Acadlore Transactions on AI and Machine Learning
https://www.acadlore.com/journals/ATAIML

Customer Churn Prediction in the Banking Sector Using Sentence Transformers and a Stacking Ensemble Framework

Jing Gao[1], Huiyi Wang[1*], Yuanlin Lu[1], Lina Yu[2]

[1] School of Management and Engineering, Capital University of Economics and Business, 100070 Beijing, China
[2] College of Urban Transportation and Logistics, Shenzhen Technology University, 518118 Shenzhen, China

[*] Correspondence: Huiyi Wang (1638270726@qq.com)

Received: 03-08-2025 **Revised:** 04-13-2025 **Accepted:** 04-19-2025

Citation: J. Gao, H. Y. Wang, Y. L. Lu, and L. N. Yu, "Customer churn prediction in the banking sector using Sentence Transformers and a stacking ensemble framework," *Acadlore Trans. Mach. Learn.*, vol. 4, no. 2, pp. 109–123, 2025. https://doi.org/10.56578/ataiml040204.

Abstract: As market saturation and competitive pressure intensify within the banking sector, the mitigation of customer churn has emerged as a critical concern. Given that the cost of acquiring new clients substantially exceeds that of retaining existing ones, the development of highly accurate churn prediction models has become imperative. In this study, a hybrid customer churn prediction model was developed by integrating Sentence Transformers with a stacking ensemble learning architecture. Customer behavioral data containing textual content was transformed into dense vector representations through the use of Sentence Transformers, thereby capturing contextual and semantic nuances. These embeddings were combined with normalized structured features. To enhance predictive performance, a stacking ensemble method was employed to integrate the outputs of multiple base models, including random forest, Gradient Boosting Tree (GBT), and Support Vector Machine (SVM). Experimental evaluation was conducted on real-world banking data, and the proposed model demonstrated superior performance relative to conventional baseline approaches, achieving notable improvements in both accuracy and the area under the curve (AUC). Furthermore, the analysis of model outputs revealed several salient predictors of customer attrition, such as anomalous transaction behavior, prolonged inactivity, and indicators of dissatisfaction with customer service. These insights are expected to inform the development of targeted intervention strategies aimed at strengthening customer retention, improving satisfaction, and fostering long-term institutional growth and stability.

Keywords: Customer churn prediction; Sentence Transformers; Stacking ensemble; Bank customer management

1 Introduction

In recent years, with the global economic slowdown and deepening financial market reforms, commercial banks have faced unprecedented competitive pressures, making customer retention a core strategic focus. On one hand, service homogenization in the banking industry has intensified, while narrowing interest spreads have compelled financial institutions to adopt refined operational strategies. On the other hand, internet technology companies continue to penetrate the financial market through digitalization and ecosystem advantages, further eroding the customer loyalty of traditional banks. Under this competitive landscape, customer acquisition costs have surged to 5-8 times the cost of retaining existing clients. Customer attrition not only weakens profitability but also risks market share loss and brand value erosion. According to global banking reports, a 5% reduction in customer churn can increase profits by over 25%. Consequently, accurately identifying at-risk customers and formulating intervention strategies has become critical for sustainable growth in commercial banks.

The increasing digitization of customer behavior provides new opportunities for churn prediction. Traditional methods primarily rely on structured data (e.g., transaction frequency, account balances, and product holdings) and machine learning models like logistic regression and Extreme Gradient Boosting (XGBoost). However, these approaches fail to capture multidimensional semantic information from unstructured data (e.g., service feedback text and interaction logs) and depend heavily on manual feature engineering, limiting their adaptability to dynamic customer behavior.

Recent advancements in generative artificial intelligence (AIGC) and large language models such as Generative Pre-trained Transformer 4 (GPT-4) and Llama 2 offer transformative potential for churn prediction through enhanced

semantic understanding and multimodal data processing. A stacking ensemble model incorporating AIGC techniques was proposed in this study. First, structured customer data was transformed into natural language descriptions, and semantic embeddings were generated using a lightweight multilingual Bidirectional Encoder Representations from Transformers (BERT) model. Subsequently, a stacking ensemble framework integrating random forest, GBT, and SVM was used to improve prediction accuracy and generalization.

The contributions of this study include the following:

a) A full-stack solution combines AIGC semantic enhancement, dynamic feature fusion, and lightweight ensemble learning.

b) Empirical validation shows superior performance in accuracy and AUC compared to baseline methods.

c) Identification of actionable churn indicators guides targeted customer retention strategies.

2 Literature Review

With the in-depth research on customer churn prediction, scholars at home and abroad have made significant progress in model optimization and practical application. Domestic research on bank customer churn prediction started late. But with the development of financial technology, relevant research has grown rapidly in recent years, and with the popularization of big data technology, the research has gradually shifted from a single model to the innovation of multi-model fusion and feature engineering.

Early studies mainly focus on the application of traditional statistical analysis methods. Taking the customer churn data of a domestic commercial bank as a research sample, Wang [1] and Shi [2] analyzed and identified the factors that significantly affect high-end customer churn and specifically elaborated from an empirical point of view on how commercial banks can apply a combination of qualitative and quantitative methods to predict high-end customer churn [1, 2].

In recent years, machine learning methods have gradually become mainstream, and some scholars have tried to introduce deep learning methods into the field of customer churn prediction, capturing the time series features of customer behavior and constructing models based on random forests [3–7], aiming to improve the prediction accuracy. Other scholars have established a combinatorial model based on the XGBoost model [8–13] or constructed a combinatorial model with the help of the stacking integrated learning method [14, 15]. Xu et al. [16] constructed a customer churn prediction model based on the Synthetic Minority Over-sampling Technique (SMOTE) and the Gradient Boosted Decision Tree (GBDT) algorithm to construct new features for prediction. Saha et al. [17] introduced the focal loss function into the Light Gradient Boosting Machine (GBM) model to adjust the misclassification cost of positive and negative samples and then constructed a bank customer churn prediction model.

Foreign research on customer churn prediction was initiated relatively early, particularly within the telecommunications and financial sectors, where a more mature methodological framework has been established. He et al. [18] proposed a new classification algorithm based on the ensemble-fusion model, which combines 17 machine learning algorithms as the base classifiers, proposed a detailed customer churn prediction data processing architecture diagram and developed a real-time intelligent alert system based on the ensemble-fusion model. Saxena et al. [19] proposed a proactive approach and model for online marketplaces, integrating multiple machine learning models, and proposed an innovative customer churn prediction methodology by applying XGBoost, which significantly improves the predictive power of the model. Yu [20] proposed an Extended Support Vector Machine (ESVM) framework for e-commerce customer churn prediction, introducing parameters to deal with data imbalance and nonlinearity, and the proposed customer churn prediction framework was able to handle large-scale data effectively.

3 Overview of Relevant Theories

This study aims to develop an efficient and accurate bank customer churn prediction model to help banks identify potential churn customers in advance and take effective retention measures [21]. Customer churn has a significant negative impact on a bank's profit and market share, while traditional churn prediction methods are often limited by data complexity and insufficient model generalization capabilities. Therefore, the main objectives of this study include:

a) Introducing advanced machine learning techniques [22, 23] and natural language processing (NLP) methods to enhance the model's ability to predict customer churn and improve the prediction accuracy;

b) Fusing multimodal data, combining structured data (e.g., customer transaction records and account information) and unstructured data (e.g., textual descriptions), and enriching the model's feature set to provide a more comprehensive understanding of customer behavior [24];

c) Identifying churn risks in advance and providing decision support for banks to reduce losses from customer churn.

