ENGINEERING
W Booth School of Engineering Practice and Technology
McMaster University

Final Project

Machine Learning Approach to Bank Customer Retention

Course Number and Name:
SEP 785: Machine Learning
Semester, Year, and Group Number:
2025 Fall

Name of Student:	**Name of Instructor and TA:**
Foram Brahmbhatt(400651023) Savan Sachpara(400676965)	Sayyed Faridoddin Afzali Mahdi Astaraki

Table of Contents

List of Figures

List of Tables

Abstract

Customer churn represents a considerable financial threat to banks, as it is far more expensive to lose current customers than to gain new ones. This project outlines a comprehensive machine learning process to forecast customer churn utilizing a public dataset comprising 10,000 customers. The process combines Exploratory Data Analysis (EDA), K-Means segmentation with PCA support, and two supervised models, Random Forest and XGBoost.

PCA and clustering identified four unique customer segments exhibiting significant behavioral trends, featuring a high-risk group of older clients with moderate balances (36% churn rate) and a medium-risk group of high-balance yet inactive customers (29% churn). These insights emphasize groups that could gain the most from focused retention strategies.

The same preprocessing pipeline was used to train the random forest and XGBoost. After tuning hyperparameters, random forest achieved a recall of 0.6683. And F1 score of 0.6051. Which was better than XGBoost(recall 0.4398, F1 0.5559) file identifying churners. A cost-sensitive evaluation confirmed that the random forest is optimal for business purposes. As ignoring churners leads to greater expenses than making unnecessary retention offers.

Introduction

In a banking industry that is highly competitive and with ever-increasing asset costs, customer retention is foremost. The predicting of customer churn has turned into a strategic crucial to allow the business to engage in proactive retention efforts. Traditional statistical models often fail to capture all nonlinear dependencies and interactions in the data from customers' actions (He, 2009) leading to the adoption of ML techniques.

The next stage of the project was to build a complete machine learning workflow to predict which bank customers are likely to leave. This included feature engineering, customer segmentation, and using both Random Forest and XGBoost for churn prediction. The goal was to provide clear insights that can help the bank make better business decisions.

- ## Background

  Bank churn is generally caused by a customer's dissatisfaction, low engagement with the bank, financial stress or the customer not getting relevant bank products. If a bank knows and can forecast its churn, then it can design up-interventions like product suggestions, personalized contact or giving the customer monetary motive for staying.

- ## Problem Statement

  The major task is to spot with precision the customers who have a high probability of leaving (Churn) based on their demographic, financial, and behavioral characteristics. The problem is to use different kinds of data, deal with classes having unbalanced numbers, and find the nonlinear relationship that exists among features.

- ## Objectives

  - To carry out an exploratory analysis that shows while at revealing the relationships between features and the characteristics of churn.
  - To apply PCA + K-Means for customer segmentation and then interpret the segments according to their likelihood of churn.
  - To create models of classification (Random Forest and XGBoost) capable of predicting the best churn(model).
  - To apply cross-validation, hyperparameter tuning, and robust metrics as means for model evaluation.
  - To conduct an insights extraction for customer retention strategies that can directly be worked on.

- ## Report Structure

  The report follows the usual stages of an ML project-namely data explanation, process, trials, results, analysis, and conclusions.

Data

1. Data Description

The dataset comprises 10,000 customers with 14 features across demographic, behavioural, and financial categories. The target variable Churn indicates whether a customer exited the bank. (kumar, n.d.)

Feature	Type	Description
CreditScore	Numeric	Customer's credit rating
Geography	Categorical	France / Germany / Spain
Gender	Categorical	Male / Female
Age	Numeric	Customer age
Tenure	Numeric	Years with the bank
Balance	Numeric	Account balance
NumOfProducts	Numeric	Number of products held
HasCrCard	Binary	Credit card (1=Yes)
IsActiveMember	Binary	Active membership status
EstimatedSalary	Numeric	Annual estimated salary
Churn	Binary	1 = Exited, 0 = Stayed

Table 1 Data Description

2. Preprocessing steps

- **Missing values:** Dataset cleanliness assured by absence of missing entries.
- **Duplicates:** Earlier to modelling duplicates were removed.
- **Categorical Encoding:** One-Hot Encoding used for Gender and Geography.
- **Feature Scaling:** Numerical features scaled for clustering (PCA + KMeans).
- **Class Imbalance:** 20% Churn rate; resolved by threshold tuning and selection of a model.

Methodology

1. Unsupervised Learning: Customer Segmentation via K-Means

The first phase of the methodology focused on using unsupervised learning to capture intrinsic customer groups. Based on the idea that combining clients with comparable financial and behavioural profiles would produce a powerful synthetic predictor, we applied K-Means Clustering to the standardized numerical features. (Pedregosa, 2011)

- **Algorithm:** K-Means partitions n observations into k clusters.
- **Clustering Features:** CreditScore, Age, Tenure, Balance, NumOfProducts, EstimatedSalary, and IsActiveMember.
- **Optimal k:** Using elbow method, k=4 was selected, and the cluster membership was treated as a new categorical feature..

2. Preprocessing Pipeline:

- The project required structured preprocessing to make sure the models were being evaluated on inputs that were clean, consistent, and comparable.
- Numerical features would be standardized using StandardScaler to prevent scale-driven bias, while categorical fields would be encoded with OneHotEncoder so the models could interpret them correctly.
- Using a ColumnTransformer allowed these steps to run together, ensuring the data entering each model reflected the patterns uncovered during the EDA stage.
- With preprocessing defined, the cleaned features flowed directly into the classifier through a unified Pipeline. This design ensured that models were trained and tested (80/20 train-test-split) under identical conditions, reducing data leakage and improving fairness in the comparison of the models.
- It also allowed cross-validation(5-fold StratifiedKFold) and hyperparameter tuning in order to evaluate both preprocessing and model behavior as one system and hence supported a more reliable analysis of which model performed best on the churn prediction task.

3. Model selection

Two ensemble models based on trees are chosen due to their effective handling of mixed feature types, ability to accommodate non-linear relationships (Pedregosa, 2011), and consistent performance on moderately imbalanced datasets like this churn data (in which churn represents about 20% of the entire customer base). (He, 2009)

Random Forest Classifier

Random Forest was selected as the baseline ensemble model due to its stability, interpretability, and resistance to overfitting (Pedregosa, 2011). Its design of aggregating numerous decision trees makes it ideal for datasets with interrelated numerical attributes such as Balance, Age, and CreditScore. It effectively manages noisy variables, which was crucial after noticing during EDA that various predictors displayed overlapping distributions for churners and non-churners. Random Forest can be a robust benchmark due to its competitive performance requiring little adjustment.

