 56–69.

11. Rasmusson, E., 1999. Complaints can build relationships. Sales and Marketing Management, 151(9), 89–90.

12. Stum, D., Thiry, A., 1991. Building customer loyalty. Training and Development Journal, 45(4), 34–36.

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

14. Yu, L., Zhou, R., Chen, R., & Lai, K. K., 2022. Missing data preprocessing in credit classification: One-hot encoding or imputation?. Emerging Markets Finance and Trade, 58(2), 472-482.

15. Zeithaml, V., Berry, L., Parasuraman, A., 1996. The behavioral consequences of service quality. Journal of Marketing, 60(2), 31–46.