Journal of Predictive Analytics and Intelligent Data
Volume 2, Issue 2, May-August 2026
ISSN: 3139-2210 (Online)

Predictive Modeling and Risk Scoring for Bank Customer Churn

Gouri Nandkumar Salgar

ABSTRACT

Customer churn is a significant challenge for retail banking institutions, as customer loss can adversely affect revenue stability, customer lifetime value, and opportunities for cross-selling and upselling. This study presents a machine-learning-based customer churn prediction and risk-scoring system using a European bank customer dataset comprising 10,000 records. The proposed workflow includes data preprocessing, exploratory data analysis, feature engineering, categorical encoding, model development, evaluation, and deployment through an interactive Streamlit application. Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting were evaluated using Accuracy, Precision, Recall, F1-score, and ROC-AUC. Gradient Boosting achieved the best performance, with 87.10% accuracy and 86.69% ROC-AUC. Feature-importance analysis identified Age, Num Of Products, Is Active Member, Balance, and Geography_Germany as influential features. The deployed application generates customer-level churn probabilities and risk categories to support proactive, data-driven customer retention strategies.

KEYWORDS: *Customer churn, Machine learning, Predictive analytics, Gradient boosting, Banking analytics.*

INTRODUCTION

Customer retention is an important factor in the financial services industry. Banks invest significant resources in acquiring and maintaining customers, making customer churn a major business concern. When customers leave a bank, the institution may lose recurring revenue, customer lifetime value, cross-selling opportunities, and long-term business relationships.

Traditional churn analysis generally examines historical customer behavior after churn has occurred. Although such analysis provides useful insights, it provides limited opportunity for intervention. Predictive modeling addresses this limitation by estimating the probability that an individual customer will churn before the event occurs. Machine learning can analyze multiple customer attributes simultaneously and identify nonlinear patterns associated with churn.

This work develops a predictive churn intelligence system that generates customer-level churn probabilities, classifies customers into risk categories, identifies influential predictive features, and provides an interactive interface for scenario analysis.

PROBLEM STATEMENT AND OBJECTIVES

Despite having access to extensive customer-level data, banks may lack effective systems for identifying customers at high risk of leaving. This can result in reactive retention activities, broad customer targeting, and inefficient use of marketing resources.

The primary objectives are to develop an accurate churn prediction model, generate individual churn probabilities, identify important churn drivers, compare multiple machine-learning algorithms, and integrate the selected model into an interactive application. Secondary objectives include improving the identification of potential churners, reducing unnecessary retention interventions, improving interpretability, and enabling scenario-based risk analysis.

DATASET AND PREPROCESSING

The study uses a European bank customer dataset containing 10,000 records and 14 original attributes: Year, Customer ID, and Surname were excluded from predictive modeling because they were not used as predictive input variables. Customer ID and Surname are identifiers, while Year did not contribute to the modeling workflow. The target variable is exited, where 0 represents a retained customer and 1 represents a churned customer.

MANTECH Publications
Journal of Predictive Analytics and Intelligent Data
Volume 2, Issue 2, May-August 2026
ISSN: 3139-2210 (Online)

Data inspection was performed to examine data types, missing values, and feature characteristics. CustomerId and Surname were excluded from predictive modeling because they are identification attributes rather than meaningful behavioral predictors. Geography and Gender were encoded using one-hot encoding. The resulting model input contained 15 features after feature engineering and categorical encoding.