XGBoost Classifier

XGBoost was chosen as the second model because its boosting approach can pick up patterns that Random Forest might miss. Since churn cases are the minority group, XGBoost helps the model pay more attention to customers who are harder to classify correctly. It also has built-in regularization and efficient tree-level optimization, which make it strong at handling unbalanced data. The main goal was to see whether this extra complexity would actually lead to better churn prediction compared to Random Forest (Chen, 2016).

Experiments and Findings

1. Exploratory Data Analysis (EDA)

- **Data Description and Preparation**:
  - The initial dataset contained 10,000 customer entries and 14 features.
- **Data Quality**
  - The dataset is found to be of good quality, with no missing values identified across any of the columns.
- **Cleaning and Transformation**

  The following steps performed to prepare the data for analysis:
  - **Irrelevant Columns Dropped**: Identifier columns such as RowNumber, CustomerId, and the non-influential Surname were removed successfully.
  - **Target Renaming:** The target variable Exited was renamed to Churn for clear understanding.
  - **Encoding:** Categorical features (Gender, Geography) were converted to a numerical format for correlation analysis (e.g., in the Correlation Heatmap). For Gender, Male was mapped to 1 and Female was mapped to 0. For geography one-hot encoding was used.

2. Exploratory Data Analysis (EDA) Findings

- **Target Variable Distribution:** The analysis shows a clear class imbalance, with only about one in five customers having churned.
  - **Churn Rate:** The percentage of customers who have exited the bank is 20.37%.
  - **Distribution:** 7,963 customers stayed (79.63%) and 2,037 customers exited (20.37%).

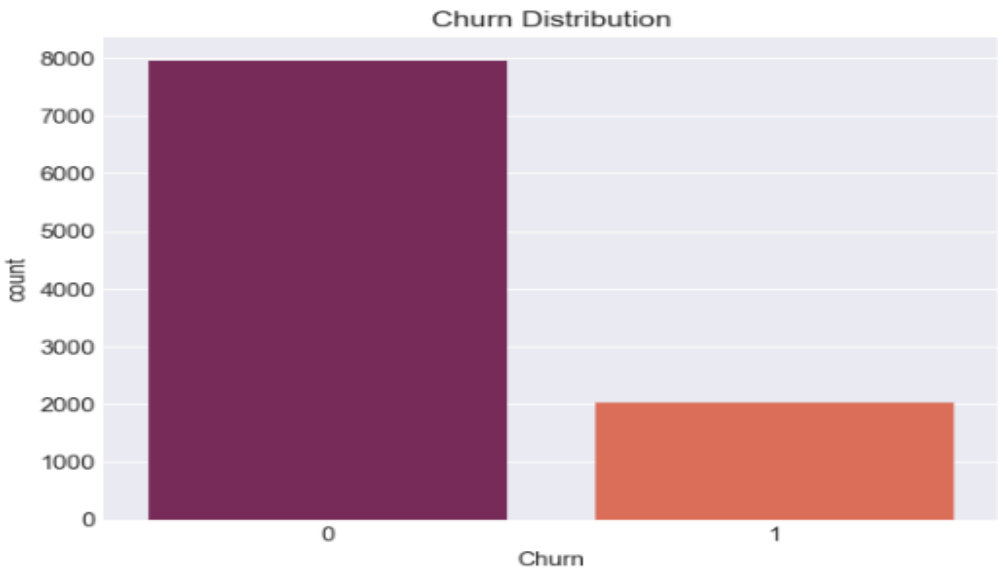

Figure 1 Churn Distribution

- **Impact of Categorical Features**

The Exploratory Data Analysis (EDA) here focuses on examining how the main categorical variables Geography and Gender affect the likelihood of a customer leaving the bank.
The low baseline (20.37%) highlights the need for metrics appropriate for imbalanced classification, like ROC-AUC, instead of more accuracy, to avoid deceptive performance evaluations in the following machine learning stage.

- **Impact of Numerical and Binary Features**

Age Distribution vs Churn: Examining the density plot (Age Distribution by Churn Status) shows a clear trend where customers between the ages of 40 and 65 have a notably higher chance of churning; whereas the customers in the age group of 25-40 have a lower chance of churning.

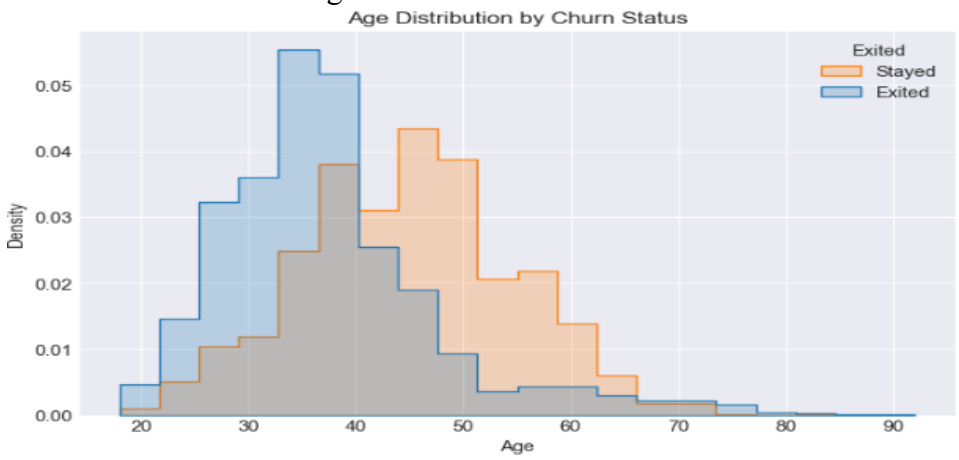

Figure 2 Age Distribution by Churn State

Age vs. Balance: The general trend of customers who have exited (Churned) closely resembles the trend of customers who have remained when examined solely through these two features. Additional factors, like NumberOfProducts or IsActiveMember, are probably necessary to distinguish the high-risk customers within these distributions.

