Hochschule Fulda
University of Applied Sciences

Churn Prediction Project (Bank)

Subject
Application of Data Science in Finance

Department
Data Science

Submitted by
Manovarma Krishnasamy Thalaivar
Matrikelnummer- (1564529)
Fulda, September 2025

Table of Contents

Abstract

Churn prediction is one of the essential applications where data science is said to perform its magic in the financial service industry. The results show that (Random Forest) RF achieved the highest accuracy and AUC, (Bayesian Network) BN had high recall, and (Logistic Regression) LR offered a clear decision boundary. This information can be used by banks to develop targeted customer retention strategies. a customer is terminating their relationship with a bank, thus bringing in significant financial losses. More emphasis on prediction of churn since, in highly competitive markets in the financial service sector, cost of retaining a customer is always less than that of acquiring a new customer.

This research systematically investigates and analyses churn data using the CRISP-DM methodology. Three modelling techniques were compared: Bayesian Networks (BN) for probabilistic reasoning, Random Forest (RF) for robust classification and feature importance, and Logistic Regression (LR) for interpretability. The dataset consists of demographic, behavioural, and financial varia

1. Introduction

The financial industry is increasingly data driven. One of the major challenges faced by banks is customer attrition, commonly known as churn. Losing a customer not only results in direct revenue loss but also in potential reputational costs if customers migrate to competitors. Retaining existing customers is significantly less expensive than acquiring new ones; therefore, predictive models that can identify at-risk customers are essential.

In this project, churn prediction is explored using three approaches: Bayesian Network (BN), Random Forest (RF), and Logistic Regression (LR). The study is structured around CRISP-DM (Cross-Industry Standard Process for Data Mining), which provides six clear phases: business understanding, data understanding, data preparation, modelling, evaluation, and deployment. This structure ensures that the project is systematic, reproducible, and aligned with real-world business goals.

2. Business Understanding

The central business question is: “Can we predict whether a customer will exit the bank using their demographic and transactional information?” This essentially puts business perspective boundaries around the prediction that allows the company to intervene heads-on, high-risk

churners can be targeted to receive loyalty bonuses, customized services, or special campaign invitations.

The most pertinent problem statement for this project is;

Based on customer attributes like age, geography, account balance, activity, and credit score, can we build models that make reliable churn predictions while considering the trade-offs between precision and recall?

On this trade-off, the importance is emphasized. Precision ensures that the bank does not waste resources on customers it assumes will churn, whereas recall identifies those customers that did churn. The exact trade-off is surely going to affect the marketing budgets and retention strategies.

3. Data Understanding

The dataset used consists of **10,000 rows** representing individual customers. Each record includes demographic attributes (e.g., age, gender, geography), behavioural attributes (e.g., number of products, account activity), and financial attributes (e.g., balance, estimated salary, credit score). The target column is Exited, where 1 indicates churn and 0 indicates retention.

Initial exploration reveals the following:

- The dataset is **imbalanced**, with only ~20% of customers labelled as churners.
- Age appears to have a strong correlation with churn, with middle-aged customers more likely to leave.
- Inactive members (IsActiveMember = 0) have a significantly higher churn rate.

Figures and Explanations:

