OPEN ACCESS JOURNALS
Indiana Journal of Multidisciplinary Research
Abbriviate Tittle- Ind J Multi Res
ISSN *(Online)*- 2583-3820
Journal Homepage Link- https://indianapublications.com/Journals/IJMR
DOI: https://doi.org/10.5281/zenodo.12667386

INDIANA PUBLICATIONS
PRODUCTIVE AND QUALITY RESEARCH

Research Article

Volume-04|Issue-03|2024

Banking Churn Prediction with Random Forest and Discrete Trees in ML

Varun P N[*1], Yaswanth V[2], Y Taraka Tapasvi[3], Vishant Kulkarni[4], Prof. Padmasree N[5]

[1,2,4,3,5]Computer Science Engineering RVITM, Bengaluru, India.

Article History
Received: 20.05.2024
Accepted: 05.06.2024
Published: 30.06.2024

Citation
Varun, P. N., Yaswanth, V., Tapasvi, Y. T., Kulkarni, V. & Padmasree, N. (2024). Banking Churn Prediction with Random Forest and Discrete Trees in ML. *Indiana Journal of Multidisciplinary Research*, *4*(3), 49-53.

Abstract: The project revolves around employing advanced Machine Learning techniques, particularly Random Forests and Discrete Trees, for predicting customer churn within the banking sector. By leveraging these sophisticated algorithms, the aim of the project is to develop a robust model that accurately identifies potential churners among bank customers. In the competitive landscape of the banking industry, customer retention is paramount. This project delves into predictive analytics to address this challenge, focusing on customer churn prediction. The utilization of Random Forests and Discrete Trees allows for the exploration of diverse data attributes, including transaction history, demographics, and account activities, aiming to uncover significant patterns and indicators associated with bank customer churn. The workflow involves meticulous data preprocessing, exploration, and feature engineering to ensure the dataset's readiness for model implementation. Through the utilization of Random Forests and Discrete Trees algorithms. The project seeks to harness the strengths of ensemble methods and decision trees to accurately predict and classify customers who might potentially churn. The project's success will contribute to enhancing banking strategies, enabling proactive identification of customers at risk of churning. Ultimately. The goal is to empower banks with a predictive tool that aids in the formulation and implementation of effective customer retention strategies, thereby reducing customer attrition and fostering sustained growth within the banking industry.

Keywords: Bank Churn , Machine learning, Random Forests, Discrete Trees , Banking Industry

INTRODUCTION

Churning refers to the process where a customer switches from one company to another. Not only does churning result in a loss of income, but it also has other negative effects on operations, particularly on Customer Relationship Management (CRM), which is crucial in banking. Establishing long-term relationships with customers is a priority for banks, and reducing churn can contribute to expanding their customer base. The service provider's challenges are found in the behaviour of the customer and their expectations. In the dynamic realm of schooling, the mixing of superior technology has become vital to meet the evolving desires of present- day lecture rooms. Recognizing the constraints of traditional attendance systems, this challenge proposes a comprehensive answer — the real-time student Attendance machine with Emotion and feedback evaluation.

This advanced knowledge is leading to changes in purchase behaviour. This is a big challenge for current service providers to think innovatively to reach their expectations. Private sectors need to recognize customers Liu and Shih strengthen this argument in their paper by indicating that increasing pressures on companies to develop new and innovative ideas in marketing, to meet customer expectations and increaseloyalty and retention. Some customers might be keeping their relationship status null that means they willkeep their account status inactive. By keeping this account inactive it might be the customer transferring their relationship with another bank. There are different types of customers are in the bank.

Banks encounter difficulties in delivering higher service quality to middle-class customers who demand lower fees, enhanced quality, and innovative policies. Meeting these expectations is essential for providingtimely and cost-effective services, which are key in preventing customer churn. Retaining existing customers is critical and necessitates efficient systems for early churn prediction. This paper suggests leveraging machine learning, specifically artificial neural networks (ANN), to forecast churn within the banking sector, thereby enhancing retention strategies.

