BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

Customer Analytics using Data Science: A Review of Segmentation, Churn Prediction, Recommendation and Explainable AI Techniques

[1]Pranav Gosavi, [2]Neelam jain, [3]Dr. Vishal Puranik

[1]Student, [2]Assistant Professor – Seminar Guide, [3]Assistant Professor – Seminar Guide
[1]Department of Artificial Intelligence and Data Science,
[1]Ajeenkya DY Patil School of Engineering, Pune, India

Abstract

The information age is dominated by massive data collection. Data on what customers buy, use and write are some examples. Manually processing and analyzing this volume of data is infeasible. Therefore, Data Science techniques are utilized to extract the value and patterns in the data. Customer Analytics, the science of analyzing customer behavior, is a classic example of this. We outline the data preprocessing, feature engineering, customer segmentation, churn prediction, recommendation systems, and Explainable AI (SHAP and LIME) steps of this process. We examine several recent papers in the domain of churn prediction with application in ecommerce, banking, telecommunications, and retail sectors. Churn prediction and Ensemble methods such as Random Forest and XGBoost are the standard in industry. While there is a wealth of literature on churn prediction using data from a single sector, there are many remaining challenges. These include data privacy, multi source data integration, bias, and cold-start problems. There is a need to get rapid, actionable insights. As a review paper, we don't build a system to test our hypotheses.

Keywords

Customer analytics, Customer segmentation, Churn prediction, Data science, Explainable AI, Machine learning, Personalization, Predictive analytics, Recommendation systems, Retention, Sentiment analysis.

I. INTRODUCTION

With the advent of online purchases and mobile app and wallet transactions, and online review systems, a vast array of consumer data is being generated. Consumer data collected and processed enables us to understand consumer behavioral patterns, their likes and dislikes and their potential future actions. This data is leveraged to better customer satisfaction, enhance company products and optimize the effectiveness of company marketing strategies.

Consumer data is rapidly increasing because of e-commerce, mobile applications, and digital payments. Now simple spreadsheets and reports are insufficient for analyzing customer data and for finding hidden customer behavior and trends. Customer data analysis is where data science comes in to the picture to address the need for a solution that combines data preprocessing, statistics, visualization, and machine learning.

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

There are three different types of customer analytics. They are known as descriptive, predictive, and prescriptive analytics. Descriptive analytics gives you insight into what has already occurred, for instance last month's sales. Predictive analytics gives you the potential future sales and also provides you with insight into the potential churn rate of customers. Prescriptive analytics recommends to you the best path for your customers. ML algorithms are able to cluster your customers to target and recommend products for them and also are able to predict the customers that have a high probability of churning.

Finally, consumer reviews are processed and analyzed using NLP and ML algorithms. In order to get high-quality and impactful analysis of the consumer data, the models and the data itself must be of a high quality and must also account for privacy concerns and regulations. Also, many of the algorithms within the machine learning models that are used in customer analytics function as a "black box" to the end user. To overcome this problem, tools have been developed to explain the algorithms within the models and the resulting output. Some examples are SHAP [9] and LIME [10].

The focus of this paper is to study customer analytics and the tools used in its process, including data collection and preparation, and visualization. Furthermore, this paper provides a review of various ML approaches that are utilized for the purposes of segmentation, prediction of churn and recommendation systems.

II. BACKGROUND

A. Data Types for Customers

Data sources for customers are numerous, and the main types are:

- Transactional Data: Purchase details including payment methods, orders, and order values, as well as discounts.
- Behavioural Data: This is the user's actions on an app and includes the pages that are viewed, the user's searches, clicks and even the user's app activity.
- Demographic and Profile Data: Details like age, location and type of account.
- Service-Usage Data: The telecom sector uses this data type for contract length, service usage and the type of customer service account.
- Feedback Data: Ratings and reviews are included in this data.

