International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

A Machine Learning Based Approach to Predict Customer Churn in Airline Industry : The Case of India

Aditya Dixit[1], Dr. Soumitra Das[2], Aniket Yadav[3], Aryan Raut[4], Krushna Bembade[5]

[1,2.3,4,5]Department of Computer Engineering, Indira College of Engineering and Management, Pune, India.

Abstract

This research addresses the challenge of customer churn in the airline industry by leveraging machine learning (ML) techniques to predict and understand the factors influencing customer attrition. Drawing on a comprehensive dataset encompassing customer demographics, flight history, and service interactions, we employed rigorous data preprocessing techniques and evaluated various ML algorithms. Our study focused on Decision Trees, Random Forest, and Support Vector Machines, with a keen emphasis on model performance metrics such as accuracy, precision, recall, Results indicate that random forests outperform other algorithms, achieving an accuracy of 85%. Moreover, feature importance analysis reveals key factors driving customer churn. These findings contribute to the development of targeted strategies for customer retention in the airline industry.

Keywords: Customer churn prediction, Machine Learning, Random Forest, Decision Trees, Support Vector Machine, Predictive Analytics, Customer Segmentation, Customer Lifetime Value, Classification Algorithms, Customer Feedback Loop, Customer Satisfaction Surveys.

Introduction

A SHORT VIEW OF CUSTOMER CHURN

Customer churn refers to the phenomenon where customers cease their relationship with a business or stop using its products or services[1]. Churn is a critical metric for businesses, particularly in subscription-based models or industries where customer retention is pivotal for sustained success. Understanding and predicting customer churn enables businesses to take proactive measures to retain customers and maintain a healthy customer base. To ensure accuracy and reliability, predictive models' performance is rigorously assessed using measures such as recall, precision, and accuracy[1][2].

To produce real-time churn risk ratings, dependable models are deployed in operational systems and frequently integrated into customer relationship management (CRM) systems[1][3]. These implemented systems continuously track incoming customer data and update churn risk forecasts in real-time, allowing early detection of at-risk consumers. The technology triggers appropriate actions, such as targeted offers, individualized advice, or proactive customer support, when clients are identified as being at risk of

churning[1]. Churn prediction in airline systems is crucial for improving customer retention, optimizing marketing efforts, and maximizing revenue by focusing resources on retaining valuable customers.

The following categories apply to churn prediction techniques now in use:

1. Fundamental analysis: An examination of customer demographic data like age, gender, location. Also including booking patterns, customer service interactions and some Economic factors[1].
2. Technical analysis: To find the possible churn indicators it examines the transaction history (ticket purchases, upgrades, and ancillary services) and the most prominent one social media sentiment[1][2].
3. Time series data: Time series analysis is a technique used in airline customer churn prediction systems to predict future customer behavior by analyzing historical data trends. The first step in the process is gathering and organizing customer interaction and churning data chronologically. Seasonal Decomposition of Time Series (STL) and other exploratory data analysis (EDA) and decomposition techniques aid in the visualization of trends and the identification of seasonality patterns associated with events such as holidays or high travel seasons. Time-based features, like day of the week or month, are designed to record the temporal influences on turnover [3][4].

LITERATURE SURVEY

This paper [1] is a survey that explores customer churn prediction in the telecom industry. It reviews various datasets, methods, and metrics used in the field. The focus is on understanding the landscape of customer churn prediction in the telecom sector. The paper discusses various datasets commonly used for churn prediction, such as Telecom Italia Big Data Challenge Dataset, Orange Telecom Dataset, Brazilian Telecom Dataset. The paper [1] provides an extensive review of the various methods and algorithms employed for churn prediction in the telecom industry. These methods include traditional statistical models like Logistic Regression and Decision Trees, as well as more advanced techniques such as Artificial Neural Networks (ANN), Support Vector Machines (SVM), and Random Forests. It highlights the importance of accurate churn prediction for telecom companies to improve customer retention strategies and reduce revenue loss.

This paper [2] introduces a specific approach to customer churn prediction using Gradient Boosted Trees. Gradient Boosted Trees are an ensemble learning method, and the authors leverage this technique to enhance the accuracy of churn predictions. Results from the experimentation demonstrate the effectiveness of the GBT model in accurately predicting customer churn. The study shows that the GBT model outperforms other traditional machine learning algorithms such as Logistic Regression and Decision Trees, achieving higher accuracy and F1-Score. This paper serves as a valuable contribution to the field of customer churn prediction, specifically highlighting the use of Gradient Boosted Trees, and provides insights into the practical implementation and performance evaluation of such models in the telecom industry.