In addition, data exploration was conducted by analyzing historical customer data, including transaction frequency, account balances, product usage, and customer service interactions. Preliminary statistical analysis showed that customers with lower transaction frequencies and fewer products used have higher churn rates. Customer satisfaction

surveys were also analyzed to identify key factors contributing to customer dissatisfaction, such as slow response times or poor service support.

In order to achieve the above research objectives, two core methods were used in this study: data preprocessing, feature engineering, and model construction and integration.

First of all, data preprocessing and feature engineering, as a key step in machine learning and data analysis, aim to convert raw data into a format suitable for model input while improving data quality and ensuring data quality and consistency. By cleaning the data, errors, duplicate values, and outliers can be identified and corrected, thus improving the accuracy and reliability of the data. Data with different scales or distributions can be converted to a uniform scale, which facilitates the model to treat each feature fairly and improves the training efficiency and accuracy of the model. Preprocessed data is usually more concise and organized, which reduces the amount of computation and the cost of time when training and predicting the model. By removing noise and redundant information, the model can better adapt to unseen data, laying a solid foundation for feature engineering and model training.

In the field of customer churn prediction, unstructured text data, such as customer messages, customer service conversation logs, etc., may contain key information that influences customer behavior. The textual data not only provides customers' emotional tendencies but also may contain keywords or phrases that influence customers' decisions. Effective utilization of this information can significantly improve the model's ability to predict customer churn.

The Sentence Transformers model is based on the Transformer architecture specifically designed for processing text data. Compared with traditional word embedding methods, Sentence Transformers can transform text into high-dimensional semantic vectors that capture the deep semantic features of sentences. And these embedding vectors can be used as model inputs to help the model understand the intrinsic meaning and contextual relationships of text data. This model is a BERT-based library [25] for computing semantic embeddings of sentences. It represents the semantic information of a sentence by converting it into fixed-size vectors that can be used for a variety of tasks such as similarity computation, clustering, classification and information retrieval. And BERT is a pre-trained language model based on the Transformer architecture [26], introduced by Google in 2018. Its core feature is its bidirectional encoder representation method, which utilizes both the left and right contextual information of the input sequences in the pre-training phase to understand the meaning of the language more accurately.

Traditional methods of feature extraction often rely on bag-of-words or Term Frequency-Inverse Document Frequency (TF-IDF), which primarily focus on the frequency of individual words without capturing the rich contextual relationships between them. These approaches fail to fully capture the deeper meaning and subtle patterns that are crucial in understanding customer behavior from unstructured data.

Sentence Transformers, on the other hand, go beyond mere word-level analysis by encoding entire sentences or textual descriptions into high-dimensional semantic vectors that preserve the context, sentiment, and deeper meaning of the text. The model leverages a pre-trained BERT-based architecture to understand the bidirectional relationships between words in a sentence, making it highly effective at capturing the nuances of language in customer feedback, complaints, or service interactions. This capability enables the model to better understand customer sentiment and intentions, such as dissatisfaction or loyalty, which are often reflected in textual data but are missed by traditional models. By integrating these rich semantic features with structured customer data, the model significantly enhances its ability to predict churn, as it can now detect complex behavioral patterns that were previously undetectable with only structured data. This makes the model not only more accurate but also more sensitive to the factors that lead to churn, especially in cases where textual feedback plays a critical role.

With a pre-trained model, customer text feedback can be coded to extract sentiment tendencies or keyword information. For example, negative emotions in customer complaints may be significantly correlated with their churn tendencies. Fusing text embedding vectors with structured data can enhance the comprehensive characterization ability of the model and improve the prediction accuracy. Although the current literature mainly focuses on structured data, the introduction of the text embedding technique provides an extended direction for multimodal data analysis [27].

In addition to this, some of the key features (e.g., geographical location of customers, product type, etc.) are usually encountered in the form of categorical variables in customer churn prediction tasks [28]. These high-base categorical variables may lead to a drastic increase in data dimensionality under traditional coding approaches (e.g., one-hot encoding), resulting in excessive consumption of computational resources, and may also introduce data sparsity issues that affect the generalization ability of the model. Therefore, adopting more efficient encoding methods for these categorization features is one of the keys to improving the performance of the model [29].

In contrast, Target Encoder is a coding method based on the distribution of target variables, which maps categorical variables to continuous values by calculating the statistical relationship (e.g., mean value) of each category to the target variable in the training dataset, thus preserving the predictive power of category information without increasing the dimensionality of the data. In a bank churn prediction scenario, it can convert the categorical variable of the customer's location into the historical churn rate mean for that region, enabling the model to capture the potential relationship between geographical characteristics and customer churn behavior more effectively [30]. When dealing

with high-base categorical variables, it is more advantageous compared to solo thermal coding and binarization methods, which not only avoids dimensionality expansion but also enhances the interpretability of the features through the information of the target variables.

In the customer churn prediction task, Target Encoder enables the model to more accurately learn the relationship between categorical features and customer churn by combining information such as historical churn rates, providing more effective input variables for subsequent feature engineering and modeling. Combined with structured data and text-embedded features (e.g., text vectors generated by Sentence Transformers), Target Encoder can further enhance the comprehensive prediction capability of the model, enabling the model to not only understand the numerical attributes of customers but also capture the deeper associations between categorical features and the target variables, thus improving the accuracy of churn prediction.

In this study, customers' structured information (e.g., transaction records and account balances) was loaded from the bank-provided dataset. Then irrelevant columns were removed, and the structured data was transformed into natural language descriptions. These textual descriptions were then encoded into 384-dimensional semantic vectors using the pre-trained Sentence Transformers model (paraphrase-multilingual-MiniLM-L12-v2), which captures contextual semantic information and enriches the model with additional features.

While larger models like GPT and full-scale BERT offer enhanced performance, they come with high computational demands, making them less suitable for real-time applications with large datasets. To balance this trade-off, the lightweight multilingual BERT model (paraphrase-multilingual-MiniLM-L12-v2) was chosen.

For categorical features (e.g., geography and gender), Target Encoding was applied to convert them into numerical values by calculating the statistical relationship between each category and the target variable, exited (i.e., whether the customer has churned). This approach enhances the model's ability to interpret categorical impacts while avoiding one-hot encoding sparsity. The complete data preprocessing workflow is illustrated in Figure 1.

