CUSTOMER CHURN PREDICTION

[1] B. Sai Vivek, [2] K. Abhishek Kumar, [3] T. Harsha Vardhan
Department of Computer Science and Engineering,
Amrita School of Computing,
Amrita Vishwa Vidyapeetham,
Bengaluru,Karnataka.
[1] bl.en.u4aie22107@bl.students.amrita.edu , [2] bl.en.u4aie22120@bl.students.amrita.edu , [3] bl.en.u4aie22159@bl.students.amrita.edu

***Abstract*— This Customer Churn Prediction is a vital strategy for businesses facing the challenges of customer attentions in today's highly competitive market. By using the latest and advanced machine learning algorithms like Logistic Regression, Random Forest and more companies can accurately forecast customer churn, enabling them to address customer intention and attrition, optimize resource allocations, improve the customer satisfaction, it also increases the company's revenue, and lead the company to the profits and gain a competitive advantage. This approach allows businesses to identify customers at risk of leaving, implement targeted retention strategies, and reduce the cost of hiring the new customers, ultimately fostering loyalty and long-term success in a great-evolving business landscape.**

I. INTRODUCTION

Churning refers to the process in which a customer leaves one company and switches to another. This not only results in a loss of income but also has negative implications for overall operations, particularly in terms of Customer Relationship Management (CRM). Establishing long-term relationships with customers is crucial for institutions as they aim to expand their customer base. Service providers face challenges related to customer behavior and their evolving expectations. The present generation, which is generally more educated than previous ones, has higher demands for connectivity, innovation, and diverse policy options. This advanced knowledge has led to changes in consumer purchasing behavior, presenting a significant challenge for service providers to think creatively and meet these expectations.

Customers may easily move their relationships from one bank to another. Some customers may keep their relationship status null, which signifies their account status is inactive. By leaving this account dormant, the consumer may be moving their connection to another bank. There are several categories of consumers in the bank. Farmers are one of the banks' most important customers; they may expect lower monthly charges because their income is modest. Businesspeople are also essential consumers since they do a large number of transactions with large sums of money. These consumers will anticipate higher levels of service excellence. Middle-class clients were one of the most significant segments; in almost every bank, these people outnumber other types of customers. These individuals will anticipate lower monthly fees, improved service quality, and new policies. Keeping multiple sorts of clients is therefore difficult. They must consider clients and their wants in order to overcome these problems and provide quality service on time and within budget to customers. Maintaining a strong working relationship with them is also a huge problem for them. If they do not overcome these eight difficulties, they may have churn. Recruiting a new client is more expensive and difficult than retaining existing consumers. Customers holding, on the other hand, are often more costly since they have already earned the trust and loyalty of existing customers. As a result, the requirement for a system that can successfully forecast client attrition in the early phases is critical for any banking institution.

II. PURPOSE

Churn prediction refers to the process of identifying customers who are likely to leave a company and switch to another. The purpose of using churn prediction is to proactively address customer attrition and reduce the negative impact it has on a company's operations, particularly in terms of Customer Relationship Management (CRM). The ability to establish and maintain long-term relationships with customers is crucial for businesses aiming to expand their customer base. In today's highly competitive market, customers have higher demands for connectivity, innovation, and diverse policy options. This, coupled with changes in consumer purchasing behavior, presents significant challenges for service providers. Customer churn not only results in a loss of income but also entails the expense and difficulty of acquiring new customers. It is more cost effective to retain existing customers by accurately predicting their likelihood of churn and implementing targeted retention strategies.

III. LITERATURE SURVEY

The paper "Customer Churn Prediction using Machine Learning: Subscription Renewal on OTT Platforms" by O.R. Devi, S.K. Pothini, M.P. Kumari, S.V., and U.N.S. Charan focuses on predicting customer churn in Over-The-Top (OTT) platforms using machine learning. The researchers employed various algorithms including Logistic Regression, Decision Tree, Random Forest, Support Vector Machines (SVM), and Artificial Neural Networks (ANN) to train and evaluate predictive models. The study utilized a dataset collected from an OTT platform, comprising customer information such as demographics, viewing patterns, and subscription details.