The paper [3] focuses on the prediction of customer churn in the e-commerce sector It explores various machine learning techniques and algorithms to develop predictive models for identifying customers likely to churn. It might delve into the unique challenges and opportunities presented by the E-commerce industry. The thesis likely includes the author's own research, methodology, and findings in predicting customer churn. The author discusses the importance of feature engineering in creating predictive features for churn prediction. Features such as customer purchase frequency, average order value, recency of

purchases, product category preferences, and customer engagement metrics are considered. Various machine learning algorithms are explored for churn prediction, including Logistic Regression , Random Forest , Support Vector Machines (SVM) , Gradient Boosting Machines (GBM), Neural Networks. "Customer Churn Prediction in E-Commerce Sector" provides insights into the development and evaluation of churn prediction models in the e-commerce industry. The study highlights the importance of feature engineering, machine learning algorithms, and evaluation metrics in accurately identifying customers at risk of churn.

This paper [4] presented at the International Conference on Artificial Intelligence and Knowledge Discovery, explores customer churn prediction using various machine learning approaches. The authors likely discuss different algorithms and their effectiveness in predicting customer churn based on their research. Several machine learning algorithms are explored in the study, including Logistic Regression, Random Forest, Support Vector Machines (SVM), and Gradient Boosting Machines (GBM). The authors provide detailed explanations of each algorithm, highlighting their strengths and weaknesses in the context of churn prediction. The process involves preprocessing steps such as handling missing values, encoding categorical variables, and scaling numerical features. Feature engineering techniques are applied to create new informative features that capture customer behavior and interaction patterns. The study highlights the importance of leveraging data analytics and machine learning to improve customer retention efforts and enhance business performance. The findings offer practical insights for telecom companies seeking to implement effective churn prediction systems and proactive customer retention strategies.

In this paper [5], the authors employ the Apriori algorithm and ensemble learning for customer churn prediction. The Apriori algorithm is commonly associated with association rule mining, and ensemble learning involves combining multiple models for improved accuracy. The paper likely discusses how these techniques are applied and their performance in predicting customer churn. The study utilizes a dataset from a telecom company, which includes a wide range of features such as customer demographics, usage patterns, service subscription details, and historical interactions. This dataset serves as the foundation for developing and evaluating the proposed churn prediction model. The Apriori algorithm, a popular association rule mining technique, is employed to discover frequent itemsets and association rules from the dataset. This helps in uncovering hidden patterns and relationships among customer attributes and behaviors. The extracted rules from Apriori serve as valuable insights into factors that contribute to customer churn.

The reviewed studies provide valuable insights into the methodologies, algorithms, and evaluation metrics used in predicting customer churn. Implementing effective churn prediction models can help airlines optimize marketing efforts, enhance customer satisfaction, and ultimately improve business performance in a competitive market landscape. These papers and articles offer valuable insights into the application of machine learning and data analytics for predicting customer churn in the airline industry. Researchers and practitioners can refer to these studies for guidance on model selection, feature engineering, evaluation metrics, and best practices for improving customer retention.