Maintaining a balance between addressing customer needs and ensuring operational efficiency is key. By focusing on reliability and budget-friendly services, banks can reduce customer dissatisfaction and churn rates. In conclusion, addressing diverse customer needs, particularly in the middle-class segment, is vital for banks. Implementing predictive analytics can proactively identify churn risks, enhancing service quality and retention efforts.

Random forests offer a valuable feature selection indicator, with scikit-learn providing a variable that shows the relative important feature in the prediction. This score helps in selecting the most important features and excluding the least important ones for model building.

For regression and classification problems, a supervised machine learning technique called a decision

***Corresponding Author:** Varun P N

tree is commonly employed. This technique involves recursively partitioning the dataset into subsets based on the most significant feature at each node, creating a tree-like structure. The leaf nodes of the tree representeither numerical values (in regression) or classes (in classification). Decision trees are effective for variable selection and provide a clear understanding of the explanatory power of the variables within the dataset.

LITERATURE SURVEY

The telecommunications sector is placing a growing emphasis on customer retention over customer acquisition, acknowledging the reduced costs linked with keeping existing customers. Consequently, effective churn management becomes paramount, necessitating comprehensive customer analytics frameworks[1]. In the gaming industry, there is a rising trend towards implementing customer relationshipmanagement models to proactively manage churn issues. This strategic approach recognizes that preventingchurn is often more cost-effective than acquiring new customers, highlighting the critical importance of churn analysis in this sector. Churn prediction models in online games must maximize both accuracy and expected profit from churn prevention. Given the skewed distribution of customer lifetime value, focusing on loyal customers with tailored benefits and adjusting prediction thresholds for higher profitability are recommended strategies[2]. Churn prediction in online freemium games is crucial because of the significantrevenue impact caused by users leaving unexpectedly. A case study conducted on The Settlers Online demonstrated effective churn detection using different labelling approaches and machine learning algorithms.By extracting features from game data and utilizing predictive classifiers, high prediction accuracies were achieved, with random forests showing particularly strong performance. These findings offer valuable insights for game companies and researchers conducting similar churn prediction studies[3]. Customer churn has become a significant challenge across various industries, necessitating advanced customer behavior analysis using diverse data types. Hard data, derived from devices and programs, can be effectivelymodelled using supervised machine learning algorithms like decision trees. Conversely, soft data, which is more subjective in nature, can be modelled using unsupervised algorithms such as K-means clustering. By integrating these data types, industries like banking can enhance the development of dynamic and efficient customer relationship management systems[4]. In the online gaming market, the proliferation of free-to- play and service-based models has introduced challenges like customer churn. Predictive models leveraging behavioral data have proven effective in addressing churn. An approach centered on frequency analysis forfeature representation derived from login records has shown promise, surpassing traditional methods like RFM. Particularly, the time-frequency plane domain analysis holds potential for significantly boosting profits from retention campaigns in online gaming.[5]

MATERIALS AND METHODS

Problem Statement

Customer churn, also known as customer attrition, refers to the situation where customers discontinue their relationship with a business by no longer purchasing its products or services. The percentage of customers thatstop using a company's products or services during a given period is referred to as the customer churn rate. Addressing the causes of customer churn can significantly impact a company's bottom line and reputation.

