A Project Report

On

CHURN ANALYSIS IN BANKS

BY

KUSH DESAI

2021B4A73158H

f20213158@hyderabad.bits-pilani.ac.in

Under the supervision of

PROF. ARUNA MALAPATI

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE PILANI (RAJASTHAN)

HYDERABAD CAMPUS

(APRIL 2024)

Abstract

Churn prediction means detecting which customers are likely to leave a service or cancel a subscription to a service. Customer churn is essential because, most of the time, acquiring new customers costs more than selling to existing customers. This paper addresses the significance and challenge of accurate churn modeling in the context of decision-making for businesses. We employ three widely used classifiers, decision tree, random forest, and logistic regression, to analyze a dataset encompassing diverse features relevant to churn prediction. Through extensive experimentation, we compare the performance of these classifiers in terms of precision, accuracy, recall, and f1-score. Our results reveal the efficacy of random forest in capturing complex relationships within the data, yielding superior predictive performance compared to decision trees and logistic regression classifiers coming last. These findings contribute to advancing churn prediction methodologies and offer valuable insights for businesses seeking to optimize customer retention strategies.

Keywords: Churn modeling, Classification, Decision tree, Random forest, Logistic regression.

1. Introduction

Customer churn happens when customers cease their relationship with a company, and it is a significant challenge across various industries, including banking, telecommunications, and e-commerce. Predicting churn is critical for businesses as it directly impacts revenue and profitability. Customer attrition in the banking industry results from clients closing their accounts or ceasing to use a specific bank's services. To keep their finances stable and reputation intact, banks must effectively comprehend and manage customer attrition. Customer attrition can have a substantial financial impact on banks, potentially leading to a loss of income for various banking services. As such, building and maintaining long-term client relationships is extremely beneficial to banks. Banks can identify clients in danger of leaving and implement retention measures by receiving information about attrition patterns. This strategy increases the average lifetime value of customers and supports bank profitability. The primary objective of this research is to develop accurate churn prediction models for a bank using machine learning techniques and to compare the performance of decision tree, random forest, and logistic regression classifiers.

Churn prediction is intriguing and important due to its potential to guide strategic decision-making for customer retention. Understanding the factors influencing churn can enable banks to implement targeted interventions, thereby minimizing customer attrition and maximizing revenue. Customer churn is essential because, most of the time, acquiring new customers costs more than selling to existing customers. This is the metric that determines the success or failure of a business. Successful customer retention increases the customer’s average lifetime value, making all future sales more valuable and improving unit margins.

However, accurately predicting churn is inherently challenging due to the complex and multifaceted nature of customer behavior. Its complexity arises from the interplay of numerous factors influencing customer behavior. Naive approaches often fail to capture the intricate relationships between various customer attributes and churn, resulting in suboptimal predictive performance.

Prior attempts at churn prediction have primarily relied on traditional statistical methods or simplistic machine learning algorithms. However, these approaches often struggle to handle large and diverse datasets effectively, leading to limited predictive accuracy and interpretability. In contrast, this research utilizes decision tree, random forest, and logistic regression classifiers, to develop robust churn prediction models. By leveraging the flexibility and scalability of these algorithms, we aim to overcome the limitations of previous approaches and achieve superior predictive performance.

The key components of our approach involve preprocessing the dataset to handle outliers, feature selection to identify the most relevant predictors of churn, normalization of features, and model evaluation using appropriate evaluation metrics such as accuracy, precision, recall, and F1-score. We also conduct extensive experiments to compare the performance of decision tree, random forest, and logistic regression classifiers on the dataset. The results of our study provide valuable insights into the effectiveness of different machine learning techniques for churn prediction and offer practical implications for businesses seeking to improve customer retention strategies.