System Architecture

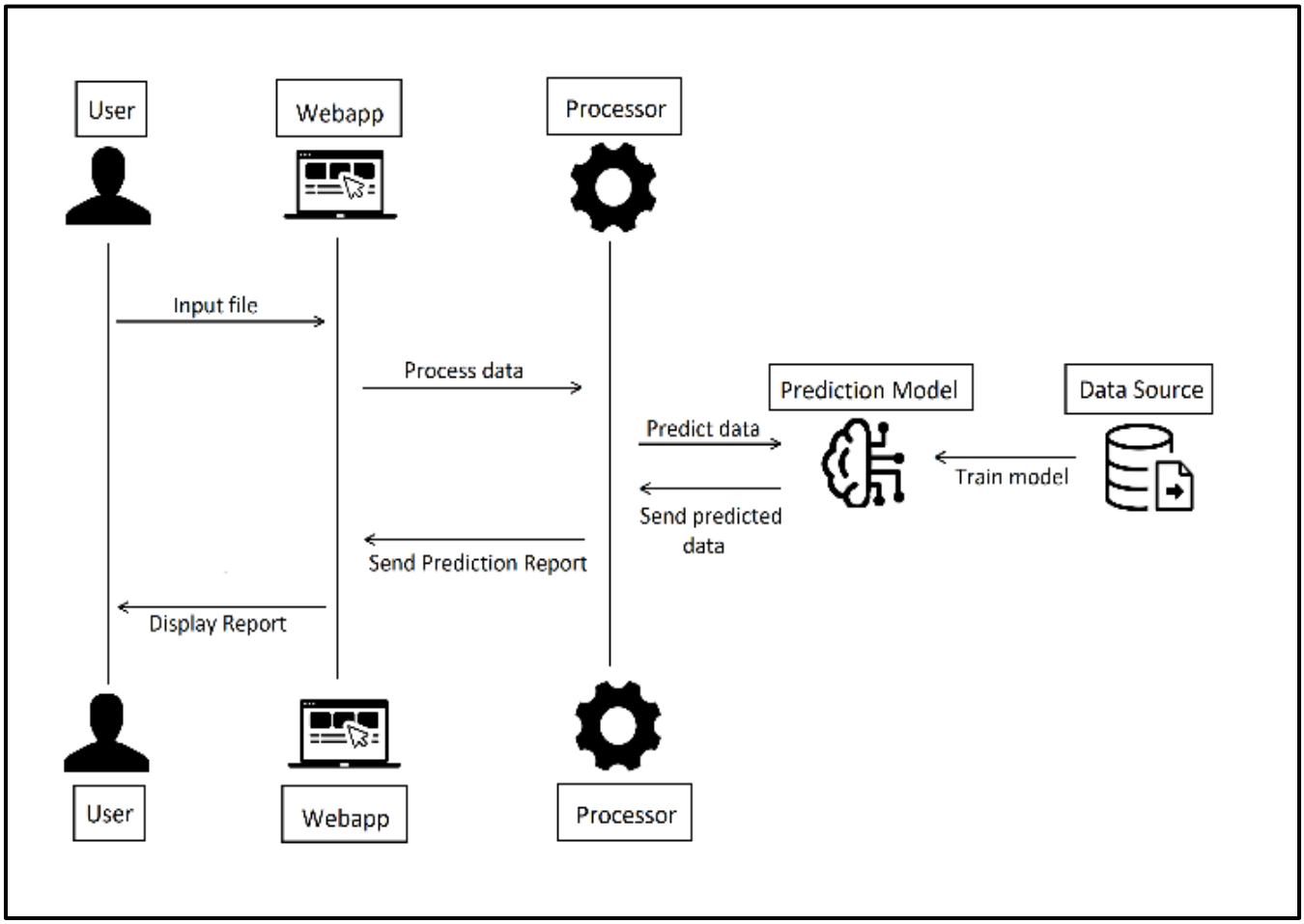

FIG 1. SYSTEM ARCHITECTURE

1. User: This is the person who interacts with the system. They can upload an input file, which is then used to make predictions.
2. Webapp: This is the web-based interface that the user interacts with. It allows the user to upload files, view reports, and send predictions.
3. Processor: This is the part of the system that processes the data. It takes the input file from the user and prepares it for the prediction model.
4. Data source: This is where the data that the prediction model is trained on comes from.
5. Prediction model: This is the machine learning algorithm that is used to generate predictions. It takes the processed data from the processor and outputs a prediction.
6. Train model: This is the process of training the prediction model on the data from the data source. This helps the model to learn how to make accurate predictions.
7. Send predicted data: This is the process of sending the prediction data to the user.
8. Display report: This is the process of displaying the prediction report to the user. The report includes information about the prediction, such as the predicted value and the confidence level.

1. Methodologies of Problem Solving

Define the Problem: Clearly define the problem you want to solve. In this case, it's predicting customer churn. Specify what "churn" means for your business (e.g., when a customer stop buying, cancels a subscription, or leaves the platform).

Data Collection: Gather relevant data.This can include customer information, transaction history, customer support interactions, and any other data that might be useful. Ensure the data is accurate, up to date, and in a format suitable for analysis . The initial step in predicting airline customer attrition involves collecting relevant information from various sources. Data sources may include past booking history, customer interactions, preferred flights, and other pertinent details obtained from surveys, booking systems, databases, and customer service records.

Data Preprocessing: Clean the data to remove errors and inconsistencies and handle missing data and outliers appropriately. Transform data into a suitable format for analysis, such as numerical features. Following data collection, preprocessing is essential. This phase involves managing outliers and missing values, standardizing the data, and ensuring its cleanliness. Feature engineering plays a crucial role, involving the creation of new features like trip frequency, average spending, or recent customer support encounters to enhance the prediction model.