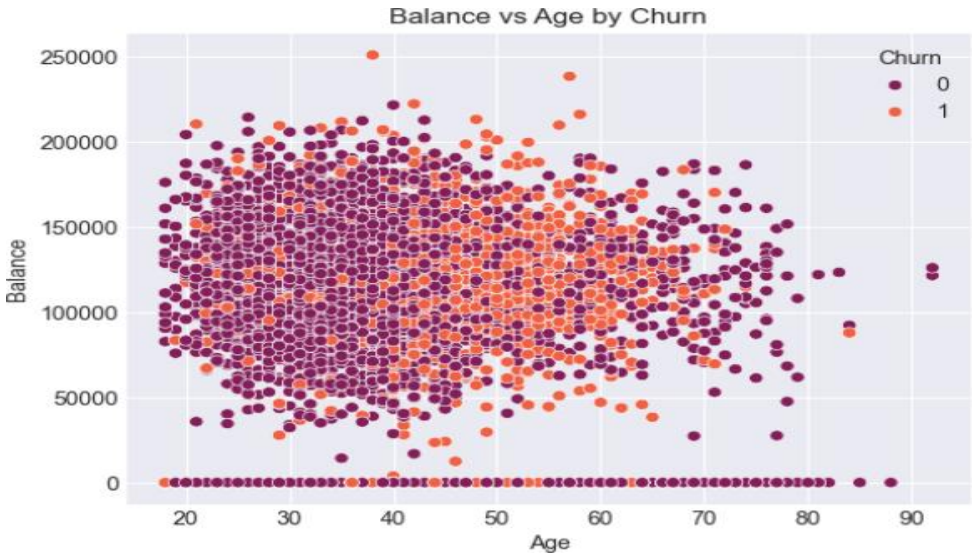

Figure 3 Balance vs Age by Churn

3. Correlation Analysis:

- The absolute quantitative relationships between all of the given numerical features and the Churn target variable were successfully determined by analysing the correlation heatmap for the encoded data.
- Strongest Positive Correlation: Age (shows that older clients have a higher likelihood of churning).
- Strongest Negative Correlations: IsActiveMember and NumberOfProducts (suggests that customers with a greater number of products and active membership are less likely to churn).
- Moderate Correlation: Churn is also notably correlated with Balance and Geography

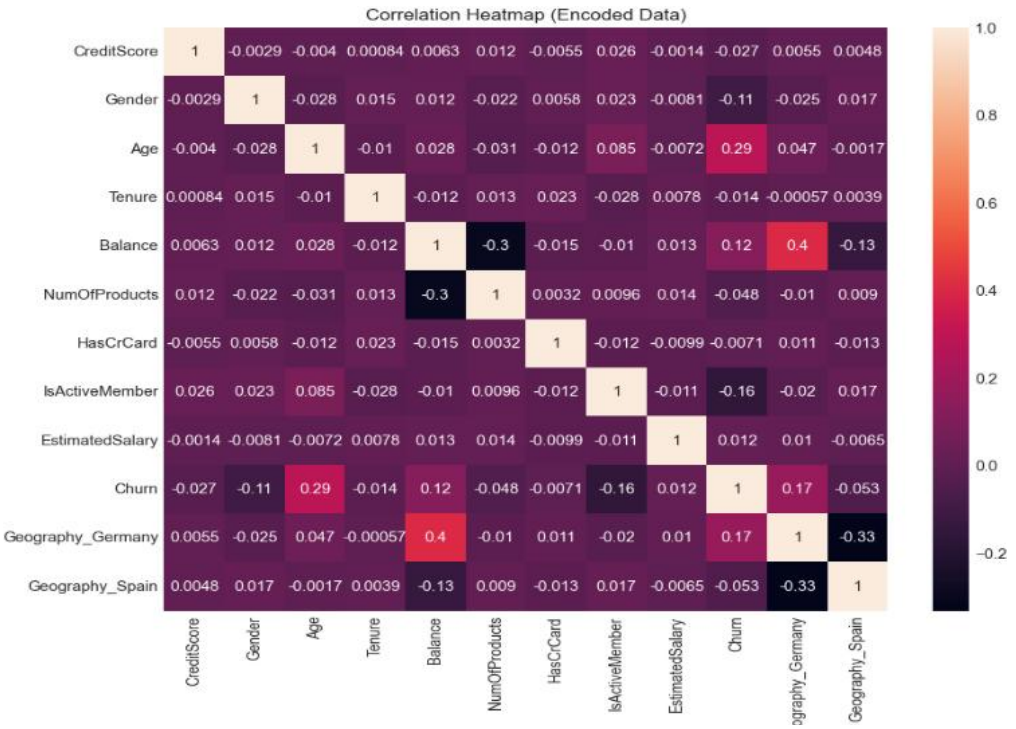

Figure 4 Correlation Heatmap (Encoded Data)

4. Customer Segmentation (Using K-means)

- **Data Pre-processing and Feature Engineering**
  The first step of the analysis was to take the processed and encoded dataset which was created from the initial EDA phase.

  - **Feature Set for Segmentation**
    The feature set included 7 columns:
    features = ['CreditScore', 'Age', 'Tenure', 'Balance', 'NumOfProducts', 'IsActiveMember', 'EstimatedSalary']

- **Feature Scaling (Standardization)**
  To make certain that all features had an equal impact on distance computations in clustering (and for PCA preparation), the data was standardized through the use of the StandardScaler. Standardization modifies the data to achieve a mean of 0 and a standard deviation of 1.

- **Dimensionality Reduction:**
  Principal Component Analysis (PCA) was performed to simplify the dataset, while preserving the majority of its essential information (Pedregosa, 2011).

  - **Variance Explained:** The cumulative explained variance with 4 components was 62% which increased to 77% with 5 components. Resulting in the final choice of 'n_components = 6' which effectively decreased the 7 features to 6 Principal Components (PCs) cumulatively explaining 90% of the variance.

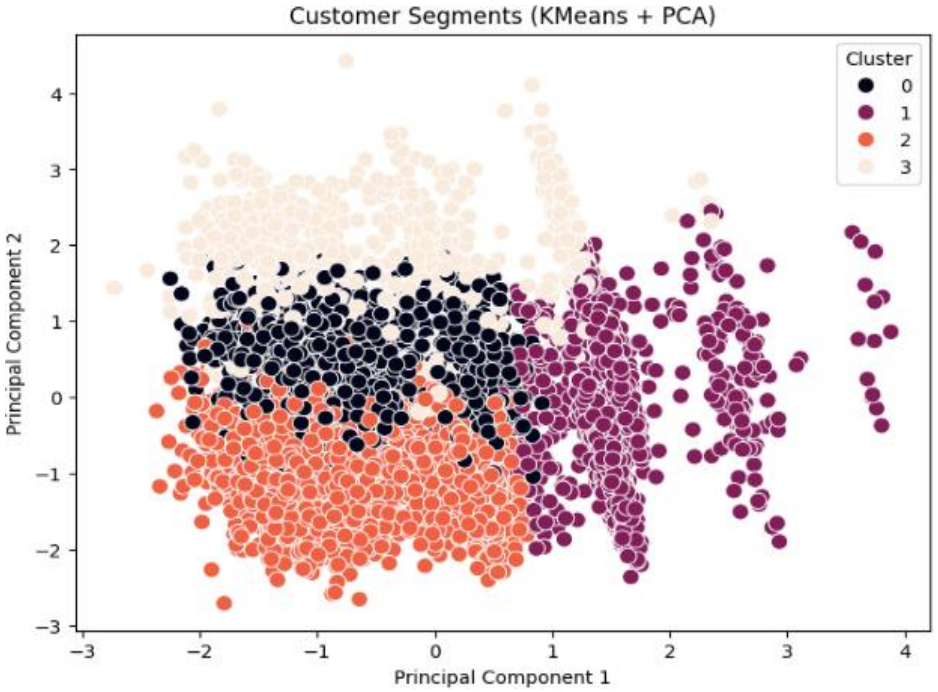

Figure 5 Customer Segments (KMeans + PCA)

- **Segment Profiling**

The PCA + KMeans model designates the customers into four distinct segments according to their financial behavior and engagement. The identified segments not only provide a clearer picture of the different types of customers such as those who are faithful, those who are likely to quit, and those whom the bank could still reach with certain measures, but also the possible measures to be taken by the bank.