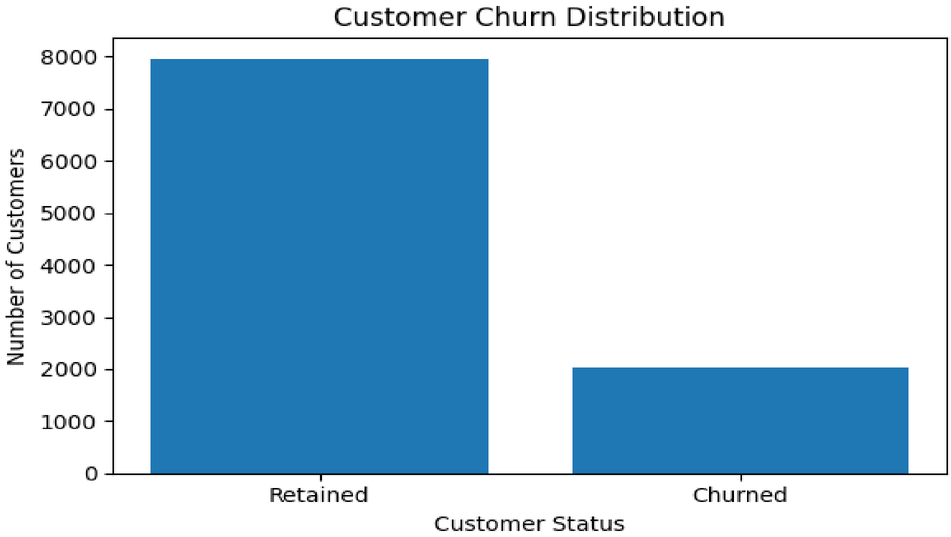

Figure 1: Customer churn distribution.

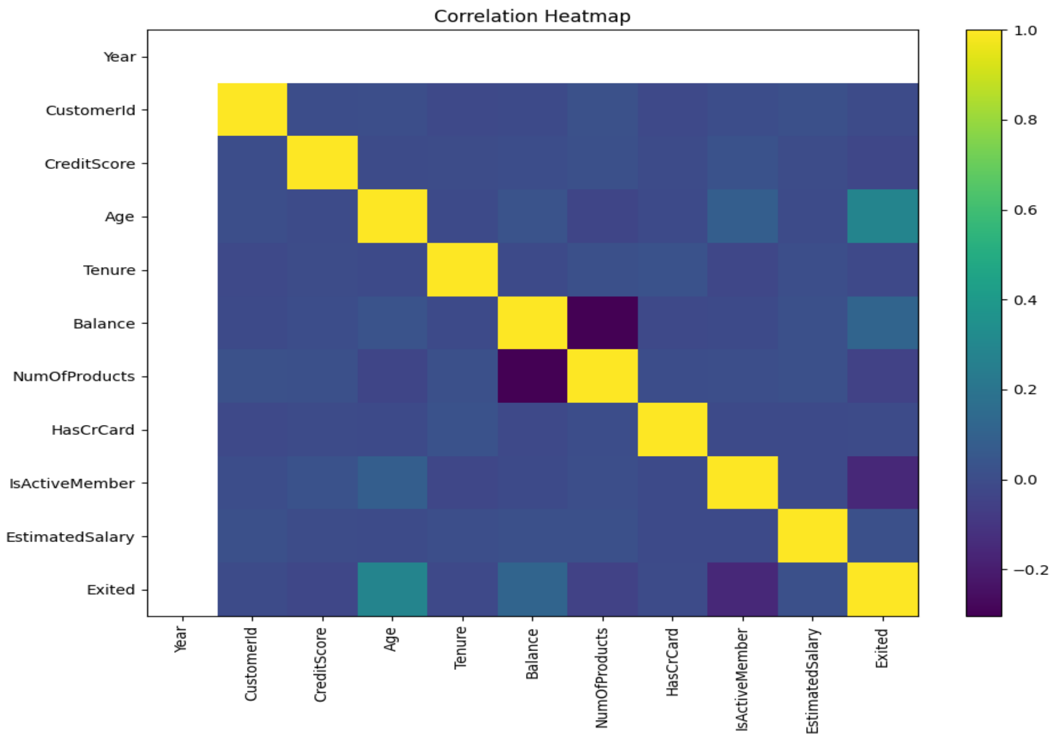

Figure 2: Correlation heatmap of numerical features.

FEATURE ENGINEERING

Four derived variables were created to represent customer relationships and engagement. Balance Salary Ratio was calculated as Balance divided by Estimated Salary plus one. Product Density was calculated as NumOfProducts divided by Tenure plus one. Age Tenure was calculated as the product of Age and Tenure. Engagement Score combined Has Cr Card, Is Active Member, and Num of Products. These features were designed to capture financial position, product utilization, relationship duration, and engagement.