Each data type used in isolation is limited in its capacity to describe the user experience. Transactions do not reflect customer satisfaction. Though combining all the data sources provides a comprehensive view, the combination raises data privacy concerns.

B. Customer Lifecycle and Retention

Customers tend to go through the phases of acquisition, growth, retention, and sometimes, loss. As retaining customers is more profitable than attracting new customers, the field of churn analysis has seen the development of a wide range of customer retention strategies and models [1], [5].

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

III. DATA PREPROCESSING AND FEATURE ENGINEERING

A. Data Preprocessing

Real customer data is messy, has duplicate entries, values of different ranges and missing values. Data Preprocessing helps in cleaning and preparing the data for the input model. Below mentioned table summarizes the most common data preprocessing techniques used.

TABLE I. COMMON PREPROCESSING STEPS

Technique	Purpose	Example
Duplicate removal	Avoid double counting customer.	Saving same order twice.
Missing values Handling	Fill/Drop.	Fill missing age value with the median.
Treating Outliers	Reduced influence of extreme values.	Huge one-time purchase.
Encoding	Converting text values to numbers.	Contract Type: One Hot Encoding.
Scaling	Bringing values to similar range.	Normalize Tenure/Income.
Balancing Classes	Handle cases of lesser Churners.	Class Weights/Oversampling.

Scaling is important in Distance-based methods. Class imbalance is a very common challenge in Customer Churn. A good model that makes a “No Churn” prediction will have an acceptable accuracy rate [1], [2].

B. Feature Engineering and Selection

In Feature Engineering, we create new features from existing features in our data. The most common technique is RFM which stands for Recency, Frequency, and Monetary value. Another example is Average Basket Size and many more like the tenure, complaints count, and changes in the usage of a service in the last few months. Following this is the removal of features with similar data and information.

IV. EXPLORATORY DATA ANALYSIS

Before creating a model, it’s best to learn from the data. To gain an understanding of the data, different statistical methods and charts are used. Histograms reveal data distributions, box plots indicate outliers, and correlation matrices display how variables change in tandem. Trends can also be visualized through time using a line graph. Line graphs help distinguish trends such as peak times in sales or periods of lower activity. Exploratory data analysis can assist with selecting which data features to include, and can often indicate the reasoning behind customer churn. For example, shorter customer contracts have higher customer churn. Dashboards allow managers to see data results, even those who don’t have a technical background.

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

V. CUSTOMER SEGMENTATION

For the sake of customer management, segmentation puts customer groups in silos that can then be approached with different methods. Since segmentation relies on grouping, unsupervised clustering techniques are utilized. K-Means is most common because of its simple design. Initially, k clusters are determined. Customers are then assigned to the cluster with the closest centroid. Then the cluster centroids are updated to the center of the assigned customer points. This process repeats until the clustering becomes stable.

$$J = \Sigma_i \, \Sigma_{x \in Ci} \, \|x - \mu_i\|^2$$

The objective function J measures how much the cluster members deviate from the assigned cluster centroid. The most important variable is k. Most people decide on a value of k by using either the elbow method or by the silhouette score. Hierarchical clustering or DBSCAN may be used instead of K-Means if clusters do not form circles and have many noise points.

Table II shows RFM inputs. This example is illustrative.

TABLE II. ILLUSTRATIVE EXAMPLE OF RFM SEGMENTS

Segment	Recency	Frequency	Spending
Loyal customers	Recent	High	High
Occasional buyers	Medium	Low	Low
At-risk customers	Long ago	Was high	Was high
New customers	Very recent	Very low	Low

Segmentation gives a better understanding of how customers should be approached. For example, rewarding loyal customers with points, while using different tactics for new or at-risk customers.

VI. CHURN PREDICTION AND RETENTION

Customer retention through churn prediction has become a standard predictive analytics tool. Churn is when customers stop paying. Using supervised learning techniques, we train the model with previous data where we define a binary label: retained (did not churn) or churned. We will define customer probability score to predict the propensity of churn for our current customer base. Retention personnel will focus their outreach programs on the highest probability customers. Fig. 1 shows a general customer analytics pipeline.