Work Folw

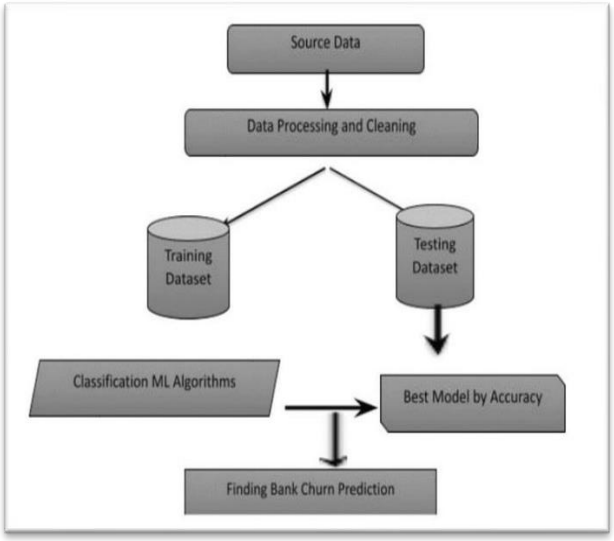


Figure 1: Work Flow

Problem Approach

1. Data Collection and Preprocessing: Collect the data of the bank customers in terms of the transactions and their interactions. Later, preprocess by cleaning, handling the missing values, selecting the features and normalization of the data.
2. Feature Engineering: Utilization of more advanced high-dimensional feature engineering used to identify the meaningful features from the bank dataset. It will analyze the crucial patterns and behavior's featuresthat determine the churn of the customers.
3. Model Development and Training: Utilize the machine learning models such as Random Forest and Discretetree which are tailored for the banking churn performing tasks. Establish the model
4. as mentioned above and train the model on the processed data. Cross-validation methods would be used to test and improve the model to get the best optimal-determined performance.
5. Model Evaluation and Comparison: After training the models, evaluate their performances by calculating theaccuracy, precision, recall, the F1-score, among others. Secondly, it would fine-tune and compare the performances of the different models to ascertain which model yields the best

***Corresponding Author:** Varun P N

results in predicting the churn in banking.

6. Interpretability Analysis: After training models to predict customer churn, interpretability analysis helps understand what explains why or how customers churn. After model deployment, the result of the interpretabilityanalysis goes to interpretation to explain the model's decisions.
7. Validation and Deployment: Finally, the trained model would be deployed and validated using new datasets or through online experiments. It can be deployed as a decision support system by the banking institutions to helpthem with real decision-making processes.
8. Finding the Best Model by Accuracy: After iterating through various models and testing them, choosethe model that yields the highest accuracy in predicting churn.

The technologies or libraries used in the design are:

Python: Python is a dynamically semantic high-level programming language that is object-oriented and interpreted. Its abundance of pre-built data structures, along with dynamic typing and dynamic binding, makeit an appealing language for both scripting and linking existing components and quickly developing new applications. Because of its straightforward and basic syntax, Python promotes readability, which lowers software maintenance costs. Python's support for packages and modules encourages code reuse and programmodularity. For all major platforms, the Python Interpreter and large standard library are freely redistributable and available for free in source or binary form.

Pandas: It's a Popular python based data analysis toolkit which can be imported using import pandas as PD. It represents a diverse range of utilities, ranging from parsing multiple file-formats to converting an entire data table into a NumPy matrix array.

NumPy: It is a fundamental Python library used extensively in data science for efficient array manipulation and numerical operations. It provides support for multidimensional arrays and matrices, along with a wide range of mathematical functions, making it essential for tasks involving linear algebra, Fourier transforms, and array-based computations.

Seaborn: It is a statistical data visualization library that operates on top of Matplotlib and integrates seamlessly with Pandas data structures. It simplifies the creation of complex visualizations by offering high- level abstractions and easy-to-use plotting functions. Seaborn is particularly useful for generating informative and visually appealing plots from Pandas Data Frames.

Matplotlib: It is a versatile Python library for creating static, interactive, and animated visualizations. It supports a diverse range of plot types, including histograms, scatter plots, bar charts, and more. Matplotlib is widely used in data analysis and presentation, providing researchers and data scientists with powerful tools for visualizing data effectively.

Scikit-learn: (or sklearn) is a comprehensive machine learning library for Python that offers a standardized interface for implementing various machine learning algorithms. It includes tools for tasks such as classification, regression, clustering, and dimensionality reduction. Scikit-learn is built on top of NumPy, SciPy, and Matplotlib, providing a user-friendly environment for building and evaluating machine learningmodels**.**

Keras: It's a neural network framework built on TensorFlow, CNTK, and Theano, providing a higher abstraction level for neural network development. Using Keras in deep learning allows for easy and fast prototyping as well as running seamlessly on CPU and GPU. This framework is written in Python code which is easy to debug and allows ease for extensibility.

