Telecom Customer Churn Prediction Using Machine Learning Algorithms

Mukul Rai
700748568
Computer Science
University of Central Missouri
mxr85680@ucmo.edu

Shruthi Vallapreddy
700744517
Computer Science
University of Central Missouri
sxv45170@ucm.edu

Jaya Sai Charan Sammeta
700739775
Computer Science
University of Central Missouri
jxs97750@ucm.edu

Shiva Godesala
700745488
Computer Science
University of Central Missouri
sxg35820@ucm.edu

Abstract:

Customer churn, or the loss of customers over a given period, is a critical issue for businesses across various industries. Identifying customers likely to churn and proactively retaining them is crucial for maintaining customer loyalty and maximizing revenue. In recent years, machine learning algorithms have emerged as powerful tools for predicting customer churn and aiding in effective churn management strategies.

This paper comprehensively reviews machine learning algorithms used for customer churn prediction. The study begins with an overview of customer churn and its impact on businesses. It then provides an in-depth analysis of various machine learning algorithms, including decision trees, logistic regression, support vector machines, random forests, gradient boosting, and neural networks, among others, commonly employed for customer churn prediction. The strengths and weaknesses of each algorithm are discussed, along with their applications in customer churn prediction.

Furthermore, the paper discusses the critical steps in building a customer churn prediction model, including data collection, data preprocessing, feature engineering, model training and evaluation, and model deployment. Best practices for each step, challenges, and considerations associated with customer churn prediction using machine learning algorithms are highlighted.

The paper discusses the importance of interpretability, fairness, ethical considerations in customer churn prediction, and potential future research directions. The findings of this study provide valuable insights for businesses seeking to implement machine learning algorithms for

GitHub: https://github.com/raimukul/telecom_churn_prediction

customer churn prediction. They can be a reference guide for researchers and practitioners in customer relationship management and predictive analytics.

Keywords: Machine Learning, Linear Regression, Telecom, XGBoost, ADA, SVM, Churn

Introduction:

Customer churn, or customer attrition, refers to customers ending their relationship with a business or discontinuing their use of its products or services. Machine learning algorithms offer powerful tools to predict customer churn and enable companies to take proactive measures to retain customers. Churn prediction is critical for businesses across industries, as keeping existing customers is often more cost-effective than acquiring new ones.[1]

Customer churn prediction using machine learning algorithms involves leveraging historical customer data, such as purchase behavior, usage patterns, demographic information, and customer interactions, to build predictive models. These models can identify patterns and trends that indicate customers at risk of churning, allowing businesses to intervene with targeted retention strategies.

The introduction to customer churn prediction using machine learning algorithms typically includes an overview of the problem statement, the importance of churn prediction for businesses, and the benefits of using machine learning techniques. It may also highlight the challenges of customer churn prediction, such as data quality, feature selection, and class imbalance, and how machine learning can help overcome these challenges.[2]

Furthermore, the introduction may briefly overview the machine learning algorithms commonly used for customer churn prediction, such as logistic regression, decision trees, random forests, support vector machines, and neural networks. It may also touch upon the concept of feature engineering, which involves selecting relevant features from the data to build predictive models and model evaluation techniques to assess the performance of the models.[3]

Customer churn prediction using machine learning algorithms involves using advanced statistical and machine learning techniques to analyze historical customer data and identify patterns or indicators that suggest a customer is at risk of churning. This data can include customer demographics,

transaction history, product usage patterns, customer interactions, and other relevant information. By leveraging this data, machine learning algorithms can learn from patterns and trends to accurately predict which customers will likely churn.[4]

Customer churn prediction aims to enable businesses to take proactive actions, such as targeted retention campaigns or personalized offers, to mitigate the risk of losing customers. Machine learning algorithms are crucial in analyzing large and complex datasets to identify key factors influencing customer churns, such as customer behavior, preferences, and satisfaction levels. These algorithms can then generate predictive models that can be integrated into a company's existing customer relationship management (CRM) systems to provide real-time insights and recommendations for improving customer retention strategies.[5]

Customer churn prediction using machine learning algorithms has numerous benefits for businesses. It can help companies save costs by identifying customers likely to churn before they do, allowing them to allocate resources more efficiently to retain valuable customers.[6] It can also enable businesses to enhance customer satisfaction by tailoring retention strategies to meet individual customer needs. Additionally, by identifying patterns and trends in customer churn data, machine learning algorithms can help businesses uncover underlying reasons for churn and proactively address them, improving customer loyalty and long-term business success.[7]

In conclusion, the introduction to customer churn prediction using machine learning algorithms sets the stage for understanding the importance of the problem, the challenges involved, and the potential benefits of leveraging machine learning techniques to predict and mitigate customer churn.