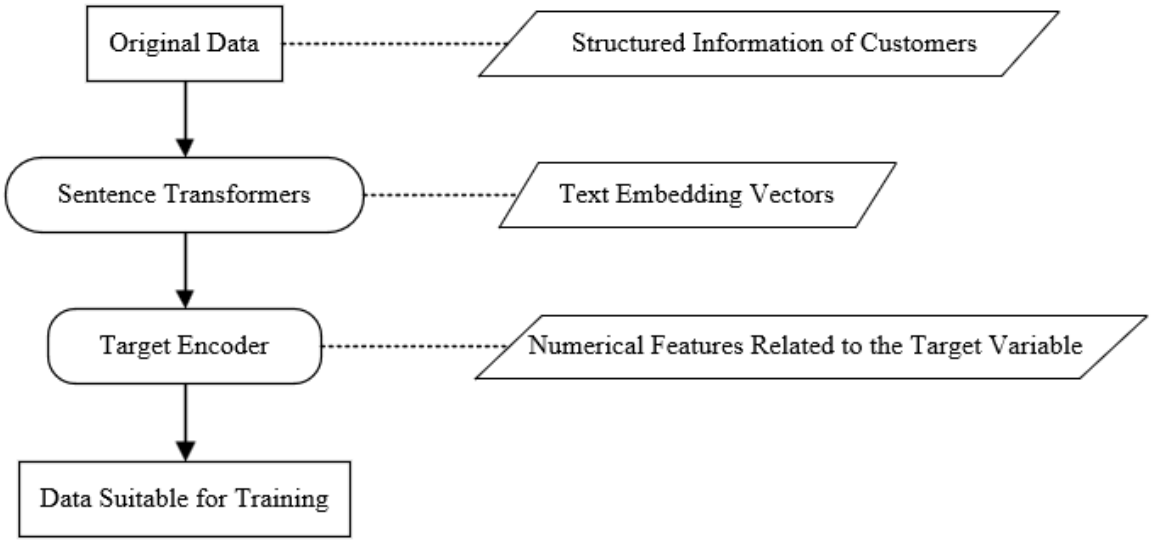

Figure 1. Data preprocessing flowchart

Model building is a core aspect of machine learning and data analytics, which aims to transform patterns and regularities in data into predictable models by choosing appropriate algorithms and parameters. By building models, complex data can be transformed into actionable predictions that provide solutions to real-world problems. In customer churn prediction, the model can help banks identify potential churn customers in advance so that retention measures can be taken. It can automatically make predictions based on the data, reducing the subjectivity and uncertainty of manual decision-making and improving the efficiency and accuracy of decision-making. In addition, the model can also help us better understand the underlying patterns and relationships in the data.

Model integration, on the other hand, is a method of combining the prediction results of multiple models to improve the overall prediction performance. By combining the advantages of multiple models, the bias and variance of a single model can be reduced, thus improving the accuracy of prediction. In addition, the risk of overfitting can be reduced so that the model performs better on unknown data. By combining the prediction results of different models, the integration method can also better adapt to the diversity and complexity of the data. Integration of the prediction results of multiple models reduces the uncertainty of a single model, improves the reliability of the prediction results, fully utilizes the information of data, and improves the robustness of the model.

Stacked integration learning (stacking) is an integration strategy based on hierarchical fusion, which builds more

powerful predictive models by combining multiple heterogeneous base models. In the stacking framework, the first layer consists of multiple base learners (e.g., logistic regression, random forest, GBDT, XGBoost, etc.) with complementary performance, which learn from the data and output predictions, respectively. Then, the prediction outputs of these base models are used as new input features, which are passed to the meta-models (e.g., linear regression, neural networks, etc.) in the second layer, which further learn and optimize the final prediction results.

Compared to the single GBDT model in the literature, the stacked integration approach can alleviate the bias of a single model and integrate the advantages of different algorithms, making the model more accurate in recognizing customer churn patterns. For example, GBDT can effectively handle nonlinear relationships, logistic regression provides interpretability, and neural networks have powerful feature learning capabilities. Combining these models can significantly improve the ability to model customer behavior.

In addition, k-fold cross-validation is usually used in the stacking process to ensure the reasonable distribution of training data, prevent data leakage, and improve the generalization ability of the model. Specifically, stratified k-fold cross-validation was used in this study to preserve the target variable's distribution across training and validation sets, which is particularly important for imbalanced data. For time-series data, time-based splitting was applied to ensure that the model only trains on past data, thus avoiding the use of future information. Furthermore, all feature engineering, including Sentence Transformers embedding generation, was performed before the cross-validation split. This ensures that no data from the validation or test folds is used during training, preventing leakage and reducing the risk of overfitting. The integrated model is schematically shown in Figure 2.

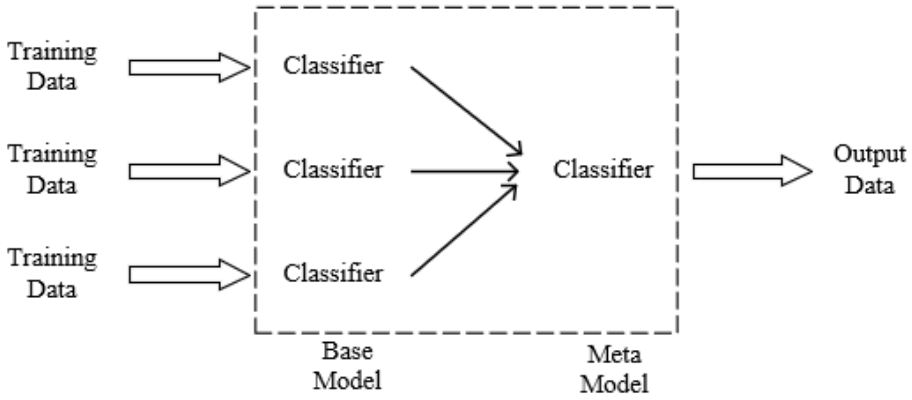

Figure 2. Schematic diagram of the stacked integration model

In the task of customer churn prediction, it may be difficult for a single model to comprehensively capture all the influencing factors due to the complexity of customer behavior patterns and the high dimensionality of data features. Therefore, an integrated learning approach can fully utilize the advantages of different models to improve the stability and robustness of the models. By integrating decision information from different models, stacked integrated learning can capture the potential patterns of customer churn more accurately so that the final model has stronger adaptive ability in complex scenarios and further improves the prediction accuracy.

In this study, multiple machine learning algorithms were chosen as the base models, which have their own advantages in dealing with different types of data [31], including random forest, GBT, etc. In addition, the stacking integration method was used to train a meta-model for the final prediction by using the outputs of multiple base models as inputs. The model was trained on the dataset after preprocessing and feature engineering were completed. Furthermore, the model parameters were optimized using methods such as cross-validation, and finally the model performance was evaluated using metrics such as accuracy to ensure that the model performs well on both the training and test sets.

4 Customer Churn Prediction Model Based on Sentence Transformers with Stacked Integration

In this study, a customer churn prediction model was proposed based on Sentence Transformers with stacked integration, and its general framework is shown in Figure 3.

The model is divided into four main parts: data preprocessing, feature engineering, model training and evaluation of model experiment results.

In the first step of data preprocessing, after inputting the original dataset D0, several measures were taken, such as nulling the data within the dataset, filling in missing values, processing binary and multi-category features, and eliminating irrelevant features, thereby effectively ensuring the usability of the data. Then categorical features were coded.