FEATURES

The features used in the proposed approach are:

Customer Satisfaction Surveys: Customer satisfaction surveys and Net Promoter Score (NPS) ratings are essential tools for businesses to gauge customer sentiment and identify potential churn risks. These surveys provide direct feedback from customers regarding their experiences, preferences, and levels of satisfactionwith products or services. NPS specifically measures customer loyalty and likelihood to recommend a business to others. Low NPS scores or negative survey responses can indicate dissatisfaction and highlight customers who may be considering switching to competitors. By analyzing survey data, businesses can takeproactive steps to address issues, enhance customer experiences, and implement targeted retention strategies aimed at reducing churn and fostering stronger customer relationships.

Account Information: Details related to the customer's account, such as account type, account age, balance, transaction history, and the number of transactions, are important indicators of engagement andsatisfaction.

Customer Interaction: Features related to customer interactions with the bank, such as the frequency of visits to branches, calls to customer service, or engagement with online banking platforms, can help assess customer engagement and satisfaction levels.

Graphical User Interface (GUI): The system features a graphical user interface developed using the interlibrary,providing users with a visually appealing and interactive experience.

RESULTS AND ANALYSIS

***Corresponding Author:** Varun P N

RESULTS BY APPLYING CLASSIFIERS DIRECTLY

Classifier	Accuracy(%)	Accuracy After oversampling(%)
KNN	81.65	81.37
SVM	79.63	70.36
DT	78.99	91.98
RF	85.18	95.74

Following the completion of preprocessing, the data will be in its operational form. For the remaining study, the ten features derived from preprocessing are utilized. Of it, 70% will be utilized for training, while the remaining 30% will be randomly used for testing. In addition to the designated feature selection techniques, the classifiers themselves will be employed. Additionally, each model is assessed based on the accuracy thatresults from a 10-fold cross-validation. Additionally, a random confusion matrix for every model was created. When employing various feature selection techniques, classifiers function differently. The next paragraphs will cover the features used for each feature selection method as well as the specific classifier parameter information.

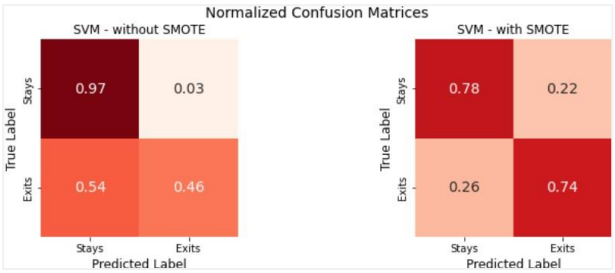
The k-value for KNN is set to 5. That is, while classifying the new data, the five closest neighbors are taken into account. Sometimes the accuracy increases when the number of neighbors is less than 5, and vice versa.But choosing fewer neighbors is not a smart idea because the data is being classified at random. However, the results notably diminish when the number of neighbors exceeds five. As a result, 5 is chosen as the value of k (where the accuracy and the change are optimum). Euclidean distance is also the unit of measurement employed. The linear kernel function (LSVM) is utilized for SVM. The forest's tree count in the case of RFis fixed at 100.

RESULTS AFTER RELIEFF FEATURE SELECTION

Classifier	Accuracy(%)	Accuracy After oversampling(%)
KNN	82.15	80.99
SVM	79.63	69.53
DT	77.61	90.74
RF	81.75	92.19

The above table displays the outcomes of several categorization strategies both with and without oversampling, or feature selection. The results demonstrate that the accuracy of the DT and RF classifiers rose as a result of oversampling, whereas KNN accuracy remained same and SVM accuracy decreased. These findings suggest that SVM is not a good choice for large data sets.

The top features selected by the relief technique are the quantity of items, age, balance, tenure, gender, and having a Cr Card. The accuracy of the different classifiers is shown in Table VI. While SVM accuracy remained unchanged, KNN accuracy increased when compared to KNN without feature selection. In contrast, DT and RF accuracy somewhat decreased in comparison to previous models.



The confusion matrix compares SVM and SVM with SMOTE for customer churn prediction. SVM without SMOTE correctly predicted 97% of staying customers but only 54% of existing customers. In contrast, SVMwith SMOTE improved predictions, with accuracies of 78% for staying and 74% for exiting customers.

SMOTE effectively addresses class imbalance by creating synthetic samples of the minority class, enhancingthe model's overall performance in predicting customer churn.