2. Related Work

The analyses of two research papers show churn prediction in the banking sector using various classification techniques and feature selection methods. The first study[3] compares supervised classifiers such as random forest, artificial neural networks (ANN), and decision trees, emphasizing their performance and interpretability. It identifies key features like age, number of products, and customer activity level as significant predictors of churn, providing practical insights for banks to prioritize retention strategies. However, it lacks statistical significance inference and generalization beyond European banks. Conversely, the second study[4] delves into classifier performance with oversampling and feature selection methods like MRMR and Relief. It highlights the effectiveness of oversampling in mitigating data imbalance for tree classifiers but notes its adverse impact on SVM. While feature selection enhances KNN accuracy, it slightly reduces performance for DT and RF. Both studies contribute valuable insights into churn prediction, but gaps remain in statistical inference, model generalization, and the scalability of classifiers, particularly in handling large datasets. Additionally, the studies could benefit from exploring the impact of external factors like economic conditions and regulatory changes on churn behavior. However, they offer a solid foundation for further research to refine churn prediction models and optimize customer retention strategies in the banking industry.

3. Methodology

The problem addressed in this research is churn prediction, aiming to develop accurate models to forecast customer churn in banks. To solve this problem, we require a dataset containing historical customer data encompassing various attributes such as customer information, purchasing behavior, service usage, and interaction history. This information can be obtained from the bank's database.

For this study, we utilize a publicly available dataset[2] on Kaggle from the banking industry, commonly used for churn prediction research. The dataset consists of anonymized customer information, including demographic details, service subscriptions, account-related details, geographic data, and churn status. Its properties include categorical and numerical features, outliers, and class imbalance, reflecting real-world challenges in churn prediction tasks.

The selected approach for churn prediction involves employing three machine learning classifiers: decision tree, random forest, and logistic regression. We chose these algorithms due to their versatility, scalability, and interpretability, making them suitable for handling diverse datasets and providing insights into the factors driving churn. Additionally, these classifiers have been widely used in various projects for churn prediction research and have demonstrated competitive performance in previous studies.

The methodology involves several steps: data preprocessing, feature selection, model training, and evaluation. In the data preprocessing stage, we handle outliers, binarize categorical features, normalize some numerical attributes, and apply sampling techniques to ensure data quality and consistency. Feature selection techniques such as information gain or feature elimination are then applied to identify the most relevant predictors of churn. Subsequently, the selected features are used to train decision tree, random forest, and logistic regression models using appropriate algorithms. Finally, the trained models are evaluated using performance metrics such as accuracy, precision, recall, and F1-score to assess their predictive performance and generalization ability.

This approach offers several advantages, including a comprehensive analysis of multiple classifiers tailored to the banking industry, feature selection to enhance model interpretability and performance, and thorough evaluation using diverse performance metrics. By leveraging these techniques on banking data, we aim to develop robust churn prediction models that enable banks to actively address customer attrition and optimize customer retention strategies effectively.

4. Experiments

4.1 Dataset

The dataset is the details of the customers in a company.
The initial dataset has the following attributes: CustomerId, Surname, CreditScore, Geography, Gender, Age, Tenure, Balance, NumOfProducts, HasCrCard, IsActiveMember, EstimatedSalary, Exited.

EDA was performed to understand dataset variables, clean the data, and analyze relationships between variables. The data preprocessing included removing redundant features like 'CustomerId' and 'Surname', normalization of the 'Age' attribute, binarization of categorical attributes 'Geography' and 'Gender', feature Selection based on the correlation matrix, and sampling the dataset using the oversampling technique. Then, from sklearn's model_selection, I used the train_test_split function to split the dataset into 70% for training and 30% for testing.

