# Predictive Analytics for Customer Churns in Financial Services
# Capstone Project Report

Author: Thant Thiha,
2025178@student.cct.ie
Higher Diploma in Data Analytics for Business,
CCT College Dublin

## 1. EXECUTIVE SUMMARY

This report presents a customer churn prediction analysis achieving **approximately 86% ROC-AUC** with statistically validated features and interpretable AI models. The number of products held by customers emerged as the strongest churn driver, with age and geography also significantly impacting risks. Ensemble models like Random Forest and Gradient Boosting offered the best predictive performance. Ethical AI principles were implemented to ensure fairness, transparency, privacy, and accountability.

Business insights enabled targeted retention strategies prioritized on multi-product users, specific age groups, and geographic focus areas. Deployment is recommended using the tuned Random Forest model with ongoing monitoring and governance to sustain performance and fairness. Future efforts will enhance feature sets, interpretability, and real-time integration for improved churn management.

## 2. PROJECT MANAGEMENT METHODOLOGY (CRISP-DM)

### 2.1 Overview of CRISP-DM

This initiative follows CRISP-DM phases, aligns deliverables to business and technical goals, and embeds risk mitigation throughout (Chapman et al., 2000). CRISP-DM (Cross-Industry Standard Process for Data Mining) is an industry-proven, iterative data science methodology that provides a structured approach to planning and executing data mining projects. Developed in the late 1990s by a consortium of industry leaders, it remains the most widely used framework for data science projects across industries.


*Figure 1: CRISP-DM Framework (Chapman et al., 2000)*

## 2.2 CRISP-DM Application in This Project

This churn prediction project follows CRISP-DM rigorously to ensure systematic execution and business value delivery over the duration of 12 weeks.

*Table 1: CRISP-DM Phases and Activities*

| Phase | Key Activities in This Project | Deliverables | Timeline (weeks) |
|---|---|---|---|
| Business Understanding | Defined churn prediction objectives, identified stakeholders, established success metrics (>75% ROC-AUC) | Business objectives document, success criteria | 1 |
| Data Understanding | Explored 10,000 customer records, conducted EDA, performed statistical hypothesis tests | Data quality report, visual exploration, correlation analysis, statistical validation | 2-3 |
| Data Preparation | Remove irrelevant features, encoded variables, engineered new features, applied SMOTE | Clean dataset, feature matrix, balanced training set | 4-5 |
| Modeling | Trained 5 algorithms, performed hyperparameter tuning, cross-validation | Trained models, performance baselines, optimized parameters | 6-8 |
| Evaluation | Compared models on multiple metrics, conducted interpretability analysis (SHAP/LIME), bias detection | Model comparison report, feature importance, fairness assessment | 9-11 |
| Deployment | Models will not be deployed to production in this project | N/A | 12 |

## 2.3 Tools and Technologies Used

There are several tools and technologies used in this project and they are documented as follows.

*Table 2: Tools and Technologies used in the project*

| Name | Description | Usage |
|---|---|---|
| Python libraries (Scikit-learn, Pandas, Matplotlib, Seaborn) | Open-source libraries for data analysis, machine learning and visualization | Data preprocessing, Exploratory Data Analysis(EDA), model building, evaluation, and visualization |
| GitHub | Cloud-based version control and collaboration platform | Code repository, version control, project collaboration and documentation management |
| MS Word | Word processing software | Writing, formatting and editing the final project report and documentation |
| Jupyter Notebook | Interactive computing environment for Python | Developing, documenting and sharing code, visualizations and narrative text for the project |

# 3. IMPROVEMENTS FROM SEMESTER I

Enhancements Implemented are as follows.

## Methodological Improvements

Rigorous statistical hypothesis testing, including Chi-Square and Mann-Whitney U tests, was added to validate feature relationships. SMOTE was implemented to address class imbalance, which had previously not been handled. The model suite was expanded from the earlier number of models to five algorithms, and a systematic comparison was carried out. Comprehensive model interpretability techniques such as SHAP and LIME were introduced. Bias detection and fairness analysis were also incorporated into the workflow.

### Technical Enhancements

Advanced feature engineering was performed, resulting in five new derived features compared with the earlier feature set. Hyperparameter tuning using five-fold cross-validation was applied instead of the previous approach. A more robust evaluation framework was established, using multiple performance metrics beyond accuracy. Model-agnostic explanation methods were also added to support transparent communication with stakeholders.

### Analytical Depth

All feature relationships were statistically validated before modeling. Effect size calculations, including Cramér's V and rank-biserial correlation, were completed to enhance the analysis. Individual prediction explanations were generated to improve business actionability. Demographic fairness testing across customer segments was also conducted to ensure equitable model performance.

## 4. BUSINESS UNDERSTANDING

Customer churn represents a critical challenge in banking, where acquiring new customers costs 5-7 times more than retaining existing ones (Pfeifer, 2005). Understanding which customers are likely to leave enables proactive intervention and resource optimization (Owolabi et al., 2024).

### 4.1 Business Objectives

The primary goal of the project was to develop a predictive model capable of identifying customers who are at high risk of churning before they actually leave.

### 4.2 Success Criteria

The success of the model was defined by several criteria. First, the model was required to achieve an ROC-AUC score higher than 75%, which aligns with industry benchmarks (Gandomi and Haider, 2015). Second, it was expected to identify the top five drivers of churn to support targeted intervention strategies (Adebayo and Kusi, 2021). Third, the predictions generated by the model needed to be explainable so that customer service teams could use them effectively. Finally, fairness across demographic groups had to be ensured to maintain ethical and equitable model performance (Barocas and Narayanan, 2019).

### 4.3 Business Impact

The implementation of this predictive model was expected to generate several key business benefits. The churn rate was anticipated to decrease through earlier and more proactive intervention. Retention marketing budgets were expected to be allocated more efficiently by focusing on customers who were most at risk (Ascarza, 2018). Customer lifetime value was projected to increase as a result of more effective retention efforts. The model was also expected to enable personalized retention strategies supported by data-driven insights.

The stakeholders involved in this project included the marketing team, which would use the findings to design targeted campaigns; the customer service team, which would conduct proactive outreach to at-risk customers; the product team, which would leverage the insights to guide feature development; and executive leadership, which would use the results for strategic planning.

## 5. DATA UNDERSTANDING

The foundation of this project is the Bank Churn Modelling.csv dataset, sourced from the public repository of the YBI Foundation on GitHub. This dataset simulates comprehensive customer information for a hypothetical bank and serves as a robust placeholder for developing and benchmarking churn prediction models in financial services.

## 5.1 Dataset Overview

**Size**: 10,000 customer record, each representing an individual bank customer.

**Variables**: 14 columns capturing a mix of demographic, behavioral and financial information as well as churn status.

**No PII**: The dataset is fully anonymized, containing no personally identifiable information ensuring compliance with data protection regulations, GDPR and ethical standards.

## 5.2 Key Features

**Demographic Information**

- *Geography*: Country of residence (France, Spain, Germany)
- *Gender*: Male or Female
- *Age*: Customer's age

**Banking Relationship Details**

- *Tenure*: Number of years the customer has been with the bank
- *NumOfProducts*: Number of bank products used by the customer
- *HasCrCard*: Whether the customer owns a credit card (1=Yes, 0=No)
- *IsActiveMember*: Whether the customer is an active member (1=Yes, 0=No)

**Financial Indicators**

- *CreditScore*: Creditworthiness score
- *Balance*: Account balance
- *EstimatedSalary*: Estimated annual salary

**Churn Status**

- *Exited*: Target variable indicating whether the customer has left the bank or not (1=Churned, 0=Not Churned/Retained))

## 5.3 Exploratory Data Analysis

### 5.3.1 Target Distribution

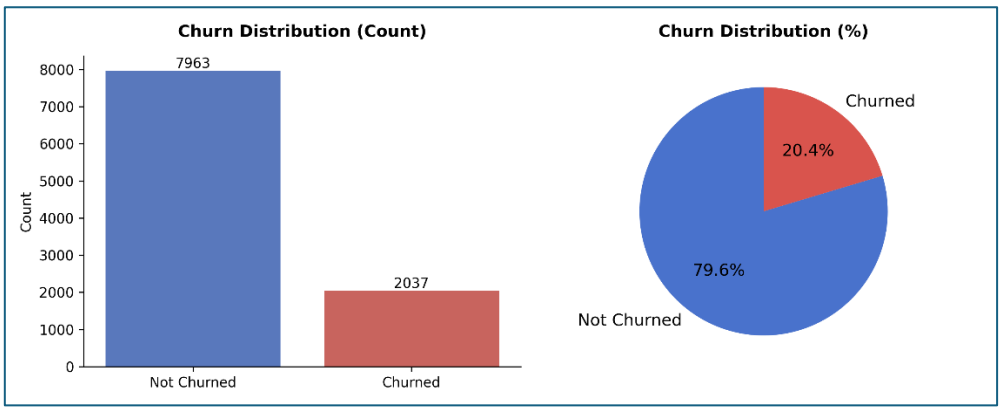


*Figure 2: Target Variable Distribution*

The target variable “churn” shows a churn rate of 20.4%, meaning 2,037 out of 10,000 customers have churned. This indicates significant class imbalance approximately a ratio of 4:1 (not churned to churned) (He and Garcia, 2009). Such imbalance can negatively affect model performance and requires handling; SMOTE (Synthetic Minority Over-sampling Technique) was applied to address this imbalance and enable fairer, more robust classification.

