International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 2, Mar-Apr-2026, ISSN (Online): 2395-566X

Comparative Study of Statistical Models for Customer Churn Classification

Jyoti Gupta, Ayush Patel, Siddharth Prabhudesai, Rahul Neve
Dept. of Information Technology Thakur College of Engineering & Technology University of Mumbai.

Abstract- Customer churn prediction plays a vital role in helping businesses retain customers and minimize revenue loss in competitive markets. This study focuses on developing a predictive framework to identify customers who are likely to discontinue a service based on historical data. The dataset used in this project consists of customer demographic, behavioral, and financial attributes, which are preprocessed and transformed through feature engineering techniques to improve model performance. Multiple machine learning classification models are implemented and evaluated to determine their effectiveness in predicting churn. To address the issue of class imbalance, appropriate techniques are applied to ensure fair model training. The models are assessed using key performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC, providing a comprehensive comparison of their predictive capabilities. The analysis highlights the importance of factors such as customer tenure, service usage patterns, and billing characteristics in influencing churn behavior. The results demonstrate that machine learning models can effectively capture underlying patterns in customer data and provide reliable predictions. This study offers valuable insights into churn prediction and presents a data-driven approach that can support businesses in designing targeted customer retention strategies.

Keywords- Customer Churn, Machine Learning, Classification, Predictive Modeling, Data Preprocessing, Feature Engineering, Class Imbalance, SMOTE, Customer Retention, Data Analytics.

I. INTRODUCTION

In today's highly competitive business environment, organizations are increasingly focused on retaining their existing customers, as customer acquisition is often more costly and resource-intensive than customer retention. Customer churn, which refers to the loss of customers over a given period, poses a significant challenge for businesses across various industries. Understanding and predicting churn behavior has therefore become essential for maintaining profitability and sustaining long-term growth.

With the rapid growth of data availability, companies now collect extensive information about their customers, including demographic details, service usage patterns, and transaction histories. This has opened the door to data-driven decision-making, where machine learning techniques can be applied to analyze large datasets and uncover meaningful patterns. By leveraging these techniques, businesses can identify customers who are at a higher risk of leaving and take proactive measures to improve retention through targeted strategies and personalized interventions.

However, predicting customer churn is not a straightforward task, as it involves complex relationships between multiple variables and often includes challenges such as class imbalance and noisy data. Different machine learning models vary in their ability to capture these patterns, making it important to evaluate and compare their performance in a systematic manner. Selecting the most suitable model requires careful consideration of both predictive accuracy and practical applicability.

This study aims to develop and evaluate multiple classification models for customer churn prediction using a structured dataset containing customer-related attributes. The research focuses on preprocessing the data, handling class imbalance, and assessing model performance using standard evaluation metrics. By analyzing the results, the study seeks to identify key factors influencing churn and demonstrate how predictive modeling can support businesses in making informed, data-driven decisions to enhance customer retention.

II. LITERATURE SURVEY

Customer churn prediction has been a widely studied problem in the fields of data science and business analytics due to its direct impact on customer retention and organizational profitability. Early approaches to churn prediction primarily relied on traditional statistical methods, with Logistic Regression being one of the most commonly used techniques. Its popularity stems from its simplicity, interpretability, and ability to provide probabilistic outputs. However, as datasets became larger and more complex, the limitations of linear models in capturing non-linear relationships became increasingly evident.

To address these limitations, machine learning techniques such as Decision Trees and ensemble methods gained prominence. Decision Trees offer a rule-based structure that is easy to interpret and can capture non-linear patterns in data. However, they are prone to overfitting, especially when dealing with high-dimensional datasets. This led to the development of ensemble methods such as Random Forest, which combines multiple decision trees to improve generalization and reduce variance. Random Forest has been widely recognized for its robustness and strong performance in classification tasks, including churn prediction.