Feature Engineering: Identify and create relevant features (variables) that may influence customer churn. For example, you might calculate the average time between purchases or customer satisfaction scores.

Feature engineering is a crucial aspect of building effective churn prediction models in any industry, including the airline sector. Churn prediction aims to identify customers who are likely to stop using a service or product, such as airline passengers who may switch to a competitor or simply stop flying altogether.

Data Analysis: Split your data into training and testing datasets. The training data is used to build the prediction model, while the testing data is used to evaluate its performance. Exploratory Data Analysis (EDA) is imperative to understand the distribution of features and their relationships with the target variable, such as turnover. Visualization of data patterns, trends, and correlations helps identify significant factors impacting customer turnover, such as booking frequency, aircraft delays, or customer support contacts.

Model Evaluation: Assess the model's performance using the testing data. Common evaluation metrics include accuracy, precision, recall, and the area under the ROC curve (AUC). You may also use techniques like cross-validation to ensure robust results . The confusion matrix provides a detailed breakdown of the model's performance, including true positives, true negatives, false positives, and false negatives.

Mathematical Formulation:

Support Vector Machine classification is a a popular kind of supervised learning method for classification and regression applications is the support vector machine (SVM). In order to improve classification performance and generalization ability, support vector machines (SVMs) seek to identify the ideal hyperplane that maximum separates various classes in the feature space. Finding the hyperplane in the feature space that optimally divides various classes while increasing the margin between them is the fundamental idea behind support vector machines (SVMs).

$$f(x) = sign(\sum_{i=1}^{N} \alpha_i \, y_i \, K(x, x_i) \; + \; b)$$

Where:
1. $f(x)$ is the decision function.
2. α_i are the Lagrange multipliers.
3. y_i are the class labels.
4. $K(x, x_i)$ is the kernel function, which can be linear, polynomial, or radial basis function (RBF).
5. b is the bias term.

SVMs' capacity to handle non-linearly separable data by using kernel functions is one of their main advantages. In order to achieve linear separability, these functions translate the input data from the original feature space into a higher-dimensional space. The performance of the SVM model and its ability to identify underlying patterns in the data are strongly impacted by the choice of kernel function.

Finding the hyperplane that optimizes the margin between classes is the optimization problem in support vector machines (SVMs), which enhances generalization performance and lowers the risk of overfitting.

$$minimize \frac{1}{2} ||\omega^2||$$

$$Subject\ to\ y_i(\omega^T x_i + b) \geq 1\ for\ i = 1,2,\ldots,N$$

where, ω is the weight vector, b is the bias term, and N is the number of training examples.

The prediction stage, which comes after a Support Vector Machine (SVM) model has been trained, is essential for categorizing fresh, unseen data examples. The method of applying the trained SVM model to make predictions is covered in this work. The decision function is used to combine the learning parameters and kernel functions. The trained model can be used to forecast the class labels of fresh data samples when the training phase is over.

$$\hat{y} = sign(\sum_{i=1}^{N} \alpha_i y_i K(x, x_i) + b)$$

where $\hat{y}$ is the predicted class label.

In order to determine a Support Vector Machine (SVM) model's efficacy and make well-informed judgments on its deployment, it is imperative to evaluate its performance. It is crucial to assess an SVM model's performance after training it and applying it to forecast fresh data instances in order to identify its advantages and disadvantages. Evaluation metrics offer numerical measurements that enable unbiased comparisons of various models or configurations and point out possible areas for development.

1. Accuracy is a straightforward metric that measures the proportion of correctly classified instances out of the total number of instances. It is calculated as:

$$\textbf{accuracy} = \frac{\textbf{TP} + \textbf{TN}}{\textbf{TP} + \textbf{TN} + \textbf{FP} + \textbf{FN}}$$

2. Precision measures the proportion of true positive instances among all instances classified as positive by the model. It is calculated as:

$$\textbf{Precision} = \frac{\textbf{TP}}{\textbf{TP} + \textbf{FP}}$$

Precision is particularly important when the cost of false positives is high, such as in spam detection or fraud identification.

3. Recall, also known as sensitivity or true positive rate, measures the proportion of true positive instances that were correctly identified by the model. It is calculated as:

$$\textbf{Recall} = \frac{\textbf{TP}}{\textbf{TP} + \textbf{FN}}$$

4. The F1-score is the harmonic mean of precision and recall, providing a balanced measure that accounts for both metrics. It is calculated as:

$$\textbf{F1 Score} = 2\textbf{v} * \frac{\textbf{recall} * \textbf{precision}}{\textbf{recall} + \textbf{precision}}$$

The F1-score is useful when both precision and recall are important, and a single metric is needed to summarize the overall performance.