The evaluation of model performance involved metrics such as accuracy, precision, recall, and F1-score. However, the work had some limitations. The dataset used was historical, lacking real time data. The article did not provide details about the preprocessing techniques and feature engineering methods employed, which could impact model performance. Future research in this area could involve exploring advanced algorithms like Gradient Boosting and Deep Learning to further improve predictive accuracy. Incorporating external data sources, such as social media sentiment analysis, could enhance churn prediction models on OTT platforms [1]. The paper "Customer Churn Prediction Using Machine Learning Approaches" by R. Srinivasan, D. Rajeswari, and G. Elangovan addresses customer churn prediction using machine learning. The authors explored multiple algorithms, including Decision Tree, Random Forest, K-Nearest Neighbors (KNN), and Naive Bayes, to develop predictive models. The study utilized a dataset comprising customer information from a telecom company, including features such as customer demographics, service usage patterns, and billing details. Additionally, the authors did not mention the specific preprocessing techniques or feature selection methods used, which may affect the model's performance. Future research in this field could involve exploring more advanced machine learning techniques, such as Support Vector Machines (SVM) or Gradient Boosting, to enhance the accuracy of churn prediction models. Furthermore, incorporating additional features like customer sentiment analysis or social media data could provide valuable insights for better churn prediction [2]. The paper "Customer Churn Prediction Using Machine Learning: Commercial Bank of Ethiopia" by M.H. Seid and M.M. Woldeyohannis focuses on customer churn prediction in the context of the Commercial Bank of Ethiopia, utilizing machine learning techniques. The authors employed various algorithms, including Logistic Regression, Decision Tree, Random Forest, and Artificial Neural Networks (ANN), to develop predictive models. The study utilized a dataset obtained from the Commercial Bank of Ethiopia, consisting of customer-related information such as demographics, transaction history, and account details [3]. The paper "Customer Churn Prediction Using Machine Learning" by V. Agarwal, S. Taware, S.A. Yadav, D. Gangodkar, A. Rao, and V.K. Srivastav discusses customer churn prediction using machine learning techniques. The authors employed various algorithms, including Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, and Support Vector Machines (SVM), to build predictive models. To evaluate the performance of the models, the authors used metrics such as accuracy, precision, recall, and F1-score. One drawback identified in their work was the imbalance in the dataset, with a majority of customers being non-churners. This imbalance could affect the performance of the churn prediction models. In terms of future scope, the authors suggested exploring ensemble methods, such as stacking or bagging, to further enhance the accuracy of churn predictions of the models. Additionally, incorporating customer sentiment analysis or social media data could provide valuable insights for improved churn prediction [4]. The paper "Customer Churn Prediction using Machine Learning" by R.K. Peddarapu, S. Ameena, S. Yashaswini, N. Shreshta, and M. PurnaSahithi focuses on customer churn prediction using machine learning. The authors explored several algorithms, including Logistic Regression, Decision Tree, Random Forest, and Artificial Neural Networks (ANN), to develop predictive models. The study of a dataset obtained from a telecommunications company, comprising customer information such as demographics, usage patterns, and subscription details. A drawback mentioned in their work was the imbalance in the dataset, with a disproportionate number of non-churners compared to churners. This imbalance could affect the accuracy of churn prediction models. They also recommended incorporating additional features, such as customer feedback or sentiment analysis, for better churn prediction. Furthermore, considering temporal aspects of customer behavior and incorporating time- series analysis techniques could enhance the predictive capabilities of churn models [5]. The paper "A Survey on Customer Churn Prediction using Machine Learning and data mining Techniques in E-commerce" by P. Gopal and N.B. MohdNawi provides a comprehensive overview of customer churn prediction in e-commerce using machine learning and data mining techniques. The authors conducted a survey of various studies in the field to analyze the algorithms, datasets, metrics, drawbacks, and future scope. Different datasets from e-commerce domains were utilized in the surveyed studies, encompassing customer information, purchase history, demographics, and behavioral data. The evaluation metrics employed in the studies varied but commonly included accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC). The authors identified several drawbacks in the existing works, such as imbalanced datasets, limited feature selection techniques, lack of interpretability, and insufficient consideration of temporal dynamics. Future research opportunities were identified, including the exploration of ensemble models, deep learning techniques, and hybrid models. The authors also suggested incorporating external data sources like social media data and customer reviews to enhance churn prediction accuracy [6]. The paper “Customer Churn Prediction Using Machine Learning Methods: A Comparative Analysis" by H. Karamollaoğlu, İ. Yücedağ, and İ.A. Doğru presents a comparative analysis of machine learning methods for customer churn prediction. The authors investigated and compared the performance of various algorithms including Logistic Regression, Decision Tree, Random Forest, Support Vector Machines (SVM), and Artificial Neural Networks (ANN). The article highlighted certain drawbacks in their work, including the lack of consideration for imbalanced datasets, the absence of feature selection techniques, and limited exploration of advanced machine learning algorithms. These limitations could affect the accuracy and robustness of the churn prediction models. In terms of future scope, the authors suggested addressing the imbalanced dataset issue through techniques like oversampling or under sampling. They also recommended incorporating feature selection methods to improve model performance and exploring advanced algorithms such as Gradient Boosting and Deep Learning for enhanced churn prediction accuracy. Additionally, the integration of customer sentiment analysis or social media