- **Segment Analysis**

Cluster 0 - High Balance & Very Active (Low Churn: 13%)
These clients maintain substantial balances and actively use their accounts on a regular basis. They are faithful and secure in the long run. The financial institution should persist in providing premium or personalized services to these customers in order to keep them.
Cluster 1 - Many Products, Low Balance (Low Churn: 12%)
This group consumes a larger variety of products compared to others, however, their account balances are not that high. They look content with their situation in general. The bank could motivate them to develop their savings or investments.

Cluster 2 - High Balance but Inactive (Medium Churn: 29%)
Their bank balance may be high, but on the contrary, they hardly use or communicate with the bank. The inactivity might be the reason for the higher churn in this case. These customers will require either re-engagement campaigns or targeted personal outreach.

Cluster 3 - Older Customers with Medium Balances (High Churn: 36%)
The oldest group is also the one that is most likely to exit. They might be changing their banks because of retirement-related matters or more advantageous benefits. Marketing of retirement products or loyalty offers might be effective in preventing their exit.

Below is the table of mean value per cluster for each feature:

Feature	Cluster 0	Cluster 1	Cluster 2	Cluster 3
CreditScore	653.3	650.46	647.28	653.22
Gender	0.56	0.54	0.54	0.53
Age	35.30	35.92	37.47	59.88
Tenure	4.84	5.14	5.08	4.94
Balance	107,773.22	9,540.17	105,903.15	75,891.74
NumOfProd ucts	1.29	2.13	1.27	1.43
HasCrCard	0.70	0.71	0.71	0.70
IsActiveMember	1.00	0.49	0.00	0.83

Estimated Sal ary	100,774.29	99,575.45	101,699.25	94,919.59
Churn	0.13	0.12	0.29	0.36
Geograph y_ Germany	0.3362	0.0503	0.3473	0.2599
Geograph y_ Spain	0.2243	0.3147	0.2081	0.2554

Table 2 Mean Value for Feature per Cluster

5. Identifying the Best Number of Clusters:

The Elbow Method was used to determine the optimal quantity of clusters K. This technique graphs the inertia (total squared distances of samples to their nearest cluster centre) for various values of K.

The plot showed a distinct "elbow" at K=4, indicating that adding more clusters beyond this point provides diminishing returns in reducing inertia.

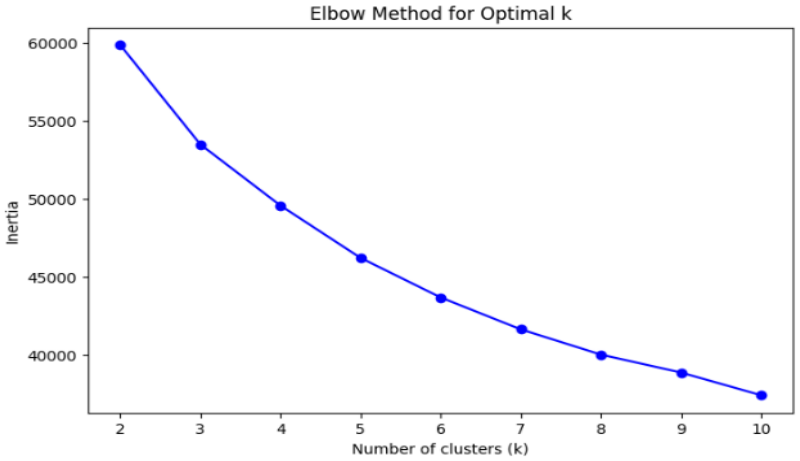

Figure 7 Elbow Method for Optimal K

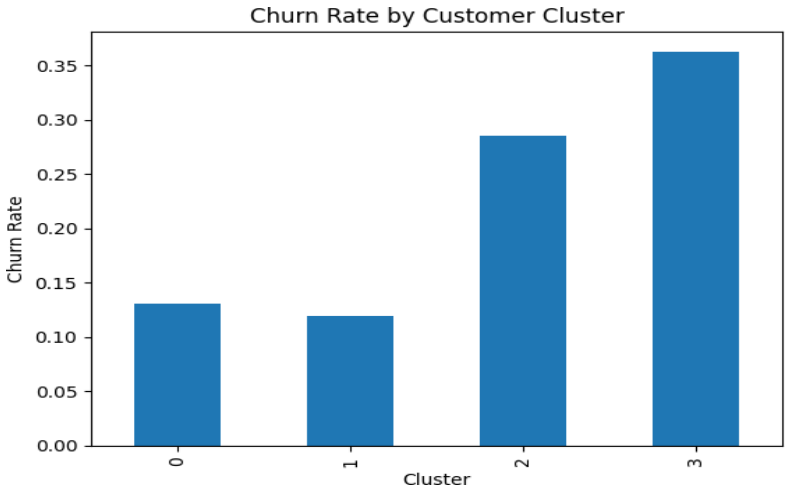

Figure 6 Churn Rate by Customer Cluster

6. Random Forest Classifier Performance

The Random Forest model is developed as one of the primary approaches for predicting customer churn in the bank dataset. As customer churn impacts both revenue and retention costs, early identification is necessary. Random Forest was chosen because it can effectively capture nonlinear relationships, reduce overfitting by ensemble learning, and offer interpretable feature importance, making it a strong baseline model for comparison.

- **Data Preparation and Preprocessing**

  The analysis used the bank_churn_clustered.csv dataset generated during the segmentation phase. An 80/20 stratified train-test split was performed, creating 8,000 training samples and 2,000 testing samples while preserving the natural 80/20 churn imbalance. Preprocessing was done with a pipeline that included StandardScaler for numerical features and OneHotEncoder for categorical variables. All transformations were fit only on the training set to prevent data leakage (Pedregosa, 2011). Since churners comprised only about 20% of the data, the baseline model was trained with class_weight='balanced' to offset the imbalance (He, 2009).

