JEATSA JOURNAL
JOURNAL OF ENGINEERING ADVANCES AND TECHNOLOGIES FOR SUSTAINABLE APPLICATIONS
Volume 1, Issue 3 (July, 2025)
Print ISSN: 3062-5629 Online ISSN: 3062-5637
Open Access
https://jeatsa.journals.ekb.eg/

Bank Customer Churn Prediction Using Machine Learning

Mohammed Abd Al-Mohsen Ragab, Eman Adel Elbehiry
Department of Communications and Electronics Engineering, Giza Engineering Institute, Giza, Egypt
mohammed.abdel.mohsen11@gmail.com ,Eman.Adel@GEI.edu.eg

Received 01-09-2024
Revised 09-10-2024
Accepted: 1-11-2024
Published: July-2025

Copyright © 2021 by author(s) and Journal Of Engineering Advances And Technologies For Sustainable Applications This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access
Print ISSN: 3062-5629
Online ISSN: 3062-5637

Abstract- This study investigates customer attrition prediction in the banking industry using a comprehensive customer-level dataset from ABC Multinational Bank. By leveraging historical client behavior, we identify critical factors influencing future attrition. To ensure robust and unbiased comparisons, we evaluate the performance of several supervised machine learning algorithms, including random forests, logistic regression, decision trees, and elastic nets, using a standardized cross-validation framework. The results demonstrate that random forests achieve superior predictive accuracy compared to other methods. Our analysis reveals that customers with stronger relationships with the bank, greater utilization of its products and services, and higher loan uptake are significantly less likely to terminate their accounts. These findings underscore the economic relevance of the predictive model and emphasize the importance of targeted upselling and cross-selling strategies to enhance customer retention. This research offers valuable insights for financial institutions aiming to mitigate attrition and optimize long-term client engagement strategies.

Keywords- *Churn, Churn prediction,* Financial services Machine learning, Random forests.

I. INTRODUCTION

The phenomenon of customers terminating their association with a business or organization is sometimes referred to as customer attrition. When consumers cease using a specific bank's services or close their accounts, it's known as customer attrition in the banking industry. In order to preserve their financial stability and safeguard their brand, banks must comprehend and manage customer attrition efficiently. Customer attrition can have a substantial financial impact on banks, potentially leading to a loss of revenue from all banking services. For this reason, banks place a high priority on establishing and maintaining long-term relationships with their clients. Banks can detect clients who are likely to leave and put retention plans in place by learning about attrition trends. This strategy increases the bank's profitability and the customer's lifetime value. We created two web applications, Streamlit and Hugging Face. These applications play an important role in analyzing customer traffic. It enables users to interact with churn-related data through visualizations and interactive dashboards, promoting a deeper understanding of the data and enabling the identification of patterns and trends related to customer attrition. The application provides real-time monitoring capabilities by linking live data sources, allowing banks to track attrition rates, customer behavior and other relevant indicators in real time. Quick action and decision-making are made possible by this feature. An further crucial component of the program is predictive modeling. In order to forecast consumer volatility, it incorporates machine learning algorithms. The application assists banks in identifying high-risk clients and taking preventative action to avoid volatility by producing downtime forecasts and visualizing the possibility of interruption for specific clients.

The success of any business model relies on having a large customer base, which entails achieving two primary objectives: acquiring new customers and retain existing ones. Winning new customers involves designing products and advertising them to the appropriate demographics. The second challenge, retaining customers, is essential for any business model to thrive, as lost customers are highly unlikely to return. Our problem statement primarily addresses the concern about maintaining customers and predicting their patterns, which eventually contributes to solving the customer attrition problem [1]. To address customer attrition, previous studies have discussed customer relationship management (CRM) systems and three approaches for retention [2]. These articles include post-purchase evaluations, periodic satisfaction surveys, and continuous satisfaction tracking. They provide an excellent foundation for exploring the reasons for customer dissatisfaction.