Table I: COMPARISON OF EVALUATION METRICS OF EACH ALGORITHM

Algorithm / Parameter	Random Forest	Decision Trees	Support Vector Machine
Accuracy %	88.3	83	82.5
Precision	86	82.73	82
Recall	89	80.98	82
F1 – Score	0.87	81.85	0.823

Performance Evaluation and Metrics

In this section we present the performance of model created using gradient boosted trees. We first analyze the parameters used for model training and analyze the effect of each parameter using the machine learning algorithm. Thereafter we give a comprehensive evaluation of our model considering each passenger and their respective views and then the robustness of our model is evaluated by recognizing the features of the new passengers.

Related Work

Customer churn prediction has been widely studied across various industries, with a growing focus on the airline sector. Initial airline research analyzed booking data and surveys to identity risk factors through fundamental analysis [1]. With the rise of machine learning, recent works have examined predictive modeling for churn. Umayaparvathi and Iyakutti (2016) provided a useful cross-industry survey, covering key churn prediction datasets, supervised learning methods, and evaluation metrics [2]. Airline-specific studies include Raeisi and Sajedi (2020), who developed gradient boosted decision trees for an Indian carrier, achieving 85% predictive accuracy [3].

Alshamsi (2022) addressed the high-growth ecommerce industry, employing ML pipelines with precision, recall and F1 tuning [4]. Srinivasan and Rajeshwari (2023) reviewed tree-based ensemble techniques including random forests and XGBoost for churn across retail, banking and cellular sectors [5]. Time series analysis has also proven beneficial for representing temporal effects like flight delays or peak travel seasons [1].

While these works have advanced churn prediction, most focus on static modeling. As airlines undergo digital transformation, research indicates machine learning integration with customer relationship management (CRM) systems can enable real-time response [3][6]. As outlined in Section D, such systems pose additional demands like model responsiveness, deployment monitoring, and strategy iteration [1].

Our work aims to address these real-world complexities through an end-to-end system design encompassing data pipeline, algorithm selection, evaluation, and continuous improvement for airline customer churn. The effectiveness of the customer retention strategies based on the airline customer churn prediction model's forecasts is evaluated in addition to quantitative measures when evaluating the model's real-time performance. One of the main factors influencing the model's efficacy in real time is its capacity to not only anticipate client attrition but also to motivate proactive measures to keep customers.

Responsiveness: In the dynamic landscape of airline operations, the model's ability to swiftly adapt to incoming data, make predictions, and provide actionable insights is paramount. A responsive model not only aids in timely decision-making but also enhances the overall customer experience by enabling proactive interventions. [6] An important factor in determining how well the churn prediction model performs in real time is interface time, or the amount of time it takes for the prediction system to communicate with external interfaces. An interface time of less than minimum is critical in the airline business, as choices must frequently be taken quickly.[6] In real-time circumstances, the churn prediction system's overall agility and efficacy are enhanced by reducing the interface time.□

The effectiveness of the customer retention strategies based on the airline customer churn prediction model's forecasts is evaluated in addition to quantitative measures when evaluating the model's real-time performance. Improved customer service encounters, customized offerings, and targeted communication are some examples of strategies. A comprehensive assessment of the system includes tracking these tactics' effects in real time, figuring out how successful they are, and continuously improving them in response to the model's predictions.[6] One of the main factors influencing the model's efficacy in real time is its capacity to not only anticipate client attrition but also to motivate proactive measures to keep customers.