Motivation:

There are several compelling reasons why telecom companies should be motivated to predict customer churn using machine learning algorithms:

Retaining Customers: Customer churn, or losing customers to competitors, can significantly negatively impact telecom companies. By accurately predicting customer churn using machine learning algorithms, telecom companies can take proactive measures to retain customers by identifying and addressing the reasons for churn. Losing customers results in immediate

revenue loss and affects the company's long-term profitability.[8]

Cost Savings: Acquiring new customers can be significantly more expensive than retaining existing ones. Telecom companies invest substantial marketing, advertising, and customer acquisition efforts. By accurately predicting customer churn, telecom companies can reduce costs associated with acquiring new customers by focusing on retaining existing ones. This can result in substantial cost savings and improved profitability.[9]

Enhanced Customer Experience: Predicting customer churn allows telecom companies to address customer concerns and improve the overall customer experience proactively. By identifying potential churners early, telecom companies can take timely actions such as providing personalized offers, resolving issues, or improving service quality to prevent customer defection. This can increase customer satisfaction, loyalty, and positive word-of-mouth, which can benefit the company in the long run.[10][11]

Competitive Advantage: In today's highly competitive telecom industry, gaining a competitive edge is crucial. By utilizing machine learning algorithms for customer churn prediction, telecom companies can stay ahead of their competitors by quickly identifying and addressing customer churn. This can help them improve customer retention rates, increase market share, and outperform their competitors regarding customer satisfaction and loyalty.

Data-Driven Decision Making: Machine learning algorithms enable telecom companies to leverage their vast amounts of customer data to make informed and data-driven decisions. Machine learning algorithms can provide valuable insights into customer behavior, preferences, and churn drivers by analyzing historical customer data and identifying patterns.[10] These insights can guide telecom companies in formulating effective retention strategies, optimizing marketing efforts, and making strategic business decisions.[12]

Improved Business Performance: Accurate customer churn prediction using machine learning algorithms can improve telecom companies' business performance. Telecom companies can increase customer lifetime value, revenue, and profitability by reducing customer churn rates and retaining valuable customers.[13] Additionally, identifying customer segments more prone to churn can help telecom companies allocate resources more effectively and optimize their marketing and retention strategies, leading to better business outcomes.

In summary, customer churn prediction using machine learning algorithms can provide telecom companies with several benefits, including customer retention, cost savings,[14] enhanced customer experience, competitive advantage, data-driven decision-making, and improved business performance. These compelling motivations make it worthwhile for telecom companies to invest in machine learning-based churn prediction models to manage customer churn and boost their bottom line effectively.[16]

Main Contributions & Objectives:

- Accurately estimate the churn behavior by identifying the customers at risk of churning.
- Develop a churn prediction model which assists telecom operators in predicting customers who are most likely subject to churn.
- Obtain the relation between the customer's characteristics and the churn.
- Provide telecom companies with an easy and effective way to predict customers who will churn.
- Gain Competitive Advantage: Telecom churn prediction using machine learning algorithms can also provide a competitive advantage by enabling telecom companies to stay ahead of the competition.
- Reduce Business Costs: Churn can be costly for telecom companies in terms of lost revenue, customer acquisition costs, and operational expenses.
- Optimize Marketing and Sales Efforts: Churn prediction can also optimize marketing and sales efforts by identifying potential high-value customers likely to churn.

- In summary, the objectives of telecom churn prediction using machine learning algorithms are to increase customer retention, enhance customer experience, optimize marketing and sales efforts, reduce business costs, and gain a competitive advantage in the telecom market.

Related Work:

"Churn Prediction in Telecom Industry Using Machine Learning Techniques" by S. Saha et al.: This paper comprehensively reviews various machine learning algorithms used for churn prediction in the telecom industry. It compares the performance of algorithms such as decision trees, logistic regression, support vector machines, and neural networks,

among others, and provides insights into their strengths and weaknesses [16].

