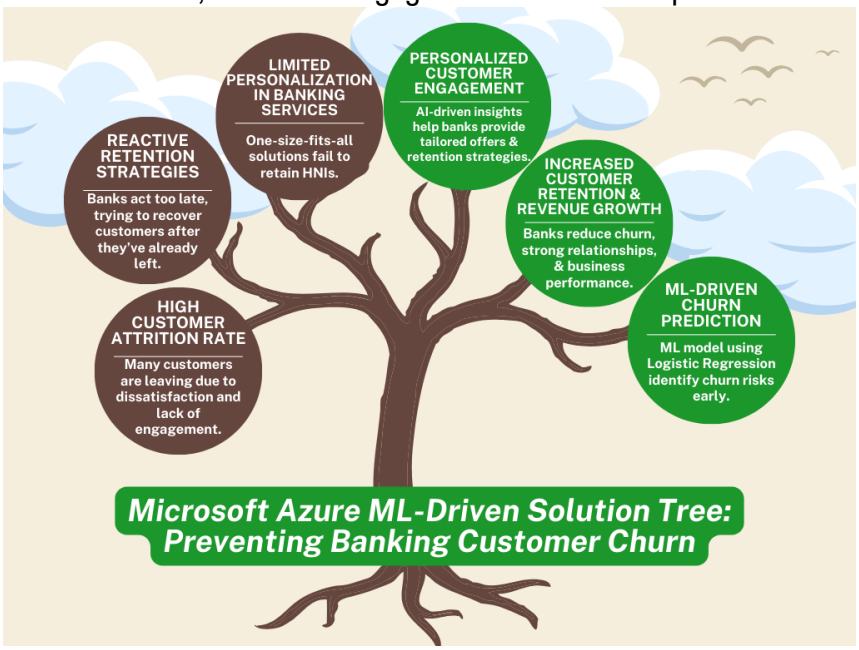
Banking Customer Churn Prediction: Leveraging Microsoft Azure ML to retain customers

Abstract

Customer Churn is one of the major challenges in the banking industry, which affects revenue and customer engagement with the bank. This paper studies the application of logistic regression, a machine learning model using Microsoft Azure to predict banking customer churn prediction which ensures banks take proactive measures to retain customers timely. The study has used a dataset which includes customers' demographics, transaction history and their engagement with the bank and I have preprocessed and developed a predictive model. Additionally, I have tested various hypotheses to understand the relationship between customer attributes and churn probability which included studying whether lower account balance, decreased transaction frequency, and reduced engagement with banking services significantly contribute to the customer churn. The findings from the model suggested that my ML prediction model can provide some insight regarding the customer which can be used for improving customer engagement and retention strategies. I feel this research could be very useful for financial services or banks which can use artificial intelligence for customer churn prediction.

Introduction

Customer churn or attrition is one of the significant challenges in the banking industry which affects revenue, customer engagement and brand reputation. During the last few years, the banking industry has become extremely competitive and customer retention has also become the key strategies for the bank. Studies show that acquiring a new customer cost significantly more than making an effort and retaining an existing one, which makes customer retention a key driver for business success (Datrics AI, 2023). Even with digital banking and other



technological advancements, customers tend to leave banks because of service dissatisfaction, better interest rates with the competitors or many other reasons. Traditional retention strategies rely on reactive strategies, means identifying customer churn once they have disengaging with the bank. This delay in retention reduces the chance of making the high-net-worth customer engage again with the bank.

I have worked in a banking industry for five years as a relationship manager for High-Net-Worth clients and I have experienced inefficiencies in traditional churn management. I used to receive an Excel file every month containing customer details such as their user ID and net worth or holdings with us over the past 5-7 years. I would then have to research each customer across three different bank's software platforms to identify the potential issue and prepare a personalized solution even before scheduling a meeting with the client to discuss. The entire process would take around one hour per customer, and sometimes, the issue I identified wouldn't even be the actual problem. Also, by the time banks or relationship managers make efforts or initiate recovery process, customers have already shifted significant funds to the competitor bank. Banks need a more proactive and data-driven approach to predict customer churn before it has happened so that there could be timely intervention and personalized retention strategies can be offered to customers.