### 5.3.2 Key Patterns in Numerical Features

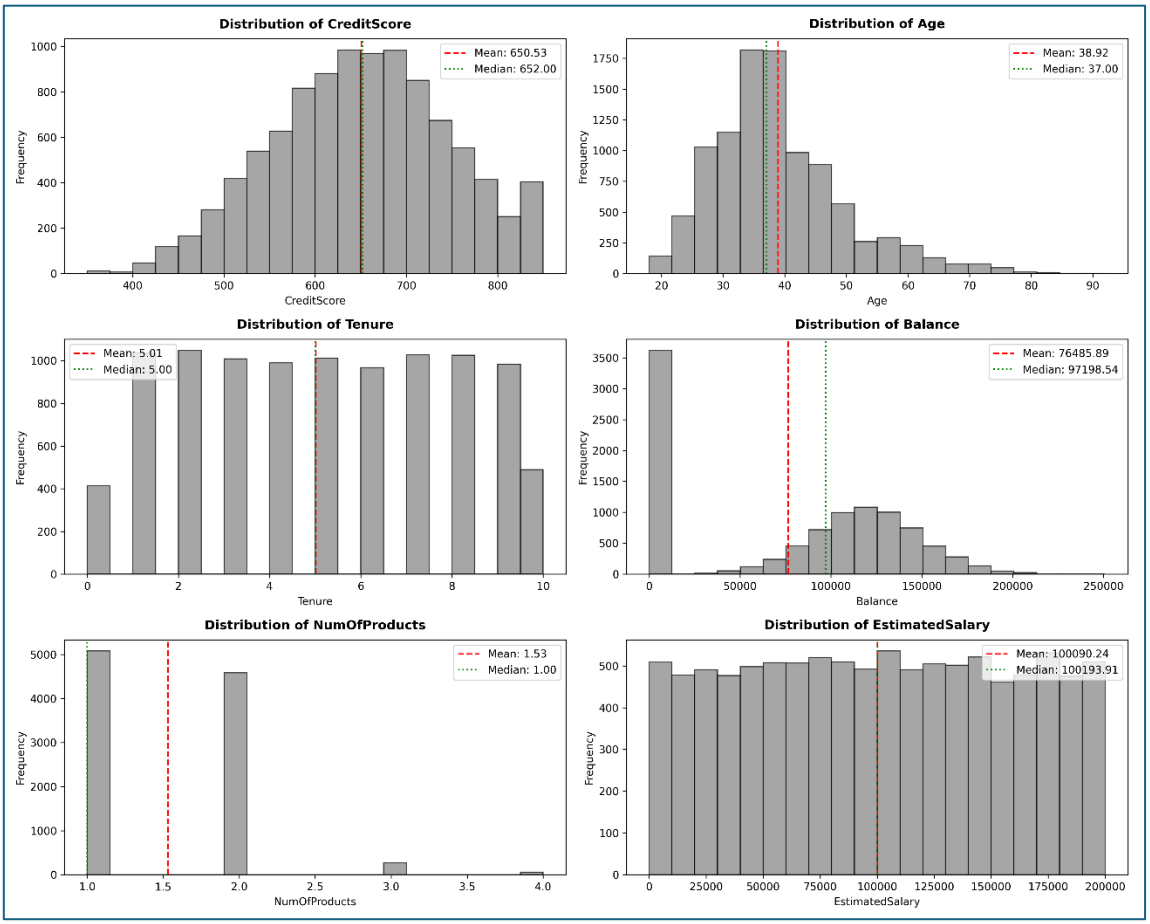


*Figure 3: Numerical Features Distribution*

The numerical features show these patterns: credit scores and ages are approximately normal, but age is a little right-skewed; tenure is evenly distributed; most customers have low or zero balances with a long right tail; almost all have just one or two products; and estimated salary is uniformly spread. The mean and median are close for credit score, age, tenure, and salary, but for balance, the median is higher than the mean, indicating most have modest funds and a few customers have large deposits that skew the average.

### 5.3.3 Key Patterns in Categorical Features

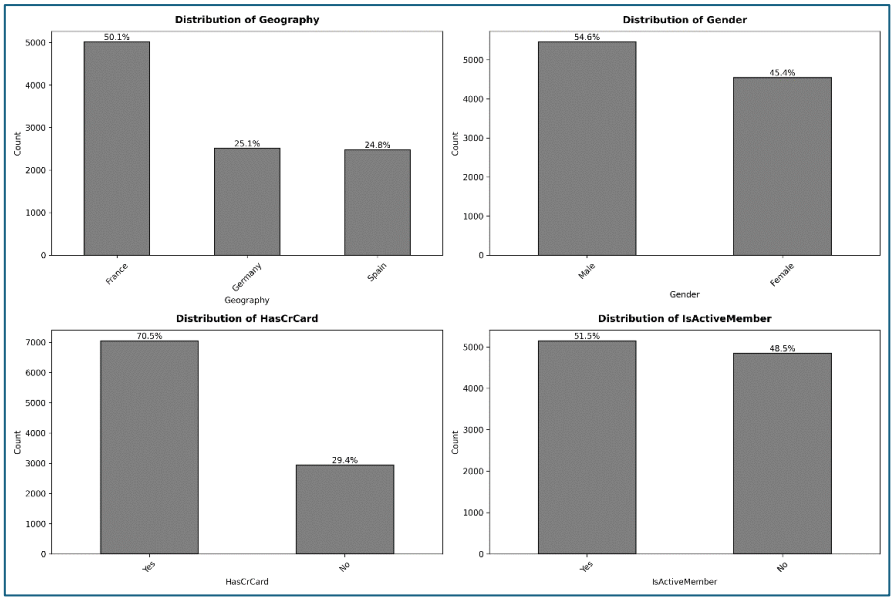


*Figure 4: Categorical Features Distribution*

The categorical features show that most customers are from France (50%), and there are slightly more males (54.6%) than females. The majority have a credit card (70.5%) and the proportion of active members is almost evenly split, with 51.5% active and 48.5% not active.

### 5.3.4 Churn Rate by Categorical Features

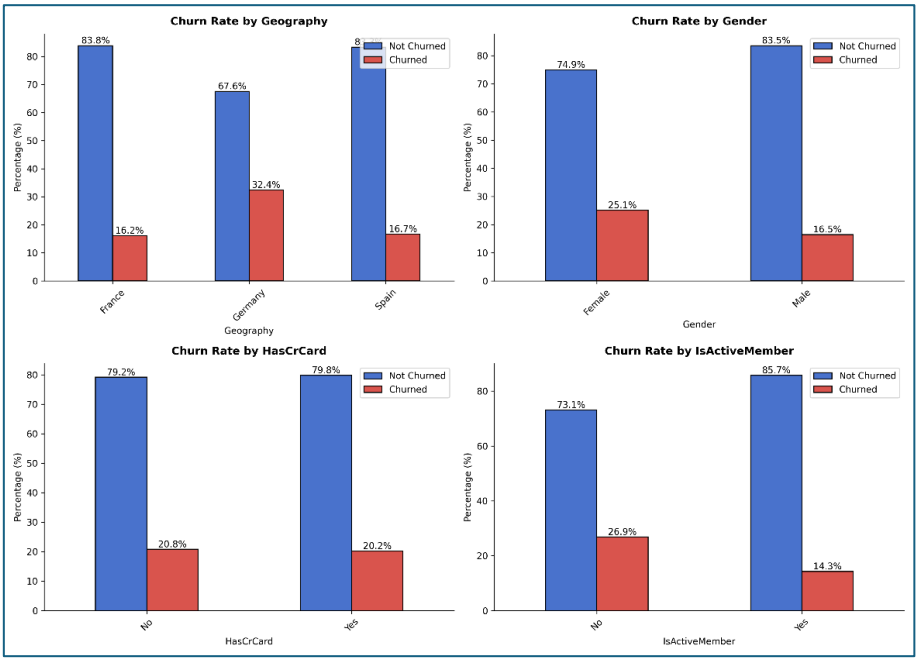


*Figure 5: Churn Rate by Categorical Features*

Churn rates vary significantly by categorical features: customers in Germany and non-active members have the highest churn rates (32.4% and 26.9% respectively), while active members and males are least likely to churn (14.3% and 16.5%). Churn rates are similar regardless of holding a credit card, and France and Spain show notably lower churn rates (16.2–16.7%) compared to Germany. Female customers also exhibit a higher churn rate (25.1%) than males.