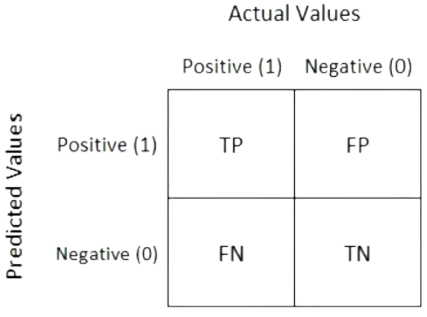
The final pre-processed data had the following attributes: CreditScore, Age, Tenure, Balance, NumOfProducts, HasCrCard, IsActiveMember, EstimatedSalary, Exited, Geography_France, Geography_Germany, Geography_Spain, Gender_Female, Gender_Male.
The shape of X_train, X_test, y_train, y_test are (10235, 12), (4387, 12), (10235,1), (4387,1), respectively, where X is all the final attributes except the target variable, and Y is the target variable.
Correlation Matrix:

	CreditScore	Geography_France	Geography_Germany	Geography_Spain	Gender_Female	Gender_Male	Age	Tenure	Balance	NumOfProducts	HasCrCard	EstimatedSalary
CreditScore	1.000000	0.007716	0.000095	0.009040	0.006918	0.006918	0.005389	0.032429	0.013483	0.003089	0.000321	0.029473
Geography_France	0.007716	1.000000	0.617386	0.510035	0.014259	0.014259	0.053332	0.004171	0.259010	0.004913	0.006247	0.005010
Geography_Germany	0.000095	0.617386	1.000000	0.361761	0.028860	0.028860	0.076789	0.008652	0.414827	0.018095	0.023111	0.004235
Geography_Spain	0.009040	0.510035	0.361761	1.000000	0.014658	0.014658	0.020759	0.004517	0.146629	0.013963	0.017867	0.010568
Gender_Female	0.006918	0.014259	0.028860	0.014658	1.000000	1.000000	0.049077	0.016005	0.021431	0.020988	0.001364	0.032355
Gender_Male	0.006918	0.014259	0.028860	0.014658	1.000000	1.000000	0.049077	0.016005	0.021431	0.020988	0.001364	0.032355
Age	0.005389	0.053332	0.076789	0.020759	0.049077	0.049077	1.000000	0.026877	0.032190	0.013652	0.003963	0.004625
Tenure	0.032429	0.004171	0.008652	0.004517	0.016005	0.016005	0.026877	1.000000	0.000359	0.023327	0.026236	0.007232
Balance	0.013483	0.259010	0.414827	0.146629	0.021431	0.021431	0.032190	0.000359	1.000000	0.179410	0.001821	0.010897
NumOfProducts	0.003089	0.004913	0.018095	0.013963	0.020988	0.020988	0.013652	0.023327	0.179410	1.000000	0.009774	0.038009
HasCrCard	0.000321	0.006247	0.023111	0.017867	0.001364	0.001364	0.003963	0.026236	0.001821	0.009774	1.000000	0.032562
EstimatedSalary	0.029473	0.005010	0.004235	0.010568	0.032355	0.032355	0.004625	0.007232	0.010897	0.038009	0.032562	1.000000

4.2 Evaluation Method

We base the evaluation of ML classification models on the confusion matrix, which is used to evaluate the performance of a classification model. We can compute valuable metrics such as accuracy, precision, recall, and F-score.

Confusion matrix: performance measurement for machine learning classification problem where output can be two or more classes. It is a table with 4 different combinations of predicted and actual values.



TP(true positive) = client exited the company as predicted
FP(false positive) = client was predicted to exit but stayed TN(true negative) = client instead stayed as predicted
FN(false negative) = client was predicted to stay but exited instead

Accuracy: This is the fraction of predictions our model got right.

$$Accuracy = \frac{Number_of CorrectPredictions}{TotalNumber_of Predictions} = \frac{TP + FN}{TP + FP + TN + FN}$$

Precision: This is a ratio of correct positive predictions to the total predicted positives. In other words, the percentage that is truly positive out of all the positive classes predicted.

$$Precision/Positivepredictedvalue(P) = \frac{TP}{TP + FP}$$

Recall: Out of the total positive actual values, the percent of 1's correctly classified.

$$Recall(R) = \frac{TP}{TP + FN}$$

F-Score: This combines the precision and recall scores of a model. The accuracy metric computes how often a model made a correct prediction across the entire dataset. It is the weighted average score of the recall and precision.