- **Initial Model Performance (Before Tuning)**

  The baseline Random Forest did a great job on the majority class, accurately classifying 97% of the non-churners. However, this model had limited ability in the detection of churners, only managing to have a recall of 41% for the minority class. It therefore missed most of the churning, representing a severe issue in business. The model achieved an average ROC-AUC of 0.8564 over five-fold cross-validation, with low variability, therefore showing stable but consistently insufficient performance for churn prediction.

	precision	recall	f1-score	support
0	0.86	0.97	0.91	1593
1	0.75	0.41	0.53	407
accuracy			0.85	2000
macro avg	0.81	0.69	0.72	2000
weighted avg	0.84	0.85	0.83	2000

Figure 8 Classification Report for Random Forest

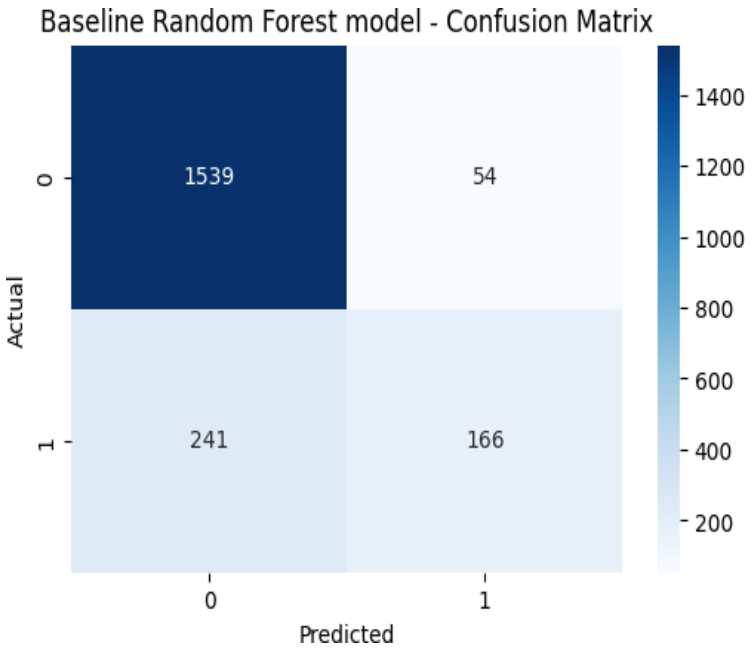

Figure 9 Baseline Random Forest model – confusion matrix

- **Hyperparameter Tuning**

To improve performance further, RandomizedSearchCV was used with ROC-AUC as the metric for optimization. It considered ten random combinations of parameters with five-fold stratified cross-validation. The optimal configuration included 300 trees, max_depth=20, max_features='sqrt', among other tuned values. This optimized model had a cross-validated ROC-AUC of 0.8576, which represents a small yet significant lift from the baseline

Hyperparameter	**Values Tested**
classifier__n_estimators	100, 200, 300
classifier__max_depth	None, 5, 10, 20
classifier__min_samples_split	2, 5, 10
classifier__min_samples_leaf	1, 2, 4
classifier__max_features	'sqrt', 'log2', None
classifier__bootstrap	True, False

Table 3 Hyperparameters for Random Forest

.

- **Final Tuned Model Performance**

  The tuning yielded a large improvement in the model's ability to detect churn, according to evaluation on the test set. Recall for churners went up from 41 to 67%, so now the model identifies nearly seven of ten customers likely to leave. While precision dropped from 76% to 55%, this trade-off is acceptable; false positives are low-cost compared to the financial impact from failing to catch actual churners. The model continued to perform well overall, with a 0.61 F1 for churn and ROC-AUC of 0.8473. The resulting cost analysis clearly demonstrated that catching more churners provides much higher savings than any extra retention offers due to the false positives.

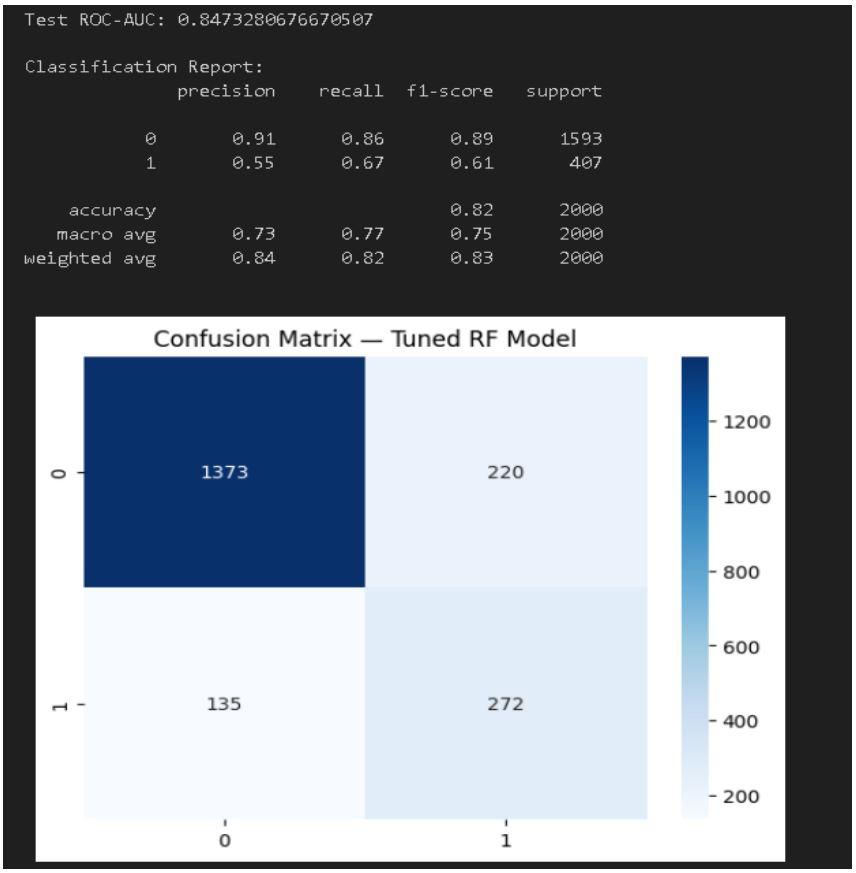

Figure 10 Confusion Matrix - Tuned RF Model

- **Threshold Optimization Strategy**

  Random Forest did not require explicit threshold tuning beyond hyperparameter optimization because it was trained with class_weight='balanced', which inherently shifts the decision boundary during model construction to prioritize minority class detection **(He, 2009)**. The results of this training-time strategy are comparable to those of test-time threshold adjustment. To enhance minority class recall, however, XGBoost's boosting mechanism necessitates explicit post-hoc threshold tuning (changed from 0.5 to 0.35). Each algorithm's native toolkit is reflected in the various tuning strategies: XGBoost uses test-time threshold adjustment, whereas Random Forest uses training-time weighting to address

class imbalance. In addition to allowing for fair algorithmic comparison and representing Random Forest's natural operating point, evaluating Random Forest at the default threshold shows that Random Forest requires less tuning complexity in production deployment.

- **Feature Importance and Partial Dependence Insights**

  Feature importance analysis confirms Age as the most influential predictor, followed by Number of Products, Balance, Salary, and Credit Score. Tenure, IsActiveMember, Gender, and HasCrCard have very little importance for the prediction.