Another important model used in classification problems is the Support Vector Machine (SVM), which is particularly effective in handling high-dimensional data. SVM works by finding an optimal hyperplane that maximizes the margin between different classes. Although it can provide high accuracy, its performance is sensitive to the choice of kernel and parameters, and it may require significant computational resources for large datasets.

A major challenge in churn prediction is class imbalance, where the number of non-churn customers significantly exceeds churn customers. This imbalance can bias models toward the majority class, reducing their ability to correctly identify churn cases. Techniques such as the Synthetic Minority Over-sampling Technique (SMOTE) have been introduced to address this issue by generating synthetic samples for the minority class, thereby improving model performance.

In recent years, research has emphasized the importance of feature engineering and data preprocessing in improving predictive performance. Studies have shown that variables such as customer tenure, billing patterns, and service usage behavior are strong indicators of churn. Additionally, the use of multiple evaluation metrics, including precision, recall, F1-score, and ROC-AUC, has been recommended to provide a comprehensive assessment of model effectiveness, particularly in imbalanced datasets.

Overall, existing literature highlights that no single model universally outperforms others in all scenarios. Instead, the effectiveness of a model depends on the nature of the dataset, feature representation, and evaluation criteria. This motivates the need for a comparative analysis of multiple models under consistent conditions, which forms the basis of this study.

III. METHODOLOGY

This study follows a systematic approach for developing and evaluating machine learning models for customer churn prediction. The methodology involves data preprocessing, feature transformation, handling class imbalance, model training, and evaluation using statistical metrics.

1) Problem Formulation

Customer churn prediction is formulated as a binary classification problem, where the target variable Ytakes two possible values:

Y=1→ Customer churned

Y=0→ Customer retained

Given a feature vector X=(x_1,x_2,...,x_n), the objective is to learn a function:

f(X)→Y

that predicts the probability of a customer churning.

2) Data Preprocessing

The dataset is preprocessed to ensure quality and consistency before model training.

- Missing values are handled using appropriate imputation techniques.
- Categorical variables are encoded into numerical form using one-hot encoding.
- Numerical features are scaled using standardization:

$$X_{scaled} = \frac{X - \mu}{\sigma}$$

where $\mu$is the mean and $\sigma$is the standard deviation.

3) Handling Class Imbalance

Since churn datasets are typically imbalanced, the Synthetic Minority Over-sampling Technique (SMOTE) is used.

SMOTE generates synthetic samples using nearest neighbors:

$$x_{new} = x_i + \lambda(x_{nn} - x_i)$$

where:

- x_i= minority class sample
- x_{nn}= nearest neighbor
- $\lambda \in [0,1]$is a random value

4) Model Development

The following classification models are implemented:

A. Logistic Regression

Logistic Regression models the probability of churn using the sigmoid function:

$$P(Y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}}$$

B. Decision Tree

Decision Trees split data based on impurity measures such as Gini Index:

$$Gini = 1 - \sum p_i^2$$

where $p_i$represents class probabilities.

C. Random Forest

Random Forest is an ensemble of decision trees trained using bootstrap sampling. The final prediction is obtained by majority voting:

$$\hat{Y} = \text{mode}(Y_1, Y_2, \ldots, Y_n)$$

D. Naive Bayes

Naive Bayes is a probabilistic classifier based on Bayes' Theorem, assuming independence between features:

$$P(Y \mid X) = \frac{P(X \mid Y) \cdot P(Y)}{P(X)}$$

It calculates the posterior probability of a class given the input features.

E. K-Nearest Neighbors (KNN)

KNN is a non-parametric, instance-based learning algorithm that classifies a data point based on the majority class among its $k$nearest neighbors:

$$\hat{Y} = \text{mode}(Y_1, Y_2, \ldots, Y_k)$$

Distance between data points is commonly calculated using Euclidean distance:

$$d(x, y) = \sqrt{\sum (x_i - y_i)^2}$$

5) Model Evaluation Metrics

To evaluate model performance, the following metrics are used:

Accuracy

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision

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

Recall

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

F1-Score

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

ROC-AUC

ROC curve plots True Positive Rate vs False Positive Rate:

$$TPR = \frac{TP}{TP + FN}, FPR = \frac{FP}{FP + TN}$$

6) Experimental Setup

1. Dataset is split into training and testing sets (e.g., 80:20).
2. SMOTE is applied only on training data.
3. Models are trained and evaluated on unseen test data.

