**PREDICTING CREDIT CARD CUSTOMER SEGEMENTATION**

Lahari Kotapati Reethu Bhargavi Sajjala Dheerajkumar Tadiparthy

*Department of CSE,* *Department of CSE,* *Department of CSE,*

*SRM University-AP, India* *SRM University-AP, India* *SRM University-AP, India*

*lahari\_kotapati@srmap.edu.in*  *bhargavi\_sajjala@srmap.edu.in dheerajkumar\_t@srmap.edu.in*

Beecha Venkata Naga Hareesh

*Department of CSE,*

*SRM University-AP, India*

*nagahareesh\_beecha@srmap.edu.in*

**1.ABSTARCT**

This study proposes a novel approach for predicting credit card customer segmentation by combining binning techniques, decision tree analysis, and Naive Bayes classification. The methodology involves preprocessing the dataset using binning to simplify and enhance data analysis. A decision tree is then employed to create a transparent and interpretable model for customer segmentation based on key features. Additionally, Naive Bayes classification is integrated to improve predictive accuracy. The research, conducted on a real-world credit card dataset, demonstrates the effectiveness of the combined approach in accurately segmenting customers. The findings contribute to advancing credit card customer segmentation methodologies, providing valuable insights for targeted marketing and risk management in the financial industry.

**2.INTRODUCTION**

The dataset under review is a significant resource for customer information in the context of consumer credit card portfolios. It has been carefully gathered to assist analysts in forecasting client attrition. This dataset provides a thorough profile of each consumer by including a wide range of demographic information, such as age, gender, marital status, and income category. Furthermore, it explores the complexities of the client's connection with the credit card issuer, taking into account factors like the type of card, the number of months it has been open, and inactive times. With information on total revolving balance, credit limit, average open-to-buy rate, and important analytical metrics like the total amount of change from quarter 4 to quarter 1, average implementation ratio, and a Naive Bayes classifier attrition flag, these insights into customers' spending behavior as they approach the decision to quit are especially interesting.

Within the scope of this research, our goal is to extract useful insights from the abundance of data present in the dataset by utilizing predictive analytics approaches, particularly decision tree analysis and the Naive Bayes algorithm. Specifically, our method relies heavily on the use of binning techniques to help classify continuous variables into discrete intervals. This preprocessing step improves our models' interpretability and efficacy. Our goal as we work through this information is to provide a comprehensive knowledge of the factors using an upcoming departure or long-term account stability in addition to predicting client attrition. Using decision trees and Naive Bayes provides a strong analytical framework for both strategic portfolio management and meeting the demands of individual consumers. Essentially, the goal of this project is to enable decision-makers to navigate the complexities of consumer behavior inside a credit card portfolio by utilizing the predictive capabilities of data analytics.

**3.LITERATURE REVIEW**

The study applied K-means clustering to categorize a Chinese bank's credit card customers. Using data mining methods, including decision trees, the research identified distinct criteria for high-quality, potential high-quality, common, and unfavorable customers. The decision tree provided actionable insights for targeted marketing, emphasizing the importance of data mining technology in credit card marketing for gaining a competitive advantage.[1]

The paper tackles credit card churn prediction using an ensemble system with majority voting, incorporating various classifiers. Employing techniques such as SMOTE for handling the highly unbalanced dataset, the study achieves notable results with 92.37% sensitivity, 91.40% specificity, and 91.90% overall accuracy through tenfold cross-validation. Feature selection using Classification and Regression Tree (CART) identifies key predictor variables. The decision tree J48 generates 'if-then' rules serving as an early warning system. In conclusion, the ensemble system, particularly when applied to SMOTED data, proves effective in credit card churn prediction, emphasizing the significance of decision tree-generated rules for proactive management and prediction in the banking sector.[2]