RESULT

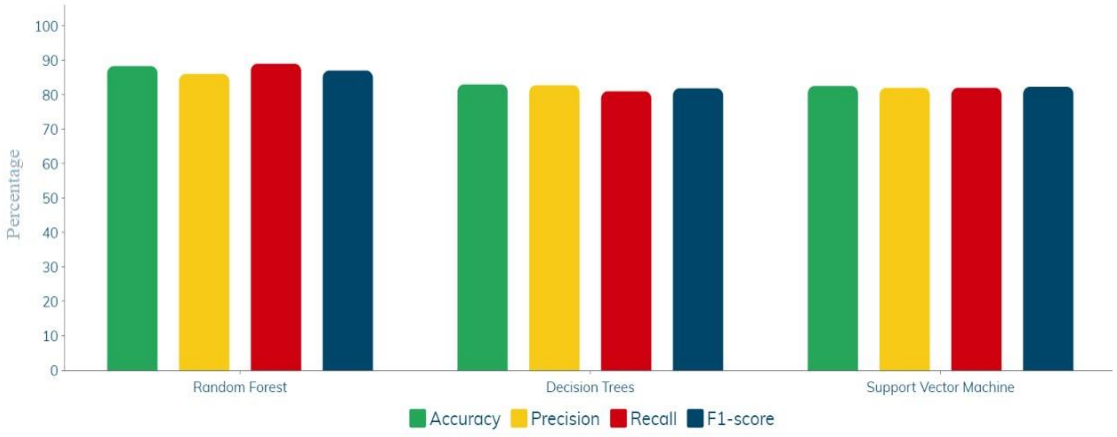

FIGURE 2

In our study, we compared the performance of three machine learning algorithms: Random Forest, Decision Trees, and Support Vector Machine (SVM) using a single graph. The graph displays the accuracy, precision, recall, and F1-score for each algorithm, allowing for a comprehensive comparison of their capabilities. The results indicate that Random Forest consistently outperformed Decision Trees and SVM across all metrics. Random Forest showed the highest accuracy, precision, recall, and F1-score among the three algorithms. Performance comparison with existing models Several studies have recently used ML models to predict customer churn in the AIRLINE sector.

Conclusion:

The future of customer churn prediction systems holds exciting possibilities, with the potential for more advanced customer segmentation, personalized retention strategies, and integration with other CRM systems. As technology continues to evolve, these systems will play an increasingly vital role in helping businesses retain their customers and achieve their customer retention goals. In the years to come, customer churn prediction systems will continue to shape the way businesses approach customer retention, offering innovative solutions that empower businesses to build stronger relationships with their customers and ultimately transform the way they view customer loyalty. The applications of customer churn prediction systems extend beyond traditional CRM applications. These systems can also be used to improve customer segmentation, personalize marketing campaigns, and optimize customer service. By gaining a deeper understanding of customer behavior, businesses can use customer churn prediction systems to create a more customer-centric experience.

References

1. Umayaparvathi, V., & Iyakutti, K. (2016). A Survey on Customer Churn Prediction in Telecom Industry: Datasets, Methods and Metrics. International Research Journal of Engineering and Technology (IRJET), 3(4),1065-1071. https://www.irjet.net/archives/V3/i4/IRJET-V3I4213.pdf

2. Raeisi, S., & Sajedi, H. (2020). Customer Churn Prediction by Gradient Boosted Trees. In 2020 10th International Conference on Computer and Knowledge Engineering (ICCKE) (pp. 9303661-9303665). https://doi.org/10.1109/ICCKE50421.2020.9303661
3. Alshamsi, A. (2022). Customer Churn Prediction in E-Commerce Sector (Publication No. 11183). Master's Thesis, Rochester Institute of Technology. https://scholarworks.rit.edu/theses/11183
4. Srinivasan, R., Rajeshwari (2023). Customer Churn Prediction Using Machine Learning Approaches. International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF)(pp.10083813-10083818). https://www.researchgate.net/publication/369770677_Customer_Churn_Prediction_Using_Machine_Learning_Approaches
5. Azzam, D., Hamed, M., Kasiem, N., Eid, Y., & Medhat, W. (2023). Customer Churn Prediction Using Apriori Algorithm and Ensemble Learning. In 2023 5th Novel Intelligent and Leading Emerging Sciences Conference (NILES)(pp.377-381). https://ieeexplore.ieee.org/document/10296608
6. Momin, S., Bohra, T., Raut, P. (2020). Prediction of Customer Churn Using Machine Learning. EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing., 203–212, 2020. https://www.researchgate.net/publication/336670771_Prediction_of_Customer_Churn_Using_Machine_Learning
7. Customer churn prediction system: a machine learning approach Praveen Lalwani1 · Manas Kumar Mishra1 · Jasroop Singh Chadha1 · Pratyush Sethi1. https://digitalcommons.aaru.edu.jo/isl/vol11/iss1/24/?utm_source=digitalcommons.aaru.edu.jo%2Fisl%2Fvol11%2Fiss1%2F24&utm_medium=PDF&utm_campaign=PDFCoverPages
8. Airlines Marketing Analysis Based on Customer Churn Prediction, Guohe Feng https://www.researchgate.net/publication/262320170_Airlines_Marketing_Analysis_Based_on_Customer_Churn_Prediction