METHODOLOGY

A stratified train-test strategy was used to preserve the class distribution of the target variable. The dataset was divided into training and testing subsets using an 80:20 stratified split with random state = 42, resulting in 8,000 training records and 2,000 testing records. Four algorithms were evaluated. Logistic Regression was used as an interpretable baseline. Decision Tree was used to model nonlinear relationships. Random Forest provided an ensemble tree-based benchmark. Gradient Boosting was evaluated as an advanced ensemble method.

The models were evaluated using Accuracy, Precision, Recall, F1-score, and ROC-AUC. Accuracy measures overall correctness, Precision measures the proportion of predicted churners that were actually churners, Recall measures the proportion of actual churners identified by the model, F1-score balances Precision and Recall, and ROC-AUC measures discrimination across classification thresholds.

EXPERIMENTAL RESULTS

Table: 1

Model	Accuracy	Precision	Recall	F1	ROC-AUC
Gradient Boosting	87.10%	79.45%	49.39%	60.91%	86.69%
Random Forest	86.50%	78.19%	46.68%	58.46%	85.35%

Logistic Regression	80.85%	59.52%	18.43%	28.14%	77.42%
Decision Tree	78.05%	46.28%	48.89%	47.55%	67.20%

Gradient Boosting produced the strongest results across all reported metrics. Its ROC-AUC of 86.69% indicates strong discrimination between churned and retained customers. Its precision of 79.45% indicates that the majority of customers flagged as churners were actual churners in the test set. Recall was 49.39%, indicating that further threshold optimization could be considered when the business objective prioritizes capturing more actual churners.

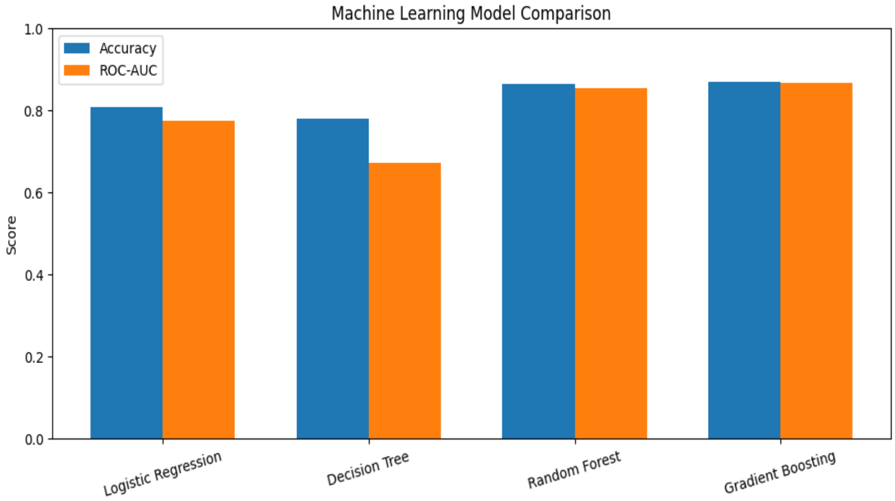

Figure 3: Comparison of machine learning models using Accuracy and ROC-AUC.

MODEL EXPLAINABILITY

Feature importance was extracted from the final Gradient Boosting model. Age was the most influential feature with an importance of 0.381661, followed by NumOfProducts at 0.289727 and IsActiveMember at 0.098882. Balance, Geography_Germany, EngagementScore, and BalanceSalaryRatio contributed 0.062397, 0.052418, 0.030995, and 0.025045, respectively.

Table: 2

Rank	Feature	Importance
1	Age	0.381661
2	Num Of Products	0.289727

3	IsActiveMember	0.098882
4	Balance	0.062397
5	Geography_Germany	0.052418
6	EngagementScore	0.030995
7	BalanceSalaryRatio	0.025045
8	CreditScore	0.017139
9	EstimatedSalary	0.014191
10	Gender_Male	0.013554

The feature-importance values indicate the relative contribution of each variable to the model's predictions; they do not establish causality or indicate whether a higher feature value increases or decreases churn. Directional interpretation would require methods such as SHAP or partial-dependence analysis.