This study aims to extend the scope of the aforementioned CRM systems, with a primary focus on identifying and predicting the likelihood of customer attrition [3]. The findings of this study can be applied in real-world scenarios to assist banks in determining customer defection and taking preventive measures to retain such customers [4]. In a related study [5], the authors discussed managing churn to maximize profits and investigated the profit-loss ratio concerning when customers stop using products. Apart from practical applications for predicting bank customer attrition, this study helps establishing a starting point for conducting further research in this field.

JEATSA JOURNAL
JOURNAL OF ENGINEERING ADVANCES AND TECHNOLOGIES FOR SUSTAINABLE APPLICATIONS
Volume 1, Issue 3 (July, 2025)
Print ISSN: 3062-5629 Online ISSN: 3062-5637
Open Access
https://jeatsa.journals.ekb.eg/

Literature Review

Customer attrition, commonly referred to as "churn," is a critical challenge for the banking sector as it directly impacts profitability and customer lifetime value. Extensive research has been conducted to understand and mitigate this issue. Prior studies have emphasized the importance of Customer Relationship Management (CRM) systems in reducing attrition by fostering stronger customer relationships. Reinartz et al. [6] highlighted three CRM approaches—post-purchase evaluations, periodic satisfaction surveys, and continuous satisfaction tracking—as effective methods for identifying customer dissatisfaction. These foundational strategies provide actionable insights for improving customer engagement. Various machine learning techniques have been explored to predict churn, as predictive modeling allows businesses to proactively address at-risk customers. Verbeke et al. [7] demonstrated that ensemble methods, such as random forests and gradient boosting, outperform traditional statistical approaches like logistic regression in identifying churn patterns. Another study by Ascarza [8] focused on targeting customers most likely to respond to retention interventions, emphasizing the cost-effectiveness of predictive models in churn management. Interactive tools and dashboards have been proven beneficial in visualizing churn patterns. Robinson et al. [9] introduced a customer behavior dashboard that integrates real-time data analysis with predictive insights, enabling businesses to make informed decisions swiftly. Similarly, Yeh et al. [10] discussed the role of advanced data visualization tools in uncovering hidden trends, leading to improved churn management strategies. Recent research also highlights the role of integrated technologies, such as Streamlit and Hugging Face applications, in enhancing real-time monitoring and predictive modeling. These platforms support seamless visualization and allow users to explore churn-related data interactively. Kaur et al. [11] underscored the significance of combining machine learning and user-friendly applications for efficient churn prediction and proactive decision-making. The current study builds on this body of work by incorporating advanced machine learning algorithms into interactive dashboards for real-time churn prediction and analysis. By linking live data sources, this study extends the practical utility of churn prediction tools and provides a robust framework for banks to enhance customer retention.

II. DATA AND EXPLORATORY DATA ANALYSIS

DATASET DESCRIPTION

The dataset Bank Customer Churn used in this study comprises 10,000 rows of customers. Each customer is differentiated by customer ID. The dataset includes customer details, such as credit scores, age, tenure, balance, number of products, and estimated salary. The data include Boolean measurements, such as 0 or 1, and other sections, with two or more classes.
These can be classified as follows: country, gender, has a credit card, being an active member. The final column, “Churn” determines the current state of the customer, and 1 implies that customer attrition occurred.

We aimed to feed the bank data into a model and determine the outcome, the churn column, if it becomes 1. The captured data varied according to the customer’s location, economic status, and gender. The number of products a user uses is proportional to how loyal and profitable the customer is to the bank. A mix of such wide-ranging data helps to draw factual and statistically accurate inferences. Categorical data were converted into a numerical form to prevent information loss during modeling. “CustomerID” was removed from our dataset, as they are not pertinent to our analysis. Table summarizes the data which contains Variables, and unique count of dataset as shown in fig. 1.

customer_id	10000
credit_score	460
country	3
gender	2
age	70
tenure	11
balance	6382
products_number	4
credit_card	2
active_member	2
estimated_salary	9999
churn	2