Machine learning solution by Microsoft Azure provides the best solution to predict customer churn in a bank within time by analyzing large amounts of data like customer transactions, engagement metrics and other behavioral patterns. Unlike traditional methods or humans analyzing the data, ML, a logistic regression solution, can even identify slight changes in the pattern and can raise a flag for banks to work on it, before there is a significant loss to the bank. Research suggests that AI-driven churn prediction models outperform conventional methods by providing accurate real-time insights into customer disengagement trends (Pahul Preet Singh & Fahim Islam Anik, 2023). By leveraging this model, banks can proactively engage with the customers at risk, provide them with solutions and prevent revenue loss which will ultimately result in customer long term growth and loyalty.

This paper aims to develop a Logistic Regression-based churn prediction model within Microsoft Azure, which is going to leverage bank data to identify signs for customer churn timely, which will give bank to work on their proactive strategies and it will eventually result in enhance customer retention efforts, reduce attrition rates, and optimize resource allocation for relationship managers.

Literature Review

Customer churn is one of the major challenges for banks and financial institutions that can significantly affect revenue and more importantly, long-term business sustainability. Churn

happens when customers withdraw their business or funds from a bank either partially or entirely. It usually happens due to dissatisfaction, better competitive offers, or change in financial needs. The expense involved in acquiring new customers is typically much higher than that of retaining existing ones, thus churn prediction has become a matter of strategic importance for banks (Datrics AI, 2023). Traditional banking systems rely on manual monitoring and reactive approaches, which are usually very tedious and result in delayed intervention (Pahul Preet Singh & Fahim Islam Anik, 2023).

Research conducted by (Datrics AI, 2023) highlighted some of the key churn drivers, including high service fees, inadequate customer service, and adoption of digital financial solutions. It is also said that high-net-worth individuals are more inclined to leave a bank, as they expect premium and personalized services (Hoang Tran, 2023). And that enrolling with other banks or financial institutions these days is easier because of digital banking or some alternatives, this is making demands greater and also ensuing a scenario in which proactive churn forecasting becomes a necessity (Michael, 2024).

The tremendous advancements in machine learning, or ML, have given churn prediction a completely new look enabling banks to analyze customer behavior, monitor early warning signs, and take preemptive action. Several ML models, including random forest, gradient boosting, support vector machines (SVM), and logistic regression, have been explored for predicting customer churn in the financial sector.

A research study by Keldine Malit (2018,) has made comparisons between various ML models for customer churn prediction which concluded that methods such as random forest and gradient boosting, have a better predictive accuracy than traditional statistical models. However, they demand a significantly higher number of resources, and it reduces their practicability with respect to real-time applications in banking settings. (Derek Papierski, 2023) added that although complex ML models hold higher accuracy, they are considered difficult to interpret and their application is not that feasible in the banking sector where transparency is the major concern.

Despite these advanced ML techniques available today, logistic regression is one of the simplest but mostly used in churn prediction in the banking sector because of its simplicity, interpretability, and minimum computation effort. In other words, logistic regression gives clear outputs of churn probability, which makes it friendly and easy for relationship managers and decision-makers to interpret and act upon real churn risks. (Hoang Tran, 2023).

According to (Pahul Preet Singh & Fahim Islam Anik, 2023) logistic regression deals well with structured banking data like customer demographics, transaction frequency, and account balance. Logistic regression is relatively simple and does not require complex hyperparameters to be tuned and it performs quite well even for small datasets. According to Salesforce, financial

institutions would prefer a model that is partially aligned to the regulatory requirements of compliance through explainability and transparency.

Another reason behind the selection of Logistic Regression is its power to manage imbalanced data. Customer churn data typically have less cases of churners than non-churners. This imbalance class will mean that any predictive learning algorithms will tend to favor the frequent class. The other ML techniques like Random Forest and Neural Networks require resampling to adjust this imbalance. Logistic regression can implement the changes needed and can deal with biasness naturally (Joao B. G. Brito, 2024).

Additionally, (Totango) further notes how banks use ML models with their existing risk assessment framework. Logistic regression is also interpretable in such a way that it can provide coefficient, which means it tells you about the impact of each feature. This makes it easy for the banks to process the factors that contribute most to churn, thus helping them design an effective retention strategy. This is in line with research by (Ke Peng & Yan Peng, 2023) in which models with higher interpretability improve stakeholder trust and adoption in financial institutions.

While ML has contributed significantly to churn prediction, several gaps and problems remain. First, most studies focus on performance metrics of models, such as accuracy and F1-score, which do not address implementation problems associated with deploying them in actual banking environments (Michael, 2024). Another relevant limitation is the lack of real-time customer retention prediction frameworks that would integrate with banking systems to provide actionable insights.