$$F - Score = \frac{2PR}{(P + R)}$$

4.3 Experimental setup

Decision trees are a nonparametric supervised learning method for classification. Decision trees are valued in churn modeling for their interpretability, capturing nonlinear relationships and interactions, ability to handle mixed data types, scalability, and efficiency. They provide a straightforward representation of decision-making processes, accommodate various data types without extensive preprocessing, and are relatively efficient for analyzing large datasets. However, they can suffer from overfitting if not properly controlled. Techniques like pruning and limiting tree depth help eliminate this issue.

The random forest classifier is a versatile and robust machine learning model for classification notable for its ensemble learning approach. It builds multiple decision trees during training and combines their predictions to improve accuracy and mitigate overfitting. Each tree is trained on a random subset of the data, ensuring diversity among the trees. Random forest handles high-dimensional data well and is resistant to overfitting, making it suitable for churn modeling. It provides estimates of feature importance, aiding in understanding the underlying data patterns. While individual trees may lack interpretability, random forest provides insights into feature importance, aiding in understanding churn behavior

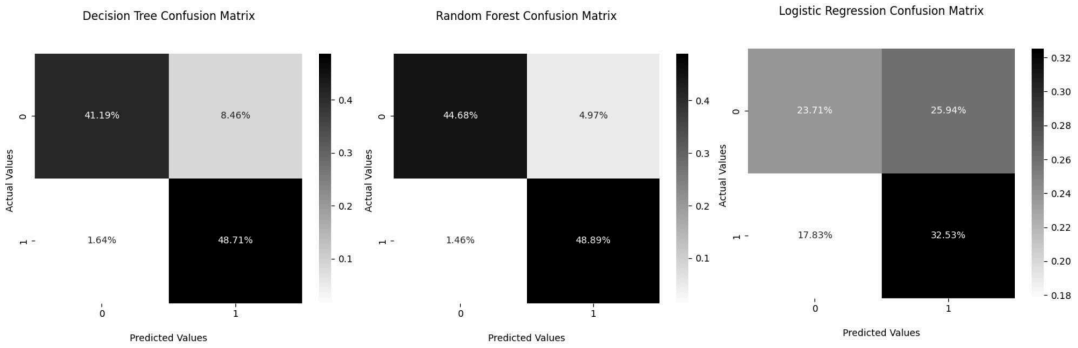
and informing effective retention strategies.

Logistic regression is a fundamental statistical model used for binary classification tasks. It employs a linear classification model to estimate probability based on predictor variables, yielding probabilities rather than discrete class labels for nuanced interpretation. Logistic regression uses a logistic function to map input variables to a binary outcome, typically representing the likelihood of a target class label. Robust to noise and outliers, it efficiently handles data variability, while its assumption of linearity accommodates complex relationships through interactions and nonlinear transformations. Moreover, logistic regression's computational efficiency makes it suitable for analyzing large datasets.

5. Results

In churn analysis, the results suggest that random forest classification is the most effective method for predicting customer churn in this scenario. Its higher accuracy, precision, recall, and F-score indicate better overall performance in identifying churn instances and minimizing false positives and false negatives. Decision trees also offer a competitive performance but may be slightly less robust than random forest due to their single-tree nature. While providing interpretable results, logistic regression lacks the predictive power demonstrated by decision trees and random forest in this churn analysis task. Therefore, businesses seeking accurate churn prediction models should prioritize using random forest classification based on these outcomes.