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

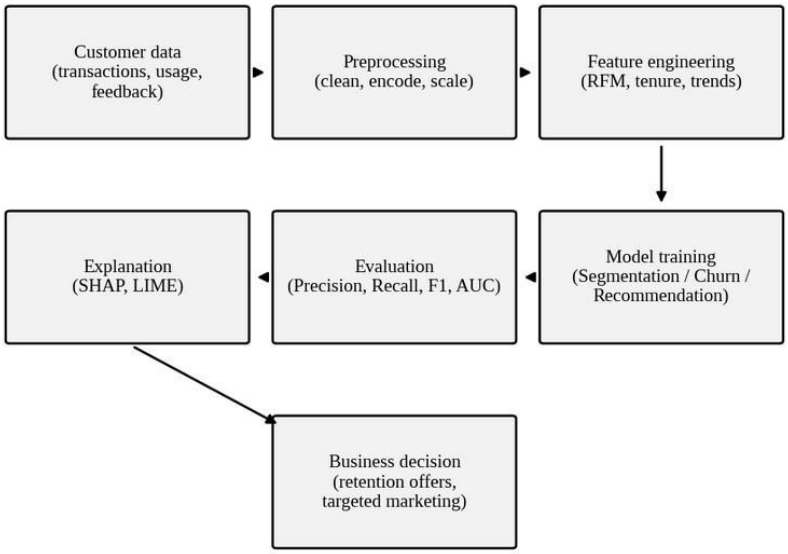


Fig. 1. General customer analytics pipeline.

A. Models Used

There are a number of techniques used for building churn prediction models. A comparison of models found most often in our research can be seen in Table III.

TABLE III. COMMON CHURN PREDICTION MODELS

Model	Advantage	Limitation
Logistic Regression	Fast, simple, easy to interpret	Not good at learning complex relationships
Decision Tree	Easiest for creating simple rules	Can be overfit to data
Random Forest	More accurate than single decision tree	Harder to interpret
XGBoost [8]	Good accuracy on data in a tabular format	Requires tuning and has poor explainability
LSTM [2]	Good for time sequential data	Need a larger data set and greater computational resources

B. Evaluation Metrics

Because accuracy is misleading with imbalanced data, other metrics must be utilized. For a particular class (e.g., churned customers), TP represents the number of true positive instances, FP is the number of false positives, and FN is the number of false negatives. Then we define:

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

$$\text{Precision} = TP / (TP + FP) \quad (2)$$

$$\text{Recall} = TP / (TP + FN) \quad (3)$$

$$F1 = 2 \times \text{Precision} \times \text{Recall} / (\text{Precision} + \text{Recall}) \quad (4)$$

Precision informs the user on the number of customers actually churned who were identified by the prediction as churned, whereas recall provides the number of true churned customers who were detected by the prediction. Also used extensively is the Area Under the Curve (AUC) of the ROC. The balance of missing a customer who should have been classified as churned vs. the cost of making an offer to a customer who was actually not going to churn is often a difficult one to estimate. Churn prediction models, like all machine learning models, must be tested for generalization across different data sets and/or time periods.

VII. SUGGESTIONS AND CUSTOMIZATION

Most sites use recommendation systems. Amazon uses recommendation systems to help customers find things to buy and Amazon video uses them to help customers decide what to watch. Recommendation systems can fall into one of three main categories.

- Content based recommendation: Recommends things the customer has shown an interest in based on properties like category, brand, etc.
- Collaborative filtering recommendation: Recommends things other users have shown an interest in.
- Hybrid methods: Combine elements from the first two to make recommendations.

Most recommendation systems suffer from the cold start problem. A new customer has no profile. A new product has no ratings. The systems can't give suggestions to new customers or products until a user interacts with the system. Recommendation systems can give suggestion based on general popularity until new data comes in or by giving new users a quick profile setup to gather information.