Second, the majority of studies do not address sentiment analysis together with the structured banking data. Given that finance decisions strongly rely on emotional judgment customers made by collecting transactional data along with their feedback through emails, surveys, and tickets, the churn prediction models might be improved (Pahul Preet Singh & Fahim Islam Anik, 2023).

Finally, there are few studies that discuss the long-term effects of AI-based churn rescue mechanisms. Most of the studies illustrate how ML models predict churn, with very low studies having observed the effects of such predictions in real banking situations where the churn rate was actually reduced (Hoang Tran, 2023).

The literature review indicates that customer churn prediction is an integral part of banking strategy, and ML has transformed banking strategies for retention. Some such models like Random Forest and Gradient Boosting have very high accuracy; however, they are complex, computationally expensive and most importantly lack interpretability, which makes them relatively non-applicable in banking where product transparency is of topmost priority. Logistic regression is still preferred due to its explainability, brainless implementation, and adherence to standards by the banking regulator. This research provides an answer to existing gaps by focusing on building

a practical, basic, scalable, and interpretable churn prediction model through ML Logistic Regression in Microsoft Azure.

Methodology

Problem Statement

Customer retention is among the biggest challenges in the banking industry. Losing customers means losing revenue, both in the short and long run. Determining why customers leave is a top priority for banks if they want to act on retention strategies in a timely manner. This study aims to develop a machine-learning tool using Microsoft Azure to facilitate churn prediction by working on customer attributes, financial behavioral aspects, and banking engagement metrics. The patterns of customer transactions and banking habits are tracked through the model to take preventive measures before a customer has an intention to leave. The aim is also to build a strong machine-learning model that would predict churn while providing insights to build better customer retention strategies.

Formulation of Hypotheses

I have formulated four hypotheses based on my literature review for investigating the factors influencing customer churn. The first hypothesis explains that customers with higher balance rates on their accounts would be less likely to churn, as customers with high balances would be expected to maintain a closer financial relationship with the bank, leading to a lower chance of leaving. The second hypothesis suggested that customers using multiple products will have a lower rate of churn, because using more financial services mutes a customer with more than one product, like a loan, a savings account, and a credit card, giving this customer some kind of dependency on the bank. The third studies the relationship of active membership status with lower churning rates, assuming that such customers will remain in close contact with their banks and will be less likely to go elsewhere. The fourth and last hypothesis finds out whether customers with higher credit scores would have lower churn, assuming a good credit rating reflects both financial stability and a long-term banking relationship. These hypotheses guided the selection of features for model building and evaluation.

Data Collection and Overview

The data set used in this study was downloaded from Kaggle and is Bank Customer Churn Dataset of 10,000 customers who are account holders at ABC Multistate Bank. The dataset includes 12 features which represent demographic, financial, and customer engagement. Some key features included credit score, age, balance, number of products used, active membership, country, gender, tenure, credit card ownership, and estimated salary. The target variable of this study is churn, represented by a binary variable where a value of 1 means the customer left the

bank and 0 means they did not leave. The dataset gave a complete picture of customer banking behavior which would facilitate predictive churn modeling. (Figure 1)

Figure 1: Dataset Overview

```
# Display first few rows to verify loading
df.head()
```

[7] ✓

```
{'infer_column_types': 'False', 'activity': 'to_pandas_dataframe'}
{'infer_column_types': 'False', 'activity': 'to_pandas_dataframe', 'activityApp': 'TabularDataset'}
```

	customer_id	credit_score	country	gender	age	tenure	balance	products_number	credit_card	active_member	estimated_salary	churn
0	15634602	619	France	Female	42	2	0.00	1	1	1	101348.88	1
1	15647311	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
2	15619304	502	France	Female	42	8	159660.80	3	1	0	113931.57	1
3	15701354	699	France	Female	39	1	0.00	2	0	0	93826.63	0
4	15737888	850	Spain	Female	43	2	125510.82	1	1	1	79084.10	0

Data Pre-Processing

Before the training of the machine learning model, data pre-processing was done to maintain data quality and consistency of the data. It involves dealing with empty values, like categorical list fields like gender and country which had missing values. It is one of the major steps as it would affect the accuracy of the model. Feature selection (Figure 2) was carried out to align the data with the hypotheses, out of which five essential features such as credit score, age, balance, number of products used, and active membership were used for further study. Those features were selected due to their strong influence over churn and to provide the capacity for model interpretability and accuracy. The numerical features were standardized to ensure consistency in numerical data, while categorical features were converted into numerical forms applying one-hot encoding. Once all necessary pre-processing steps were completed, the data was set for training the model.