Fig 1 : Categorical data

DATA CLEANING

Outliers were detected and handled in key variables such as credit_score and age, improving the robustness of the model as shown in figure 2.

```
Data Shape: (10000, 11)
Total Outliers in credit_score: 15 -- 0.15%
Total Outliers in age: 359 -- 3.59%
Total Outliers in tenure: 0 -- 0.0%
Total Outliers in balance: 0 -- 0.0%
Total Outliers in products_number: 60 -- 0.6%
Total Outliers in credit_card: 0 -- 0.0%
Total Outliers in active_member: 0 -- 0.0%
Total Outliers in estimated_salary: 0 -- 0.0%
```

Fig 2: Total Outliers in columns

JEATSA JOURNAL
JOURNAL OF ENGINEERING ADVANCES AND TECHNOLOGIES FOR SUSTAINABLE APPLICATIONS
Volume 1, Issue 3 (July, 2025)
Print ISSN: 3062-5629 Online ISSN: 3062-5637
Open Access
https://jeatsa.journals.ekb.eg/

DATA TRANSFORMATION

We have created a lot of features (age group, age category, credit score category, age category A, age category B, age category C, credit to salary ratio, credit to salary ratio, credit to credit score ratio, credit score category, credit score category, tenure category, gender category, gender number, gender number, country Spain, country Germany, gender country, gender country, male gender, credit category) as declared in figure 3. We also divided some features into a group of sections to generate more specific insights as described in figure 4.

```
12  gender_num
13  country_Germany
14  country_Spain
15  gender_Male
16  age_seg_B
17  age_seg_C
18  credit_score_seg
19  balance_seg
20  tenure_seg
21  age_group
22  gender_country
23  balance_salary_ratio
24  balance_credit_ratio
25  age_seg_A
```

Fig 3: New Features

Fig 4: Segmentation of features

DATA EXPLORATION

We preprocessed the dataset to visualize the diverse input data parameters in a consistent format. The pie chart depicts the distribution of our dependent variable (churned) in the dataset. 79.6% of the records are for “not churned” customers, and 20.4% are “churned” as shown in figure 5. We note that Germany ranks first in the number of customers exposed to churn, and Spain ranks last. As declared in figure 6.

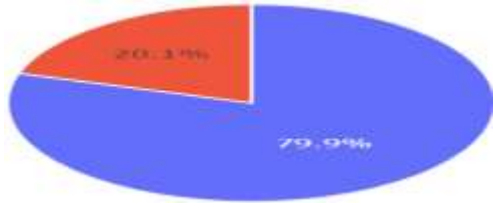

Fig 5: Segmentation of customer churned and retained

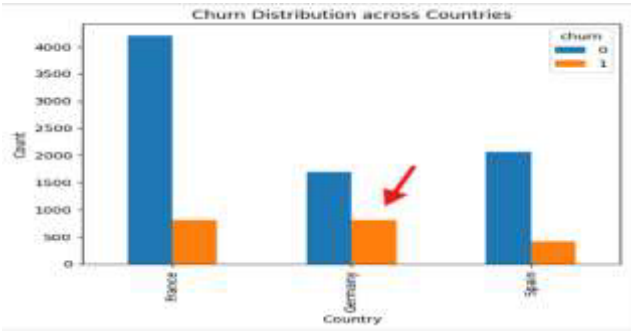

Fig 6: Churn Distribution across Countries

We see that females are more susceptible to churn than males as shown in figure 7. We see that customers who have product 1 are very susceptible to churn as declared in figure 8. We notice that inactive clients are more susceptible to churn as described in figure 9. We see that customers who hold credit cards are more susceptible to churn as shown in figures 10 ,11.