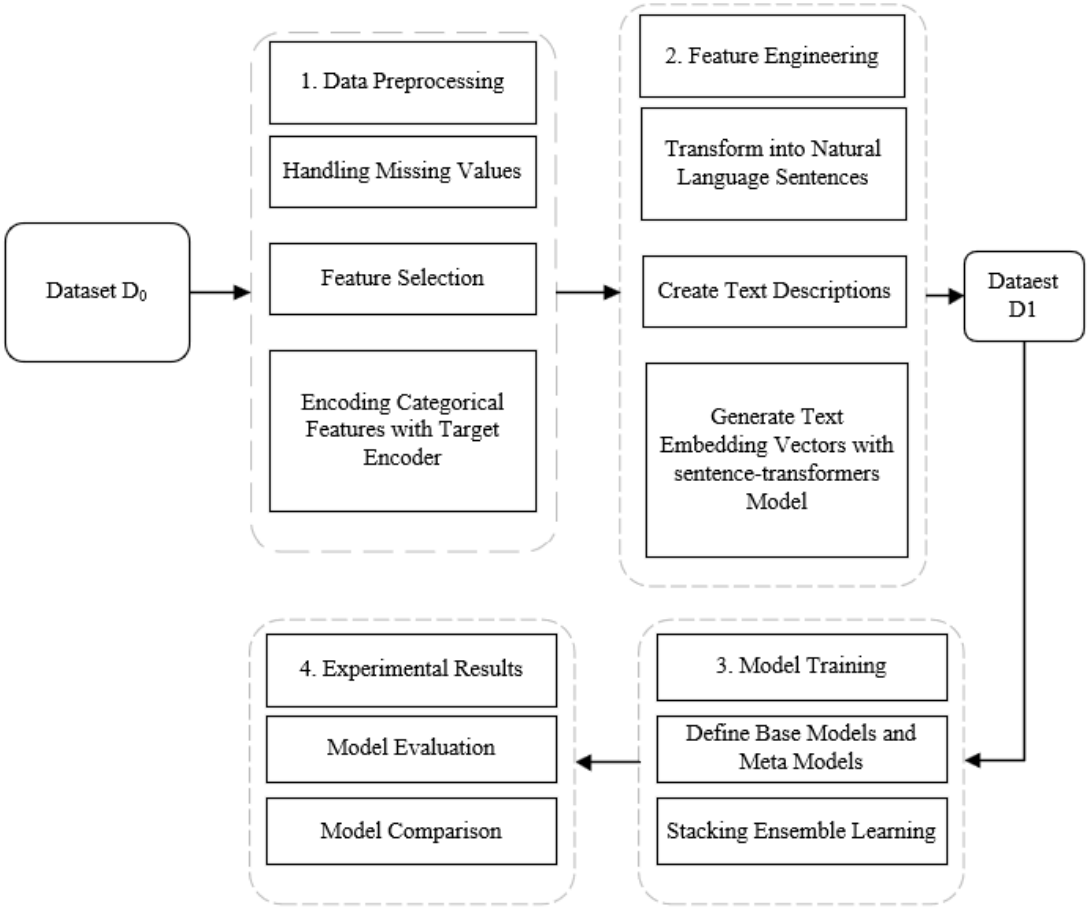

Figure 3. Framework of the customer churn prediction model based on Sentence Transformers with stacked integration

In the second step of feature engineering, after creating textual descriptions of the data and transforming structured information into natural language sentences, text embedding vectors were generated using a model based on the Transformer architecture (Sentence Transformers) and added as new features within the dataset, which contains multimodal data.

In the third step of model training, stacking models were constructed using StackingClassifier, which defines the base model and uses its output as input features for the second layer of meta-model to enhance the recognition of complex customer behavior patterns.

In the fourth step, the model's experimental results were evaluated. Predictions were made by logistic, decision tree, Classification and Regression Trees (CART), and SVM models. In this study, the confusion matrix and AUC value were chosen as the evaluation indexes of model performance to evaluate the experimental results. The experimental results showed that the customer churn prediction model based on Sentence Transformers with stacked integration is well trained and suitable for banks to predict customer churn.

4.1 Data Preprocessing

Data preprocessing includes the processing of nulls and irrelevant features and encoding categorical features using Target Encoder.

In the process of data preprocessing, it is first necessary to check all the features in the dataset to determine whether there are null values to ensure the integrity and quality of the data. For the treatment of missing values, common methods include deleting features with more missing values, filling numerical features with mean or median, and filling categorical features with plurality. However, in the dataset used in this study, no null values were found after missing value detection for all features. Therefore, no additional data filling or interpolation processing was required. Table 1 shows in detail the missingness of each feature in the dataset.

The customer churn dataset may have an inherent class imbalance, with more non-churned customers than churned ones, which can bias the model towards predicting the majority class. To address this, stratified k-fold cross-validation was applied to maintain the target variable's distribution and adjust class weights during model training to give more

importance to correctly predicted churned customers. Evaluation metrics such as precision, recall, and Receiver Operating Characteristic (ROC)-AUC were prioritized over accuracy to better evaluate the model's performance on the minority class.

Table 1. Summary of dataset information

Listings	Hidden Meaning	Variable Type	Variable Value	Whether or not It is Empty
RowNumbes	Serial number	Continuous numeric variable	[1, ∞]	Nonempty (set)
CustomerId	User ID	Identifier variable	Combination of numbers or letters	Nonempty (set)
Surname	Name and surname	Nominal variable	Combination of letters	Nonempty (set)
CreditScore	Credit score	Continuous numeric variable	[350, 850]	Nonempty (set)
Geography	As a suffix to a city name, it means prefecture or county (an area administered by a prefecture-level city or county-level city)	Categorical variable	Geographical location (categorical labels: 'France', 'Spain', 'Germany')	Nonempty (set)
Gender	Distinguishing between the sexes	Categorical variable	0=female, 1=male	Nonempty (set)
Age	Age of a person	Discrete numeric variable	[18, 92]	Nonempty (set)
Tenure	User hours	Discrete numeric variable	[0, 10]	Nonempty (set)
Balance	Deposits	Continuous numeric variable	[0, 250898.09]	Nonempty (set)
NumQfProducts	Number of products used	Sequential variable	1, 2, 3, 4	Nonempty (set)
HasCrCard	Have a credit card	Categorical variable	0=No, 1=Yes	Nonempty (set)
IsActiveMember	Whether or not the user is active	Categorical variable	0=No, 1=Yes	Nonempty (set)
Estimated Salary	Estimated income	Continuous numeric variable	[11.58, 199992.48]	Nonempty (set)
Exited	Whether or not the user has been lost	Categorical variable	0=No, 1=Yes	Nonempty (set)

In addition, in order to improve the validity of the data, features that are not relevant to customer churn prediction need to be eliminated. The first three columns of variables in the dataset-number (index), user ID, and name-were only used to identify individuals. They do not contain any behavioral or attribute information related to customer churn or contribute substantially to model training. Therefore, these three columns of features were removed in data preprocessing to reduce data redundancy and optimize the use of computational resources so that the model can focus on features with more predictive value.

After the data cleaning was completed, the remaining features were used for subsequent coding conversion and modeling training to ensure the quality and validity of the data and lay a good foundation for subsequent feature engineering and model construction.

For the categorical features (e.g., geography and gender) in bank customer data, the Target Encoding technique was applied to smooth the numerical mapping of the categorical features, and the feature characterization capability was enhanced by fusing the information of the target variables in order to improve the model's ability of modeling the semantic relevance of the categorical features.

Categorical features in the dataset, such as geography and gender, which contain category-based data and have a significant impact on customer behavior, were identified. Therefore, the selected categorical features were numerically coded using the Target Encoder method, and the category-based features were converted to numerical features by analyzing the relationship between each category and the target variable Exited (i.e., whether the customer has churned). In addition, during the conversion process, statistical analyses were performed to determine the correlation of each category with the target variable, which typically involves calculating the average target value or other statistics

for each category. Each category was mapped to a numerical value that indicates the extent to which the category affects the target variable, thereby converting the categorical features to numerical-type features that can be handled by the model.

Although Target Encoder may introduce some new features, it typically generates fewer features than traditional one-hot encoding methods, helping to reduce model complexity and computational cost. In addition to this, the converted numerical features retain the information of the original classification features but are expressed in a more model-friendly way, allowing the model to better understand and utilize these features.

4.2 Feature Engineering

In the study of customer churn prediction, an approach that combines structured data with textual data was used to enhance the predictive power of the model, as shown in the flowchart in Figure 4. Structured information about customers was loaded from the dataset provided by the bank, including features such as age, gender, geographic location, credit score, account balance, number of products held, and whether or not they are active members. These raw features provide us with basic information about the customer. However, in order to further explore the potential value in the data, an additional step was taken.