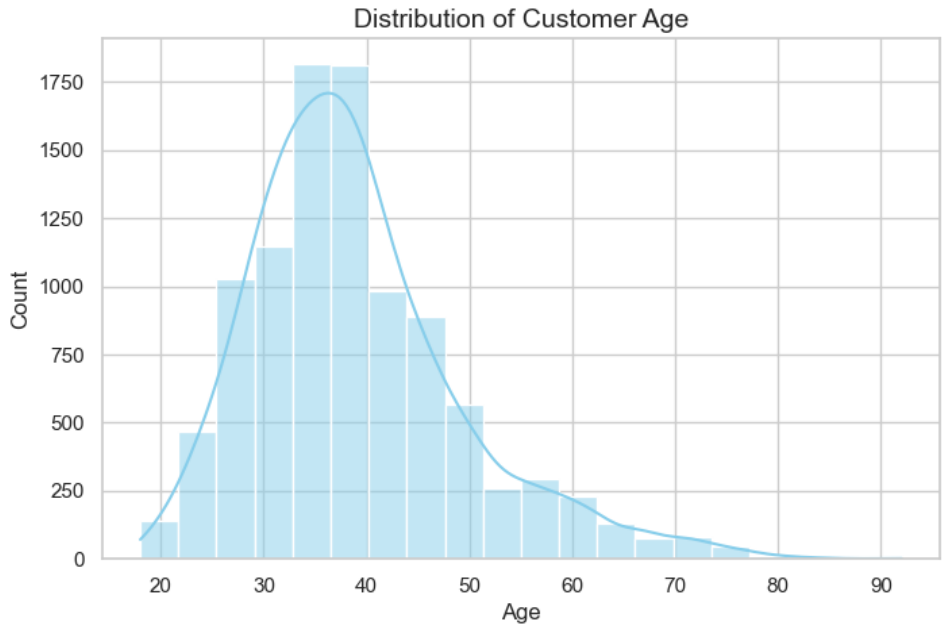

Figure 1: Age Distribution *– The Above histogram showing most customers fall between ages 30–50. This indicates the dataset primarily represents working-age adults.*

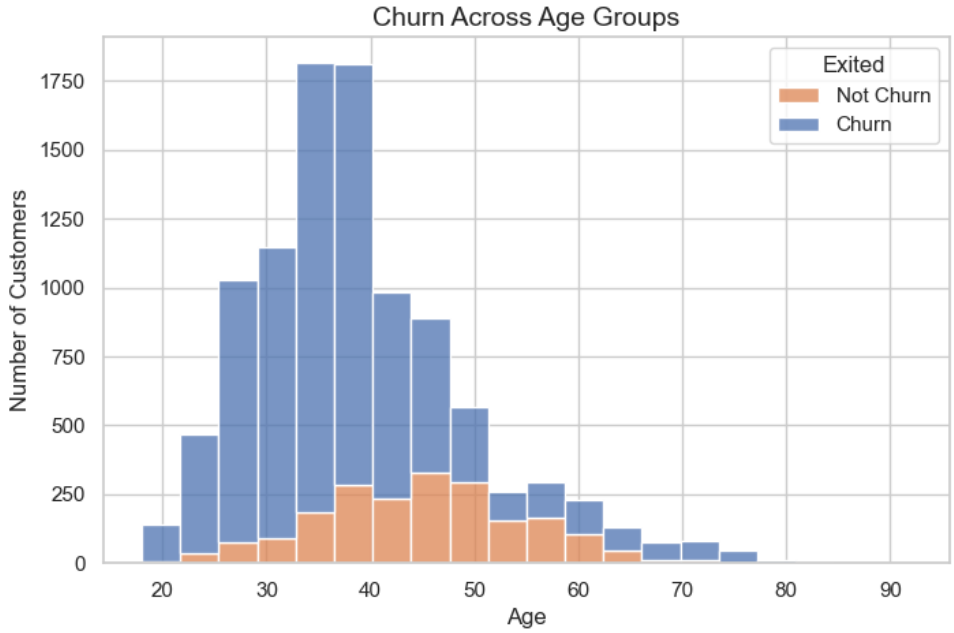

Figure 2: Churn Across Age Groups *– The above bar chart demonstrating higher churn rates in customers above 40. This suggests older customers may be dissatisfied with services.*