This study focuses on customer churn in the credit card industry, employing rough set theory (RST) and a flow network graph to predict and understand patterns in a Taiwanese commercial bank. Analyzing a dataset of 21,000 customers, the research identifies rules related to churn based on demographic, psychographic, and transactional variables. The combined model, integrating RST and a flow network graph, effectively predicts churn and provides valuable insights for marketing strategy decisions. The model's advantage lies in expressing data through easily understandable decision rules. The study suggests future research explore longer period data, extend the model's applicability to different countries, and compare its effectiveness with alternative approaches.[3]

This paper develops a credit card customer churn prediction model using a feature-selection method and five machine learning models. The C5 tree model outperforms others, identifying the top three important variables as total transaction count, total revolving balance on the credit card, and change in transaction count. Merging categorical variables improves model performance, emphasizing the importance of variable selection. The study concludes that early prediction models using machine learning can assist banks in retaining customers. Future work should explore optimal independent variables and new machine learning models for improved accuracy and efficiency in churn prediction.[4]

**4.DATASET DESCRIPTION**

This dataset contains a wealth of customer information collected from within a consumer credit card portfolio, with the aim of helping analysts predict customer attrition. It includes comprehensive demographic details such as age, gender, marital status and income category, as well as insight into each customer’s relationship with the credit card provider such as the card type, number of months on book and inactive periods. Additionally, it holds key data about customers’ spending behavior drawing closer to their churn decision such as total revolving balance, credit limit, average open to buy rate and analyzable metrics like total amount of change from quarter 4 to quarter 1, average utilization ratio and Naive Bayes classifier attrition flag (Card category is combined with contacts count in 12months period alongside dependent count plus education level & months inactive). Faced with this set of useful predicted data points across multiple variables capture up-to-date information that can determine long term account stability or an impending departure therefore offering us an equipped understanding when seeking to manage a portfolio or serve individual customers.

**5.DATA PREPROCESSING**

In data mining, binning, also known as discretization or bucketing, is a preprocessing technique used to transform continuous numerical variables into discrete intervals or bins. The primary goal of binning is to simplify the complexity of the data and make it more manageable for analysis. This process involves grouping a range of numerical values into a smaller number of discrete "bins" or categories.

The benefits of binning in data mining include:

1. Simplification: Binning simplifies the representation of continuous data, making it easier to understand and analyze.

2. Noise Reduction: Binning can help reduce the impact of outliers or noisy data points by grouping them into a common category.

3. Improved Performance: Some machine learning algorithms may perform better when dealing with categorical or discrete data, and binning facilitates this transformation.

4. Interpretability: Binned data is often more interpretable, especially when communicating findings to non-technical stakeholders.

5. Feature Engineering: Binning can be part of feature engineering, where the transformed data may reveal patterns or relationships that were not apparent in the original continuous form.

**6.IMPLEMENTATION**

We implemented a methodical approach to identify trends and forecast customer attrition in our extensive study on credit card customer segmentation. The first step was gathering a dataset that included a wide range of characteristics associated with credit card users. This dataset, whose source, dimensions, and complexity were described, provided the framework for our future analytical efforts. Using statistical and visual tools, we carefully engaged in exploratory data analysis (EDA) to determine the distribution and properties of the data. Histograms, box plots, and other visualization tools were used to examine key attributes and reveal any possible relationships with the goal variable, the attired flag.

With the EDA providing us with new insights, we proceeded into the complex world of data preprocessing. In order to discretize continuous data, a complex approach to binning techniques was used, taking into account missing values and outliers. Encoding of the categorical variables and a sensible division of the dataset into training and testing sets were essential for a reliable model evaluation. The process of feature engineering involved adding new features and scaling them appropriately to further improve the dataset in preparation for future modelling.

The creation of a correlation matrix, which revealed the correlations between variables, was a crucial step in our research. This matrix served as our predictive modelling journey's compass, highlighting characteristics that were closely related to the attrition flag.

We explored the subject of machine learning after improving and enriching our dataset. The interpretability and effectiveness of decision tree and naive Bayes algorithms in classification problems led to their selection. These models were intensively trained on the assigned training set, and the predicted performance was optimized by fine-tuning the hyperparameters. The model was evaluated using a variety of metrics, including F1-score, accuracy, recall, and precision. A thorough investigation of the confusion matrix added to the depth of our analysis by offering a more granular understanding of the models' predictive power.