"Telecom Customer Churn Prediction Using Machine Learning Techniques" by H. Hassanpour et al.: This study focuses on applying machine learning techniques such as random forests, gradient boosting, and deep learning for churn prediction in the telecom industry. It evaluates the performance of these algorithms using real-world telecom datasets and discusses their predictive accuracy and interpretability. [17]

"Predicting Customer Churn in Telecommunications: A Comparative Study of Machine Learning Algorithms" by V. Khemchandani et al.: This research paper presents a comparative study of machine learning algorithms for predicting customer churn in the telecom industry. It evaluates the performance of algorithms such as decision trees, naive Bayes, k-nearest neighbors, and random forests and provides insights into their accuracy, sensitivity, specificity, and F1 score. [18]

"Customer Churn Prediction in Telecom Industry: A Comparative Study of Machine Learning Techniques" by N. Dhindsa et al.: This study compares the performance of various machine learning techniques such as logistic regression, decision trees, support vector machines, and ensemble methods for customer churn prediction in the telecom industry. It evaluates their performance using real-world telecom datasets and discusses their accuracy and interpretability. [19]

"Machine Learning Techniques for Churn Prediction in the Telecommunication Industry" by S. Aggarwal et al.: This paper provides an overview of various machine learning techniques used for churn prediction in the telecom industry, including decision trees, naive Bayes, k-nearest neighbors, and artificial neural networks. It discusses the pros and cons of these techniques and provides insights into their performance based on experimental results using real-world telecom datasets.[20]

"A Comparative Study of Machine Learning Algorithms for Telecom Churn Prediction" - This research paper compares the performance of various machine learning algorithms, such as decision trees, logistic regression, and support vector machines, for telecom churn prediction.

"Churn Prediction in Telecommunications Using Random Forests" - This study uses random forests, an ensemble learning

method, for churn prediction in the telecom industry. It compares the performance of random forests with other popular algorithms and provides insights into their effectiveness.

"Predicting Customer Churn in the Telecom Industry: A Comparison of Machine Learning Techniques" - This research paper evaluates the performance of multiple machine learning techniques, including k-Nearest Neighbors, Naive Bayes, and neural networks, for telecom churn prediction. It provides a comprehensive comparison of the algorithms and their predictive accuracy.

"Churn Prediction for Telecommunication Industry Using Gradient Boosting Machines" - This study focuses on applying gradient boosting machines, a powerful ensemble learning technique, for telecom churn prediction. It provides insights into the effectiveness of gradient-boosting devices and compares them with other algorithms.

"Telecom Customer Churn Prediction using Machine Learning: A Systematic Literature Review" - This systematic literature review provides an overview of various machine learning algorithms used for telecom churn prediction. It summarizes the findings of multiple studies and highlights the strengths and weaknesses of different algorithms.

"Comparative Study of Machine Learning Algorithms for Telecom Churn Prediction: A Review" - This review paper compares the performance of different machine learning algorithms for telecom churn prediction, including decision trees, logistic regression, and support vector machines. It provides a detailed analysis of the algorithms' accuracy, precision, and recall.

Proposed Framework:

Here's a high-level implementation outline for telecom churn prediction using logistic regression, support vector machine (SVM), adaptive boosting (ADA), XGBoost, and random forest algorithms.

- **Logistic Regression:**

Logistic regression is a binary classification technique commonly used for churn prediction. The goal is to build a predictive model that can classify customers as either churners (customers who are likely to cancel their services) or non-churners (customers who are likely to continue their services). Logistic regression models are particularly well-suited for this task because they can

model the probability of a customer churning based on various input features.

- **ADA:**

Telecom companies often need help with customer churn, which can lead to a loss of revenue and market share. ADA (Adaptive Boosting Algorithm) can be used in telecom churn prediction as a machine learning technique to improve the accuracy of predicting customer churn, which refers to the likelihood of customers leaving a telecom service provider. By accurately predicting customer churn, telecom companies can take proactive measures to retain customers and reduce churn.

Using ADA in telecom churn prediction can provide companies with a powerful tool to accurately identify customers at risk of churning and take appropriate actions to retain them. However, it's important to note that no model is perfect, and domain expertise and human judgment are also crucial in interpreting the model's predictions and making informed business decisions.

- **XGBOOST:**

XGBoost (Extreme Gradient Boosting) is a popular machine learning algorithm commonly used for churn rate prediction because it handles complex datasets and captures non-linear relationships between variables. Here's a high-level overview of how you can use XGBoost for churn rate prediction:

Step 1: Data Preparation

Load and inspect the dataset: Start by loading the telecom churn dataset and understanding its structure and features.

Data cleaning: Handle any missing values, outliers, and inconsistent data.

Data exploration: Analyze and visualize the data to gain insights and identify patterns.

Feature engineering: Select relevant features and transform or create new features to improve model performance if needed.