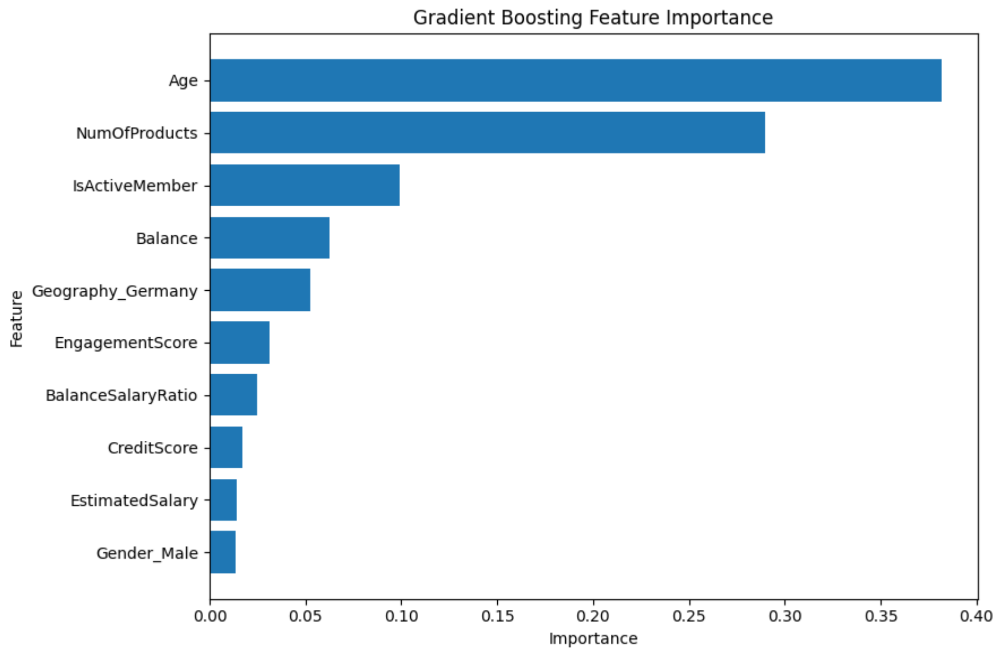

Figure 4: Feature importance obtained from the Gradient Boosting model.

CHURN RISK SCORING

The final model generates a probability between 0 and 1 for each customer. This probability is converted to a percentage and classified into three operational risk levels: Low Risk for probabilities below 0.30, Medium Risk for probabilities from 0.30 to below 0.70, and High

MANTECH Publications
Journal of Predictive Analytics and Intelligent Data
Volume 2, Issue 2, May-August 2026
ISSN: 3139-2210 (Online)

Risk for probabilities of 0.70 or greater. These thresholds are initial operational thresholds and can be optimized according to business costs and retention objectives.

STREAMLIT APPLICATION

The trained Gradient Boosting model was integrated into an interactive Stream lit application. Users can enter customer characteristics including Credit Score, Age, Tenure, Balance, Num of Products, Has Cr Card, Is Active Member, Estimated Salary, Geography, and Gender. The application performs the same feature engineering and encoding used during model development and returns churn probability, predicted class, risk category, and feature-importance information. The application therefore converts the predictive model into a practical decision-support interface for customer retention analysis.

The deployed application is available at:

https://bank-customer-churn-prediction-dm2w8rw8gfawkc27bgxzyd.streamlit.app/

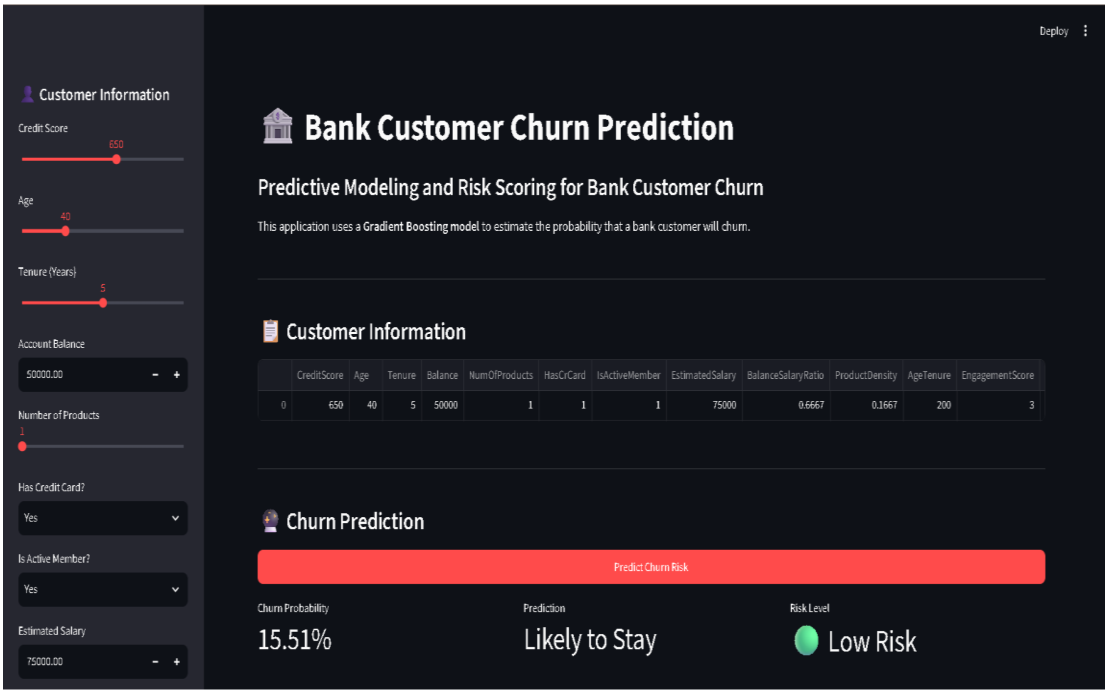

Figure 5: Customer-level churn probability and risk classification.