A diagram of a data analysis

Description automatically generated

**Fig-1 Flowchart**

**Binning:** Binning, or discretization, in data mining involves transforming continuous numerical variables into discrete categories or bins. This simplifies data representation, reduces noise, and aids in analysis. Binning methods include equal-width, equal-frequency, and custom binning. It can be unsupervised or supervised, considering the target variable. Binning may result in information loss and impact certain machine learning models. Applications include histogram construction and variable transformation. The choice of binning strategy depends on the dataset's characteristics and analysis goals. Types of binning are

Equal-Width Binning: Divides the range of values into equal-width intervals. This method may not be suitable for datasets with unevenly distributed values.

Equal-Frequency Binning: Divides the data into intervals with approximately the same number of data points in each bin. This helps handle unevenly distributed data.

**Decision Tree:** Decision Tree It is a fundamental algorithm that forms the basis for more advanced methods like Random Forest. The process entails generating a tree-like structure with a node for each internal element choice based on a specific feature and a leaf node corresponding to a predicted outcome after recursively dividing the data according to the selected characteristics. Decision Tree seeks to identify key decision points to effectively classify future stock market outcomes.

**Naïve Bayes:** Naive Bayes Naive Bayes is a good contender for stock market forecasting because to its effectiveness in handling high dimensional data and relatively quick training time, which adds important insights to the project's more comprehensive predictive modelling strategy.

**Performance Metrics:**

Let us consider J=True positive, K=True Negative, L=False Positive, M=False Negative

**Accuracy:** Calculates the proportion of times a model was accurate across the entire dataset, which is still accurate if the dataset is class balanced.

*Accuracy = (J + K)/ (J + K + L + M)*

**Precision:** a classification model's capacity to focus on relevant data items exclusively. Mathematicians divide the total true positives by the total true positives plus false positives to determine accuracy.

*Precision = (J)/ (J + L)*

**Recall:** The capacity of a model to locate all pertinent occurrences in a data source. Before using the recall formula, we need divide the total number of true positives by the sum of true positives and false negatives.

*Recall = (J)/ (J + M)*

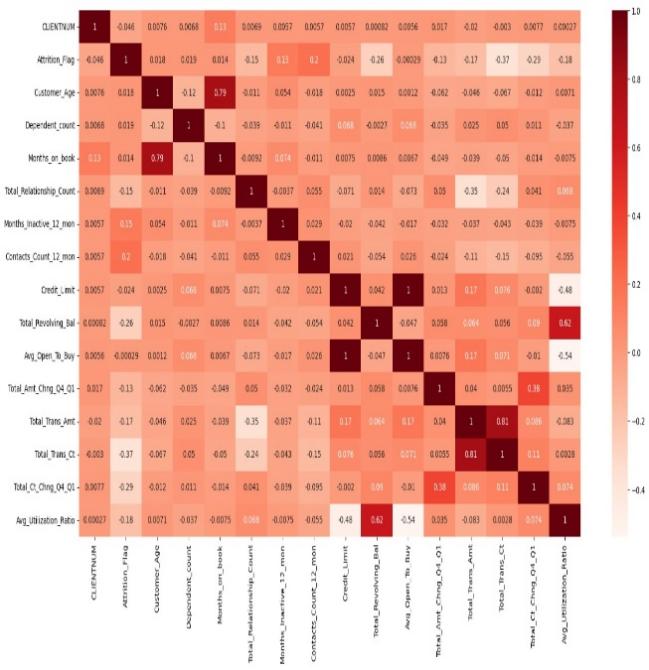
**F1 score:** A singular metric that fuses both recall and precision by employing the harmonic mean, creating a comprehensive and balanced evaluation of a model's performance in terms of identifying true positives and minimizing false positives.