  Partial dependence plots display obvious non-linear effects: the risk of churn rises sharply for customers aged 40–60, reaches its lowest point for customers with exactly two products, and goes up again for customers with very high balances. Salary and Credit Score presented weak individual effects. For Tenure, almost no relation can be seen regarding churn.

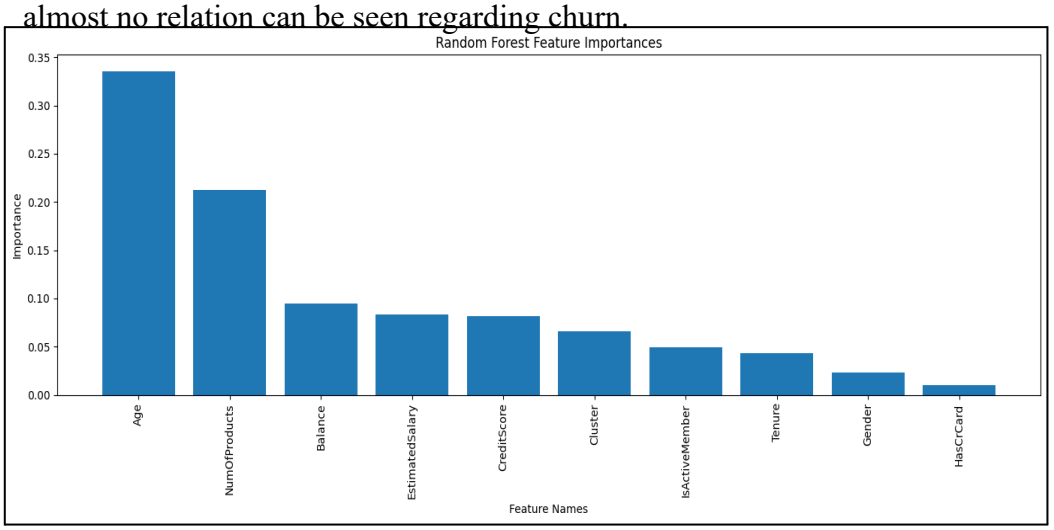

Figure 11 Random Forest Feature Importance

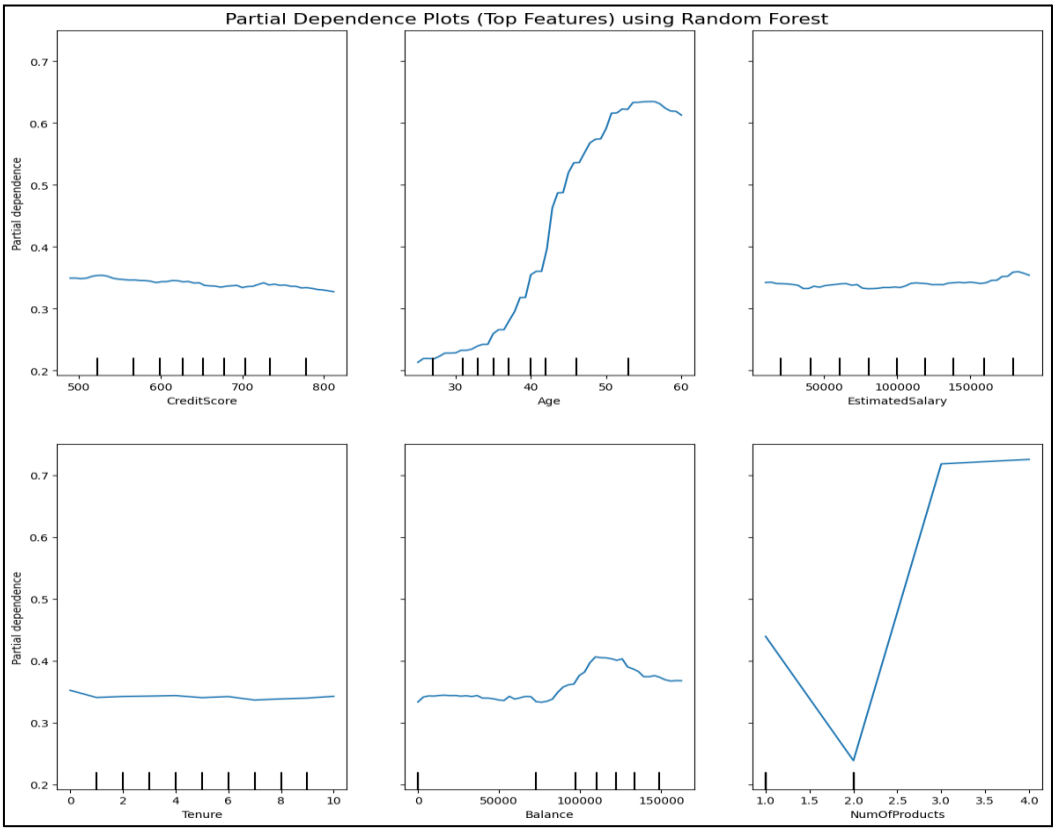

Figure 12 Random Forest Partial Dependency Plot

7. XGBoost Classifier Performance

Further we found that XGBoost is a powerful, highly efficient and flexible ensemble method as it is known for achieving high state-of-the-art results in classification types of problems (Chen, 2016), making it a strong candidate to outperform the previously established Random Forest model.

- **Model Evaluation and Performance (Initial & Tuned)**

The model had been evaluated firstly with initial parameters, then optimized using RandomizedSearchCV, and finally assessed on the dedicated test set.

- **Initial Test Set Performance (Before Tuning)**

We trained on the training data and tested on the unseen test set (Default Threshold).

The baseline XGBoost shows a very better ability to identify churners than the baseline Random Forest, with a Churn Recall of 0.46 (up from RF's 0.41). Therefore, more than half of the real churners are somehow still missed (False Negatives: 221), indicating that recall needs to be improved.
ROC-AUC: 0.8474 (Good baseline measure of class divisibility).

```
BASE XGBOOST MODEL
ROC-AUC: 0.8473697117764916
              precision    recall  f1-score   support

           0       0.87      0.96      0.91      1593
           1       0.74      0.46      0.56       407

    accuracy                           0.86      2000
   macro avg       0.80      0.71      0.74      2000
weighted avg       0.85      0.86      0.84      2000
```

Figure 13 Classification Report for XGBoost Base Model

- **Hyperparameter Tuning**

  To enhance the model's overall predictive capability, RandomizedSearchCV had been employed to explore a broad range of hyperparameters, using roc_auc as the main scoring metric.
  Included Tuning Parameters: n_estimators, learning_rate, max_depth, subsample, and colsample_bytree
  Tuned parameters and Best ROC-AUC: 0.8507

- **Tuned Model Evaluation and Threshold Adjustment**

  The tuned model performance on the test set remained stable compared to the baseline (Tuned Model Evaluation (Default Threshold: 0.5):
  ROC-AUC: 0.8540 (A minor improvement)
  Churn Recall (Class 1): 0.44 (Minimal change from baseline)

- **Final Tuned Test Set Performance (Threshold: 0.35)**

  Tuning successfully increased the overall predictive score, but Recall remained the primary challenge for the minority class at the default threshold.

```
Threshold: 0.35
              precision    recall  f1-score   support

           0       0.90      0.91      0.90      1593
           1       0.62      0.59      0.60       407

    accuracy                           0.84      2000
   macro avg       0.76      0.75      0.75      2000
weighted avg       0.84      0.84      0.84      2000
```

Figure 14 Classification Report for XGBoost(Tuned Threshold - 0.35)

Churn Recall Improved Notably: The recall for the churn class has been jumped from 0.44 to 0.59.