data could provide valuable insights for more effective churn prediction models [7]. The paper "A Machine Learning Model for Customer Churn Prediction using Cat Boost Classifier" by J. Jane Rubel Angelina, S.J. Subhashini, S. Harish Baba, P. Dheeraj Kumar Reddy, P.V. Sudheer Kumar Reddy, and K. Sameer Khan focuses on customer churn prediction using the CatBoost classifier. The authors developed a machine learning model using the CatBoost algorithm and evaluated its performance for churn prediction. The study utilized a dataset obtained from a telecom company, which included customer- related information such as demographics, service usage patterns, and billing details. The dataset was divided into training and testing sets for model evaluation. To assess the performance of the model, the authors employed metrics such as accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC). Regarding future scope, the authors suggested exploring ensemble methods, such as combining multiple classifiers, to further improve the predictive accuracy of churn models. [8]. The paper "E-Commerce Customer Churn Prediction Scheme Based on Customer Behavior Using Machine Learning" by P. Nagaraj, V. Muneeswaran, A. Dharanidharan, M. Aakash, K. Balananthanan, and C. Rajkumar presents a customer churn prediction scheme for e-commerce based on customer behavior using machine learning techniques. The authors proposed a methodology that leverages machine learning algorithms to predict customer churn based on customer behavior patterns. The study utilized a dataset from an e-commerce platform, containing customer-related information such as purchase history, browsing behavior, and transaction details. The dataset was divided into training and testing sets for model evaluation. To evaluate the performance of the churn prediction scheme, the authors employed metrics such as accuracy, precision, recall, and F1-score. While the article did not explicitly mention drawbacks in their work, potential limitations could include issues such as data quality, class imbalance in the dataset, or the choice of machine learning algorithms, which may impact the accuracy and reliability of the churn prediction scheme. In terms of future scope, the authors suggested exploring the incorporation of more advanced machine learning algorithms such as deep learning or ensemble techniques to improve the accuracy of churn prediction [9]. The paper “A Smote-Based Churn Prediction System Using Machine Learning Techniques" by A.O. Akinrotimi, R.O. Ogundokun, M.A. Mabayoje, R.A. Oyekunle, and M.O. Adebiyi presents a churn prediction system utilizing machine learning techniques and the SMOTE (Synthetic Minority Over-sampling Technique) algorithm. The authors aimed to improve the performance of churn prediction models by addressing the issue of class imbalance in the dataset. The study utilized a dataset from a telecommunications company, containing customer-related information such as demographics, service usage patterns, and billing details. The dataset was preprocessed using the SMOTE algorithm to address class imbalance, ensuring a balanced representation of churned and non-churned customers. These factors may impact the accuracy and reliability of the churn prediction system. In terms of future scope, the authors suggested exploring ensemble methods or hybrid models to further improve churn prediction accuracy [10].

IV. PROPOSED METHODOLOGY

Data Validation/ Cleaning/Preparing Process:

Importing the library packages and loading the specified dataset. To investigate the variable. Identifying data by form and type, as well as analyzing missing and duplicate values. The methods and techniques for cleaning data will differ depending on the dataset. The primary goal of data cleaning is to detect and remove errors and anomalies to increase the value of data in analytics and decision making

Exploration data analysis of visualization:

Data visualization is an important skill in applied statistics and machine learning. Statistics does indeed focus on quantitative descriptions and estimations of data. Data visualization provides an important suite of tools for gaining a qualitative understanding algorithm or the same algorithm multiple times to form a more powerful prediction model. The random forest algorithm combines multiple algorithms of the same type i.e. multiple decision trees, resulting in a forest of trees, hence the name "Random Forest". The random forest algorithm can be used for both regression and classification tasks

Data Visualization:

Data visualization techniques such as histograms, scatter plots, and correlation matrices can be used to understand the distribution and relationships of the features in the datasets. This helps to identify patterns and trends in the data that can inform the development of churn prediction models.