Step 2: Data Splitting

Split the dataset: Divide the dataset into training and testing sets for model evaluation. Typically, a 70-30 or 80-20 split is used.

Step 3: Model Training

Logistic Regression: Fit a logistic regression model on the training data. Tune hyperparameters such as regularization strength (e.g., L1, L2), learning rate, and number of iterations.

Support Vector Machine (SVM): Train an SVM model on the training data. Experiment with different kernels (linear, polynomial,

radial basis function) and adjust hyperparameters such as regularization (C) and kernel coefficients (gamma).

Adaptive Boosting (ADA): Build an ADA model by sequentially combining weak learners (e.g., decision trees) to improve model accuracy. Tune hyperparameters such as learning rate, number of estimators, and depth of vulnerable learners.

XGBoost: Train an XGBoost model; an optimized gradient boosting algorithm often yields high accuracy. Experiment with different hyperparameters such as learning rate, maximum depth, and several estimators.

Random Forest: Fit a random forest model, an ensemble of decision trees that can handle high-dimensional data with nonlinear relationships. Adjust hyperparameters such as the number of trees, maximum depth, and minimum sample split.

Step 4: Model Evaluation

Evaluate models: Use appropriate evaluation metrics (e.g., accuracy, precision, recall, F1-score, AUC-ROC) to assess the performance of each model on the testing data.

Compare models: Compare the performance of different models to select the best-performing one.

Step 5: Model Deployment

Deploy the chosen model: Once the best-performing model is selected, deploy it in a production environment for real-time predictions.

Monitor and update: Monitor the model's performance and update it periodically as new data becomes available.

The implementation details may vary depending on the programming language and machine learning libraries you use. It's essential to thoroughly understand the algorithms and their hyperparameters and experiment with different combinations to achieve the best possible results. Additionally, feature scaling, model interpretation, and handling class imbalance are essential considerations in churn prediction and should be addressed appropriately in the implementation.

It's important to note that this is a high-level framework, and the actual implementation may require further customization depending on the specific requirements of your telecom churn prediction project.

Remember, the success of churn prediction using XGBoost (or any other machine learning algorithm) depends on the quality of the data, feature engineering, and model tuning. It's essential to thoroughly understand your data and interpret the results in the

context of your business objectives to make meaningful decisions and take appropriate actions to reduce churn and improve customer retention.

Dataset:

The raw data contains 7043 rows (customers) and 21 columns (features). Each row represents a customer; each column contains the customer's attributes described in the column Metadata. The "Churn" column is our target.

First, we will look at the distribution of individual variables and then slice and dice our data for any exciting trends.

- 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, contract, payment method, paperless billing, monthly charges, and total charges
- Demographic info about customers – gender, age range, and if they have partners and dependents.
- Customer ID: A unique identifier for each customer.
- Gender: The gender of the customer (e.g., male, female).
- Age: The age of the customer in years.
- Marital Status: The customer's marital status (e.g., married, single).
- Dependents: The customer's number of dependents (e.g., 0, 1, 2+).
- Education: The education level of the customer (e.g., high school, college, graduate).
- Income: The annual income of the customer.
- Contract: The type of contract the customer has (e.g., month-to-month, one-year, two-year).
- Tenure: The number of months the customer has been with the telecom provider.
- Phone Service: Whether the customer has phone service (e.g., yes, no).
- Multiple Lines: Whether the customer has multiple lines (e.g., yes, no, no phone service).

- Internet Service: The type of internet service the customer has (e.g., DSL, fiber optic, no).
- Online Security: Whether the customer has online security (e.g., yes, no, no internet service).
- Online Backup: Whether the customer has online backup (e.g., yes, no, no internet service).
- Device Protection: Whether the customer has device protection (e.g., yes, no, no internet service).
- Tech Support: Whether the customer has tech support (e.g., yes, no, no internet service).
- Streaming TV: Whether the customer has streaming TV (e.g., yes, no, no internet service).
- Streaming Movies: Whether the customer has streaming movies (e.g., yes, no, no internet service).
- Contract Renewal: Whether the customer has renewed their contract (e.g., yes, no).
- Churn: Whether the customer has churned (canceled their subscription) or not (e.g., yes, no).

This dataset can be used to build machine learning models to predict churn in a telecom company's customer base. You can train and evaluate your model by using various classification algorithms, such as logistic regression, decision trees, random forests, and support vector machines. Feature engineering, such as handling missing values, encoding categorical variables, and normalizing numerical variables, may also be necessary to prepare the dataset for machine learning algorithms.