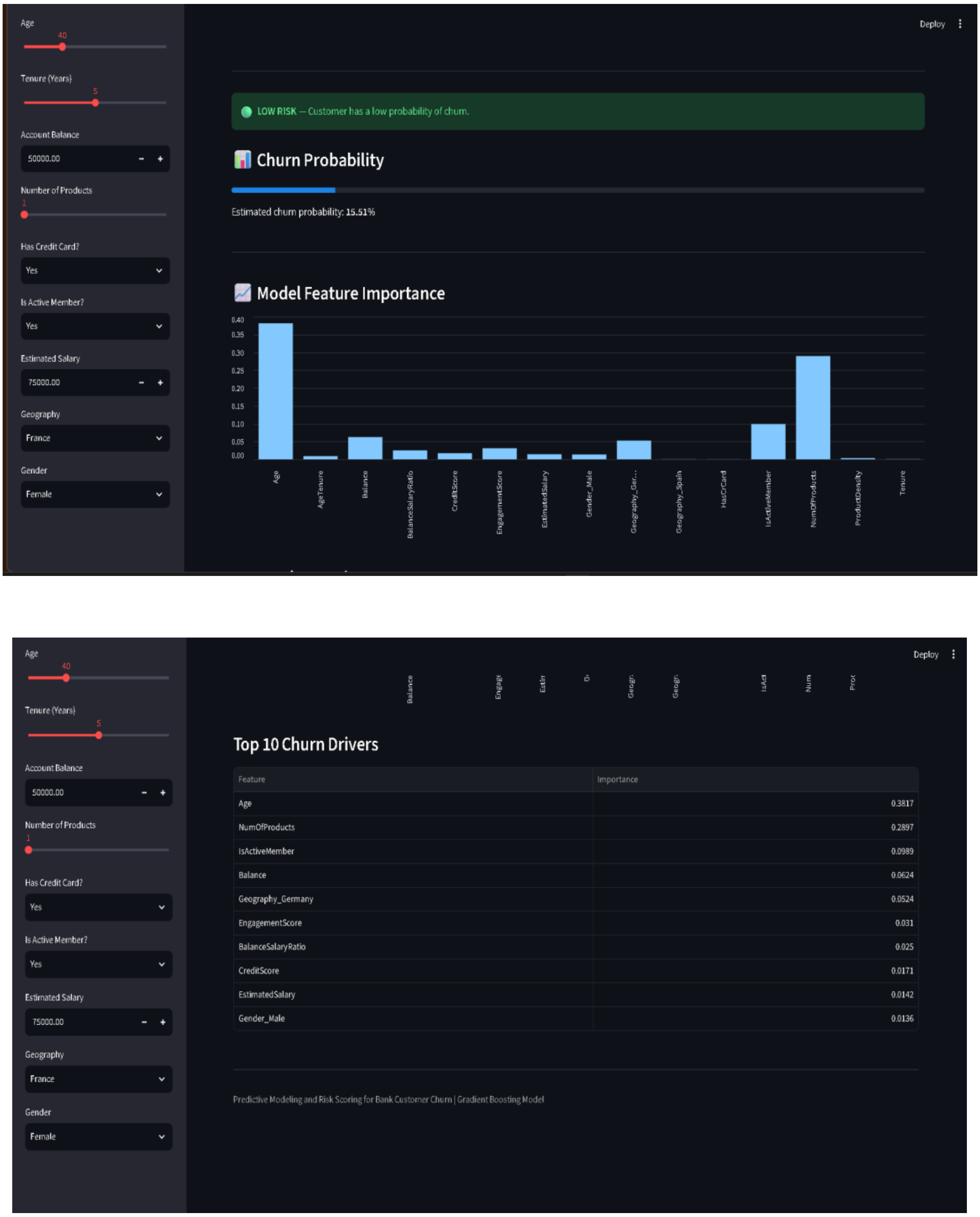

Figure 6: Streamlit-based bank customer churn prediction dashboard.

BUSINESS IMPLICATIONS

The model can support proactive retention by prioritizing customers according to predicted risk. High-risk customers can be considered for personalized engagement, product reviews, service interventions, or targeted retention offers. The relatively high importance of IsActiveMember suggests that engagement should be monitored, while the importance of

NumOfProducts indicates that product relationships deserve closer analysis. Age-specific strategies may also be considered because Age was the strongest feature in the trained model.

LIMITATIONS AND FUTURE WORK

The study is subject to several limitations. The dataset contains a limited set of customer attributes and may not represent all banking populations. Feature importance describes model behavior rather than causal relationships. The current risk thresholds are predefined and have not been optimized using financial cost functions. SHAP-based explanations were not included in the current implementation because the required package was unavailable in the development environment.

Future work can include SHAP-based individual explanations, hyperparameter optimization, cross-validation, probability calibration, cost-sensitive learning, XGBoost comparison, real-time monitoring, CRM integration, and cloud deployment. Threshold optimization based on the relative cost of missed churners and unnecessary interventions would also improve business usefulness.

CONCLUSION

This paper presented a predictive customer churn intelligence system for banking applications. The workflow combines preprocessing, feature engineering, machine-learning model comparison, risk scoring, feature-importance analysis, and Stream lit deployment. Among the evaluated algorithms, Gradient Boosting achieved the best performance, with 87.10% accuracy and 86.69% ROC-AUC. Age, Num Of Products, Is Active Member, Balance, and Geography_Germany were among the most influential predictive features. The resulting risk-scoring application provides a practical mechanism for identifying customers who may require proactive retention attention. The approach demonstrates how predictive analytics can transform churn management from a reactive activity into a data-driven decision-support process.

REFERENCES

1. L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.
2. C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.
3. J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, vol.29, no. 5, pp. 1189–1232, 2001.
4. T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. New York, NY, USA: Springer, 2009.
5. T. M. Mitchell, *Machine Learning*. New York, NY, USA: McGraw-Hill, 1997.

***Author for Correspondence**

Gouri Nandkumar Salgar

E-mail: gourisalgar56@gmail.com

Student, Department of Data Science, D. Y. Patil College of Engineering and Technology, Kolahpur.

Received Date: August 28, 2026

Accepted Date: September 02, 2026

Published Date: September 04, 2026

Citation: Gouri Nandkumar Salgar**.** Predictive Modeling and Risk Scoring for Bank Customer Churn. Journal of Predictive Analytics and Intelligent Data. 2026; 2(2): 59-68p.