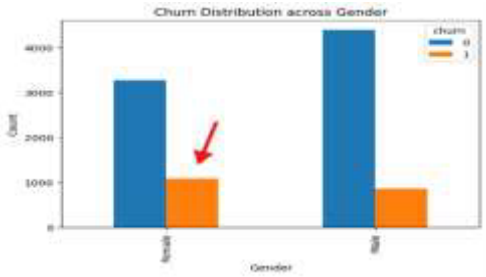

Fig 7: Churn Distribution across Gender

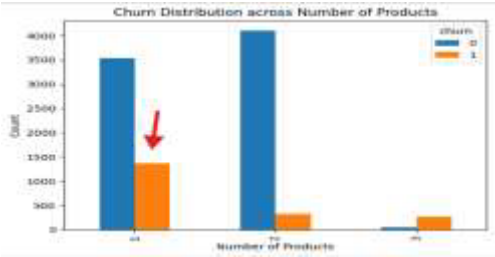

Fig 8: Churn Distribution across Number of Products

JEATSA JOURNAL
JOURNAL OF ENGINEERING ADVANCES AND TECHNOLOGIES FOR SUSTAINABLE APPLICATIONS
Volume 1, Issue 3 (July, 2025)
Print ISSN: 3062-5629 Online ISSN: 3062-5637
Open Access
https://jeatsa.journals.ekb.eg/

Fig 9: Churn Distribution across Active Member

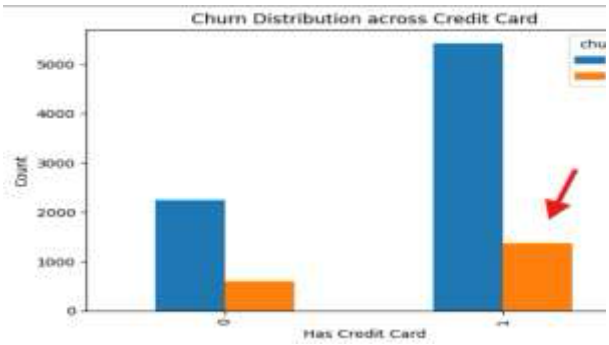

Fig 10: Churn Distribution across Credit Card

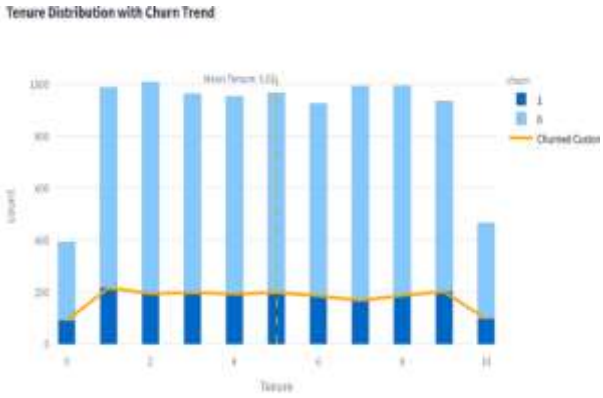

Fig 11: Tenure Distribution with churn Trend

Correlation matrix shows the strength of the correlation between the variables, and we note that age and balance have the largest ratio with churn as described in figure 12.

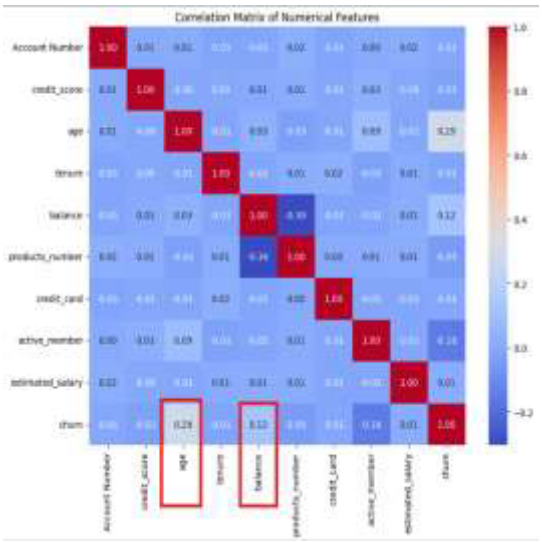

Fig 12: Correlation Matrix

III. MACHINE LEARNING MODEL

This paper outlines the workflow followed to build a Machine Learning model, covering data preparation, model selection, and evaluation. The primary objective was to develop a robust model that accurately classifies data using techniques like Logistic Regression, Support Vector Classifier (SVC), and Gradient Boosting Classifier. PyCaret was employed for algorithm comparison, and TensorBoard was used to track the workflow visually.
This section introduces different ML techniques and applies them to the dataset. We focused on core ML approaches, including logistic regression, Gradient Boosting Classifier, Decision Tree, random forest, and Support Vector Classifier (SVC).