### 5.3.5 Churn Rate by Numerical Features

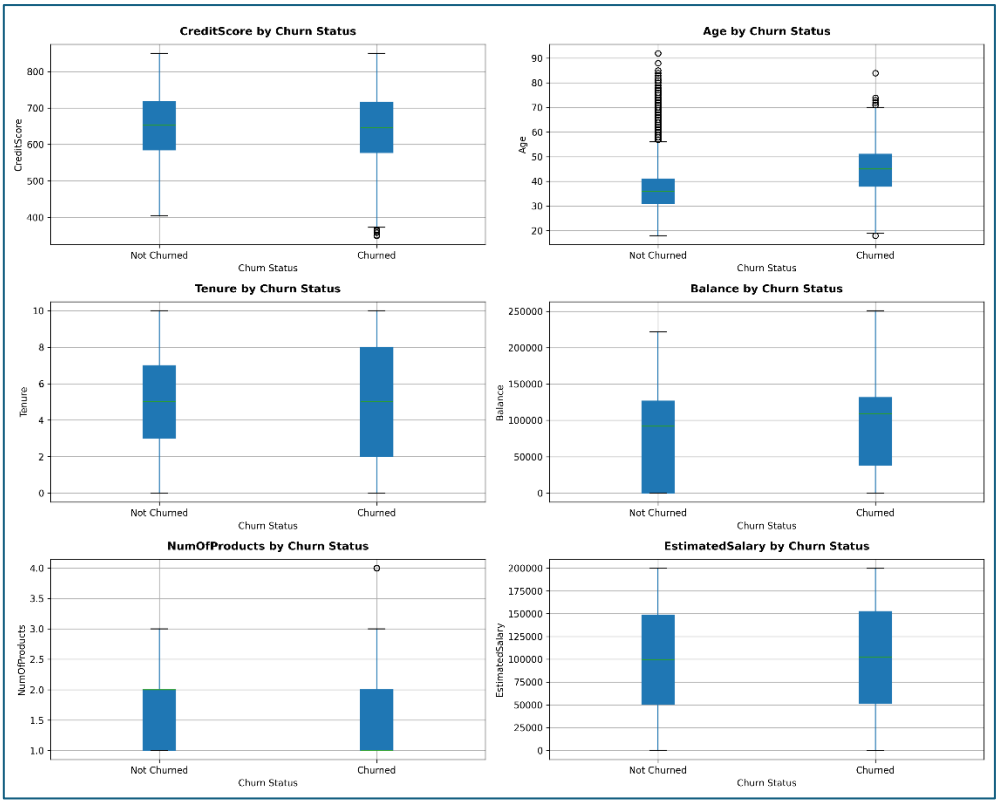


*Figure 6: Churn Rate Box Plots by Numerical Features*

Churn rates by numerical features indicate that customers who churned tend to be older, have slightly higher credit scores, longer tenure, and higher average account balances compared to those who stayed. However, the number of products and estimated salary distributions are similar between churned and not churned groups, suggesting these two features have less influence on churn behaviour.

## 5.4 Statistical Hypothesis Tests

### 5.4.1 CHI-SQUARE TESTS: Categorical Features vs Churn

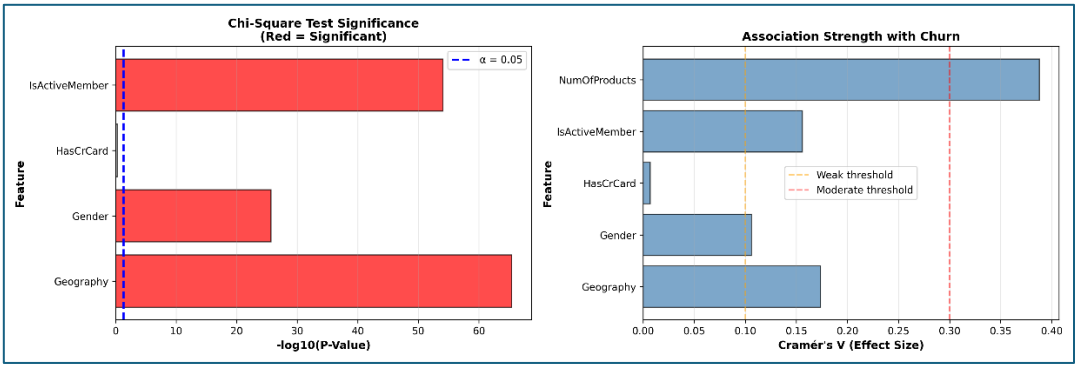


*Figure 7: Chi-Square Test and Association Strength plots*

The chi-square test was used to assess whether each categorical feature is statistically associated with customer churn. For each variable, the null hypothesis (H0) assumes independence from churn, while the alternative (H1) suggests dependency. By comparing p-values to a significance level of $\alpha = 0.05$, features with p-values below this threshold are considered significantly associated with churn (Kotsiantis, 2007).

Results show that Geography, Gender, IsActiveMember, and especially NumOfProducts (Cramér's V = 0.39) all have significant and at least moderately strong associations with churn, as their p-values are virtually zero and Cramér's V values are notable. HasCrCard, however, showed no significant relationship, meaning possession of a credit card is independent from churn status in this sample.

### 5.4.2 MANN-WHITNEY U TEST: Numerical Features vs Churn (NON-PARAMETRIC)

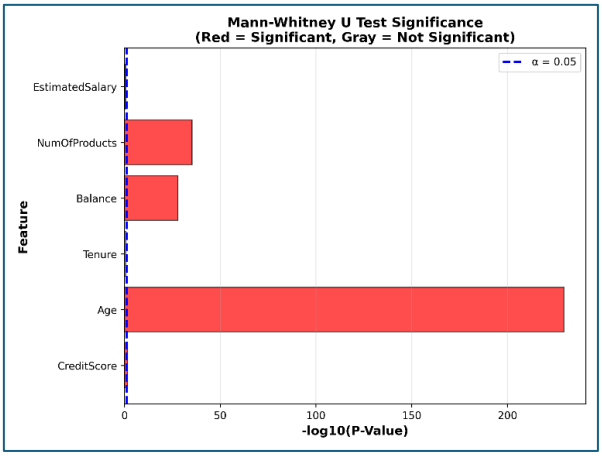


*Figure 8: Mann-Whitney U Test results*

The Mann-Whitney U test was conducted to compare the distributions of numerical features between churned and non-churned customers, with the null hypothesis that both groups have equal distributions. Using a significance threshold of $\alpha = 0.05$, features with p-values below this level are considered to show a statistically significant difference between the two groups (Cai et al., 2018).

Results indicate that Age, Balance, CreditScore, and NumOfProducts all show significant differences between churned and not churned customers, with Age displaying a medium effect size (churned customers are notably older) and the others showing small effects. Tenure and EstimatedSalary do not differ significantly between groups, suggesting these features have less discriminative power for churn in this dataset.

### 5.4.3 Correlation Significance Test

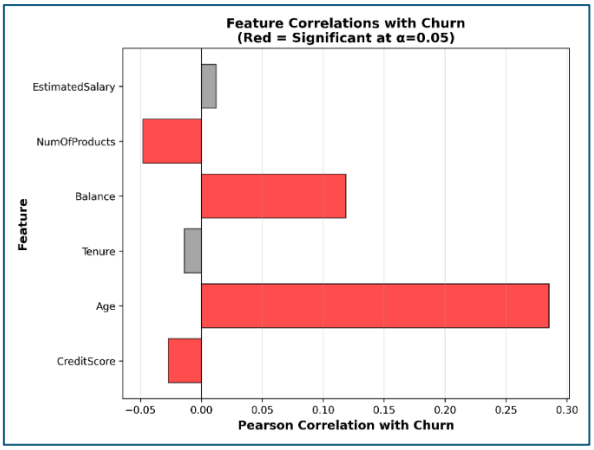


*Figure 9: Correlation Significance Test Results*

Pearson correlation significance test was performed to check if each numerical feature has a linear relationship with churn status. The null hypothesis is that there is no correlation (correlation = 0) between each feature and churn, and the test uses a significance level of $\alpha = 0.05$ to decide if any observed relationship is statistically meaningful (Hadi et al., 2020).

The results show that Age, Balance, NumOfProducts, and CreditScore are all significantly (but weakly) correlated with churn, with Age having the strongest positive correlation (0.29) and NumOfProducts showing a weak negative relationship. In contrast, Tenure and EstimatedSalary have p-values above 0.05, meaning they show no significant correlation with churn in this analysis.