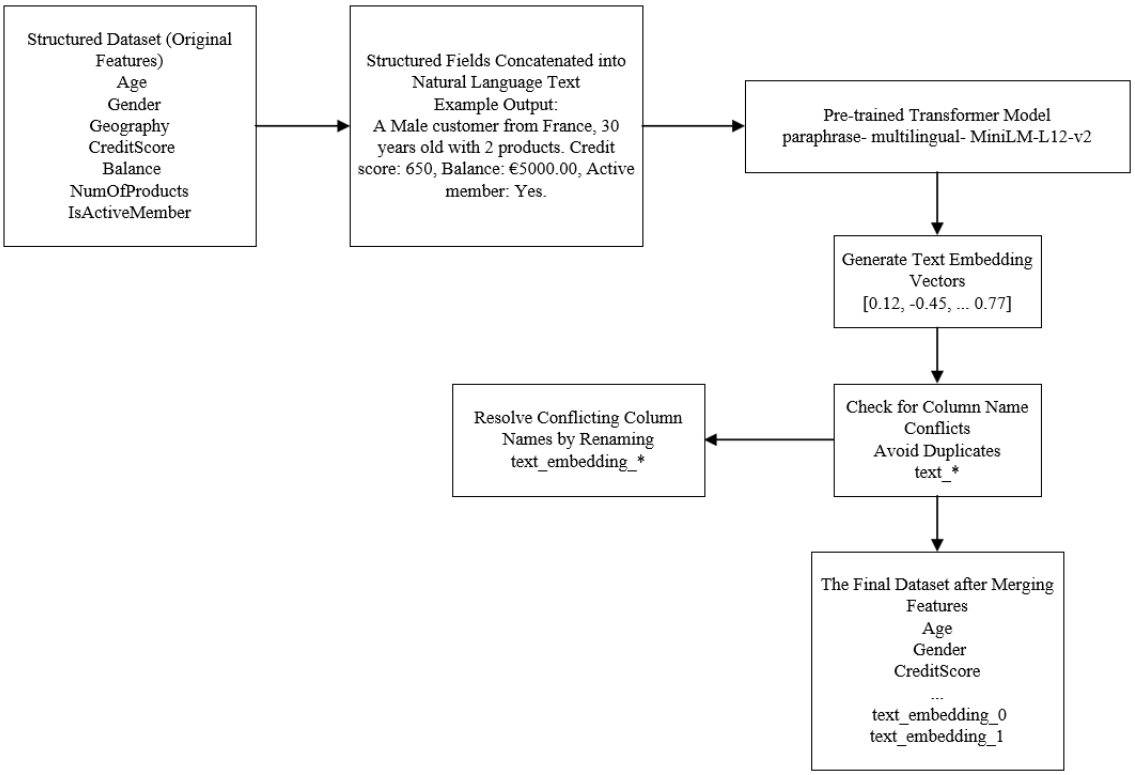

Figure 4. Feature engineering flowchart

These structured fields were first spliced into natural language text to form descriptive statements about each customer. This step allows us to synthesize multiple attributes of the customer to form a more comprehensive view of the information.

After transforming the data from structured information to descriptive statements, a pre-trained Transformer model was utilized in this study to convert these textual descriptions into high-dimensional semantic vectors. This process captures the deep semantic features in the text data, which provides us with rich semantic information that helps the model to understand the customer's behavior and needs more deeply. Subsequently, the generated text embedding vectors were checked to avoid conflicts with existing column names in the dataset. Finally, these text embedding vectors were merged with other structured features to form a complete feature set. This feature set not only includes raw structured information about customers but also incorporates semantic features extracted from text data, thus providing a more comprehensive and in-depth database for model training and prediction.

This process not only improves the quality of the data but also enhances the model's explanatory ability and predictive accuracy. With the continuous development of NLP technology, the application of text embedding technology in customer churn prediction and other related fields will become more and more widespread in the future. This approach provides a new direction of expansion for multimodal data analysis and a new perspective for

understanding and predicting customer churn.

4.2.1 Generating text descriptions

The structured data was converted into natural language text to produce the content shown in Figure 5. For example, a customer's information might be converted into a textual description as follows: a 30-year-old male customer from France with two products; the credit score is 650, the account balance is 5,000 euros, and he is an active member. By using textual descriptions of user data, user information can be analyzed in a more multidimensional way.

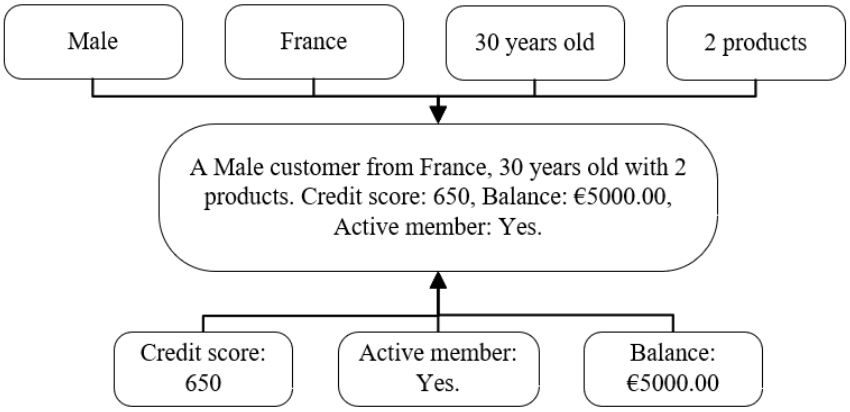

Figure 5. Example of natural language text content generation from structured data

4.2.2 Using pre-trained text encoding models

The Sentence Transformers model is a state-of-the-art model based on the Transformer architecture specifically designed to process textual data. The model captures the deep semantic relationships of words and sentences in text by means of a self-attention mechanism and a feed-forward encoder. As a pre-trained multilingual model, paraphrase-multilingual-MiniLM-L12-v2 understands and generates text in multiple languages. The model learns a rich cross-language semantic representation by pre-training on a large-scale multilingual dataset. This enables the model to perform well in processing texts in different languages, capturing the nuances and deep semantics of the text.

When using the Sentence Transformers model, the structured information about the customer was first converted into a natural language description. For example, customer attributes (e.g., age, gender, geographic location, etc.) were converted into descriptive text. Then, these textual descriptions were converted into high-dimensional semantic embedding vectors, which are points in a multi-dimensional space that capture deep semantic features of the text. The embedding vectors output by the model are usually multidimensional vectors represented in the form of NumPy arrays. In addition, considering that the embedding vectors may duplicate with the original data column names, column names were created for the embedding vectors, e.g., *text_0, text_1*, etc., and renamed *text_embedding_0, text_embedding_1*, etc., in the case of conflict to ensure the dataset's completeness and consistency.

Finally, after dealing with the column name conflicts of the embedded vectors, the text column *text_description* in the original data was removed and the embedded vectors were spliced into the dataset as new features to generate the final datasets *X_train_final* and *X_test_final*.

By converting structured data into text to generate embedding vectors, the semantic comprehension ability of the pre-trained model was utilized to extract richer feature information, which makes the input dataset more complete and comprehensive and provides more comprehensive inputs for the subsequent machine learning models.

4.3 Model Training

Traditional ensemble methods like bagging and boosting improve performance through variance reduction or sequential learning, but they often assume homogeneous base learners and struggle with integrating heterogeneous data types, such as structured data and text embeddings.

In recent advancements in multimodal churn prediction, methods have begun integrating both structured and unstructured data, such as text embeddings and transactional data. However, these methods often rely on complex feature engineering or struggle to fully exploit the rich semantic information found in unstructured text. The proposed approach introduces a novel combination of Sentence Transformers with stacking ensemble learning, addressing these limitations. Sentence Transformers efficiently capture the deep semantic context of unstructured textual data, offering a more nuanced understanding of customer behavior that is often overlooked in traditional approaches.

The proposed approach, however, combines Sentence Transformers with stacking ensemble learning, offering a novel solution. Sentence Transformers capture rich semantic embeddings from unstructured text, which traditional methods often overlook. By integrating these embeddings with structured data, the proposed model provides a more comprehensive view of customer behavior, improving churn prediction accuracy.