VIII. EXPLAINABLE AI

Black box models like Deep learning and ensemble learning models are very good at what they are built to do (accuracy), but building trust and ensuring transparency is hard. If a Telecom company decides to accept a prediction from a model, they must have a valid explanation. Explainable AI (XAI) helps with providing this transparency.

SHAP [9] uses Shapley values from game theory. SHAP assigns each feature a value that contributes to the final prediction, and the contribution value of each feature equals the difference between the prediction and the base value of the model.

$$f(x) = \varphi_0 + \Sigma\, \varphi_j x_j \quad (5)$$

where φ_0 = base value and φ_j = feature contribution of feature j. SHAP allows you to explain either an individual prediction of the model (local), or the model as a whole (global).

LIME [10] allows you to provide an explanation for an individual prediction. It accomplishes this by creating a simple model of the individual prediction data points. For example, in the churn case studies [3], [6], it identified the type of contract, customer tenure, and usage pattern as the driving features of churn. Though a model may capture these features, that model doesn't necessarily prove causation of churn.

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

IX. LITERATURE SURVEY

The studies reviewed in this paper are summarized in Table IV. In [1], Manzoor et al. summarized existing research on ML based churn prediction covering preprocessing, feature selection, and model training. Imani et al. [2] reviewed 240 studies published between 2020 and 2024 and identified the increase in the usage of hybrid/ensemble machine learning and deep learning approaches like XGBoost and LSTM. Asif et al. [3] addressed the problem of explainable churn prediction in telecom and El Attar and El-Hajj [6] surveyed the literature on the usage of ensemble models for churn prediction using SHAP to analyze features. Mamun [5] did a PRISMA based review of the 112 studies on ML for customer retention. Kakulapati et al. [7] used a combination of Random Forest, XGBoost, SHAP, and Bee Colony Optimization to solve the problem of churn prediction in retail.

TABLE IV. SUMMARY OF REVIEWED STUDIES

Study	Methodology	Key insight/limitation
Manzoor et al. (2024) [1]	Systematic Review	Preprocessing, Feature selection, Model training, Review doesn't suggest or test a new model.
Imani et al. (2025) [2]	Systematic Review	Increasing use of XGBoost, LSTM and hybrid approaches, reviewed work concentrated on telecom and retail industries.
Asif et al. (2025) [3]	Explainable AI for Churn Prediction	Predictive models are analyzed for features, research has concentrated on the telecom industry.
Mamun (2025) [5]	PRISMA based Review	Survey of predictive models, methodology of the research papers surveyed has varied significantly.
El Attar & El-Hajj (2026) [6]	Ensemble Methods	Analysis of model features, survey uses only telecom data.
Kakulapati et al. (2025) [7]	Random Forest, XGBoost and Bee Colony Optimization (Retail)	Combination of Optimization and Explainable methods, increases complexity of model and lacks generalization to non retail datasets.

From this review it is evident that, i) Machine learning has been a well researched area for predicting customer churn.

X. USE OF CUSTOMER ANALYTICS BY DOMAIN

Table V. provides a few typical examples of how customer analytics can be applied across several sectors.

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

TABLE V. USE OF CUSTOMER ANALYTICS BY DOMAIN

Domain	Sample Applications
E-Commerce	Product recommendation, personalization and buying behaviour based customer segmentation
Banking	Churn Prediction, customer segmentation, and marketing targeted product offers
Telecom	Churn prediction from data on usage, tenure and contracts. Designing retention offers.
Retail	Demand forecasting, analyzing customer reviews, assessing customer satisfaction, and developing loyalty programs.

All the sectors highlighted have commonalities. Key uses of customer analytics include customer segmentation, churn prediction, personalized marketing, sales forecasting, customer retention and customer satisfaction analysis.