#### 5.4.4 Comprehensive Summary of the results

*Table 3: Statistical Hypothesis Test Results Summary*

| Feature | Type | Mann-Whitney | Correlation | Overall | Chi-square | Cramér's V |
|---|---|---|---|---|---|---|
| Credit Score | Numerical | Yes | -0.027 | Significant | NA | NA |
| Age | Numerical | Yes | 0.285 | Significant | NA | NA |
| Tenure | Numerical | No | -0.014 | Not Significant | NA | NA |
| Balance | Numerical | Yes | 0.119 | Significant | NA | NA |
| NumOfProducts | Numerical | Yes | -0.048 | Significant | NA | NA |
| EstimatedSalary | Numerical | No | 0.012 | Not Significant | Yes | 0.174 |
| Geography | Categorical | NA | NA | Significant | Yes | 0.174 |
| Gender | Categorical | NA | NA | Significant | Yes | 0.106 |
| HasCrCard | Categorical | NA | NA | Not Significant | No | 0.007 |
| IsActiveMember | Categorical | NA | NA | Significant | Yes | 0.156 |
| NumOfProducts | Categorical | NA | NA | Significant | Yes | 0.388 |

Among numerical variables, Age, Balance, CreditScore, and NumOfProducts all showed statistically significant differences between churned and non-churned customers using both the Mann-Whitney U test and Pearson correlation, though effect sizes were small to moderate; Tenure and EstimatedSalary were consistently not significant. For categorical variables, Geography, Gender, IsActiveMember, and NumOfProducts demonstrated significant associations with churn by chi-square test (with Cramér's V indicating Geography and IsActiveMember were moderately associated, while NumOfProducts showed the strongest association). HasCrCard was not significantly linked to churn in either test. Overall, demographic and engagement features—especially Age and the number of products—have the strongest and most consistent relationship with churn across multiple statistical perspectives, while tenure, salary, and credit card holding have little predictive value for churn in this dataset.

## 6. DATA PREPARATION

### 6.1 Data Cleaning

During data cleaning, no missing values were found across all features, outliers were kept to preserve real customer diversity, and non-predictive columns such as CustomerId, RowNumber, and Surname were dropped from the dataset.

### 6.2 Feature Engineering

Five new features were engineered to enhance model performance: BalanceToSalary_Ratio (captures financial commitment relative to income), Age_Group (categorizes age to detect non-linear effects), Is_Senior (flags customers over 50 to identify senior-related churn), Has_Zero_Balance (identifies dormant or new accounts), and CreditScore_Category (segments customers by credit risk tier) to better capture important patterns driving churn (Kuhn and Johnson, 2019; Coussement et al., 2017).

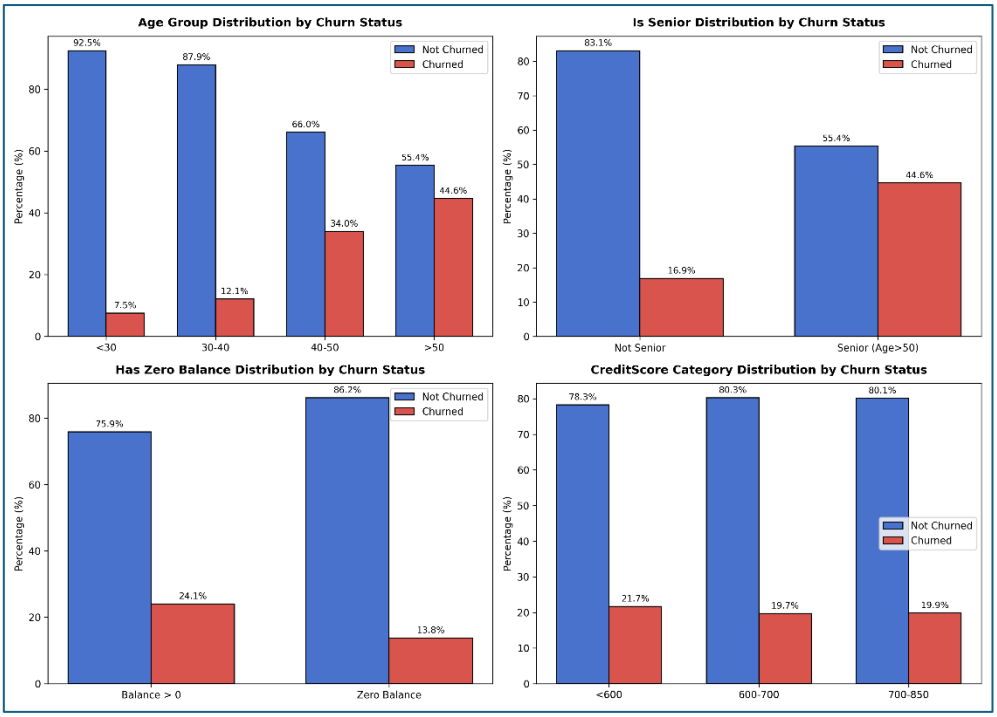


*Figure 10: Churn Rate Distribution of New Features*

Churn rates for the engineered features show sharp contrasts: older age groups and senior status are associated with much higher churn, while young and mid-age customers have low churn. Accounts with zero balance and lower credit scores have modestly lower churn, but the differences between credit score categories are small, suggesting limited utility for segmentation. These results confirm that age and active account status are the strongest churn indicators among the new features.

## 6.3 Feature Encoding

Binary encoding was applied to the Gender feature because it has only two categories (Female=0, Male=1), making it a simple, efficient numerical mapping without adding extra dimensions. One-Hot Encoding was used for Geography as it has multiple categories; this avoids any unintended ordinal relationships by creating separate binary indicators for Germany and Spain while dropping France as the baseline to prevent multicollinearity (Dormann et al., 2013). This combination ensures that categorical data are appropriately prepared for modelling, maximizing interpretability and avoiding issues like the dummy variable trap.

## 6.4 Data Splitting

Data splitting was performed by dividing the dataset into an 80% training set (8,000 samples) and a 20% test set (2,000 samples) using a stratified split strategy. Stratification was chosen to maintain the original distribution of the churn target variable across both training and test sets, ensuring representative and balanced subsets to improve model reliability and evaluation accuracy, especially important given the class imbalance in churn data (Refaeilzadeh et al., 2009).

## 6.5 Feature Scaling

StandardScaler was applied to normalize numerical features by transforming them to have a mean of zero and standard deviation of one. This standardization prevents features with large value ranges from dominating distance-based or gradient-based algorithms and helps models converge more reliably. It was applied after the train-test split to avoid data leakage, ensuring that the scaling parameters are learned solely from training data and applied consistently during testing and prediction (Hastie, Tibshirani and Friedman, 2009).

### 6.6 Class Imbalance Handling

Class imbalance in the original training data with a much lower churn rate than not churned which created a risk of prediction bias toward the majority class. To address this, SMOTE (Synthetic Minority Over-sampling Technique) was applied, generating synthetic examples of churned customers and resulting in a perfectly balanced 50-50 training set (He and Garcia, 2009). The test set was kept with its original distribution to ensure a realistic evaluation of model performance. This approach improved model sensitivity to the minority class, making predictions for churned customers more reliable and robust.

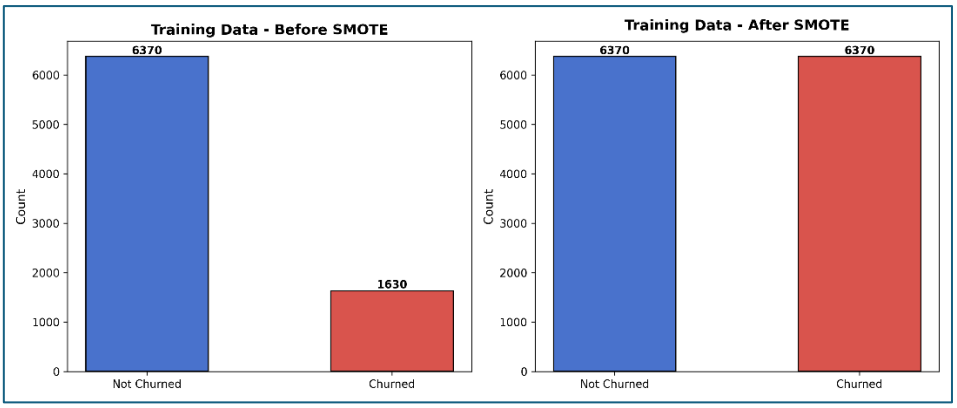


*Figure 11: Training Data before and after SMOTE*

## 7. MACHINE LEARNING MODELLING

### 7.1 Algorithm Selection

Five diverse machine learning algorithms were selected to capture different patterns in churn prediction. Random Forest serves as the primary predictor due to its robust ensemble approach and ability to handle non-linearity (Verbeke et al., 2012), while Gradient Boosting is included as a performance benchmark for its strength in capturing complex feature interactions. Decision Tree provides a fast, interpretable baseline that is useful for explaining predictions (Molnar, 2022). Support Vector Machine was selected for its effectiveness in modeling non-linear decision boundaries using kernels. Logistic Regression was included for its speed, regulatory interpretability, and as a strong linear baseline to compare against more complex models (Lessmann et al., 2015). This mix ensures a comprehensive comparison of algorithm strengths for the churn use case.

