

The Investigation of the Advancement for Machine Learning-Based Telecommunication Customer Churn Prediction

Yi Zhang^a

College of Letters & Science, University of Wisconsin Madison, Madison, U.S.A.

Keywords: Machine Learning, Customer Churn, Deep Learning.


Abstract: It is important for companies to predict customer churn, as it helps reduce losses, and machine learning is crucial to this process, making its study extremely important. Ensemble Methods, Logistic Models, Classification Trees, and Nearest Neighbor Algorithms are all traditional machine learning methods, but they have inherent drawbacks. This encompasses overfitting, a situation where models achieve high accuracy on training data but struggle to perform effectively on unseen data; model complexity, making complex models difficult to interpret and maintain; and inefficiencies in handling large-scale data, often leading to poor performance with vast amounts of data. Conversely, advancements in deep learning highlight its strengths and areas where it surpasses traditional methods. This study offers an extensive perspective on telco churn prediction and artificial intelligence, delving into possible concerns like data protection, which aims to protect user data from exploitation. To address these challenges, several solutions are proposed, including collaborating with experts to ensure the accuracy and reliability of AI models, implementing stronger security measures to protect sensitive information, and utilizing techniques to mitigate data scarcity issues. Overall, this work offers an excellent review of the telco churn prediction field, highlighting key advancements and proposing solutions to existing challenges.

1 INTRODUCTION

Telco churn is when a telecommunications company loses customers. Customers may decide to cancel their current service and switch to another company if they are dissatisfied with the current service or if another company offers better deals and packages. In the telecom industry, competition is fierce. Retaining existing customers is more cost-effective than acquiring new ones. Customer churn directly impacts a company's revenue and market position, so telecom companies pay close attention to this issue. Therefore, accurately identifying customer churn to help companies reduce losses is crucial.

In the last few years, Artificial Intelligence (AI) has gradually developed in the society and found widespread applications in the economy, including significant increases in robotics, AI startups, and patent numbers (Furman & Seamans, 2019). And at the same time, AI technology has exceeded human capabilities in several domains, including image recognition and speech recognition. AI is widely used

in search engines, health monitoring, stock analysis etc. Investment in AI is rapidly increasing worldwide, with governments actively promoting AI research. The future of AI may involve cyclical periods of growth and decline or rapid development to a level that surpasses human intelligence (De Spiegeleire et al., 2017). AI has been instrumental in the advancement of predictive analytics methods, resulting in the development of numerous predictive models. Supervised learning uses labeled data for training, such as classification trees and margin-based classifiers, and is frequently employed for categorization and regression tasks. Unsupervised learning handles unlabeled data, like Dimensionality reduction technique and k-means clustering, and is mainly used to discover patterns and structures in the data. These algorithms are applicable to different types of data and tasks, with widespread applications ranging from online shopping recommendations to photo updates on social networks (Mahesh, 2020). Machine learning analyzes data to help businesses and governments make better decision. It plays a

^a <https://orcid.org/0009-0008-0144-0573>

crucial role in cybersecurity by identifying patterns to better detect malware and cyber-attacks. Additionally, machine learning helps develop smart applications that understand user behavior and provide personalized recommendations and services (Sarker, 2021). One of the tasks that has received significant attention is predicting customer churn in the telecommunications industry. There has been extensive research on this topic in the past. A churn prediction model based on the ensemble methods algorithm has been developed by researchers from various universities and research institutions in Pakistan and South Korea. This model analyzes subscriber attrition in the telecommunications industry, identifying the causes and behavioral patterns of churn. The aim is to help telecom companies reduce customer churn, optimize their Customer Relationship Management (CRM) systems, and enhance their profitability (Ullah et al., 2019). And at the same time, the research team from Mangalam College of Engineering in India developed a churn prediction model utilizing deep learning and ensemble methods algorithms. This model is able to examine client retention issues in the telecommunications sector, identifying the reasons for churn and detecting associated behavioral patterns (Andrews et al., 2019). In recent years, many new algorithms have emerged, making it necessary to conduct a comprehensive review of this area. The rest of this paper is structured into three main sections: methods, discussion, and conclusion.