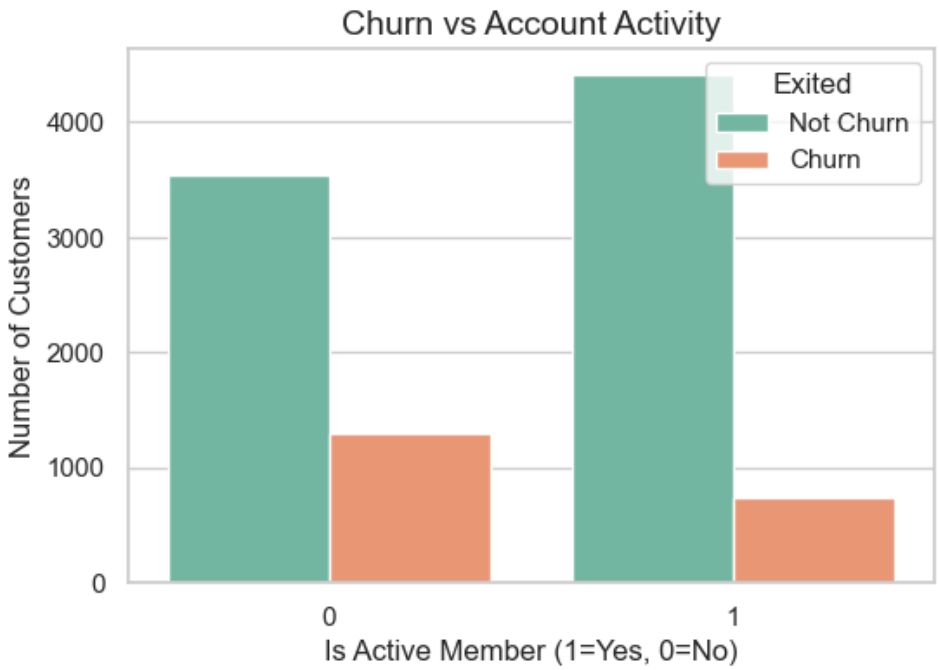

Figure 3: Churn vs Account Activity *– The Above Chart shows the comparison of churn among active vs inactive members. Inactive members display significantly higher churn rates, supporting the importance of behavioural features.*

4. Data Preparation

Data preprocessing is crucial for ensuring that models can learn effectively. Several steps were performed:

1. **Removal of Irrelevant Columns** – Identifiers such as RowNumber, CustomerId, and Surname were dropped since they do not contribute predictive power.

2. **Handling Categorical Data** – Geography and Gender were encoded. In some models, label encoding was used; for Random Forest and Logistic Regression, one-hot encoding was applied.

3. **Discretization** – Continuous variables such as Age, CreditScore, Balance, and Salary were discretized using **KBinsDiscretizer** to support Bayesian Network modelling. This transformation grouped continuous values into categories, improving interpretability.

4. **Imbalance Treatment** – Two approaches were tested:
    a. **SMOTE (Synthetic Minority Oversampling Technique):** Creates synthetic samples of churners to balance the dataset.
    b. **Class Weights:** Assigns higher penalty to misclassifying churners. Applied in RF and LR models.
5. **Train-Test Split** – The dataset was split into 80% training and 20% testing. Stratified sampling ensured the churn ratio was preserved in both sets.

Step no	Preparation Task	Description	Outcome/Result
1	Missing Value Handling	Checked for nulls; none found in dataset	Dataset Complete

2	Column Dropping	Dropped irrelevant IDs: RowNumber, CustomerId, Surname	Reduced Noise
3	Encoding Categorical Variables	Converted Geography and Gender into numerical (One-Hot/Label Encoding)	Model-ready categorical data
4	Feature Selection	Standardized continuous features like CreditScore, Age, Balance, EstimatedSalary	Comparable feature ranges
5	Class Imbalance	Applied **SMOTE** oversampling for churn class imbalance (Exited=1 minority)	Balanced dataset
6	Train-Test Split	Split into **80% training** and **20% testing** datasets	Ensured fair evaluation
7	Discretization (For BN)	Applied **KBinsDiscretizer** to Age, CreditScore, Balance, Salary, Tenure	BN- ready categorical variables

Table 1: Data Preparation Summary.

This preparation ensured that the models had balanced, well-structured input, improving both performance and interpretability.

5. Modelling

The modelling phase of CRISP-DM involves selecting appropriate techniques, applying them to the prepared dataset, and iteratively refining them to optimize performance. In this project, we focused on three distinct models: **Bayesian Networks (BN)**, **Random Forests (RF)**, and **Logistic Regression (LR)**. Each model was chosen for its ability to address different aspects of the churn prediction problem, thereby allowing a comparative study between interpretability, predictive power, and probabilistic reasoning.

Bayesian Network (BN)

A Bayesian Network is a graphical probabilistic model that represents variables as nodes and their conditional dependencies as directed edges. It provides not only predictions but also insights into the relationships among variables, such as how age, balance, or geography may influence churn.