Figure 2: Feature Selection

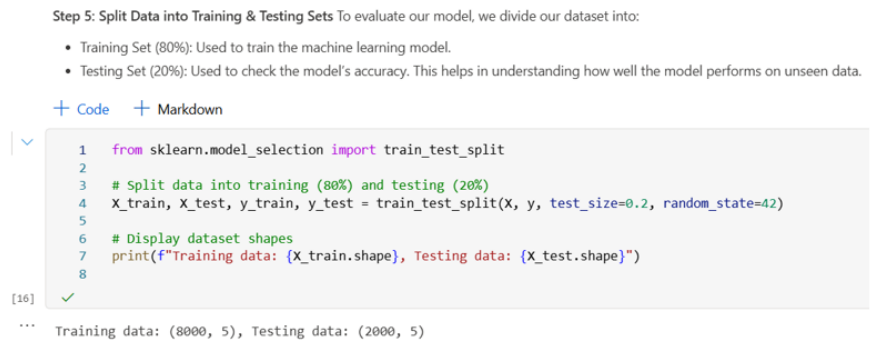
```
# Select relevant features and target variable (using correct column names)
X = df[['credit_score', 'age', 'balance', 'products_number', 'active_member']]
y = df['churn']  # Correct target column name

# Display selected features
X.head()
```

[15] ✓

	credit_score	age	balance	products_number	active_member
0	619	42	0.00	1	1
1	608	41	83807.86	1	1
2	502	42	159660.80	3	0
3	699	39	0.00	2	0
4	850	43	125510.82	1	1

Model Development

Figure 3: Split Data into Training (80%) & Testing Set (20%)



For this study, logistic regression was selected as the classification model because of its efficiency in classification problems such as churn prediction. Logistic regression was also best fitted for this analysis because it provides probabilities of predicting churn, it required less computational effort, and it was effective to use standardized numerical data. Feature engineering was performed by removing such irrelevant columns like customer id, tenure, credit card ownership, and country to reduce any issue. This way the final dataset included five independent variables and one dependent variable (churn). The dataset was split into 80:20 (Figure 3) ratio which is 80% of training data and 20% of testing data which ensured balance representation for model to be evaluated. As shown in Figure 3, the dataset is of 10,000 customers so 8000 have been used as training data and 2000 for testing data. It is this structured process that helps guarantee that the model was properly trained and optimized for churn prediction.

Model Evaluation

Evaluation of the model is extremely important to test the reliability of the model. Metrics for model evaluation were accuracy, precision, recall, and the F1-score. The Logistic Regression model was trained and considered for performance evaluation through the assessment of key classification metrics. The accuracy of the model is 81%, which indicates that the model accurately predicts customer churn in 81 out of 100 times, giving the model a good rating for customer churn prediction.

The classification report shows the model's performance with respect to both churned and non-churned customers. A precision of 0.82 and recall of 0.97 lead to an F1-score of 0.89 for non-churned customers indicates a good model, whereas it struggles for the churned customers showing a precision of 0.56, recall of 0.16, and F1-score of 0.24. This performance disparity also indicates the effect of class imbalance, as the dataset has more non-churned customers than churned which is causing a model to favor predicting the non-churned cases. Hence, overall accuracy was 81%, and the very low recall of churners denotes that it misses out on a considerable number of the actual churners. Macro-average metrics (precision: 0.69, recall: 0.56,

F1-score: 0.57) and weighted average F1-score of 0.76 further point out that the model should improve on accurately predicting churned customers.

Further Model Testing

To ensure the validity and effectiveness of the model, additional testing was done to predict customer churn. Results were displayed in tabular form for closer manual inspection of customers. The idea was to first put customer IDs through but found difficulty in pre-processing and then use credit score or age as unique identifiers. This showed the effectiveness of the model to make realistic predictions. Before generating the final predictions, the raw test data were prepared correctly so that it matched the final form of the training data. It was made sure that the test dataset included identical feature settings used during training and manually aligned columns when discrepancies existed to ensure both datasets were on the same scale. Initially, errors due to feature mismatches were corrected, such that extra features were removed in order to only keep those that were employed during model training. The test data were then scaled to fit on the training data, so the distributions of the features developed similarly. Following the preparation of the test data, we predicted churn probability for each customer using the Logistic Regression model that we trained. The final table (Figure 4) showed clearly the justified churn, adding to the trust in our model.