False Negatives Reduced: The number of missed churners dropped from 221 to 168

Trade-off: The number of false positives (non-churners incorrectly marked as churn) increased, as expected, this trade-off is often considered acceptable for retention strategies.

The XGBoost Classifier turned out to be a great model with a high ROC-AUC of 0.8540. The starting threshold placed more importance on precision, but changing it to 0.35 led to a well-matched performance with a F1-score of 0.60 and 59% Recall, thus it was more appropriate for a proactive retention system.

Discussion

- ## Model Performance Comparison

Random Forest and XGBoost models are the ones that definitely give an edge in the case of unbalanced data. The overall accuracy (85.70%) and precision (75.53%) of XGBoost are much higher than those of Random Forest (82.25% accuracy, 55.28% precision). However, if the task is to detect churners, then Random Forest is the better choice over XGBoost as it has a recall rate of 66.83% in contrast to 43.98%. In reality, Random Forest detects 52% more customers that are likely to leave.

With ROC–AUC scores of 0.8540 for XGBoost and 0.8473 for Random Forest, both models perform almost equally well when it comes to probability estimation. Both models successfully differentiate between churners and non-churners, as evidenced by the small difference (0.0067). Their different trade-offs cause them to act differently: Random Forest with class_weight='balanced' , focuses on identifying more minority-class instances, whereas XGBoost's boosting method and regularization drive it to achieve higher accuracy and reduce false positives.

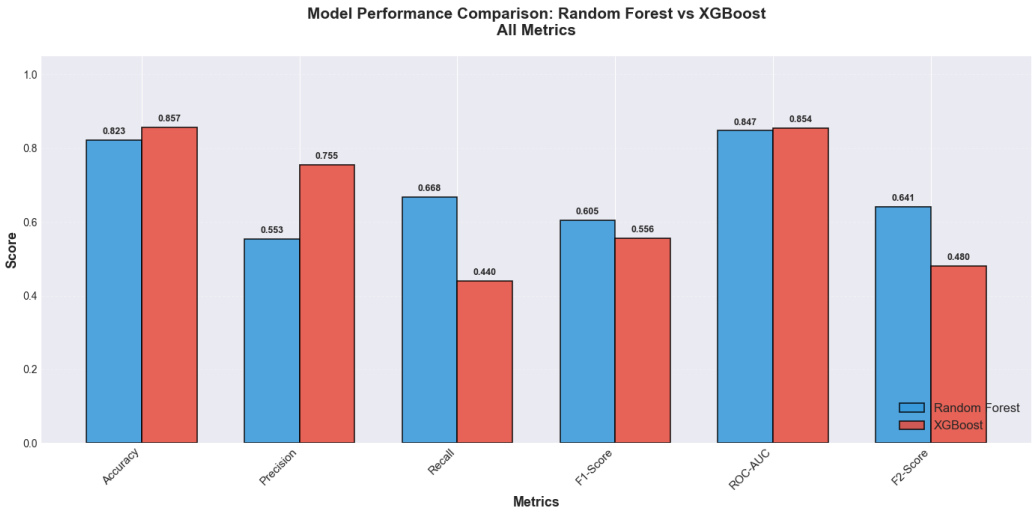

Figure 15 Model Performance Comparision: Random Forest vs XGBoost

- ## Business-Driven Model Selection

Predicting customer churn involves a cost imbalance: assuming that a false negative (missed churner) results in a 1,000 dollars revenue loss for the bank, while a false positive (unneeded retention offer) incurs a 50 dollars expense. Given this theory, recall gains greater importance than precision, rendering the F2-Score the most suitable metric.

Random Forest obtains an F2-Score of 0.6415, greatly surpassing XGBoost (0.4799). This enhancement allows Random Forest to detect around 94 more churners per 2,000 customers than XGBoost.

Cost-Benefit Analysis (Considering the previously mentioned cost framework):

Random Forest

- True positives (approximately 267) - 267,000 dollars preserved
- Inaccurate positives (approx. 216) - 10,800 dollars expense
- Total advantage: 256,200 dollars

XGBoost (Optimized, cutoff 0.35)

- True positives (approx. 176) - 176,000 dollars in savings
- Incorrect positives (approx.57) - 2,850 dollars expense
- Net gain: 173,150 dollars

Consequently, Random Forest generates 83,050 dollars additional net value, a 48% increase in ROI, based on these costs and confusion matrix metrics. This positions Random Forest as the more efficient model for practical churn management, despite XGBoost's superior accuracy and precision

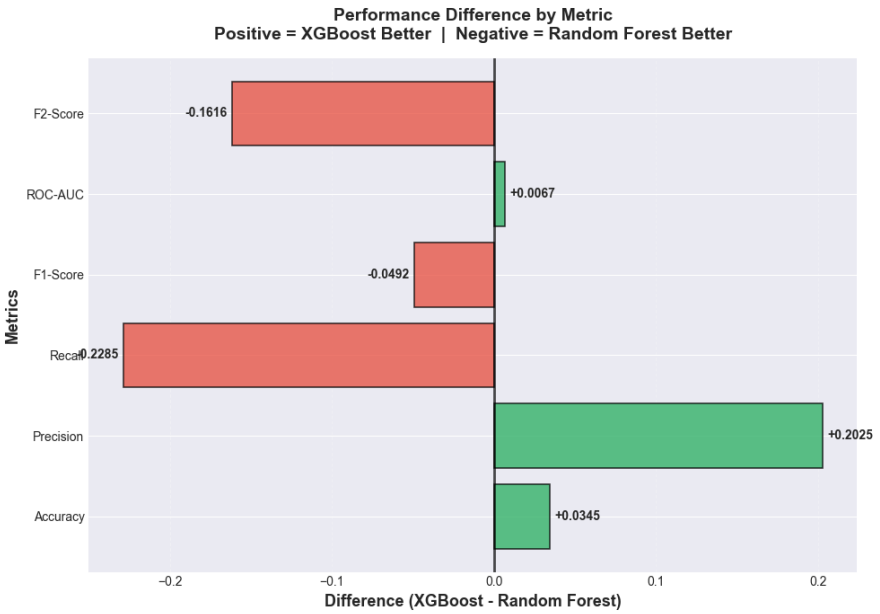

Figure 16 Performance Difference by Metric

- ## Feature Importance & Predictive Drivers

Both models emphasized the same primary predictors, indicating that these trends represent genuine customer behavior instead of being effects specific to the models. The most important factor was age (25% importance). Churn rises sharply after 40 and peaking between 50 and 60, that may be due to career changes or retirement plans.