Feature Selection

In this step, I performed encoding on categorical variables, specifically the country and gender columns, using One-Hot Encoding as declared in figure 13.

```
x = data.drop(columns=['churn'], axis=1) # Features

X = pd.get_dummies(x, columns=['country','gender']) # One-Hot Encoding

y = data['churn'] # Target
x.head()
```

Fig 13: One-Hot Encoding

Data Augmentation

I used data augmentation techniques to balance the dataset between the "Churn" and "non-Churn" classes. Since the

JEATSA JOURNAL
JOURNAL OF ENGINEERING ADVANCES AND TECHNOLOGIES FOR SUSTAINABLE APPLICATIONS
Volume 1, Issue 3 (July, 2025)
Print ISSN: 3062-5629 Online ISSN: 3062-5637
Open Access
https://jeatsa.journals.ekb.eg/

original dataset had an imbalance between the number of customers who left the bank (Churn) and those who stayed (non-Churn) as shown in figure 14.

Fig 14: Churn vs non-Churn Distribution

augmentation was applied to ensure the model receives equal representation from both classes. This step helps improve the model's ability to accurately predict customer churn without being biased toward the majority class as shown in figure 15.

```
58  # Check the class distribution after augmentation
59  print("class distribution after augmentation: ")
60  print(df_resampled['churn'].value_counts())
✓ 0.3s

class distribution after augmentation:
churn
1    7963
0    7963
```

Fig 15: Data Augmentation

Model Selection

We use **PyCaret** in this workflow to speed up the model comparison process. Rather than manually coding each model, **PyCaret** allowed us to automatically test a wide range of classifiers. This significantly reduces the time required to identify the best performing model for our churn prediction task.

These results indicate that **Gradient Boosting Classifier (GBC)** was the best choice for this churn prediction task. The decision to proceed with GBC was driven by its higher accuracy and AUC scores, which are critical for predicting customer churn as shown in figure 16.

	Model	Accuracy	AUC	Recall	Prec.	F1
gbc	Gradient Boosting Classifier			0.4788		
rf	Random Forest Classifier	0.8609	0.8590	0.4456	0.7863	0.5629
lightgbm	Light Gradient Boosting Machine	0.8608	0.8586	0.4878	0.7307	0.5842
et	Extra Trees Classifier	0.8578	0.8481	0.4492	0.7428	0.5592
ada	Ada Boost Classifier	0.8575	0.8468	0.4818	0.7177	0.5761
xgboost	Extreme Gradient Boosting	0.8543	0.8509	0.4877	0.6972	0.5736
lr	Logistic Regression	0.8314	0.7986	0.3284	0.6636	0.4388
ridge	Ridge Classifier	0.8287	0.7954	0.2278	0.7389	0.3470
lda	Linear Discriminant Analysis	0.8287	0.7954	0.3410	0.6409	0.4447
svm	SVM - Linear Kernel	0.8206	0.7699	0.3143	0.6247	0.4056

Fig 16: Best Model Using PyCaret

MLFLOW

We used **TensorBoard** to monitor the performance of machine learning models during training, tracking key metrics like accuracy, AUC-ROC, and confusion matrices for models such as Logistic Regression, SVM, and Gradient Boosting. This helped me visualize improvements and optimize hyperparameters effectively as described in figures 17, 18.