7) Data Visualization and Dashboard Development

Fig 1: System Architecture for Customer Churn Prediction

1. In addition to model development, an interactive dashboard was created to visualize customer churn patterns and key insights.
2. The dashboard was developed using Power BI to provide a user-friendly interface for exploring the dataset.
3. It includes visualizations such as:

- Churn distribution across contract types
- Churn by payment method
- Customer demographics (e.g., gender, age groups)
- Churn trends based on tenure and monthly charges.

4. Filters and slicers are incorporated to allow dynamic analysis of different customer segments.
5. The dashboard helps in identifying high-risk groups and understanding the key factors influencing churn in a more intuitive manner.

IV. RESULTS AND DISCUSSIONS

This section presents the performance evaluation of the machine learning models implemented for customer churn prediction. The models are compared using multiple evaluation metrics, including Accuracy, Precision, Recall, F1-score, to provide a comprehensive assessment of their effectiveness.

1) Model Performance Comparison
The performance results of all models are summarized in Table 1.

Table 1: Performance Comparison of Machine Learning Models

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.72410	0.617998	0.463913	0.529983
Decision Tree	0.68175	0.525368	0.526543	0.525955
Random Forest	0.74240	0.648282	0.506561	0.568726
Naive Bayes	0.72655	0.599357	0.556367	0.577063
K-Nearest Neighbors	0.70540	0.571781	0.483448	0.523917

2) Analysis of Results
From the results, it is observed that the Random Forest model achieves the highest accuracy (0.74240) and ROC-AUC score (0.795209), indicating its strong ability to distinguish between churn and non-churn customers. Its ensemble nature allows it to capture complex patterns in the data more effectively than individual models.
The Naive Bayes model shows a balanced performance, achieving the highest recall (0.556367) among all models. This indicates its effectiveness in identifying actual churn cases, which is particularly important in churn prediction where missing a churner can be costly.

The Logistic Regression model performs reasonably well, providing stable results across all metrics. However, its relatively lower recall suggests limitations in capturing non-linear relationships in the data.

The Decision Tree model shows the lowest performance in terms of accuracy and ROC-AUC, likely due to overfitting and its sensitivity to variations in the dataset.
The K-Nearest Neighbors (KNN) model provides moderate performance but does not outperform ensemble or probabilistic models. Its effectiveness is influenced by the choice of distance metric and the value of k.

A
3) Trade-off Between Precision and Recall
A key observation from the results is the trade-off between precision and recall across different models. While some models achieve higher precision, they may have lower recall, and vice versa. In the context of churn prediction, recall is particularly important as it reflects the model's ability to correctly identify customers who are likely to churn.

4) Insights from Dashboard Visualization
The dashboard analysis provides additional insights into customer behavior. It highlights that customers with lower tenure and higher monthly charges are more likely to churn. Additionally, certain payment methods and contract types show higher churn rates, indicating their influence on customer retention.
These insights complement the model results and provide a better understanding of the underlying factors contributing to churn.

5) Visual Model Comparison
To better understand the differences in model performance, a visual comparison of evaluation metrics is created using bar charts. This visualization highlights that Random Forest consistently performs better across most metrics, while Naive Bayes achieves relatively higher recall.

The graphical representation helps in quickly identifying performance gaps between models and supports easier interpretation compared to numerical tables alone.