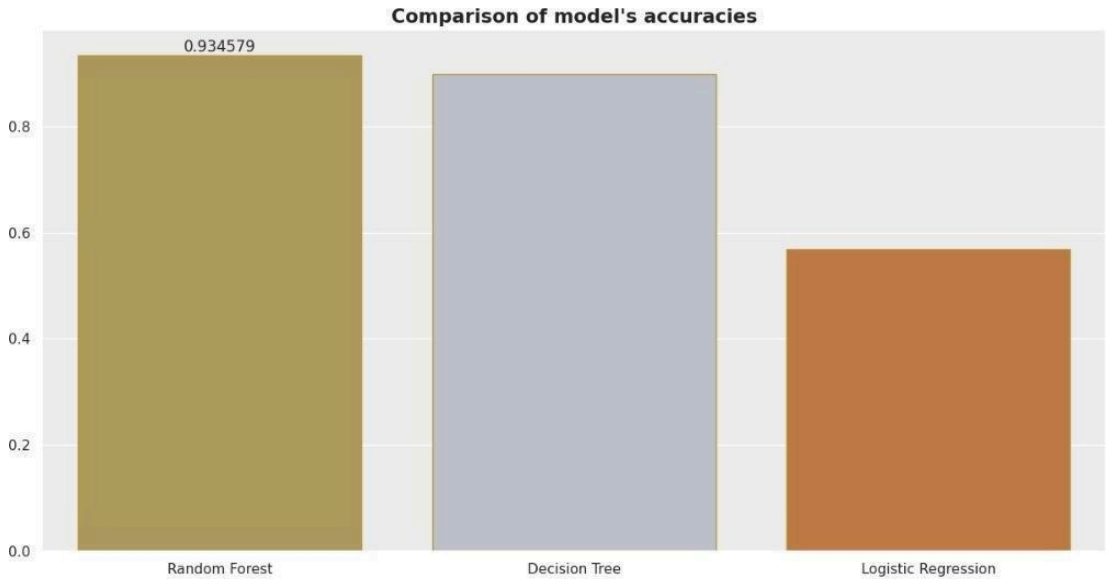
For dataset 1:

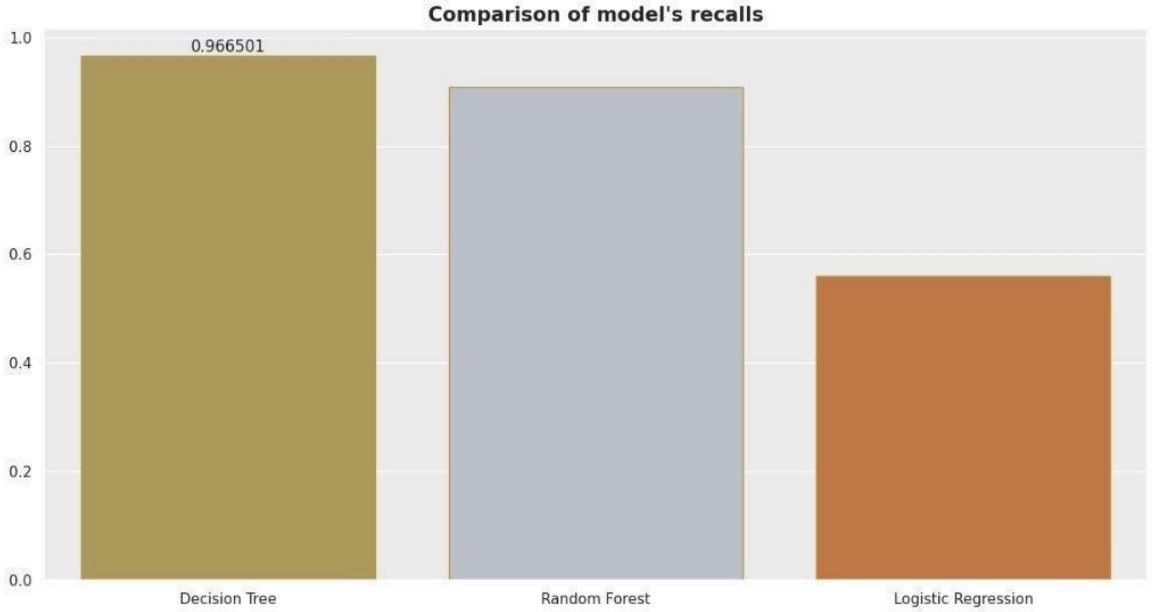


The following are the results of the evaluation metrics:

	Decision Tree Classifier	Random Forest Classifier	Logistic Regression
Accuracy:	0.898	0.937	0.563
Precision:	0.848	0.905	0.557
Recall:	0.971	0.977	0.649
F-Score:	0.905	0.939	0.599

The model with the best accuracy: Random Forest
The model with the best precision: Random Forest
The model with the best recall: Decision Tree
The model with the best f1-score: Decision Tree

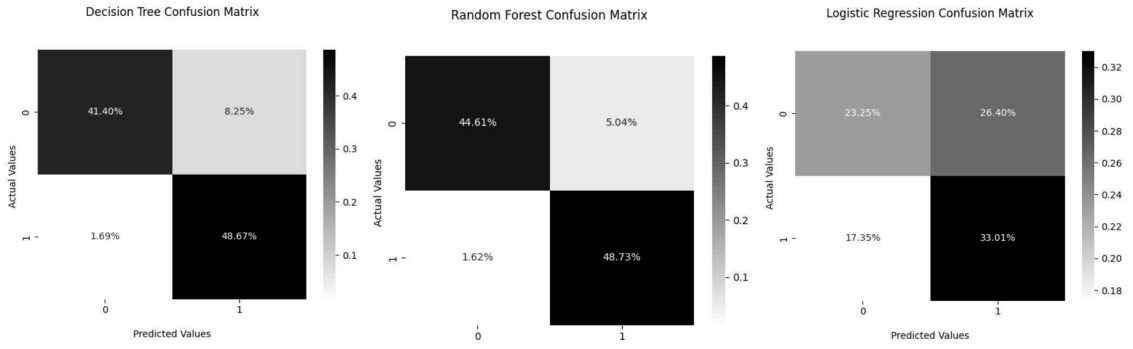




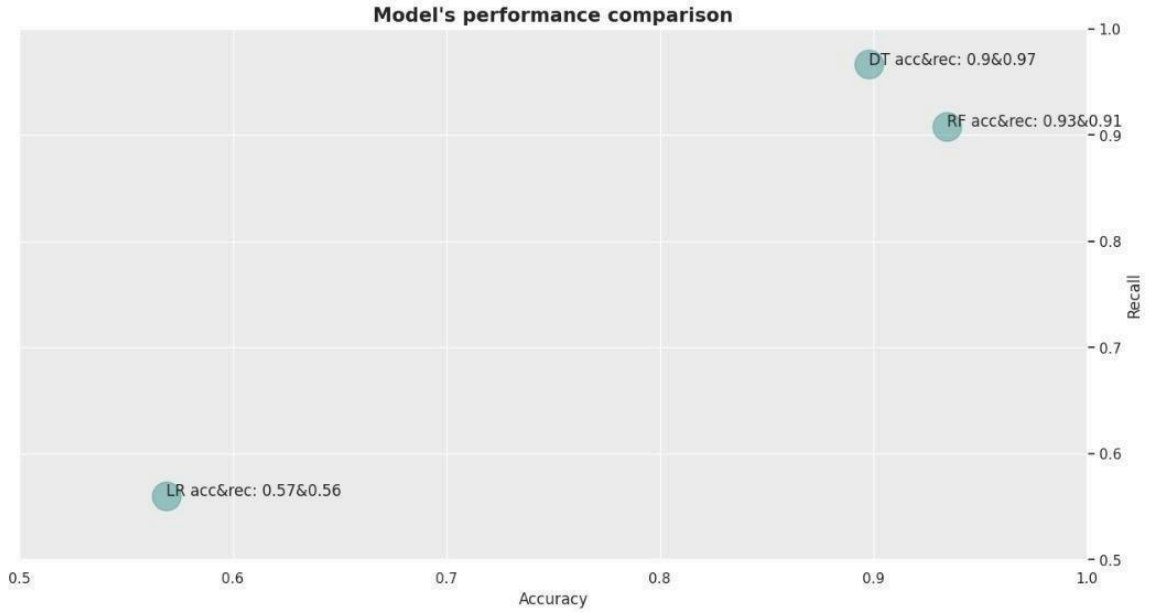
In comparing decision trees, random forest, and logistic regression classifiers with current state-of-the-art techniques for churn modeling analysis, several key considerations emerge. While traditional classifiers like decision trees and logistic regression offer simplicity and interpretability, most advance techniques such as gradient boosting machines (GBMs) like XGBoost, LightGBM, or CatBoost often outperform them in terms of predictive accuracy and scalability. GBMs excel in capturing complex relationships and nonlinear interactions within the data, making them more effective for uncovering intricate patterns. They also demonstrate robustness to noise and class imbalance, adapting weights during training to handle challenging datasets. However, GBMs may sacrifice some interpretability compared to traditional classifiers. Thus, when conducting a comparative analysis, it's essential to balance model interpretability with predictive performance and consider the specific requirements of the business application.