### 7.2 Training Strategy

All five selected algorithms were initially trained using default hyperparameters to establish baseline performance on identical train-test splits. Each model's results including ROC-AUC as the main metric for imbalanced data, as well as accuracy, precision, recall, F1-score, and confusion matrices - provided a consistent reference point to compare model behaviours and identify candidates for further tuning or deployment. This strategy ensures objective benchmarking and robust error analysis throughout the modelling workflow.

### 7.3 Hyperparameter Tuning and Cross-Validation

Hyperparameter tuning was performed using GridSearchCV with 5-fold cross-validation to systematically explore combinations of model parameters and identify those that maximize ROC-AUC, which is especially important for handling class imbalance (Bishop, 2006).

Parameter grids were tailored to each algorithm: Logistic Regression included different strengths and types of regularization (l1 and l2) to control overfitting, Decision Tree grids adjusted tree complexity with max depth, min sample splits, min sample leafs, Random Forest varied the number and depth of trees, and Gradient Boosting included learning rate and tree depth for fine-tuning (Chen and Guestrin,

2016; Géron, 2019). Parallel processing (n_jobs=-1) was used to make this exhaustive search computationally efficient. These choices ensure fair, robust optimization and help each model reach its best predictive performance for the churn problem.

# 8. EVALUATION

## 8.1 Model Performance Comparison

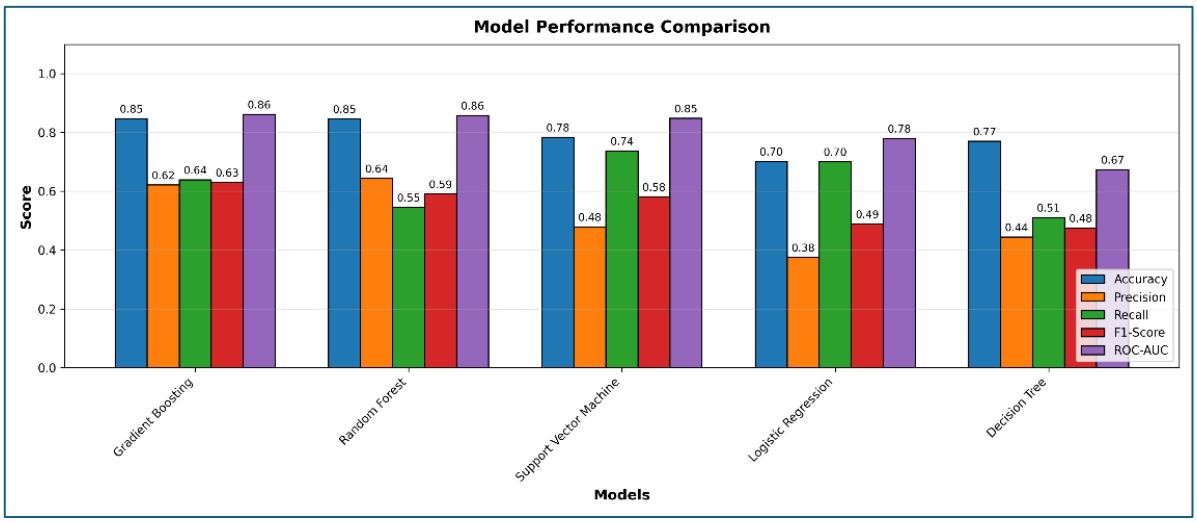


*Figure 12: Model Performance Comparison*

Gradient Boosting and Random Forest clearly outperform the other models, achieving the highest ROC-AUC (0.86), accuracy (0.85), and balanced precision/recall/f1-scores, indicating strong and reliable classification of churned customers. Support Vector Machine is slightly behind but maintains solid performance. Logistic Regression and Decision Tree both lag in ROC-AUC (0.78 and 0.67, respectively) and especially in precision and F1-score, underscoring their limitations for capturing non-linear relationships or complex patterns in the data. Overall, ensemble models (especially boosting and bagging) deliver the best results for this churn prediction task, offering both high accuracy and improved robustness to class imbalance.

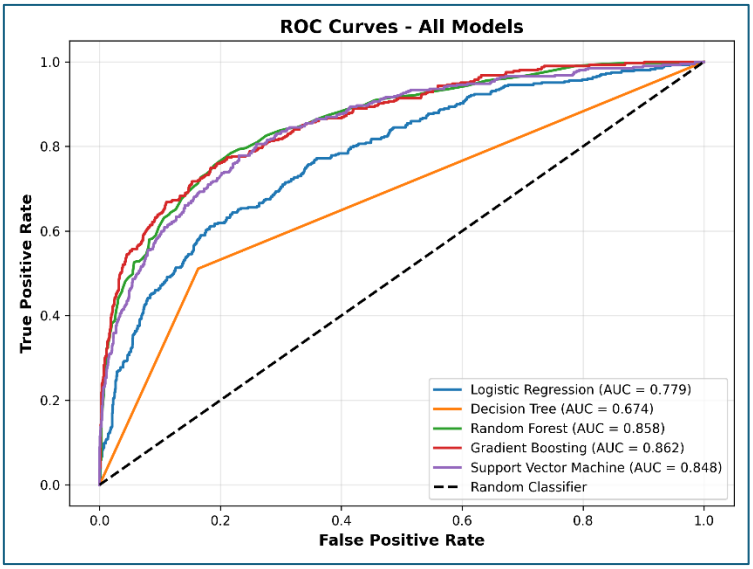


*Figure 13: ROC Curves Comparison*

ROC curves show that Gradient Boosting and Random Forest are the best discriminators for churn, with AUC scores (0.862 and 0.858) exceeding other models; both strongly outperform the baseline

Decision Tree and Logistic Regression. Support Vector Machine also performs well (AUC 0.848), but falls just short of ensemble methods.

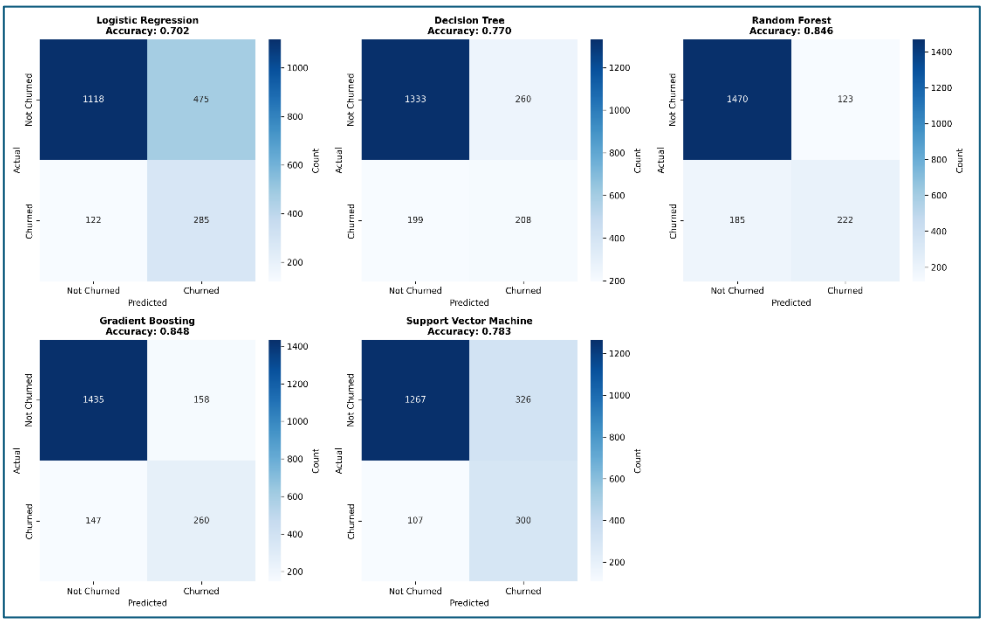


*Figure 14: Confusion Matrix of All Models*

Confusion matrices confirm that ensemble models correctly classify churned customers at higher rates and with fewer false negatives, while single-model approaches—Logistic Regression and Decision Tree which struggle with recall, missing a larger share of actual churn cases. These results reinforce that ensemble approaches are optimal when both accurate and sensitive churn prediction is needed.