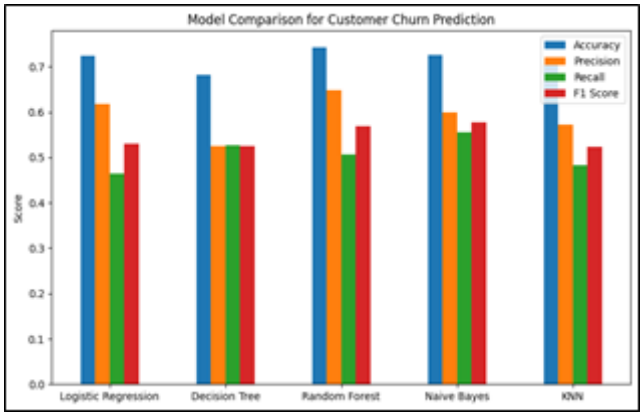

Fig 2: Visual Model Comparison

6) Insights from Tableau Dashboard

An interactive dashboard is developed using Tableau to visualize customer churn patterns and key insights. The dashboard includes multiple charts and filters that allow dynamic exploration of the dataset.

Key observations from the dashboard include:

- Customers with low tenure show a significantly higher churn rate.
- Higher monthly charges are associated with increased churn probability.
- Certain payment methods and contract types exhibit higher churn levels.
- Demographic factors such as age groups also influence churn behavior.

The dashboard enhances interpretability by transforming raw data into meaningful visual insights, enabling better understanding of customer behavior.

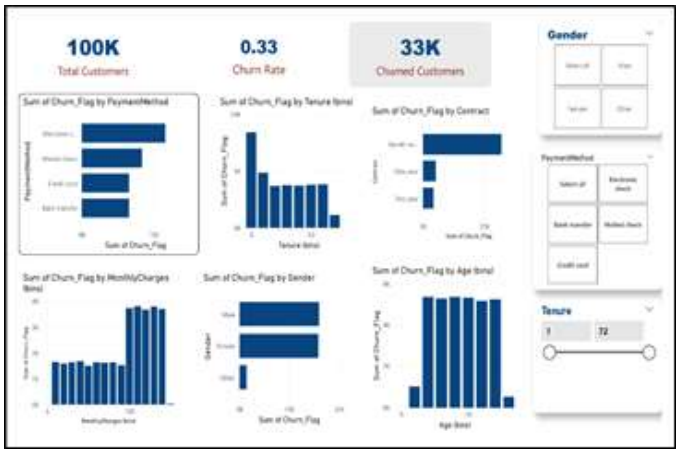

Fig 3: Tableau Dashboard for Customer Churn Prediction

7) Discussion

The results indicate that Random Forest is the best-performing model in terms of overall predictive capability, while Naive Bayes performs well in identifying churn cases due to higher recall. This highlights the importance of selecting models based on business requirements, especially when the cost of missing churn customers is high.

The inclusion of visual tools such as model comparison charts and dashboards significantly improves the interpretability of results. While numerical metrics provide quantitative evaluation, visualizations help in identifying trends, patterns, and actionable insights.

Overall, the study demonstrates that combining machine learning models with data visualization techniques results in a more effective and practical churn prediction system. This integrated approach supports data-driven decision-making and helps organizations design targeted customer retention strategies.

IV. CONCLUSION AND FUTURE WORK

1) Conclusion

This study presented a comprehensive approach to customer churn prediction using multiple machine learning classification models. The objective was to identify customers at risk of churning by analyzing demographic, behavioral, and financial data. A systematic methodology involving data preprocessing, feature engineering, class imbalance handling, and model evaluation was implemented to ensure reliable results.

The performance of five models—Logistic Regression, Decision Tree, Random Forest, Naive Bayes, and K-Nearest Neighbors—was evaluated using multiple metrics such as Accuracy, Precision, Recall, F1-score, and ROC-AUC. Among these, the Random Forest model demonstrated the best overall performance, achieving the highest accuracy and ROC-AUC score, indicating its strong capability in capturing complex patterns within the data. Naive Bayes showed relatively higher recall, making it effective in identifying churn cases, while other models provided moderate performance.

The study also highlighted the importance of using multiple evaluation metrics instead of relying solely on accuracy, especially in imbalanced datasets. Additionally, the integration of a dashboard for data visualization enhanced the

interpretability of results and provided valuable insights into customer behavior. Overall, the research demonstrates that combining machine learning techniques with data visualization tools can significantly improve decision-making in customer retention strategies.