Figure 4: Further Model Testing

✅ Predicted Churn Results:

	Credit Score	Age	Predicted Churn
0	751.0	36.0	0
1	581.0	34.0	0
2	735.0	43.0	0
3	661.0	35.0	0
4	675.0	21.0	0
5	738.0	58.0	1

Testing and Interpretation of Results of Hypotheses

The hypothesis was tested against the customer churn to understand their validity. The first hypothesis was rejected suggesting that customers who have higher account balances had a lower probability of churning. Results showed that higher balances had higher chances of being churned, which is not a popular belief in traditional banking assumptions. The second hypothesis stated that customers using multiple banking products have lower churn rates, and it was confirmed as customers who are engaged in multiple banking products are less likely to churn as it's going to be difficult for them to disengage with the bank. The third hypothesis proposed by the model was validated which is that active membership reduces churn. It was shown that actively engaged customers are usually the ones with significantly lower churn rates. The fourth hypothesis which explained that lower churn rates are caused by higher credit scores, was rejected. The study indicated that credit scores have little effect on churn.

Challenges, Limitations, and Assumptions

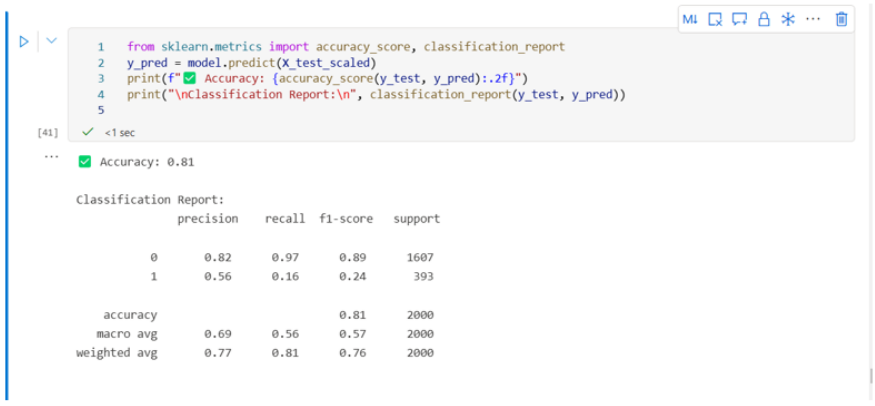
There were several challenges during the model development. One of the challenges was the feature misalignment of the test dataset from the training dataset, which had insufficient features; such situations confuse the model and lead to prediction errors. This means that proper feature selection matters when there is a need to align the two datasets correctly. Another point of concern was that missing values resulted in problems, especially in categorical features. Appropriate encoding and scaling of the features required many rounds of iteration. Further, according to initial predictions, the customer IDs were not included; this made tracking results per customer difficult. We decided to pinpoint either age or credit scores as the alternate unique identifiers for customers' choices.

The most significant limitation was the existence of imbalanced classification in the dataset. More customers did not churn, which is only 20% of customers within the database. This made it impossible for the model to generalize patterns for making predictions in favor of the minority group of customers, hence yielding lower recall for predicting already churned customers. I further assumed that all numerical features had equal importance in predicting churn, when in reality it is very possible that some factors have a non-linear effect. The other important way to provide a reasonably informative baseline is through Logistic Regression. To improve baseline performance, more challenging types of models that can be looked into include Random Forest or Gradient Boosting. Finally, it was further assumed that the dataset was an adequate reflection of customer behavior in the real world; however, there were other possible variables that were not captured in the dataset, such as customer services offered, or competitor offers.

Results

According to the report (Figure 5), the model achieved accuracy of 81%, performing generally well for classification purposes as in accurately predicting customer churn in 81 out of 100 times.