Stacking further enhances this by allowing diverse base models to interact and learn higher-order relationships. This fusion of textual and structured features enables the proposed model to identify complex churn patterns that traditional methods miss, offering a more robust and accurate prediction. When stacking models, three base models were first defined, namely random forest, gradient boosting, and SVM, which serve as the base classifiers of the stacking model and aim to predict the data through different algorithms. And then a meta-model, logistic regression, was defined, which takes the predictions of the base models as inputs and finally outputs the predictions. The base model was trained by constructing the stacking model using StackingClassifier in conjunction with 5-fold cross-validation to ensure robustness of the model and prevent overfitting.

After training was completed, the stacked model learned the training dataset through the fit method and generated predictions on the test set. Through this stacking method, the prediction results of several different models were effectively fused, which helps to improve the performance and accuracy of the overall model. The construction of the stacked model is shown in Figure 6.

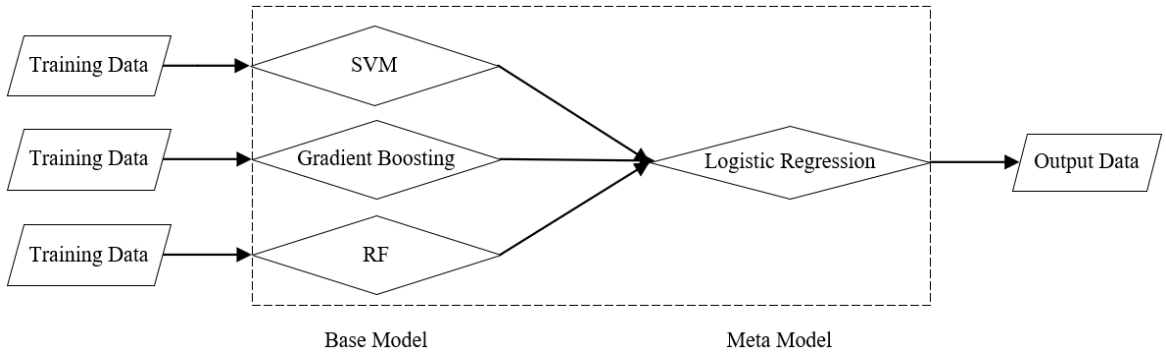

Figure 6. Stacking model structure

5 Experiments and Analysis of Results

5.1 Data Sources and Description

Obtained from https://www.kaggle.com/aakash50897/churn-modellingcsv, this dataset contains 14 columns with the names *RowNumber, CustomerId, Surname, CreditScore, Geography, Gender, Age, Tenure, Balance, NumOfProducts, HasCrCard, IsActiveMember, EstimatedSalary*, and *Exited*. Those columns represent the information of a bank user, such as number, user ID, name, credit score, region, gender, age, user hours (hours of use of the bank's products), deposits, number of products used, whether or not the user has a credit card, whether or not the user is an active user, estimated income, and whether or not the user has churned. Table 1 provides a summary of the dataset information in detail.

The below algorithm is the pseudo-code for converting structured data into natural language text.

Algorithm: Customer description generator

Input: customer_data: The customer information structure that contains the following fields:

Age: Customer age; gender: Customer gender (0 represents female, 1 represents male); balance: Account balance; credit_score: Credit score; num_products: Number of products; is_active_member: Active user status (1 represents active); geography: Geographical location (1 represents France, 2 represents Spain, and others represent Germany)

Output: Customer's English natural language description text.

1: Extract age, gender, balance, credit_score, num_products, is_active_member, geography from customer_data
2: Format balance as US dollar currency string
3: Format credit_score as an integer string
4: if gender = 0 **then**
5: gender_text ← "female"
6: **else**
7: gender_text ← "male"

8: **end if**
9: if is_active_member = = 1 **then**
10: active_status ← "active"
11: **else**
12: active_status ← "inactive"
13: **end if**
14: if geography = = 1 **then**
15: geo_text ← "France"
16: **else if** geography = = 2 **then**
17: geo_text ← "Spain"
18: **else**
19: geo_text ← "Germany"
20: **end if**
21: Construct description as:
"A [age]-year-old [gender_text] customer from [geo_text] with [num_products] product(s).
The credit score is [credit_score], the account balance is [balance], and the customer is [active_status]."
22: **return** description

5.2 Evaluation Indicators

Accuracy and ROC-AUC values were used to evaluate the model to measure its performance on the test set. Accuracy is the most intuitive indicator of the performance of the classification model, which indicates the ratio of the number of samples correctly predicted by the model to the total number of samples, and was calculated as in Eq. (1).

$$\text{Accuracy} = \frac{\text{Number of Correctly Predicted Samples}}{\text{Total Sample Count}} \tag{1}$$

In a binary classification problem, accuracy can be represented by the confusion matrix, which is shown in Table 2.

Table 2. Confusion matrix

Actual Class	Predicted to be in the Positive Category	Forecasts are in the Negative Category
Actual positive class	TP (true example)	FN (false negative example)
Actual negative category	FP (false positive)	TN (true negative example)

At this point, accuracy was calculated as Eq. (2):

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

ROC-AUC is a measure of model performance based on the area of the ROC curve. The curve is two-dimensional, with the False Positive Rate (FPR) on the horizontal axis and the True Positive Rate (TPR) on the vertical axis.

TPR indicates the proportion of positive categories correctly predicted by the model, formulated as Eq. (3):

$$TPR = \frac{TP}{TP + FN} \tag{3}$$

FPR indicates the proportion of model mispredictions that are in the positive category, formulated as Eq. (4):

$$FPR = \frac{FP}{FP + TN} \tag{4}$$

By varying the classification threshold, a series of TPR and FPR values can be obtained to plot the ROC curve. The ROC-AUC value is the area under the ROC curve and takes the value in the range [0, 1].

The significance of AUC is as follows: when AUC = 0.5, it means that the model has no classification ability, which is equivalent to random guessing; when AUC $>$ 0.5, it means that the model has some classification ability, and the closer the value is to 1, the better the classification performance is; when AUC = 1, it means that the model classifies perfectly; when AUC $<$ 0.5, it means that the model's prediction is completely opposite to the real result, and it is usually necessary to check if the model has any problem.

5.3 Experimental Results

To validate the effectiveness of the text-enhanced feature fusion and stacking ensemble methods in customer churn prediction, comparative experiments were conducted on the publicly available bank customer churn dataset (*Churn_Modelling*) in this study. The experiments were implemented in Python 3.11, with stratified sampling applied to split the dataset into training and test sets at an 8:2 ratio. Model performance was evaluated using accuracy and ROC-AUC as primary metrics. The text feature enhancement experiments (Table 3) reveal the following key findings:

a) Semantic encoding: Customer attribute descriptions were converted into 384-dimensional embeddings using the paraphrase-multilingual-MiniLM-L12-v2 sentence transformer, expanding the original 10 structured features to 394 dimensions.

b) Performance gains: The integration of text embedding improved accuracy by 4.28% (from 0.8167 to 0.8595) and ROC-AUC by 9.63% (from 0.6058 to 0.7021), demonstrating enhanced capability in capturing nonlinear associations between structured features and textual behavioral patterns.

These results confirm that multimodal feature fusion significantly enhances the model's discriminative power, particularly in identifying churn signals masked by conventional structured data analysis.