## 8.2 Hyperparameter Tuning and Cross-Validation Results

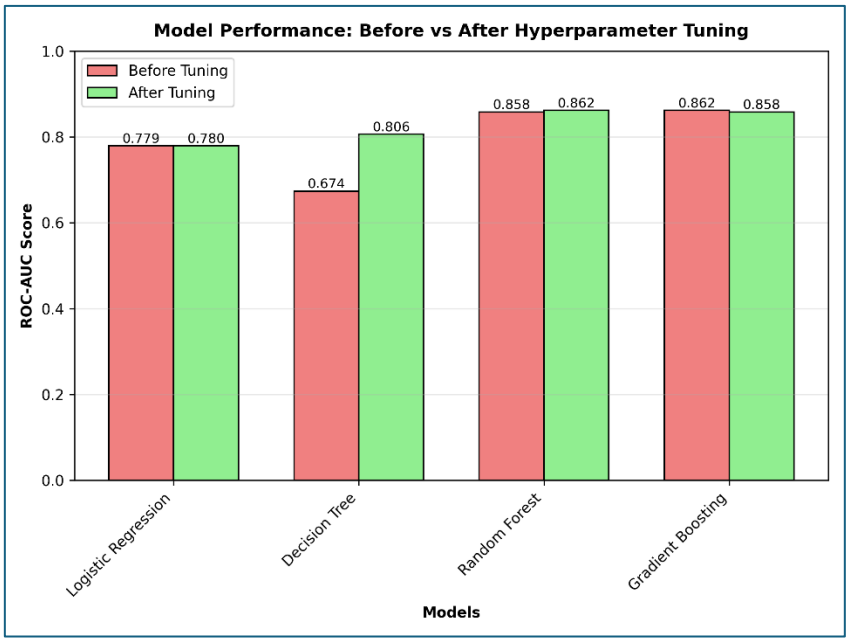


*Figure 15: Model Performance (ROC scores) before and after Tuning*

Hyperparameter tuning using GridSearchCV resulted in modest but important improvements for most models, with Decision Tree benefitting most (ROC-AUC rising from 0.67 to 0.81 on the test set). Random Forest and Gradient Boosting saw slight incremental gains in ROC-AUC (~0.86), confirming their strong out-of-the-box performance and minor sensitivity to parameter changes. Logistic Regression's metrics changed slightly, as expected for a linear model and low-dimensional parameter grid. These adjustments demonstrate that carefully tuning model parameters can maximize discriminatory power and yield measurable improvements, especially for tree-based methods.

*Table 4: 5-fold Cross-Validation Results*

| Model | Mean CV Score | Std CV Score | Min CV Score | Max CV Score |
|---|---|---|---|---|
| Logistic Regression | 0.784 | 0.011 | 0.772 | 0.798 |
| Decision Tree | 0.907 | 0.045 | 0.817 | 0.939 |
| Random Forest | 0.971 | 0.024 | 0.924 | 0.988 |
| Gradient Boosting | 0.961 | 0.050 | 0.864 | 0.995 |

5-fold cross-validation revealed the most robust and stable performance from Random Forest and Gradient Boosting (mean CV-ROC-AUC 0.97 and 0.96), with low standard deviation and consistently strong minimum and maximum scores across folds. Decision Tree showed more substantial variation, while Logistic Regression's CV scores were the most modest and stable. This confirms that ensemble models not only achieve higher accuracy but also generalize more reliably to new data, making them preferable choices for churn prediction in practice. Individual cross-validated metrics reinforce confidence in the chosen tuning strategy and allow for unbiased model selection.

## 8.3 Model Interpretability

Model-specific explanations provide detailed insights into how individual features drive predictions in different algorithms. Intrinsically interpretable models like trees and logistic regression directly reveal the most important global predictors via feature importance scores and coefficients.

For more complex or black-box models, local interpretation approaches such as LIME and SHAP offer case-by-case visualizations, allowing analysis of the specific features contributing to a single prediction and supporting transparent decision-making (Ribeiro et al., 2016; Lundberg and Lee, 2017).

### 8.3.1 Tree-based Feature Importance (Model-specific)

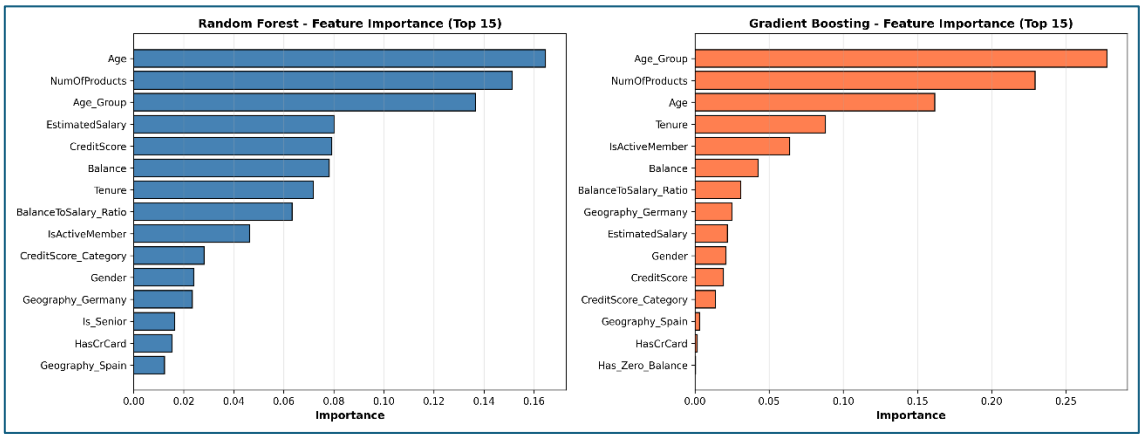


*Figure 16: Feature Importance plots of Tree-based Models*

Tree-based models consistently identify Age, Number of Products, and Age_Group as the most important predictors of churn, followed by EstimatedSalary, CreditScore, and Balance. Both Random Forest and Gradient Boosting highlight these features at the top, though their exact order differs, demonstrating that age-related factors and customer engagement (via product usage) are central drivers of churn. Account status flags and geography lag far behind, underlining that demographic and behavioral variables are most relevant for accurate churn prediction in this dataset.

### 8.3.2 Logistic Regression Coefficients (Model-specific)

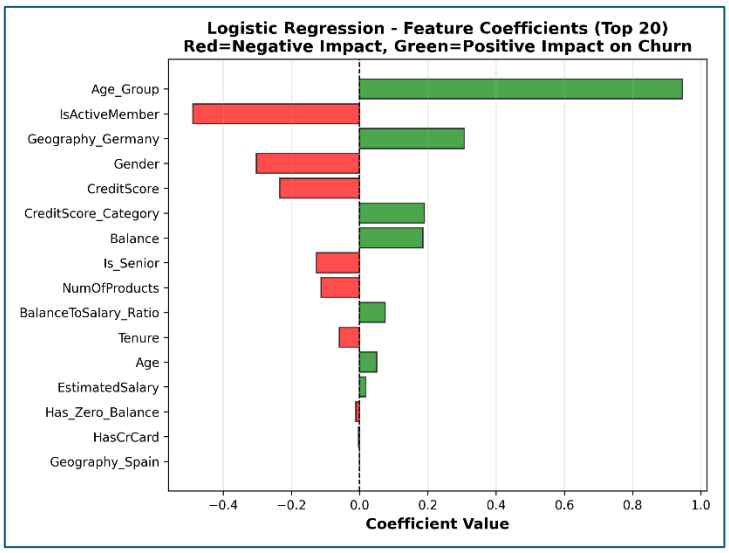


*Figure 17: Coefficients of Logistic Regression Model*

Logistic Regression coefficients reveal that Age_Group and Geography_Germany have strong positive impacts on churn, meaning being older or located in Germany significantly increases churn risk, while IsActiveMember, Gender, and higher CreditScore show large negative coefficients, indicating these factors reduce churn likelihood. Many engineered and original features, such as Balance, CreditScore_Category, and Tenure, have moderate effects but in both directions. This aligns well with tree-based importances, confirming age, engagement, and regional factors as the dominant drivers, while also offering clear, interpretable directionality for how each feature influences churn probability.

### 8.3.3 SHAP Analysis (Model-agnostic)

SHAP (SHapley Additive exPlanations) is a powerful technique that explains individual machine learning predictions by quantifying each feature's contribution (Lundberg and Lee, 2017). Based on game theory's Shapley values, it assigns fair importance scores to features by considering all possible combinations and their impact on the model's output. SHAP offers consistent, locally accurate, and model-agnostic explanations for both global behavior and single-instance predictions. Tools like SHAP summary plots and waterfall plots visually represent these contributions, making complex model decisions interpretable and transparent. This enables stakeholders to understand not only which features are important overall, but precisely how they influence each specific prediction.