2) Future Work

The following areas can be explored to enhance the current study and improve the performance of the churn prediction system:

1. Implement advanced machine learning techniques such as ensemble boosting methods or deep learning models to improve prediction accuracy.
2. Incorporate time-series or sequential data to analyze customer behavior over time and enable early churn detection.
3. Apply feature selection and dimensionality reduction techniques to improve model efficiency and reduce computational complexity.
4. Perform extensive hyperparameter tuning to further optimize model performance.
5. Explore alternative methods for handling class imbalance, such as different resampling or cost-sensitive learning approaches.
6. Develop and deploy the model in a real-time production environment for continuous churn monitoring.
7. Integrate additional data sources, such as customer feedback or external factors, to enhance prediction capabilities.
8. Improve the dashboard by adding more interactive and advanced visualizations for better decision-making.

Acknowledgement

We would like to express our sincere gratitude to our mentor, Dr. Neha Patwari, for her invaluable guidance and support throughout the course of this project. Her expertise, insights, and continuous encouragement played a significant role in the successful completion of this research work. Her guidance greatly enhanced our understanding of machine learning concepts and their practical application in customer churn prediction. The knowledge and experience shared by her helped us approach the problem more effectively and develop a structured and meaningful solution. We are truly thankful for her mentorship and constant motivation, without which this work would not have been possible.

REFERENCES

1. L. Breiman, “Random Forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
2. T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, New York, NY, USA: Springer, 2009.
3. G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning, New York, NY, USA: Springer, 2013.
4. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.
5. F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
6. C. Cortes and V. Vapnik, “Support-vector networks,” Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.
7. T. Cover and P. Hart, “Nearest neighbor pattern classification,” IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21–27, 1967.
8. H. Zhang, “The optimality of Naive Bayes,” AAAI Conference on Artificial Intelligence, 2004.s
9. I. H. Witten, E. Frank, and M. A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, 3rd ed. San Francisco, CA, USA: Morgan Kaufmann, 2011.
10. J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. Burlington, MA, USA: Elsevier, 2012.
11. W. Verbeke, D. Martens, C. Mues, and B. Baesens, “Building comprehensible customer churn prediction models,” Expert Systems with Applications, vol. 38, no. 3, pp. 2354–2364, 2011.
12. W. Verbeke, D. Martens, and B. Baesens, “Social network analysis for customer churn prediction,” Applied Soft Computing, vol. 14, pp. 431–446, 2014.
13. H. He and E. A. Garcia, “Learning from imbalanced data,” IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263–1284, 2009.
14. J. Brownlee, Machine Learning Mastery With Python, 2016.
15. M. Kuhn and K. Johnson, Applied Predictive Modeling, New York, NY, USA: Springer, 2013.
16. L. Rokach, “Ensemble-based classifiers,” Artificial Intelligence Review, vol. 33, no. 1, pp. 1–39, 2010.
17. D. J. Hand and W. E. Henley, “Statistical classification methods in consumer credit scoring,” Journal of the Royal Statistical Society, vol. 160, no. 3, pp. 523–541, 1997.

18. S. Moro, P. Cortez, and P. Rita, “A data-driven approach to predict the success of bank telemarketing,” Decision Support Systems, vol. 62, pp. 22–31, 2014.
19. T. Fawcett, “An introduction to ROC analysis,” Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874, 2006.
20. D. Dua and C. Graff, “UCI Machine Learning Repository,” University of California, Irvine, 2019. [Online]. Available: http://archive.ics.uci.edu/ml
21. Tableau Software, “Tableau Desktop,” [Online]. Available: https://www.tableau.com
22. D. Talukar, “Telco Customer Churn Dataset,” Kaggle, 2020. [Online]. Available: https://www.kaggle.com/datasets/dhrubangtalukar/telco-customer-churn-data