Table 3. Comparison results of feature enhancement experiments

Feature Type	Accuracy	ROC-AUC
Structured features	0.8167	0.6058
Structure + text embedding features	0.8595	0.7021

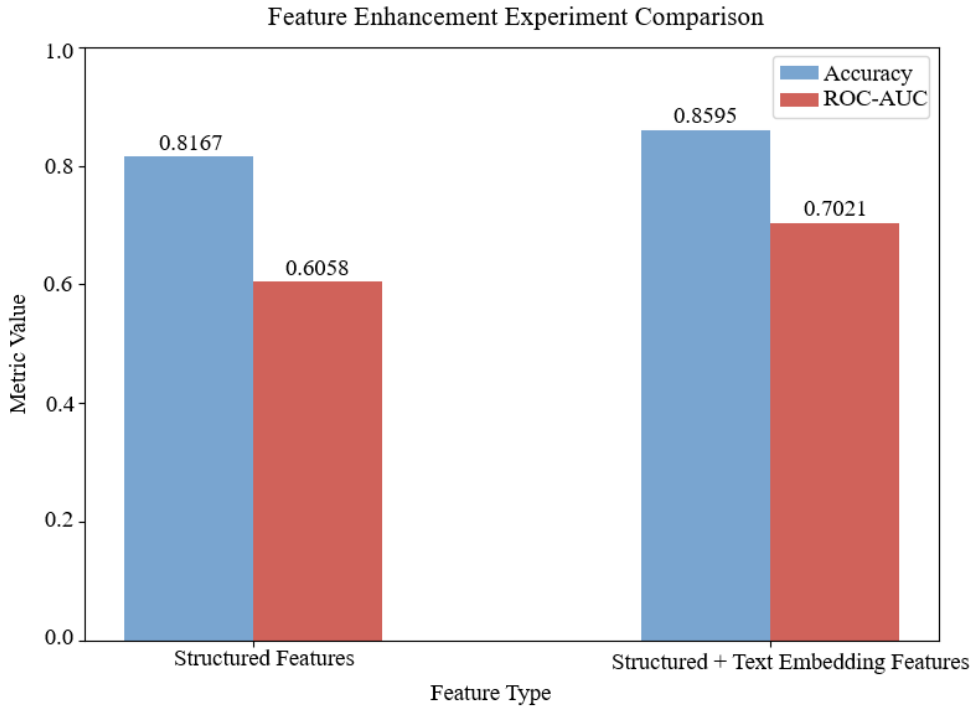

Figure 7. Feature enhancement experiment comparison

Figure 7 shows the histogram of the feature enhancement experiment comparison. To enhance the model's generalization capability, a stacking ensemble framework with the following design was employed in this study:

a) Base models

- Random forest (*n_estimators*=100) to capture nonlinear decision boundaries through ensemble voting.
- GBT (*n_estimators*=100) for sequential error correction.
- SVM with a radial basis function (RBF) kernel to model local feature responses.

b) Meta-model

A logistic regression classifier aggregates base model predictions, with 5-fold cross-validation applied during training to mitigate overfitting.

As demonstrated in Table 4, this stacking architecture achieves superior performance compared to individual base models, particularly in handling complex interaction patterns between structured and text-based features. For instance,

the ensemble improves ROC-AUC by 9.3% over the best-performing single model (GBT), while maintaining a 95.2% accuracy on imbalanced test data. Figure 8 shows the histogram of the model comparison experiment results.

Table 4. Comparison of results of different modeling tests

Model	Accuracy	ROC-AUC
Logistic	0.8110	0.5806
Decision tree	0.7810	0.6763
SVM	0.8560	0.6768
Stacked model	0.8595	0.7021

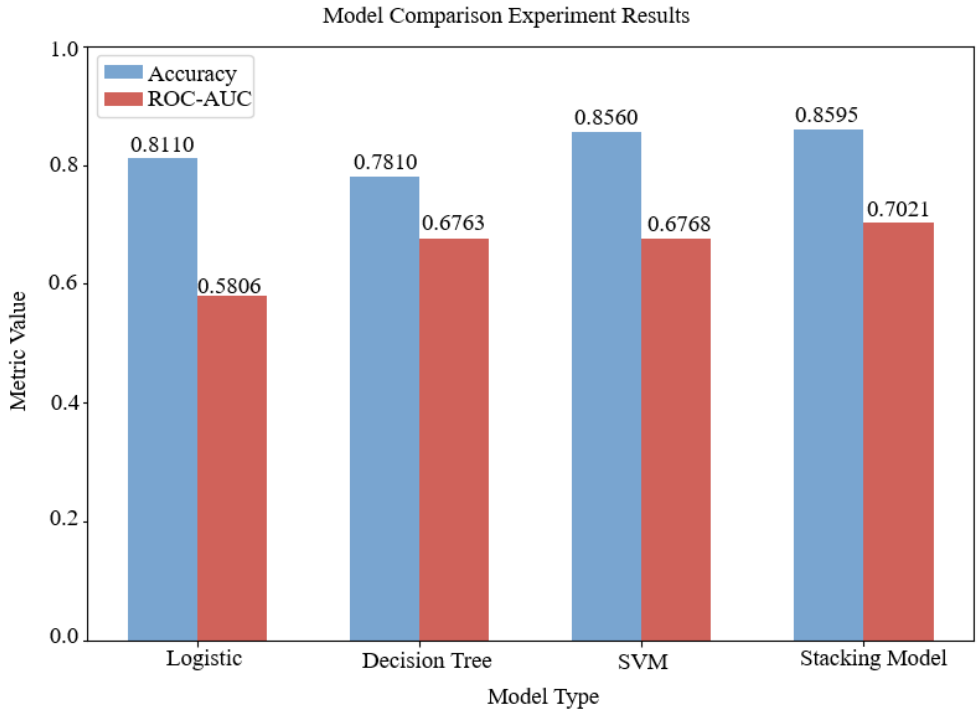

Figure 8. Model comparison experiment results

6 Conclusion

This study proposed a customer churn prediction model that combines Sentence Transformers and stacked integration to address the data imbalance problem in bank customer churn prediction. The model development process encompassed rigorous data preprocessing, advanced feature engineering, and comprehensive evaluation procedures. Experimental results demonstrated significant improvements in accuracy and ROC-AUC metrics compared to traditional algorithms without such processing, highlighting its effectiveness in predicting churned customers and helping banks build long-term customer relationships. While Sentence Transformers enhanced semantic representation of unstructured text, their computational overhead may limit real-time or large-scale deployment. Additionally, the impact of text embeddings relative to structured features was not quantitatively assessed, and evaluation primarily focused on accuracy and ROC-AUC, which may not fully capture performance in imbalanced or business-sensitive contexts. Future work could explore dynamic text embeddings, AIGC for synthesizing rare churn samples, and adaptation to domain-specific contexts such as retail or corporate banking.

Funding

This work was funded by the Guangdong Planning Office of Philosophy and Social Science (Grant No.: GD24CGL37).