XI. RESEARCH GAPS

Data privacy and protection are important due to the nature of the information being collected. In the past few years, there has been more and more data being collected by businesses from customers, which makes it easier for the data to be fragmented. Data silos often happen when data is collected from multiple disparate systems and becomes challenging to link or unify. The models are built on a training dataset that can easily become skewed. In these situations, the models tend to have biases that are difficult to correct. In the same way, businesses often struggle to gain insights from data that is not timely. Customers of a business do not generate enough data that can be used for personalization and targeted ads immediately after their first interaction. This is referred to as the "cold start" problem. Some research studies fail to generalize from the dataset of one industry to other industries. Reviews and surveys provide feedback that the company cannot infer from transaction data. This survey analyzes some of the challenges and gaps in current research, but does not present a proposed solution.

XII. TRENDS OF THE FUTURE

- Customer information-sensitive analytics which provide the data analysis you require while still safeguarding customer privacy.
- Integrating transactions and browsing and feedback from multiple customer sources.
- Analyitics to process customer activity and present the data as insights as it happens.
- Building systems to ensure models provide analysis which is fair and without bias.
- The reason behind customer recommendation must be evident.
- Incorporating review data and transactional data for customer analysis.

BODHIVRUKSHA JOURNAL OF DIVERSE DISCIPLINE
E-ISSN : 3139-1486
An Open Access, Peer-Reviewed, Multidisciplinary Scholarly Journal
Bi-monthly | Multilingual (English, Hindi, Marathi) | Academic Research
@ Volume 2, Issue 5 (September-October 2026) (Special Issue 1) ISSN: 3139-1486 (Online)

XIII. SUMMARY

Customer analytics with data science was described from the basics of data preparation and preprocessing, to the creation of customer segmentation, the prediction of churn and the development of recommender systems and model interpretability. The surveyed research suggests that ensembles and methods for Explainable AI (XAI) such as SHAP and LIME improve the predictive accuracy and interpretability of models, respectively. Performance is heavily reliant on the data quality, and many studies are constrained to specific domains. There is still a long road ahead to resolve issues related to the trade-offs of privacy, bias, the "cold start" problem, and model updates in the real world. Research should be focused on making the predictive analytics pipeline more trustworthy. Since the present study was limited to a survey of the literature, future research will entail practical deployment and analysis of a real dataset.

REFERENCES

[1] A. Manzoor, M. A. Qureshi, E. Kidney, L. Longo, "Machine Learning Techniques to Predict Customer Churn and Tips for Business People," IEEE Access, 2024.

[2] "Recent Trends, Research, and Challenges in Customer Churn Prediction and Deep Learning Techniques," Machine Learning and Knowledge Extraction, vol. 7, no. 3, 2025.

[3] "A Study on the Explainability of Artificial Intelligence (XAI) and Its Role in Customer Churn Prediction, Clustering, and Retention," International Journal of Data Science and Analytics, Springer Nature, 2025.

[4] "Retaining Customers Through Machine Learning: Ensemble Techniques to Improve Churn Prediction," ScienceDirect, 2025.

[5] M. N. H. Mamun, "Latest Research on the Use of Predictive Models and Churn Analysis in Customer Retention," SSRN, 2025.

[6] El Attar and El-Hajj, "Multiple Ensemble Methods Using SHAP Feature Interpretation for Telecom Churn," 2026. [To be updated.]

[7] Kakulapati, Hasmita Reddy, Shivputra Dhabale, "Using Random Forest and XGBoost in conjunction with SHAP and Bee Colony Algorithm for Customer Churn Prediction in Retail," 2025. [To be updated.]

[8] T. Chen, C. Guestrin, "XGBoost: Scalable Tree-Based Gradient Boosting Systems," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016): 785-794.

[9] S. M. Lundberg and S.-I. Lee, "A Unified Framework for the Interpretation of Model Predictions," NIPS, Vol 30 (2017).

[10] M. T. Ribeiro, S. Singh and C. Guestrin, "Why Should I Trust You?: Explaining the Predictions of Any Classifier," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.