This paper will firstly detail the various methods used by others to predict telecom customer churn, and then discuss the strengths and weaknesses of these methods and provide insights into future directions. Finally, this paper will summarize the key points of the entire paper.

2 METHOD

2.1 The Introduction of the Machine Learning Workflow

Machine learning algorithms typically involve the following steps: data collection, preprocessing, model building, training, and testing. First, collect a large amount of high-quality data. Then, clean and standardize the data, extract useful features, and reduce dimensionality. Next, choose the appropriate algorithm based on the specific problem, such as classification trees, and Margin-based classifier for supervised learning; k-means algorithms for unsupervised learning; and transductive support

vector machines for semi-supervised learning. Train the model with the training data, fine-tuning parameters and optimizing to enhance performance. Finally, test the model's accuracy with test data by Assessing performance indicators like correctness, specificity, sensitivity, and F1 measure on the new data. Different algorithms have their own strengths and weaknesses: supervised learning is suitable for small, well-labeled datasets, while unsupervised learning and deep learning perform better on large-scale datasets. Machine learning has numerous applications in contemporary life, from shopping recommendations to photo recognition on social networks (Mahesh, 2020).

2.2 Traditional Artificial Intelligence Techniques-based Prediction

2.2.1 Ensemble Methods

Ensemble methods is a machine learning algorithm employed for making predictions. It generates multiple classification trees and combines their results to enhance prediction accuracy (Probst et al., 2019). The ensemble methods algorithm has numerous applications in the telecommunications sector. It analyzes customer data by generating many classification trees and combines the results of these trees to make predictions. This method enables more precise identification of customers likely to churn. The findings show that the ensemble methods algorithm can accurately predict customer churn with an accuracy of 88.63%, which is extremely advantageous for telecom companies in developing strategies to retain their customers (Ullah et al., 2019). And this method is highly effective in handling large datasets, offering excellent classification accuracy and fast processing. Experimental results show that it achieves a prediction accuracy of 91.4%, making it very useful for telecom companies in developing customer retention strategies (Abdulsalam et al., 2022).

2.2.2 Logistic Model

Logistic model is a statistical technique that predicts the likelihood of a binary outcome based on one or more predictor variables. This method is frequently applied in fields such as social sciences and medical research to analyze relationships between dependent and independent variables, allowing researchers to make informed predictions and decisions (Fletcher & Islam, 2019). In telecom customer churn prediction, logistic model is extensively utilized because of its

simplicity and ease of interpretation. It provides easily understandable results and probability outputs, is suitable for mixed data types, and has high computational efficiency.

2.2.3 Classification Trees

Classification trees are non-parametric supervised learning algorithms employed for both classification and regression problems. They make decisions by recursively dividing the dataset into subsets. They offer high interpretability, can handle both continuous and discrete data, and discover non-linear relationships between attributes (Fletcher & Islam, 2019). Classification trees have extensive applications in predicting customer churn for telecom companies. By analyzing customer data and recursively partitioning it into subsets, classification trees effectively predict which customers are likely to churn. Experimental results show that classification trees excel in classification accuracy, achieving a 93% accuracy rate, which is very helpful for telecom companies in devising customer retention strategies (Hassonah et al., 2019).

2.2.4 Nearest Neighbor Algorithms

The nearest neighbor algorithms (KNN) algorithm is a non-parametric supervised learning approach that categorizes new data points by calculating the distances between them and all points in the training dataset. In telecom churn prediction, the KNN algorithm involves collecting and standardizing customer data, training the model using historical data, followed by calculating the distances between new customers and all points in the training dataset. By identifying the KNN, the algorithm predicts whether a customer will churn based on the classifications of these neighbors, helping telecom companies identify and retain potential churners (Hassonah et al., 2019).