Figure 5: Classification Report Results



```
from sklearn.metrics import accuracy_score, classification_report
y_pred = model.predict(X_test_scaled)
print(f"✅ Accuracy: {accuracy_score(y_test, y_pred):.2f}")
print("\nClassification Report:\n", classification_report(y_test, y_pred))
```

[41] ✓ <1 sec

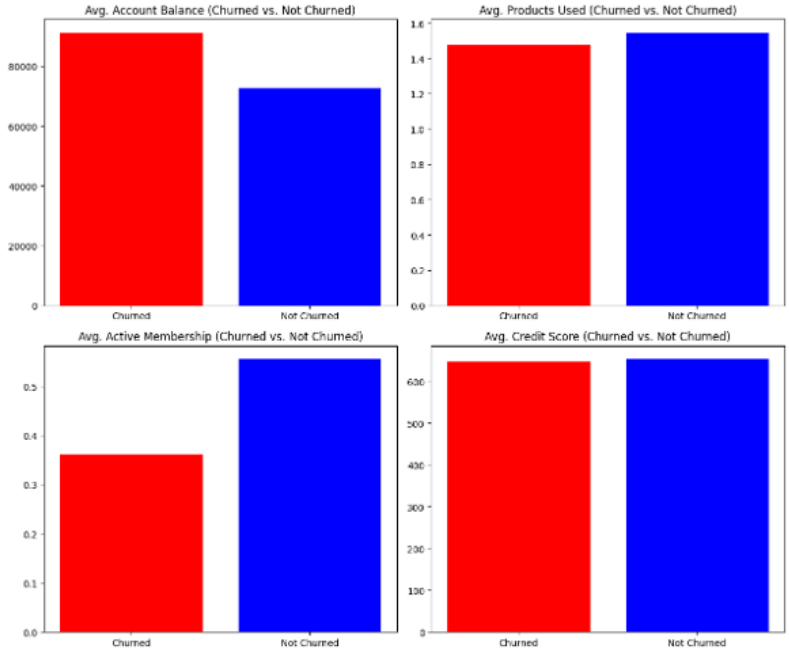
```
✅ Accuracy: 0.81

Classification Report:
              precision    recall  f1-score   support

           0       0.82      0.97      0.89      1607
           1       0.56      0.16      0.24       393

    accuracy                           0.81      2000
   macro avg       0.69      0.56      0.57      2000
weighted avg       0.77      0.81      0.76      2000
```

However, on closer inspection, the model shows a serious imbalance in churn prediction. For customers who are not churned (0), the model shows a precision of 82% and recall of 97%, thus resulting in a high F1 score of 0.89, meaning that it usually correctly identifies customers who are retained. On the other hand, for the customers who did get churned (1), precision plummets down to 56% with recall at a mere 16%,

resulting in an F1 score of 0.24, displaying an overall propensity on the model's part not to detect actual churn cases. The macro-averaged recall comes to about 56%, with a weighted average precision of 77% to support that the model is biased in favor of the majority class (non-churned customers). This imbalance makes its application in real-world scenarios very difficult, as if churn is not detected, then proactive retention strategies cannot be initiated. To increase its reliability, techniques such as oversampling, and moving the decision thresholds might be considered.

These visualizations (Figure 6) gives useful insights into customer churn behavior by comparing churned customers with non-churned customers across four hypothesis. The graph in the top-left shows that churned customers tend to have a slightly larger account balance on average than non-churned customers, meaning that account balance alone does not help stop the churn. Again, the average number of products used was practically the same for both groups, as illustrated in the top right graph, thereby making product variety not a strong predictor for churn. The graph in the lower left shows that there is a very large distinction in active membership, with non-churned customers showing exceptional engagement, reiterating the importance of continued activity in customer retention. Finally, from the analysis, we see that both don't show any significant difference in credit scores observed across both the groups, exhibiting that financial credibility doesn't seem to play a strong role in impacted churn predictions. This further shows that customer engagement and activity levels take precedence in retaining customers rather than balance scorecard or credit scores.

Figure 6: Hypothesis Testing Results



The model's performance provides valuable insights into the various factors responsible for customer churn which in itself is a major problem a banking business deals with customers' retention. The model has an overall accuracy of 81% which accurately identifies customers that churned or did not churn. However, its problem is a slightly lower recall concerning churned customers since it may not detect some of those customers truly at risk of leaving. This limitation of this model suggests that it has a strong predictive capability, but it also indicates that further

improvements like data balancing or introducing some behavioral indicators could enhance the work by better identifying those potential churners.