CONCLUSION

In conclusion, the banking sector, like any other industry, places increasing emphasis on customer engagement. Customer churn can significantly impact a bank's overall performance. Therefore, it is crucial for banks to detect potential customer churn early and take proactive measures to retain them. Various studies and methodologies exist to predict churn in the banking sector. By analyzing customer behaviour, transaction patterns, and other relevant data, banks can develop predictive models to identify customers at risk of leaving. These models enable banks to tailor retention strategies and offer targeted incentives to mitigate churn.

The objective of this research is to develop an optimal model for early client churn prediction in a bank. The study was limited by a small sample size of 10,000 records, which is relatively small compared to real commercial bank datasets. These limitations can be partially mitigated through oversampling techniques. The study evaluated the performance of RF, SVM, KNN, and Decision Tree classifiers under various conditions. When combining oversampling with the RF classifier, a superior accuracy of 95.74% was achieved. Notably, tree-based classifiers like Decision Tree and Random Forest are unaffected by feature selection methods, as reducing features tends to decrease their predictive performance. Additionally, oversampling negatively impacted SVM due to the dataset's inherent imbalance, making it challenging forSVM to handle effectively.

REFERENCES

1. Liu, D.-R., & Shih, Y.-Y. (2005). Integrating AHP and data mining for product recommendation based

***Corresponding Author:** Varun P N

on customer lifetime value. *Information & Management, 42*(3), 387–400. https://doi.org/10.1016/j.im.2004.01.008

2. Canning, G. (1982). Do a value analysis of your customer base. *Industrial Marketing Management, 11*(2), 89–93. https://doi.org/10.1016/0019-8501(82)90022-3
3. Chen, M.-S., Han, J., & Yu, P. S. (1996). Data mining: An overview from a database perspective. *IEEE Transactions on Knowledge and Data Engineering, 8*(6), 866–883. https://doi.org/10.1109/69.553155
4. Kim, M.-K., Park, M.-C., & Jeong, D.-H. (2004). The effects of customer satisfaction and switching barrier on customer loyalty in Korean mobile telecommunication services. *Telecommunications Policy, 28*(2), 145–159. https://doi.org/10.1016/j.telpol.2003.12.003
5. He, B., Shi, Y., Wan, Q., & Zhao, X. (2014). Prediction of customer attrition of commercial banks based on SVM model. *Procedia Computer Science, 31*, 423–430. https://doi.org/10.1016/j.procs.2014.05.279
6. Zoric, A. B. (2016). Predicting customer churn in the banking industry using neural networks. *Interdisciplinary Description of Complex Systems: INDECS, 14*(2), 116–124. https://doi.org/10.7906/indecs.14.2.5
7. Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in a big data platform. *Journal of Big Data, 6*(1), 28. https://doi.org/10.1186/s40537-019-0191-6
8. Raj, J., & Ananthi, V. (2019). Recurrent neural networks and nonlinear prediction in support vector machines. *Journal of Soft Computing Paradigm, 2019*, 33–40. https://doi.org/10.36548/jscp.2019.1.004
9. Tang, Y. (2013). Deep learning using linear support vector machines. *arXiv preprint arXiv:1306.0239*. https://doi.org/10.48550/arXiv.1306.0239
10. Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC Press. https://doi.org/10.1201/9781315139470
11. Kim, J. K., Song, H. S., Kim, T. S., & Kim, H. K. (2005). Detecting the change of customer behavior based on decision tree analysis. *Expert Systems, 22*(4), 193–205. https://doi.org/10.1111/j.1468-0394.2005.00304.x
12. Kandel, H. A. (2019). A comparative study of tree-based models for churn prediction: A case study in the telecommunication sector (Master's thesis). NOVA Information Management School, Campus de Campolide, Lisbon, Portugal. Available at: http://hdl.handle.net/10362/60302
13. Ali, Ö. G., & Aritürk, U. (2014). Dynamic churn prediction framework with more effective use of rare event data: The case of private banking. *Expert Systems with Applications, 41*(17), 7889–7903. https://doi.org/10.1016/j.eswa.2014.06.021

***Corresponding Author:** Varun P N