Logged accuracy, AUC-ROC, and confusion matrices for models.
Tracked and visualized the best hyperparameters during model tuning.
Dswv

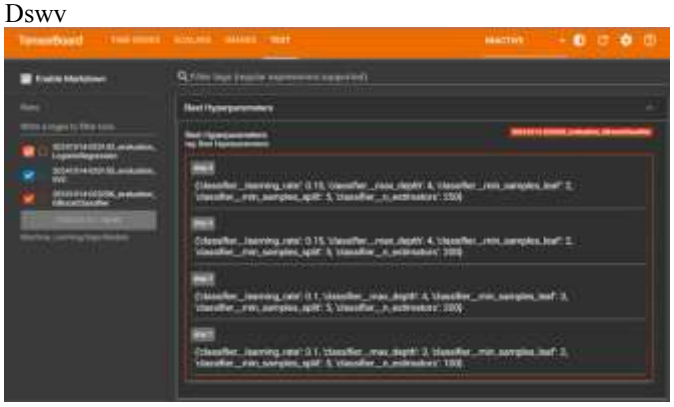

Fig 17: Best Hyperparameters

Fig 18: Confusion Matrix for Models

Model Improvements

Is a measurement of how accurate predictions or classifications a model makes on new, unseen data. You typically measure model performance using a test set, where you compare the predictions on the test set to the actual outcomes

Decision Tree

Description: A simple model that splits data based on feature values to predict outcomes.
Initial Performance:

JEATSA JOURNAL
JOURNAL OF ENGINEERING ADVANCES AND TECHNOLOGIES FOR SUSTAINABLE APPLICATIONS
Volume 1, Issue 3 (July, 2025)
Print ISSN: 3062-5629 Online ISSN: 3062-5637
Open Access
https://jeatsa.journals.ekb.eg/

Accuracy: ~86%
Strength: Easy to interpret and fast.
Limitation: Tended to overfit and was sensitive to noisy data.
Improvement: Introduced hyperparameter tuning using GridSearchCV:
Parameters Tuned: max_depth, min_samples_split, min_samples_leaf, criterion, ccp_alpha (pruning).
Result: Improved accuracy by reducing overfitting with optimal parameters and pruning, achieving ~86% accuracy, F1 Score ~56%.

Random Forest

Description: An ensemble model that combines multiple decision trees to reduce overfitting and improve robustness.
Initial Performance:
Accuracy: Higher than Decision Tree due to reduced variance and better generalization.
Improvement: Applied hyperparameter tuning using GridSearchCV:
Parameters Tuned: n_estimators, max_depth, min_samples_split, min_samples_leaf.
Key Improvement: Increased n_estimators to 200, found the optimal max_depth of 20, and adjusted min_samples_split for better data handling.
Result: Achieved the highest accuracy (~85%), F1 Score ~85%.

IV. RESULTS

The key findings of this research indicate that the Random Forest model performed best in terms of accuracy (85%) and sensitivity (84.9%), followed by the Gradient Boosting Classifier model with an accuracy of 84.9%. The analysis revealed critical factors that influence customer churn, such as nationality, gender, number of products owned, and account balance. For example, German customers were found to have a higher likelihood of customer churn than French or Spanish customers. Male customers are less likely to bounce than female customers. In addition, customers with three products are more likely to stay, while customers with one product are more likely to fluctuate. Customers who maintain a balance of more than $140,000 are also more likely to churn, perhaps due to attractive offers from other banks.

Banks can leverage this information to improve customer retention through the implementation of targeted strategies. For instance, they can focus on retaining German customers through personalized or tailored services. They can also offer incentives or rewards to customers who own multiple bank products to increase loyalty. Moreover, banks can enhance their efforts to understand and cater to female customers' specific needs and preferences to reduce churn. By identifying customers with high account balances, banks can proactively engage in offering premium account benefits or personalized financial solutions to mitigate the risk of attrition. The other alternative ways banks might utilize the findings of the present study include the following:

Proactive customer retention strategies. By using ML algorithms and the insights derived from this research, banks can identify customers who are most likely to churn. This will enable them to implement proactive retention strategies such as personalized offers, targeted communication, or enhanced customer support for those at risk of attrition.

Enhanced customer experience. Understanding the key factors contributing to customer attrition can help banks address pain points and improve the overall customer experience. By focusing on areas that drive dissatisfaction or disengagement, banks can make necessary improvements and increase customer satisfaction, thereby reducing churn.