Model Building and Evaluation:

The next step is to build churn prediction models using machine learning algorithms such as logistic regression, random forest, and gradient boosting. The models are evaluated using performance metrics such as accuracy, precision, recall, and F1-score. The best model is selected based on its predictive performance for churn prediction.

Implementation and Action:

The churn prediction model is integrated into the banking companies' systems to proactively identify and address potential customer churn. Customer retention strategies are developed and implemented based on the churn prediction insights. This can include targeted marketing campaigns, personalized offers, and improved customer service.

Continuous Monitoring and Improvement:

The performance of the churn prediction model is regularly monitored and updated as needed to adapt to changes in customer behavior and market dynamics. This involves ongoing data collection, preprocessing, and model building to ensure the model remains accurate and effective in predicting and preventing customer churn.

V. CONCLUSION

In this project, we explored the performance of various machine learning algorithms, including Random Forest, K-Nearest Neighbors (KNN), Support Vector Machines

(SVM), and Grid Search, in predicting customer churn. The results showed that Random Forest achieved the highest accuracy of 85.6%, demonstrating its robustness in capturing complex relationships and interactions between variables. While KNN and SVM also exhibited competitive accuracy rates of 83% and 85% respectively, the selection of the best algorithm should consider factors beyond accuracy alone, such as interpretability, computational complexity, and scalability, tailored to the specific needs and characteristics of the business. Additionally, the lower accuracy of Grid Search emphasizes the importance of further exploration and fine-tuning of hyperparameters to enhance the predictive power of the model. Overall, the findings highlight the superiority of Random Forest in this specific project, but the algorithm choice should be made based on a comprehensive assessment of various factors in the given business context.

VI. REFERENCES

[1] O. R. Devi, S. K. Pothini, M. P. Kumari, S. V and U. N. S. Charan, "Customer Churn Prediction using Machine Learning: Subcription Renewal on OTT Platforms," 2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC), Salem, India, 2023.

[2] R. Srinivasan, D. Rajeswari and G. Elangovan, "Customer Churn Prediction Using Machine Learning Approaches," 2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF), Chennai, India, 2023.

[3] M. H. Seid and M. M. Woldeyohannis, "Customer Churn Prediction Using Machine Learning: Commercial Bank of Ethiopia," 2022 International Conference on Information and Communication Technology for Development for Africa (ICT4DA), Bahir Dar, Ethiopia, 2022.

[4] V. Agarwal, S. Taware, S. A. Yadav, D. Gangodkar, A. Rao and V. K. Srivastav, "Customer - Churn Prediction Using Machine Learning," 2022 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS), Tashkent, Uzbekistan, 2022.

[5]R. K. Peddarapu, S. Ameena, S. Yashaswini, N. Shreshta and M. PurnaSahithi, "Customer Churn Prediction using Machine Learning," 2022 6th International Conference on Electronics, Communication and Aerospace Technology, Coimbatore, India, 2022.

[6] P. Gopal and N. B. MohdNawi, "A Survey on Customer Churn Prediction using Machine Learning and data mining Techniques in E-commerce," 2021 IEEE Asia- Pacific Conference on Computer Science and Data Engineering (CSDE), Brisbane, Australia, 2021.

[7] H. Karamollaoğlu, İ. Yücedağ and İ. A. Doğru, "Customer Churn Prediction Using Machine Learning Methods: A Comparative Analysis," 2021 6th International Conference on Computer Science and Engineering (UBMK), Ankara, Turkey, 2021.

[8] J. Jane Rubel Angelina, S. J. Subhashini, S. Harish Baba, P. Dheeraj Kumar Reddy, P.V. Sudheer Kumar Reddy and K. Sameer Khan, "A Machine Learning Model for Customer Churn Prediction using CatBoost Classifier," 2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2023.

[9] P. Nagaraj, V. Muneeswaran, A. Dharanidharan, M. Aakash, K. Balananthanan and C. Rajkumar, "E-Commerce Customer Churn Prediction Scheme Based on Customer Behaviour Using Machine Learning," 2023 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2023.

[10] A. O. Akinrotimi, R. O. Ogundokun, M. A. Mabayoje, R. A. Oyekunle and M. O. Adebiyi, "A Smote-Based Churn Prediction System Using Machine Learning Techniques," 2023 International Conference on Science, Engineering and Business for Sustainable Development Goals (SEB-SDG), Omu-Aran, Nigeria, 2023.