2.3 Deep Learning-Based Prediction

2.3.1 Artificial Neurons

The Artificial Neural Network (ANN) is a computational method that mimics the way the human brain works. It improves prediction accuracy by learning from data. ANN include several layers: input levels, hidden levels, and output levels. The intermediate layers process complex information, thereby improving the accuracy of the predictions. ANN also have significant applications in predicting telecom customer churn. By using a dataset

containing 3333 samples and 21 attributes and applying a feature selection algorithm called Relief-F, the authors identified the most useful features for prediction. These features were then used to train and test the ANN model, which achieved a prediction accuracy of 93.88%, outperforming other methods. This demonstrates that ANN is highly effective in handling complex data and accurately predicting whether customers will churn (Abdulsalam et al., 2022).

3 DISCUSSIONS

Lately, machine learning has gradually evolved from traditional methods to deep learning. Traditional machine learning methods have revealed many shortcomings, particularly when handling large and complex datasets, making these shortcomings especially prominent in today's data-driven telco churn prediction applications. Deep learning, with its powerful data processing capabilities and complex pattern recognition, enables more accurate churn predictions, helping telecom companies to more effectively devise customer retention strategies.

Firstly, when the data volume is exceptionally large, traditional machine learning methods exhibit poor computational efficiency and often struggle to handle large-scale datasets. For example, when dealing with millions or even billions of data records, traditional algorithms often have high computational complexity, leading to long processing times and an inability to meet practical application demands.

Secondly, traditional machine learning methods have high requirements for data quality. When data is missing, noisy (e.g., contains erroneous information), or incomplete, the performance of traditional methods can significantly degrade, making it difficult to handle these issues effectively. Furthermore, traditional models have limited generalization capabilities across different datasets. Many traditional models perform well on training data but show a significant drop in performance on new data.

Lastly, traditional machine learning methods have limited capability in handling complex problems. For instance, in the logistic model, it assumes a linear relationship and may not capture complex customer behaviors, making its predictive performance potentially inferior to more sophisticated algorithms.

However, Deep learning models typically have more significant advantages compared to traditional methods. Such nonlinearities can be effectively captured by them, resulting in greater accuracy, particularly when managing large datasets and

complex tasks like visual recognition, audio recognition, and language interpretation. Deep learning can handle vast amounts of data exceptionally well, and the more data it processes, the better it performs, making it particularly advantageous in big data environments. Additionally, deep learning can be applied across multiple fields, not just limited to a single domain.

In the telecommunications sector, AI is essential for forecasting customer churn. By analyzing vast amounts of customer data, AI can pinpoint those likely to churn, assess the risk, and enable telecom companies to take proactive steps to reduce churn rates, thus increasing potential profits. But there are also some limitations. AI needs a lot of data to learn and work properly. If there isn't enough data or the data quality is poor, AI's performance can suffer. Moreover, collecting and processing this data is both time-consuming and expensive.

Additionally, in most cases, AI can only provide a result without explaining why it came to that result. However, in practical applications, such as in healthcare and legal fields, decision-makers need to understand the reasons and processes behind these predictions.

Training complex AI models requires powerful computers and a lot of time. This can be a significant burden for many small companies and research institutions. AI performs well on familiar data, but it may struggle with new, unseen data. This means that AI can sometimes be unreliable in practical use.

Furthermore, highly autonomous AI systems can pose safety risks. If an AI system behaves unexpectedly or is maliciously exploited, it can lead to serious consequences. Ensuring that AI systems operate safely in all situations is a significant challenge.

Recently, with the continuous advancement of technology, some possible methods can be considered to address the existing issues mentioned above. To address the issue of AI's lack of interpretability and reliable, companies can collaborate with domain experts. These experts can provide business knowledge and feedback, helping to select important features, validate, and interpret AI model predictions, ensuring that the models align better with actual business needs and making the results more trustworthy and easier to understand. Furthermore, to address the issues of data scarcity and long training times, which can lead to high costs or render some domains impractical, transfer learning offers an effective solution. Transfer learning plays a crucial role in transferring knowledge between different yet related domains. It reduces the reliance on a large