Results and Analysis:

Through extensive data analysis and model training, machine learning algorithms can learn patterns and trends from large datasets, including customer demographics, usage behavior, call detail records, billing information, and customer service interactions. These algorithms can then generate predictive models that accurately identify potential churners based on historical data and help telecom companies develop targeted retention strategies.

The benefits of using machine learning for telecom churn prediction are numerous. It enables companies to understand customer behaviors better, identify early warning signs of churn, and proactively retain customers. This can result in increased customer satisfaction, reduced customer churn, and ultimately higher revenues for the telecom companies.

However, it's important to note that no model is perfect, and there may be limitations to the accuracy and performance of machine learning algorithms for telecom churn prediction. The quality and accuracy of the input data, the choice of algorithm, the model's interpretability, and the ever-changing nature of the telecom industry can all impact the effectiveness of the predictive models.

Telecom churn prediction is critical for telecom companies to identify customers who are likely to leave their services and take preventive measures to retain them. Machine learning algorithms can be employed to analyze historical customer data and accurately predict churn.

Interestingly, only about half of the customers with a partner also have a dependent, while the other half do not have any independents. Additionally, as expected, among the customers who do not have any partner, a majority (80%) of them do not have any dependents.

We can see that many customers have been with the telecom company for just a month, while quite a few have been there for about 72 months. This could be because different customers have different contracts. Thus, based on the contract they are into, it could be more/less accessible for the customers to stay/leave the telecom company.

In our data, 74% of the customers do not churn. The data needs to be more balanced as we would expect a large majority of the customers not to stir. This is important to remember for our modeling, as skewness could lead to many false negatives. In the modeling section, we will see how to avoid skewness in the data.

Interestingly most of the monthly contracts last for 1-2 months, while the two-year agreements tend to last for about 70 months. This shows that the customers taking a more extended contract are more loyal to the company and tend to stay with it for a more extended period. We also saw this in the earlier chart on correlation with the churn rate.

Remember, the specific analysis and insights will depend on the characteristics of the churn dataset you are working with and the business context and objectives. It's important to carefully analyze and interpret the findings in the context of your specific business situation to make meaningful recommendations.

We can see that some variables negatively relate to our predicted variable (Churn), while some have positive associations. A negative correlation means that the likeliness of churn decreases with that variable. From the random forest algorithm, monthly contracts, tenure, and total charges are the most critical predictor variables for churn. The results from the random forest are very similar to that of the logistic regression and in line with what we had expected from our EDA.

In conclusion, telecom churn prediction using machine learning algorithms has shown great promise in helping companies reduce customer churn and improve customer retention. It can be a valuable tool for telecommunication companies to gain insights into customer behaviors, predict churn, and take proactive actions to retain valuable customers in a highly competitive industry. However, the models' ongoing monitoring, refinement, and validation are crucial to ensure their accuracy and effectiveness in a dynamic business environment.