From a business perspective, it provides a predictive nature for early intervention strategies. Knowing the at-risk customers, it will help banks introduce personalized retention plans by making targeted offers or increasing customer engagement and much more. The insight from the model highlights the key factors such as active membership and account can lead to long term growth of customer relationship and bank revenue. At the end of it all, this model equips banks with data-driven interventions to deal with churn and focuses their attention on retaining high-value customers to stem revenue loss and improve customer satisfaction.

Conclusion

Customer churn is one of the biggest challenges in the banking sector today, with adverse implications for revenues, long-term development, and competitive advantage. The study uses machine-learning techniques for the prediction of churn using actual banking data. The study will also help identify the key factors for customer retention. Our analysis found that account balance, active membership status, and products usage crucially affect the determination of whether a customer will churn. The logistic regression model performed with an 81% accuracy-an indication of its power as a predictive tool. Still, the low recall for churned customers indicates that improvements need to be undertaken since some high-risk customers might be dropping out without detection. Hence, improvements should still be made along the lines of further boosting recall while ensuring it does not do much damage to precision and thus aligns the customer retention strategies.

The business relevance of this model is clearly to preserve potentially quitting customers, the banks should take proactive steps in avoiding the necessity of having to react in reverse once they leave. The insights generated in the model help personalize interventions, such as putting out certain financial products, preemptively reaching out to customers with customer service needs, and giving loyalty rewards, that can tremendously impact customer involvement and satisfaction. Also, the findings of the study suggest that the incentives to correspond with the promotion of active membership and optimization of financial product offerings could help to reduce churn. This data-driven approach gives banks the ability to keep addressing the revenue leakages, cutting off customer churn and enhancing customer loyalty simultaneously with little cost of implementation and accordingly meaningfully contributing to the banks' prolonged growth, profitability, and prosperity in competing old and new financial services industries.

The future study must focus on developing the model by including some other relevant customer behavior data, sentiment analysis from customer interactions, and external economic indicators.

In further research, advanced machine learning methods, such as ensemble methods or deep learning, could be explored to improve the actual performance, which is especially important in predicting borderline-churn-risk customers. Real-time churn prediction models could also be developed to enable banks to take automatic instant decisions based on live customer data.

The paper shows the efficiency and advantages of using the tool that I have built for predicting customer churn in the banking sector. By using these tools, financial institutions would no longer have to depend on set traditional methods but could create personalized strategies for customer retention purposes. The model will require continued refinement and AI-driven customer insights will unlock its capability for sustainable customer engagement and organizational growth.

References

Datrics AI. (2023). *Bank churn prediction using ML to retain customers*. https://www.datrics.ai/articles/bank-churn-prediction-using-ml-to-retain-customers

Pahul Preet Singh & Fahim Islam Anik (2023), Investigating customer churn in banking: a machine learning approach and visualization app for data science and management https://www.sciencedirect.com/science/article/pii/S2666764923000401

Hoang Tran (2023), Customer Churn Prediction in the Banking Sector Using Machine Learning-Based Classification Models https://www.ijikm.org/Volume18/IJIKMv18p087-105Tran8783.pdf

Michael (2024), Bank Customer Churn Prediction Using Machine Learning https://www.analyticsvidhya.com/blog/2022/09/bank-customer-churn-prediction-using-machine-learning/

Keldine Malit (2018), Kaggle - Bank Customer Churn Prediction https://www.kaggle.com/code/kmalit/bank-customer-churn-prediction

Derek Papierski (2023), Investigating Customer Churn in the Banking Industry https://medium.com/@dpapcodes/investigating-why-customers-leave-a-bank-47b41278e36c

Salesforce, Use the Retail Banking Customer Churn Prediction Dashboards https://help.salesforce.com/s/articleView?id=ind.fsc_use_churn_prediction_retail_banking_dashboard.htm&type=5

Joao B. G. Brito (2024), A framework to improve churn prediction performance in retail banking https://jfin-swufe.springeropen.com/articles/10.1186/s40854-023-00558-3

Totango, Customer retention & churn prediction https://www.totango.com/demo/live-demo

Ke Peng & Yan Peng (2023), Research on customer churn prediction and model interpretability analysis
https://pmc.ncbi.nlm.nih.gov/articles/PMC10707658/pdf/pone.0289724.pdf

Generative Pre-trained Transformer (ChatGpt): For Model Coding

Canva (For introduction image generation)

GitHub Repository link:

https://github.com/priyalrawat/BankingCustomerChurnPrediction/tree/main