amount of labeled data in the target domain, thus enhancing the performance of models within that domain. It also demonstrates broad application prospects in addressing data scarcity in real-world scenarios and optimizing machine learning model performance (Zhuang et al., 2020). Meanwhile, to address the issue of AI's lack of privacy, Federated Learning (FL) can effectively solve this problem. FL holds significant importance and has vast applications across various fields, including industrial engineering and healthcare. FL is a decentralized machine learning technique designed to enable multiple devices or organizations to train models together without disclosing raw data, thereby protecting data protection and security. Its primary goal is to address data silos and privacy issues by allowing multiple clients to participate in model training without centralized data processing. The key characteristics of FL include decentralization and data privacy protection, making it particularly important in scenarios where data privacy and decentralized data processing are crucial (Li et al., 2020).

4 CONCLUSIONS

In this study, both traditional machine learning and deep learning are analyzed in depth. Several major methods of traditional machine learning and their inherent drawbacks, such as overfitting, model complexity, and limitations in handling large-scale data, are explored. Additionally, advancements in deep learning are delved into, highlighting its strengths and areas where it surpasses traditional methods. The potential applications of AI across various domains are examined, providing a comprehensive outlook on the future of the field.

Current potential issues, including data privacy, security, algorithmic bias, and ethical concerns, are analyzed. To address these challenges, a range of solutions is proposed. These solutions include improved data processing techniques to ensure more accurate and reliable AI models, the implementation of more robust security measures to protect sensitive information, and the development of fairer algorithm designs to mitigate bias and promote ethical AI practices. Through these efforts, valuable insights and guidance for the continued development and responsible deployment of artificial intelligence technologies are aimed to be provided.

REFERENCES

- Abdulsalam, S. O., Ajao, J. F., Balogun, B. F., & Arowolo, M. O. 2022. A churn prediction system for telecommunication company using random forest and convolution neural network algorithms. *EAI Endorsed Transactions on Mobile Communications and Applications*, 7(21).
- Abdulsalam, S. O., Arowolo, M. O., Saheed, Y. K., & Afolayan, J. O. 2022. Customer churn prediction in telecommunication industry using classification and regression trees and artificial neural network algorithms. *Indonesian Journal of Electrical Engineering and Informatics (IJEI)*, 10(2), 431-440.
- Andrews, R., Zacharias, R., Antony, S., & James, M. M. 2019. Churn prediction in telecom sector using machine learning. *International Journal of Information*, 8(2).
- De Spiegeleire, S., Maas, M., & Sweijts, T. 2017. Ai – Today And Tomorrow. In *Artificial Intelligence and The Future of Defense: Strategic Implications for Small- And Medium-Sized Force Providers* (pp. 43–59). Hague Centre for Strategic Studies. <http://www.jstor.org/stable/resrep12564.8>
- Fletcher, S., & Islam, M. Z. 2019. Decision tree classification with differential privacy: A survey. *ACM Computing Surveys (CSUR)*, 52(4), 1-33.
- Furman, J., & Seamans, R. 2019. AI and the Economy. *Innovation policy and the economy*, 19(1), 161-191.
- Hassonah, M. A., Rodan, A., Al-Tamimi, A. K., & Alsakran, J. 2019, November. Churn prediction: A comparative study using knn and decision trees. In *2019 Sixth HCT Information Technology Trends (ITT)* (pp. 182-186). IEEE.
- Li, L., Fan, Y., Tse, M., & Lin, K. Y. 2020. A review of applications in federated learning. *Computers & Industrial Engineering*, 149, 106854.
- Mahesh, B. 2020. Machine learning algorithms-a review. *International Journal of Science and Research (IJSR)*. [Internet], 9(1), 381-386.
- Probst, P., Wright, M. N., & Boulesteix, A. L. 2019. Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: data mining and knowledge discovery*, 9(3), e1301.
- Sarker, I. H. 2021. Machine learning: Algorithms, real-world applications and research directions. *SN computer science*, 2(3), 160.
- 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, 60134-60149.
- Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., ... & He, Q. 2020. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1), 43-76.