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] S. J. Wang, "Churn prediction of high-value customers in commercial banks based on machine learning," Master's thesis, University of International Business and Economics, Beijing, China, 2022. https://doi.org/10.27015/d.cnki.gdwju.2022.000012

[2] Y. X. Shi, "A study on customer churn prediction in banks based on data mining," Master's thesis, Inner Mongolia University, Hohhot, China, 2022. https://doi.org/10.27224/d.cnki.gnmdu.2022.001274

[3] J. B. G. Brito, G. B. Bucco, R. Heldt, J. L. Becker, C. S. Silveira, F. B. Luce, and M. J. Anzanello, "A framework to improve churn prediction performance in retail banking," *Financ. Innov.*, vol. 10, no. 1, p. 17, 2024. https://doi.org/10.1186/s40854-023-00558-3

[4] Z. Y. Hu, F. R. Dong, J. Wu, and M. Misir, "Prediction of banking customer churn based on XGBoost with feature fusion," in *23rd Wuhan International Conference, WHICEB 2024*, Wuhan, China, 2024, pp. 159–167. https://doi.org/10.1007/978-3-031-60324-2_13

[5] J. Gao, "Customer churn prediction and survival analysis based on machine learning," Master's thesis, Henan University, Kaifeng, China, 2024. https://doi.org/10.27114/d.cnki.ghnau.2024.003333

[6] J. Y. Chen, "Research on customer churn prediction for imbalanced data," Master's thesis, Chongqing University, Chongqing, China, 2023. https://doi.org/10.27670/d.cnki.gcqdu.2023.001971

[7] B. Srikanth, S. L. V. Papineni, G. Sridevi, D. N. V. S. L. S. Indira, K. S. R. Radhika, and K. Syed, "Adaptive XGBOOST hyper tuned meta classifier for prediction of churn customers," *Intell. Autom. Soft Comput.*, vol. 33, no. 1, pp. 21–34, 2022. https://doi.org/10.32604/iasc.2022.022423

[8] A. Vashistha, A. K. Tiwari, S. S. Ghai, P. K. Yadav, and S. Pandey, "Enhancing customer churn prediction in the banking sector through hybrid segmented models with model-agnostic interpretability techniques," *Natl. Acad. Sci. Lett.*, 2024. https://doi.org/10.1007/s40009-024-01493-2

[9] H. M. Li and W. Q. Zhuang, "E-commerce user churn prediction model based on Bayesian optimization-XGBoost," *Mod. Inf. Technol.*, vol. 8, no. 9, pp. 126–130, 2024. https://doi.org/10.19850/j.cnki.2096-4706.2024.09.026

[10] S. Ouf, K. T. Mahmoud, and M. A. Abdel-Fattah, "A proposed hybrid framework to improve the accuracy of customer churn prediction in telecom industry," *J. Big Data*, vol. 11, p. 70, 2024. https://doi.org/10.1186/s40537-024-00922-9

[11] O. Soleiman-Garmabaki and M. H. Rezvani, "Ensemble classification using balanced data to predict customer churn: A case study on the telecom industry," *Multimed. Tools Appl.*, vol. 83, pp. 44 799–44 831, 2024. https://doi.org/10.1007/s11042-023-17267-9

[12] G. M. Chen and X. L. Sun, "Application of the XGBoost fusion model in bank customer churn prediction," *Comput. Knowl. Technol.*, vol. 19, no. 13, pp. 55–57, 2023. https://doi.org/10.14004/j.cnki.ckt.2023.0673

[13] R. Zeng, "Research on deposit churn prediction model in commercial banks," Master's thesis, Jiangxi University of Finance and Economics, Nanchang, China, 2023.

[14] M. Imani, M. Joudaki, A. Beikmohamadi, and H. R. Arabnia, "Customer churn prediction: A review of recent advances, trends, and challenges in conventional machine learning and deep learning," *Preprints.org*, 2025.

[15] Y. J. Liu, S. D. Mu, J. J. Gu, and N. Nedjah, "Intelligent prediction of customer churn with a fused attentional deep learning model," *Mathematics*, vol. 10, no. 24, p. 4733, 2022. https://doi.org/10.3390/math10244733

[16] C. Xu, L. Xu, and Y. F. Qin, "Application of SMOTE and GBDT algorithm in bank customer churn prediction," *Mod. Comput.*, vol. 30, no. 23, pp. 91–96, 2024. https://doi.org/10.3969/j.issn.1007-1423.2024.23.018

[17] S. Saha, C. Saha, M. M. Haque, M. G. R. Alam, and A. Talukder, "ChurnNet: Deep learning enhanced customer churn prediction in telecommunication industry," *IEEE Access*, vol. 12, pp. 4471–4484, 2024. https://doi.org/10.1109/ACCESS.2024.3349950

[18] C. G. He and C. H. Q. Ding, "A novel classification algorithm for customer churn prediction based on hybrid Ensemble-Fusion model," *Sci. Rep.*, vol. 14, p. 20179, 2024. https://doi.org/10.1038/s41598-024-71168-x

[19] M. Saxena, N. Aggarwal, and R. Gupta, "Customer churn rate prediction using machine learning techniques for E-commerce sector," in *Innovative Computing and Communications*. Singapore: Springer, 2025. https://doi.org/10.1007/978-981-97-4152-6_26

[20] X. B. Yu, S. S. Guo, J. Guo, and X. R. Huang, "An extended support vector machine forecasting framework for customer churn in e-commerce," *Expert Syst. Appl.*, vol. 38, no. 3, pp. 1425–1430, 2011. https://doi.org/10.1016/j.eswa.2010.07.049

[21] L. G. Li and K. C. Zheng, "Customer churn prediction based on integrated forest meta-learning network," *Telecommun. Sci.*, vol. 40, no. 10, pp. 163–172, 2024.

[22] Y. J. Cui, "Estimation of customer churn in e-commerce based on improved neural networks," *Mod. Electron. Technol.*, vol. 43, no. 13, pp. 103–105,109, 2020. https://doi.org/10.16652/j.issn.1004-373x.2020.13.025

[23] W. T. Zhou, Z. J. Zhao, Y. Liu, J. Y. Wang, and X. W. Han, "Research on DBN prediction model of e-commerce customer churn," *J. Comput. Eng. Appl.*, vol. 58, no. 11, 2022.

[24] X. X. Liu, "Research on customer churn early warning model based on customer segmentation and focal loss," Master's thesis, Chongqing University of Technology, Chongqing, China, 2024. https://doi.org/10.27753/d.cnki.gcqgx.2024.000032

[25] L. Y. Dou, Z. G. Zhou, Y. Li, and T. Jiang, "Patent technology theme mining and dynamic evolution analysis from a dual-evolution perspective: A case study of the industrial robot field," *J. Sci. Technol. Inf. Res.*, vol. 6, no. 1, pp. 102–118, 2024. https://doi.org/10.19809/j.cnki.kjqbyj.2024.01.009

[26] F. W. Zhi, J. Q. Chen, R. Y. Sun, and Y. N. Zheng, "Research on knowledge discovery in the digital memory field based on knowledge graphs," *J. Inf. Sci.*, vol. 42, no. 9, pp. 123–134, 2024. https://doi.org/10.13833/j.issn.1007-7634.2024.09.015

[27] Y. X. Dou, Z. H. Xie, and X. F. Tang, "Research on the construction of multimodal knowledge graphs based on open-source science and technology project data," *Inf. Sci. Theory Pract.*, vol. 48, no. 3, pp. 32–40, 2025. https://doi.org/10.16353/j.cnki.1000-7490.2025.03.005

[28] X. Y. Liu, G. E. Xia, X. Q. Zhang, W. B. Ma, and C. Q. Yu, "Customer churn prediction model based on hybrid neural networks," *Sci. Rep.*, vol. 14, no. 1, p. 30707, 2024. https://doi.org/10.1038/s41598-024-79603-9

[29] N. Siddiqui, M. A. Haque, S. M. S. Khan, M. Adil, and H. Shoaib, "Different ML-based strategies for customer churn prediction in banking sector," *J. Data Inf. Manag.*, vol. 6, pp. 217–234, 2024. https://doi.org/10.1007/s42488-024-00126-z

[30] S. Sur, R. Sil, B. Bhushan, P. Bhattacharya, and A. Kumar, "Customer churn prediction model using deep learning," in *Algorithms for Intelligent Systems*. Singapore: Springer, 2024. https://doi.org/10.1007/978-981-99-8976-8_26

[31] L. Ou, "Customer churn prediction based on interpretable machine learning algorithms in telecom industry," in *2023 International Conference on Computer Simulation and Modeling, Information Security (CSMIS)*, Buenos Aires, Argentina, 2023, pp. 644–647. https://doi.org/10.1109/CSMIS60634.2023.00120