*F1 score= 2×(precision×recall)/ (precision + recall)*

The results' interpretation was the final step in our technique. We examined the importance of the traits and uncovered how they may be used to forecast the customer turnover. The results were condensed into an integrated narrative that acknowledged any built-in shortcomings in our methodology and summarized the consequences for corporate plans.

**7.RESULTS**

Machine learning became a focal point as we improved and enriched our dataset. Decision tree and naive Bayes algorithms were chosen for their interpretability and effectiveness in classification problems. Intensive training on the assigned dataset and fine-tuning of hyperparameters optimized the models' predictive performance. Evaluation metrics, including accuracy, precision, recall, and F1-score, were employed to assess the models' effectiveness. The confusion matrix provided a granular understanding of the models' predictive power.



**Fig-2 Correlation Matrix**

To form the correlation matrix `X\_fs` is assigned the value of the DataFrame with the 'Attrition\_Flag' column excluded, effectively creating a new DataFrame that includes all features except the one indicating customer attrition. Similarly, `Y\_fs` is assigned the values of the 'Attrition\_Flag' column, representing the target variable or the outcome we aim to predict in a machine learning model. This separation is a common practice in preparing data for machine learning, where features (X) and the target variable (Y) are defined separately to facilitate model training and evaluation.

A computer code with numbers and symbols

Description automatically generated with medium confidence

**Fig-3 Decision Tree Result**

The decision tree algorithm applied in our credit card customer segmentation study achieved an accuracy of 84%, demonstrating its effectiveness in making correct predictions. With recall and F1 Score both at 84%, the model showcased a balanced ability to identify relevant occurrences and focus on pertinent data items in forecasting customer attrition and precison as 85%.

A computer screen shot of numbers

Description automatically generated

**Fig-4 Naïve Bayes Result**

The naïve bayes algorithm applied in our credit card customer segmentation study achieved an accuracy of 88%, demonstrating its effectiveness in making correct predictions. With recall and F1 Score both at 88%, the model showcased a balanced ability to identify relevant occurrences and focus on pertinent data items in forecasting customer attrition and precision as 85%.

**8.CONCLUSION**

In conclusion, our focus on the given dataset revolves around distinguishing between existing customers and those who have attired, essentially left the credit card service. Our goal is to find patterns and insights in the dataset by utilizing data mining techniques, particularly decision tree analysis and the Naive Bayes algorithm. Binning is a preprocessing technique that helps to simplify continuous data so that it may be more easily analyzed and interpreted. Our objective in starting this analytical journey is to project client attrition while also comprehending the underlying variables that influence this choice. With this information, we can better serve our customers, adjust our plans, and make well-informed decisions. Essentially, we are trying to utilize data analytics to differentiate between current and lost clients to promote a more proactive and strategic approach in the credit card industry.

**9.REFERENCES**

1.Li, W., Wu, X., Sun, Y., & Zhang, Q. (2010, December). Credit card customer segmentation and target marketing based on data mining. In 2010 International Conference on Computational Intelligence and Security (pp. 73-76). IEEE.

2.Anil Kumar, D., & Ravi, V. (2008). Predicting credit card customer churn in banks using data mining. International Journal of Data Analysis Techniques and Strategies, 1(1), 4-28.

3.Lin, C. S., Tzeng, G. H., & Chin, Y. C. (2011). Combined rough set theory and flow network graph to predict customer churn in credit card accounts. Expert Systems with Applications, 38(1), 8-15.

4.AL-Najjar, D., Al-Rousan, N., & AL-Najjar, H. (2022). Machine learning to develop credit card customer churn prediction. Journal of Theoretical and Applied Electronic Commerce Research, 17(4), 1529-1542.

5.Merikoski, M., Viitala, A., & Shafik, N. (2018). Predicting and Preventing Credit Card Default.

6.Gaganis, C., Papadimitri, P., Pasiouras, F., & Tasiou, M. (2023). Social traits and credit card default: a two-stage prediction framework. Annals of Operations Research, 325(2), 1231-1253.