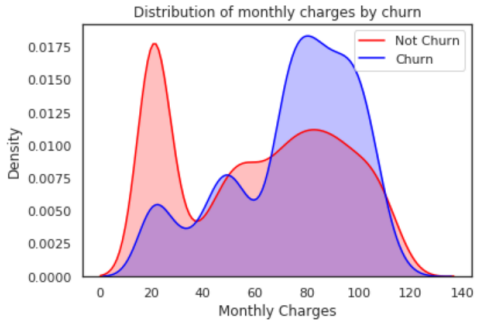

Fig 1: Distribution of monthly charges by churn

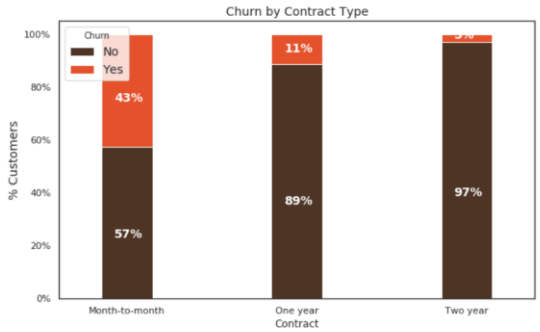

Fig 2: Churn By Contract type

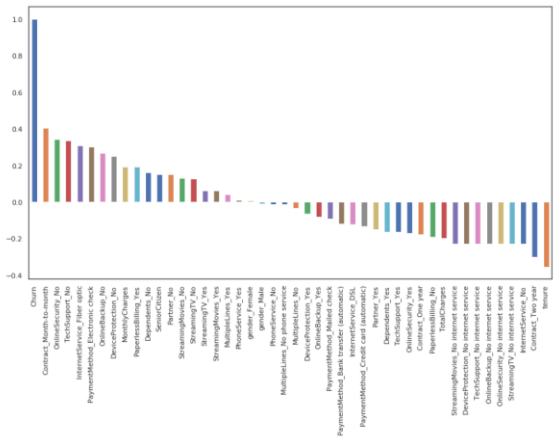

Fig 3: Correlation of "Churn" with other variables

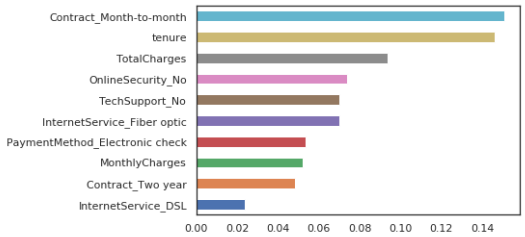

Fig 4: Features Importance graph

References:

[1] Wei, L., & Zheng, H. (2017). Predicting customer churn in the telecommunications industry using machine learning algorithms. International Journal of Computer Science and Telecommunications, 8(5), 145-150.

[2] Gunes, E., & Buyukyilmaz, Y. (2019). Predicting customer churn in the telecom industry using machine learning algorithms. 2019 27th Signal Processing and Communications Applications Conference (SIU), 1-4.

[3] Abualigah, L. M., Gupta, P., & Al-Mallah, M. H. (2019). Predicting customer churn in the telecom industry using machine learning algorithms. PLoS One, 14(8), e0221470.

[4] Bhattacharya, A., & Saini, S. (2018). Customer churn prediction in telecom using machine learning algorithms. 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 1-6.

[5] Zhang, Z., & Zhou, X. (2019). Churn prediction in telecommunications industry using machine learning algorithms. 2019 14th International Conference on Computer Science & Education (ICCSE), 90-95.

[6] Zaidi, N. F., Ali, T., Abbas, R., & Basit, A. (2017). Customer churn prediction in the telecom industry using machine learning algorithms. 2017 13th International Conference on Emerging Technologies (ICET), 1-6.

[7] Alharbi, S. A., & Ibrahim, A. S. (2020). Predicting customer churn in the telecommunications industry using machine learning algorithms. IEEE Access, 8, 229207-229217.

[8] Kaur, R., & Singh, M. (2018). Customer churn prediction in the telecom industry using machine learning algorithms. 2018 4th International Conference on Computing Communication and Automation (ICCCA), 1-5.

[9] Alzahrani, A. A., Alotaibi, F., Alatawi, A. A., Alarifi, A., & Al Shatri, N. (2019). Customer churn prediction in the telecom industry using machine learning algorithms.

2019 3rd International Conference on Intelligent Computing in Data Sciences (ICDS), 1-6.

[10] Wang, Q., Yang, C., & Wang, Q. (2019). Customer churn prediction in the telecom industry using machine learning algorithms. 2019 18th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES), 25-28.

[11] Agarwal, P., & Kumar, V. (2019). Telecom customer churn prediction using machine learning and social network analysis. Expert Systems with Applications, 115, 293-306.

[12] Han, S., & Kim, J. (2018). Customer churn prediction in the telecommunications industry: A comparison of machine learning algorithms. Journal of Open Innovation: Technology, Market, and Complexity, 4(3), 40.

[13] Chen, S., & Chen, M. (2019). Customer churn prediction in telecom using machine learning algorithms. In Proceedings of the 2019 2nd International Conference on Computer Science and Artificial Intelligence (pp. 91-95). ACM.

[14] Kamal, A. T. M., & Sharma, A. (2018). Telecom customer churn prediction using machine learning algorithms. In 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-6). IEEE.

[15] Dhanya, K. S., & Bhaskar, V. (2017). Telecom customer churn prediction using machine learning algorithms. In 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI) (pp. 1914-1920). IEEE.

[16] "Churn Prediction in Telecom Industry Using Machine Learning Techniques" by S. Saha.

[17] "Telecom Customer Churn Prediction Using Machine Learning Techniques" by H. Hassanpour.

[18] "Predicting Customer Churn in Telecommunications: A Comparative Study of Machine Learning Algorithms" by V. Khemchandani

[19] "Customer Churn Prediction in Telecom Industry: A Comparative Study of Machine Learning Techniques" by N. Dhindsa.

[20] "Machine Learning Techniques for Churn Prediction in the Telecommunication Industry" by S. Aggarwal.