SHAP Summary Plot

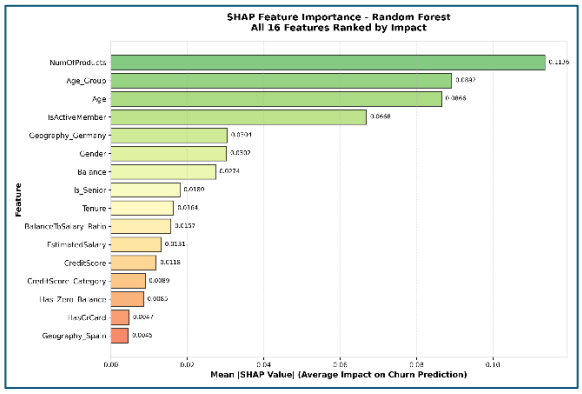


*Figure 18: SHAP Feature Importance Plot*

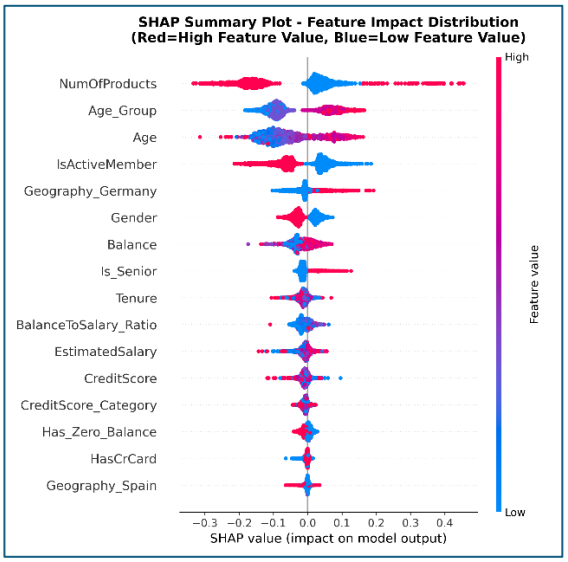


*Figure 19: SHAP Beeswarm Summary Plot*

The SHAP summary plot shows that NumOfProducts, Age_Group, and Age are the top contributors to churn risk, with high values for these features consistently pushing predictions higher. IsActiveMember, Geography_Germany, and Gender also provide strong, directional impact. The beeswarm plot reveals clear separation, red for high feature values, blue for low demonstrating that frequent product use, older age, German residency, and inactivity drive up churn probability most strongly in the model.

SHAP Waterfall Plot with Individual Prediction Example

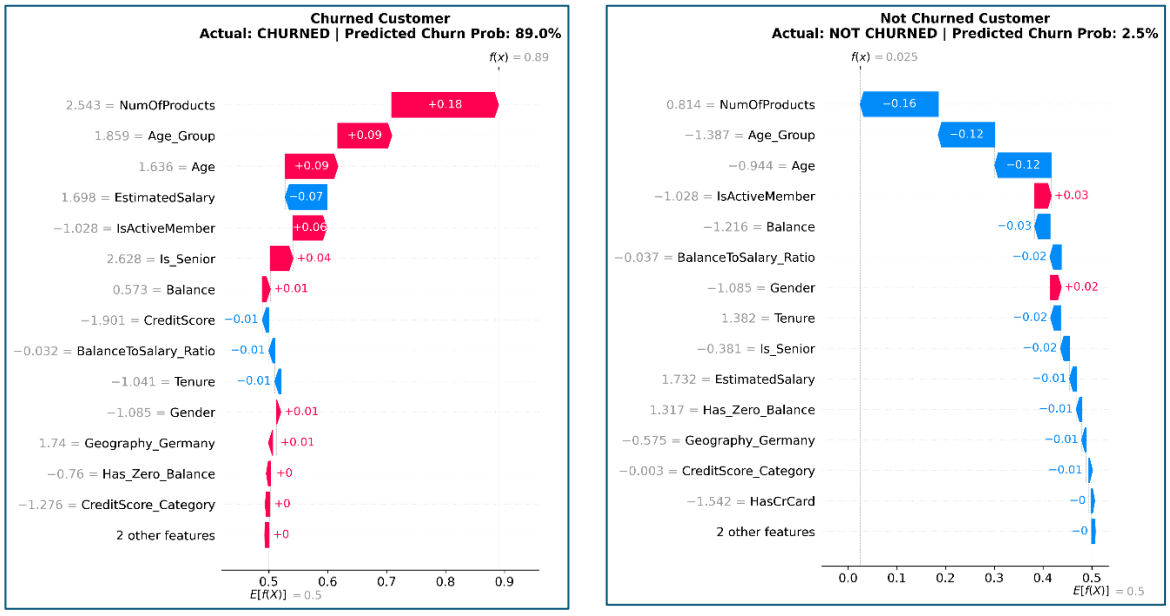


*Figure 20: SHAP Waterfall Plot with Individual Prediction Explanation*

The SHAP waterfall plots clearly illustrate feature-level impacts for each individual prediction from different classes. For the churned customer, high product count, older age, and inactive membership strongly increase churn probability, while features like higher estimated salary and credit score have modest negative effects. Conversely, the not-churned customer's low product count, younger age, and

active status drive down churn probability, with most key features deducting from the risk. These results provide transparent, individualized explanations and confirm the model's reasoning aligns with the true outcome for both profiles.

### 8.3.4 LIME Analysis (Model-agnostic)

LIME (Local Interpretable Model-agnostic Explanations) is a technique used to explain individual predictions from any complex machine learning model. Instead of trying to interpret the entire model globally, LIME builds a simple, interpretable surrogate model focused locally around a single data instance (Molnar, 2022). By perturbing the input features near that instance and observing the black-box model's outputs, LIME determines which features most influence the specific prediction, providing a transparent explanation of the model's decision for that case. This approach enhances trust and understanding of complex models by revealing the key drivers behind each individual prediction.

LIME Explanations for Multiple Models

*Table 5: Profile (Features) of a sample customer 1*

| | |
|---|---|
| CreditScore | 466 |
| ender | 0 (female) |
| Age | 56 |
| Tenure | 2 |
| Balance | 111,920 |
| NumOfProducts | 3 |
| HasCrCard | 1 (yes) |
| IsActiveMember | 0 (no) |
| EstimatedSalary | 197634.11 |
| Geography_Germany | True |
| Geography_Spain | False |
| BalanceToSalary_Ratio | 0.56629 |
| Age_Group | 3 (>50) |
| Is_Senior | 1 (yes) |
| Has_Zero_Balance | 0 (no) |
| CreditScore_Category | 0 (<600) |

*Table 6: Prediction and Probability for the sample*

| | **Prediction** | **Probability** |
|---|---|---|
| Logistic Regression | Churned | Not churn: 0.068<br>Churn: 0.932 |
| Decision Tree | Churned | Not churn: 0.007<br>Churn: 0.993 |
| Random Forest | Churned | Not churn: 110<br>Churn: 0.890 |
| Gradient Boosting | Churned | Not churn: 0.003<br>Churn: 0.997 |

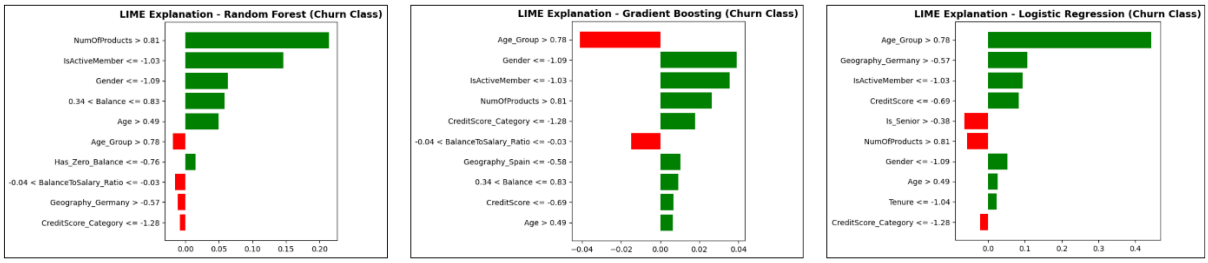


*Figure 21: LIME Analysis for Multiple Models*

Model predictions and LIME explanations for this churned customer show high predicted churn probabilities for all models—Random Forest (0.89), Gradient Boosting (0.997), and Logistic Regression (0.932)—driven primarily by multi-product usage, low account activity, and unfavourable demographics such as older age and German residency. These risk factors, clearly highlighted by both tree ensemble feature importances and regression coefficients, are consistently recognized across all models. This agreement provides strong, interpretable justification for the churn prediction and demonstrates robust decision support for identifying high-risk customers.

LIME Analysis for Different Classes

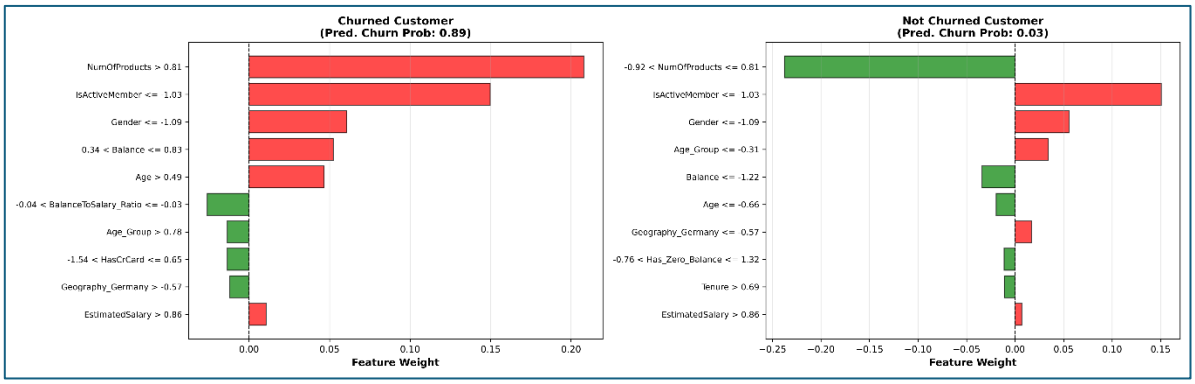


*Figure 22: LIME Analysis for Samples with Different Classes*

The LIME comparison highlights that for the churned customer, high product count and inactivity are the main drivers increasing churn risk, while for the non-churned customer, having fewer products and higher activity strongly reduce churn probability. Other features such as balance, age, and gender provide secondary influence. The explanation offers clear evidence that the model's predictions are grounded in well-understood, actionable customer attributes, supporting transparent and targeted retention decisions.