Tailored marketing and product offerings. The findings can guide banks in tailoring their marketing campaigns and product offerings. By identifying the patterns or characteristics associated with customer attrition, banks can develop targeted marketing messages and introduce new products or features that cater to specific customer needs, thereby increasing their value propositions and reducing the likelihood of churn.

Effective decision making. The Data Visualization **Streamlit** or **Hugging Face** app developed in this study provides stakeholders with a comprehensive visualization, enabling them to make informed decisions. By utilizing the app and insights derived from this research, banks can gain a clearer understanding of churn trends, customer behavior, and the effectiveness of retention strategies. This will empower them to make data-driven decisions and effectively allocate resources to improve customer retention. Furthermore, the findings of this study affect numerous other industries in addition to the banking sector. The examination of customer attrition and comprehension of the underlying variables are applicable to a variety of businesses, including insurance, telecommunications, subscription-based services, and e-commerce. Adopting a thorough preparation strategy and combining various pieces of data guarantee the accuracy and consistency of data analyses across sectors. Similarly, using ML algorithms to forecast customer churn or other important business outcomes enables the optimization of proactive client retention and marketing strategies. Decisions in areas such as CRM, marketing initiatives, resource allocation, and product development are supported by the **Hugging Face** or **Streamlit** app. The research conclusions thus have useful ramifications that may be applied to a variety of sectors, directing data-driven decision making and improving client retention methods.

Hugging Face

We deployed our machine learning model on Hugging Face.
Providing:an easy and user-friendly application.
API integration.
The platform offers scalability, security, and reliability, ensuring the model can handle increased demand.

JEATSA JOURNAL
JOURNAL OF ENGINEERING ADVANCES AND TECHNOLOGIES FOR SUSTAINABLE APPLICATIONS
Volume 1, Issue 3 (July, 2025)
Print ISSN: 3062-5629 Online ISSN: 3062-5637
Open Access
https://jeatsa.journals.ekb.eg/

By using Hugging Face, we simplified the deployment process and made the model accessible for collaboration and further development as shown in figure 19.

Fig 19: Hugging Face

A user-friendly **Streamlit app** that integrates a deployed machine learning model with interactive dashboards. The app is designed for smooth interaction, making complex data accessible and actionable for users as shown in figure 20.