Customers with exactly two products had the lowest churn, whereas people with one product (low engagement) or three or more products (possible complexity or dissatisfaction) are at a higher risk. The number of products showed a clear U-shaped trend.

Account Balance also influenced churn. Customers with high balances ($100K +) were more likely to leave, likely because financially secure clients compare alternatives more

actively. Geography revealed strong differences as well, with Germany's churn rate (32%) much higher than France's (16%), suggesting region-specific competitive issue.

The consistency of these effects across both Random Forest and XGBoost increases confidence in their validity. Surprisingly, Tenure and IsActiveMember had low importance, indicating that simply being with the bank longer or being "active" does not guarantee loyalty if other needs are unmet.

- ## Limitations & Future Directions

Despite strong performance, the models have several limitations:

1. **False positives remain high** for Random Forest (approx.. 44.7% positive rate), potentially causing over-targeting, although this is financially acceptable given the cost imbalance.
2. **Recall ceiling:** Even Random Forest misses 33% of churners (almost 129 customers), translating into significant unavoidable revenue loss.
3. **Dataset is cross-sectional**, limiting the model's ability to detect behavioural trends such as declining engagement.
4. **Limited feature space:** Additional behavioural indicators (transaction frequency, complaints, product satisfaction) could improve recall further.
5. **Threshold optimization:** Only a single adjusted threshold (0.35) was explored for XGBoost; a systematic sweep maximizing F_2 may push recall toward 70–75%.
6. **Temporal validation:** The model was evaluated across static train–test splits. Testing across different time periods is needed to detect model drift.

- ## Practical Suggestions

Given the performance and cost analysis, Random Forest is recommended for deployment. A tiered intervention strategy based on predicted churn probability can improve efficiency:

- Level 1 (Above 70%): To persuade these high-risk customers to stay, relationship managers must speak with them directly and make special offers or help.
- Level 2 (50–70%): Customers in this medium-risk range can be engaged by automated retention tactics like targeted product recommendations and emails or SMS messages.
- Level 3 (30–50%): In order to keep their relationship with the bank, lower-risk clients primarily require minimal supervision and sporadic check-ins.

Retention efforts should prioritize:

Customers aged 40–60, who show the steepest increase in churn risk.

Customers with 3+ products, where dissatisfaction may be masked by product load.

German customers, whose regional churn patterns suggest competitive or service challenges.

High-balance customers (100k+), who may be comparing financial alternatives.

Product strategy: Encourage adoption of 2-product bundles, the configuration associated with lowest churn.

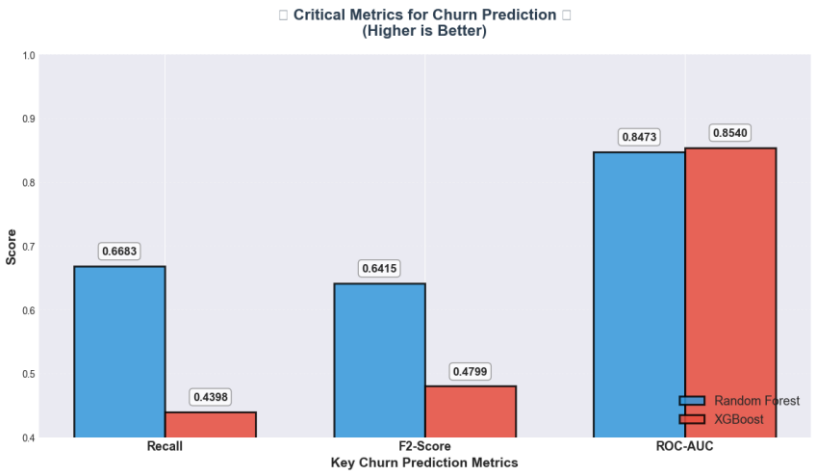

Figure 17 Critical Metrics for Churn prediction

This evaluation shows that Random Forest outperforms XGBoost for churn prediction when business impact is considered. It achieves a 34 percent higher F_2-score (0.6415 vs. 0.4799) and delivers a 48 percent greater ROI, generating 256,200 dollars in net savings compared to XGBoost's 173,150 dollars per 2,000 customers. Although XGBoost is more precise, Random Forest's higher recall (66.83 percent) is far more valuable because missing a churner costs the bank twenty times more than sending an unnecessary retention offer.

The model highlights several key churn drivers, including age, number of products, account balance, and geography. Notably, customers with three or more products show higher churn risk. It is the finding that challenges the assumption that having more products always increases loyalty.Out of 2,000 customers, Random Forest is predicted to correctly identify roughly 267 churners, adding over $80,000 in value.

Subsequent enhancements might concentrate on adjusting classification thresholds to bring recall closer to 70%, incorporating behavioral time-series features to identify early indicators of disengagement, and validating the model over various time intervals to track drift.

Conclusion:

This project combined EDA, customer segmentation, and supervised learning to understand and predict customer churn in a banking context. The segmentation analysis revealed clear behavioural groups, with the highest churn observed among older customers and inactive high-balance clients. These insights give the bank specific segments to focus on for targeted retention efforts.

Random Forest and XGBoost both models performed well, but in different ways. While Random Forest produced significantly better recall, a crucial metric when the cost of missing a churner is significantly higher than sending an unnecessary offer, XGBoost produced higher accuracy and precision. With the highest financial value under the assumed cost structure, Random Forest was found to be the best model for real-world deployment using F2-Score and a cost-benefit analysis.

Overall, the results show that churn is strongly influenced by factors such as age, number of products, account balance, and geography. Moreover , the project demonstrates how machine learning can assist the bank in taking early action and creating more focused retention tactics. The model could be enhanced in the future by incorporating behavioral time-series data, carefully modifying classification thresholds, and routinely assessing its performance to maintain its dependability over time.

References

Chen, T. &. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* doi: https://doi.org/10.1145/2939672.2939785

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), Pattern Recognition Letters. doi: https://www.sciencedirect.com/science/article/abs/pii/S016786550500303X

He, H. &. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263-1284. doi: https://ieeexplore.ieee.org/document/5128907

kumar, S. (n.d.). *Bank Customers Churn*. Retrieved from kaggle: https://www.kaggle.com/datasets/santoshd3/bank-customers

Pedregosa, F. e. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*, 2825-2830.

Verbeke, W. M. (2012). Building comprehensible customer churn prediction models with advanced rule induction techniques. *Expert Systems with Applications*. doi: https://doi.org/10.1016/j.eswa.2010.08.023