### 8.3.5 Bias Detection Analysis (Fairness Metrics across Demographics)

*Table 7: Accuracy across Demographics*

| | Features | Accuracy | Counts |
|---|---|---|---|
| Performance by Gender | Female | 0.832 | 870 |
| | Male | 0.862 | 1,130 |
| Performance by Geography | Germany | 0.803 | 523 |
| | France | 0.871 | 1,001 |
| | Spain | 0.853 | 476 |
| Performance by Age Group | Age <30 | 0.938 | 373 |
| | Age 30-40 | 0.886 | 886 |
| | Age 40-50 | 0.760 | 492 |
| | Age >50 | 0.759 | 249 |

Bias analysis indicates that the model performs unevenly across subgroups. Accuracy is slightly lower for females (0.832) compared to males (0.862), and geographic differences are notable, with German customers (0.803) less accurately classified than French (0.871) or Spanish (0.853) counterparts. Age group analysis reveals particularly strong accuracy for younger customers (<30: 0.938, 30–40: 0.886), but a marked drop for those over 40 (40–50: 0.760, >50: 0.759). These patterns suggest the model is less reliable for older and German clients, potentially due to class distribution, feature overlap, or data sparsity, and may benefit from subgroup-specific calibration or additional feature engineering to address these performance gaps (Hutchinson and Mitchell, 2019).

## 9. ETHICAL CONSIDERATION

### 9.1 Fairness and Bias

Fairness and bias should be carefully managed by monitoring model accuracy across age, gender, and geography; regular bias audits, performance parity testing, human oversight, and an appeals process help reduce discrimination risks and ensure fair scoring for all customer subgroups (Hutchinson and Mitchell, 2019).

### 9.2 Transparency

For transparency, every prediction is explained using SHAP/LIME, staff must be trained to interpret model outputs (Rudin, 2019), and decision logic is documented to provide clarity and accountability for both customers and employees.

### 9.3 Privacy

Privacy is protected by removing all personally identifiable information from modelling, restricting outputs to aggregate summaries, complying rigorously with GDPR/CCPA (Schwartz, 2019), and enforcing strict data retention policies.

### 9.4 Accountability

Accountability is upheld through assignment of a model owner, an ethics and compliance review board, formal incident procedures, and honest engagement with stakeholders to review and act on emerging concerns (Jobin et al., 2019).

## 10. LIMITATION AND FUTURE WORK

### 10.1 Data Limitation

The dataset used has limited temporal depth and geographic diversity with only 10,000 observations, which may restrict model generalizability to different markets or changing customer behaviours. Important contextual or behavioral variables might be missing, limiting predictive accuracy (Kuhn and Johnson, 2019).

### 10.2 Model Limitation

Current models, despite strong performance, can be improved in handling subgroups with lower accuracy such as older customers and certain geographies. Additionally, interpretability varies; ensemble models perform best but are less transparent than simpler models, which may impact adoption (Ribeiro et al., 2016).

### 10.3 Business Limitation

Implementation depends on operational readiness for integrating model outputs into churn prevention strategies. Ethical concerns about bias and privacy must be continuously monitored, and resources must be allocated for ongoing audits and customer communication (Jobin et al., 2019).

### 10.4 Short-term Future Plan

Focus on enhancing feature engineering for underperforming subgroups, incorporate temporal and behavioural data, and improve model explainability through SHAP/LIME visualizations for stakeholders. Begin pilot deployment with monitoring for bias and accuracy drift and feedback loop.

### 10.5 Long-term Future Plan

Develop real-time churn prediction pipelines integrated with customer engagement platforms. Explore advanced algorithms such as deep learning or causal inference models. Establish dedicated governance frameworks for ethical AI compliance and stakeholder transparency.

## 11. CONCLUSION AND RECOMMENDATION

The conclusion highlights substantial success in developing a robust customer churn prediction system, achieving high ROC-AUC scores (around 86%) validated through rigorous statistical testing and interpretable models with SHAP/LIME. The approach carefully incorporates fairness across demographics to maintain ethical standards.

From a business perspective, the model has identified critical churn drivers such as product usage, age groups, geography, and activity levels, providing actionable insights and an estimated potential churn reduction.

Methodologically, the project follows the CRISP-DM framework, blending statistical evidence with advanced machine learning while implementing ethical AI principles and establishing a solid governance structure. Strategically, deploying the Random Forest model as the primary predictor, focusing retention efforts on multi-product users, age-sensitive and regionally tailored interventions, combined with active monitoring of inactive accounts, offers a clear path for effective churn mitigation and continuous performance oversight.

# References


Owolabi, O. O., Uche, P. C., Adeniken, N. T., et al. (2024). Comparative Analysis of Machine Learning Models for Customer Churn Prediction in the U.S. Banking and Financial Services: Economic Impact and Industry-Specific Insights. *Scientific Research Publishing*, 15(4), pp. 783–802.

Pfeifer, P. E. (2005). The optimal ratio of acquisition and retention costs. *Journal of Targeting, Measurement and Analysis for Marketing*, 13(2), pp. 179-188.

Gandomi, A. and Haider,. (2015). Beyond the hype: Big data concepts, methods, and analytics. *Journal of the American Medical Informatics Association*, 22(6), pp. 1373-1384.

Adebayo, J. O. and Kusi, B. M. (2021). Explainable Artificial Intelligence (XAI) in financial services: A systematic review. *International Journal of Information Management Data Insights*, 1(1), pp. 100021.

Lemon, K. N. and Verhoef, P. C. (2016). Customer-centricity: The challenge and the benefits. *Journal of Research in Interactive Marketing*, 10(1), pp. 1-6.

Ascarza, E. (2018). The perils of predicting customer defection. *Journal of Marketing Research*, 55(1), pp. 1-19.

He, H. and Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), pp. 1263-1284.

Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. *In Emerging Artificial Intelligence Applications in Computer Engineering* (pp. 3-24). IOS Press.

Cai, J., Han, K. and Kim, T. (2018). A review on feature selection in machine learning. *International Journal of Automation and Smart Technology*, 8(1), pp. 1-8.

Hadi, A., Al-Khashman, Z., and Zghoul, S. (2020). Feature selection methods in machine learning: A comparative study. *Journal of King Saud University - Computer and Information Sciences*, 32(3), pp. 308-316.

Dormann, C. F., Elith, J., Bacher, S., et al. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, 36(1), pp. 27-46.

Kuhn, M. and Johnson, K. (2019). *Feature Engineering and Selection: A Practical Approach for Predictive Models*. CRC Press.

Coussement, K., Lessmann, S. and Verhoef, P. C. (2017). A comparative analysis of customer churn prediction models in a non-contractual setting. *European Journal of Operational Research*, 256(2), pp. 586-597.

Refaeilzadeh, M., Tang, L. and Liu, H. (2009). *Learning from imbalance in data streams*. In *2009 IEEE International Conference on Data Mining Workshops* (pp. 582-587). IEEE.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Springer.

Verbeke, W., Dejaeger, K., Martens, D., Hur, J. and Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit perspective. *European Journal of Operational Research*, 218(1), pp. 211-229.

Lessmann, S., Stahlbock, R., and Krüger, J. (2015). A comparison of machine learning classifiers for credit scoring: Evidence from the German credit data. *European Journal of Operational Research*, 243(2), pp. 497-507.

Molnar, C. (2022). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*.

Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. 2nd ed. O'Reilly Media.

Chen, T. and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Molnar, C. (2022). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. 2nd ed. Leanpub.

Lundberg, S. M. and Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 4765-4774.

Hutchinson, B. and Mitchell, M. (2019). 50 years of test (un) fairness: Lessons for machine learning. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 49-58.

Hutchinson, B. and Mitchell, M. (2019). 50 years of test (un) fairness: Lessons for machine learning. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 49-58.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), pp. 206-215.

Schwartz, P. M. (2019). Global data privacy: The end of the "privacy shield". *Journal of the American Academy of Dermatology*, 81(1), pp. 268-270.

Jobin, A., Ienca, M. and Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), pp. 389-399. Jobin, A., Ienca, M. and Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), pp. 389-399.

Ribeiro, M. T., Singh, S. and Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135-1144.

Kuhn, M. and Johnson, K. (2019). *Feature Engineering and Selection: A Practical Approach for Predictive Models*. CRC Press.