For dataset 2:
The following are the results of the evaluation metrics:

	Decision Tree Classifier	Random Forest Classifier	Logistic Regression
Accuracy:	0.901	0.933	0.569
Precision:	0.855	0.906	0.557
Recall:	0.967	0.968	0.656
F-Score:	0.907	0.936	0.601



The model with the best accuracy: Random Forest
The model with the best precision: Random Forest
The model with the best recall: Decision Tree
The model with the best f1-score: Decision Tree



Based on the outcomes of another dataset, it is evident that random forest classification produced the highest predictive performance among the three methods, achieving the highest accuracy, precision, recall, and F-score. Decision trees also performed reasonably well but were slightly outperformed by random forest regarding all metrics. On the other hand, logistic regression classification demonstrated significantly lower performance across all metrics compared to decision trees and random forest.

6. Conclusion

Churn prediction in the banking industry is a critical task aiming to develop effective models to forecast customer attrition. Applying machine learning techniques, including decision tree, random forest, and logistic regression classifiers, I analyzed a dataset comprising demographic information, transactional behavior, and service interactions. The methodology involved preprocessing the data, selecting relevant features, and training the models to predict churn. The results demonstrated that random forest classification outperformed both decision tree and logistic regression models, achieving the highest accuracy, precision, recall, and F1-score. The findings highlight the importance of ensemble methods in churn prediction tasks, particularly in complex domains such as banking, where customer behavior is multifaceted. This study contributes valuable insights into the effectiveness of different machine learning algorithms for churn modeling in the banking sector, providing banks with actionable strategies to mitigate customer attrition and optimize retention efforts.

Future research could explore the integration of more advanced feature engineering techniques, such as incorporating time-related patterns to enhance the predictive power of churn models. The ensemble methods could be further optimized by hyperparameter tuning or experimenting with different ensemble strategies. Moreover, deep learning architectures, such as recurrent neural networks or attention mechanisms, may offer new insights into capturing complex temporal dependencies and improving churn prediction accuracy in dynamic banking environments.

7. References

1. Christopher M. Bishop," Pattern Recognition and Machine Learning", New York, USA,2006
2. Churn Modelling, Kaggle, 2020. [Online]. Available: https://www.kaggle.com/datasets/shubh0799/churn-modelling/data
3. S. E. Charandabi, Prediction of Customer Churn in Banking Industry, https://www.researchgate.net/publication/342424673_Prediction_of_Customer_Churn_in_Banking_Industry
4. M. Rahman and K. Vasimalla, Machine Learning Based Customer Churn Prediction In Banking, https://www.researchgate.net/publication/348094541_Machine_Learning_Based_Customer_Churn_Prediction_In_Banking
5. M. WIRYASEPUTRA, Bank Customer Churn Prediction Using Machine Learning, https://www.analyticsvidhya.com/blog/2022/09/bank-customer-churn-prediction- using-machine-learning/