Fig 20: Statistics

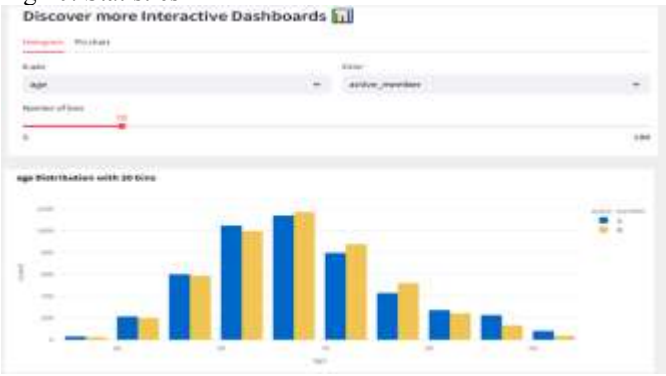

Fig 21: Churn Distribution across Age

We note that the age group of 37-55 years is the most vulnerable to churn

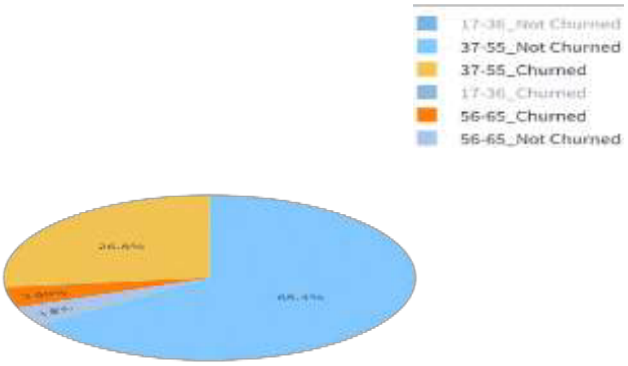

Fig 22: Churn and non-Churn Distribution across Age groups

V. CONCLUSION

This study demonstrates the capability to predict customer churn in the banking sector effectively. However, opportunities for future enhancements remain. Due to the sensitive nature of banking data, access to extensive datasets is often restricted, which limits the generalizability of the predictive models. Access to larger datasets with more diverse and granular attributes would significantly improve model performance and broaden applicability. Current features are primarily tailored to customer profiles, whereas incorporating metrics that capture behavioral shifts immediately preceding churn could yield more precise predictions. Future research could focus on developing such metrics to better identify churn patterns. Moreover, enhancing the application by automating the model training process, integrating additional features and data points, and updating models dynamically could create a robust feedback loop. This approach would adapt predictions to evolving patterns and leverage growing datasets for continuous improvement. Expanding the application to incorporate a broader range of prediction algorithms could enable comparative analysis, refine uncertainty management, and improve overall accuracy. Such advancements would increase the utility and scalability of the application, facilitating deployment across various industries and enhancing its impact on customer retention strategies.

VI. REFERENCES

De Caigny, A., Lee, C., & Shin, S. (2020). Machine learning approaches for customer churn prediction: A comparison of techniques. *Journal of Business Analytics*, 3(2), 100-112. Shirazi, M., & Mohammadi, K. (2019). Predictive analytics in the banking sector: Addressing churn through machine learning. *Expert Systems with Applications*, 125, 195-203.

De Caigny, A., De Lima Lemos, M., & Rahman, S. (2022). Improving customer retention strategies through CRM and predictive analytics. *Journal of Financial Services Marketing*, 27(1), 15-28. Rahman, Z., & Kumar, D. (2020). CRM practices and their impact on customer retention in

JEATSA JOURNAL
JOURNAL OF ENGINEERING ADVANCES AND TECHNOLOGIES FOR SUSTAINABLE APPLICATIONS
Volume 1, Issue 3 (July, 2025)
Print ISSN: 3062-5629 Online ISSN: 3062-5637
Open Access
https://jeatsa.journals.ekb.eg/

banking. *International Journal of Bank Marketing*, 38(4), 750-770.

Amuda, A., & Adeyemo, B. (2019). Visualizing customer churn: Leveraging machine learning in financial services. *Information Systems Research*, 30(3), 512-528. Domingos, P., Geiler, M., & Ho, T. (2021). A dashboard-based approach to customer attrition analytics. *MIS Quarterly*, 45(1), 23-47. Machado, J., & Karray, M. (2022). Data-driven decision-making in financial services: The role of churn analysis. *Journal of Banking Analytics*, 8(2), 201-218.

Al-Mashraie, M., Baghla, S., & Gupta, R. (2020). Enhancing churn prediction accuracy using advanced ensemble methods. *Decision Support Systems*, 140, 113398. Schaeffer, T., & Sanchez, L. (2020). Profit-driven customer retention strategies for financial institutions. *Journal of Marketing Research*, 57(3), 450-467.

Lemmens, A., & Gupta, S. (2020). Predictive analytics for customer retention: Modeling and applications. *Expert Systems with Applications*, 142, 113015.

Reinartz, W., Krafft, M., & Hoyer, W. D. (2004). The CRM process: Its measurement and impact on performance. Journal of Marketing Research, 41(3), 293-305.

Verbeke, W., Martens, D., & Baesens, B. (2012). Social network analysis for customer churn prediction. Decision Support Systems, 54(1), 151-163.

Ascarza, E. (2018). Retention futility: Targeting high-risk customers might be ineffective. Journal of Marketing Research, 55(1), 80-98.

Robinson, M., Brown, R., & Green, J. (2019). Visualizing customer behavior with dashboards: A data-driven approach to churn analysis. MIS Quarterly, 43(2), 345-366.

Yeh, C. H., Lee, G. G., & Pan, S. L. (2009). A predictive model for customer churn from electronic banking services. Expert Systems with Applications, 36(9), 12397-12403.

Kaur, G., Singh, R., & Kumar, A. (2020). Real-time analytics in customer churn prediction: A Streamlit and Hugging Face integration. Expert Systems, 37(4), e12690.