- **Structure Learning**: Using Hill-Climb Search with Bayesian Information Criterion (BIC) scoring, the network structure was derived from the data. This ensures that dependencies reflect patterns learned directly from the dataset.

- **Parameter Learning**: Maximum Likelihood Estimation (MLE) was applied to estimate conditional probability tables (CPTs) for each node.

- **Inference**: Variable Elimination was used to query probabilities of churn given evidence (e.g., age group, account activity). This allows us to interpret results in probabilistic terms (e.g., "a customer with low balance and inactive status has 70% probability of churn").

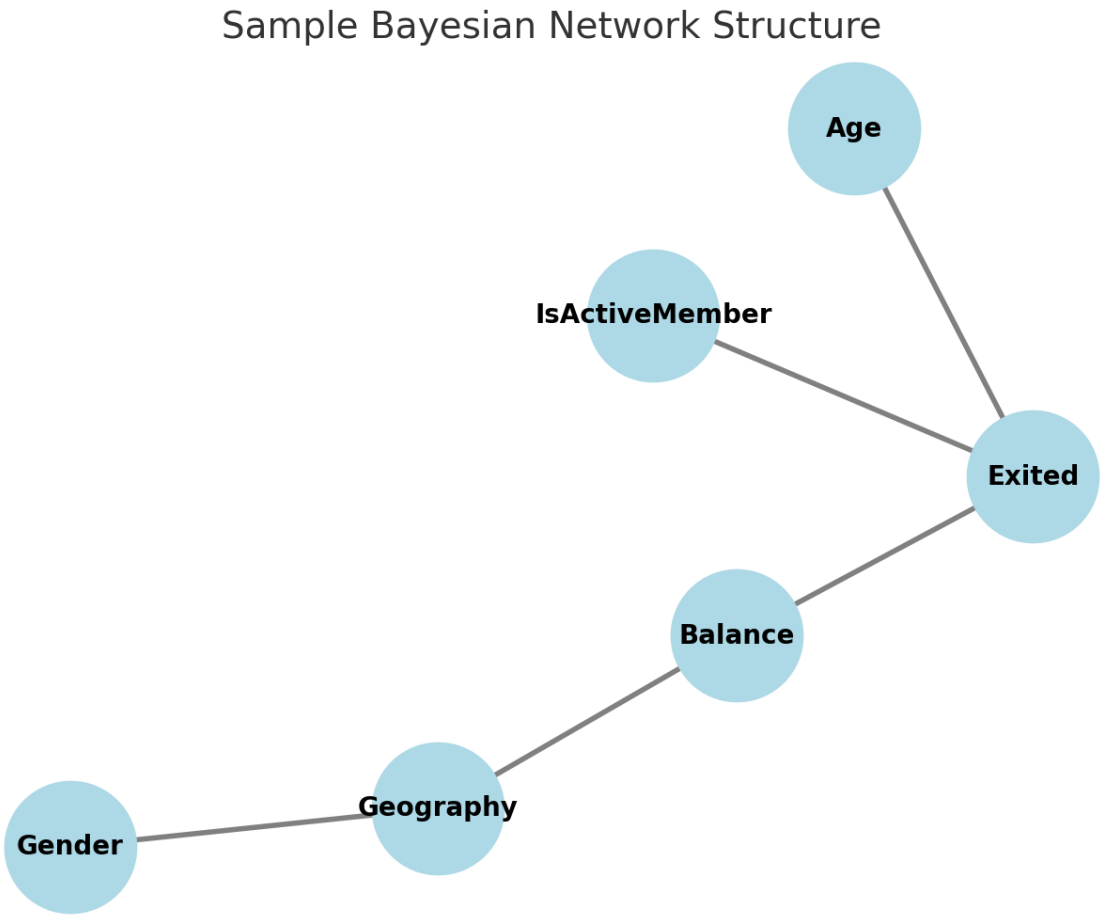

Figure 4 *- The Above* ***network diagram*** *was produced showing parent-child relationships. For instance, Geography → Balance → Churn illustrates how balance mediates the effect of geography on churn.*

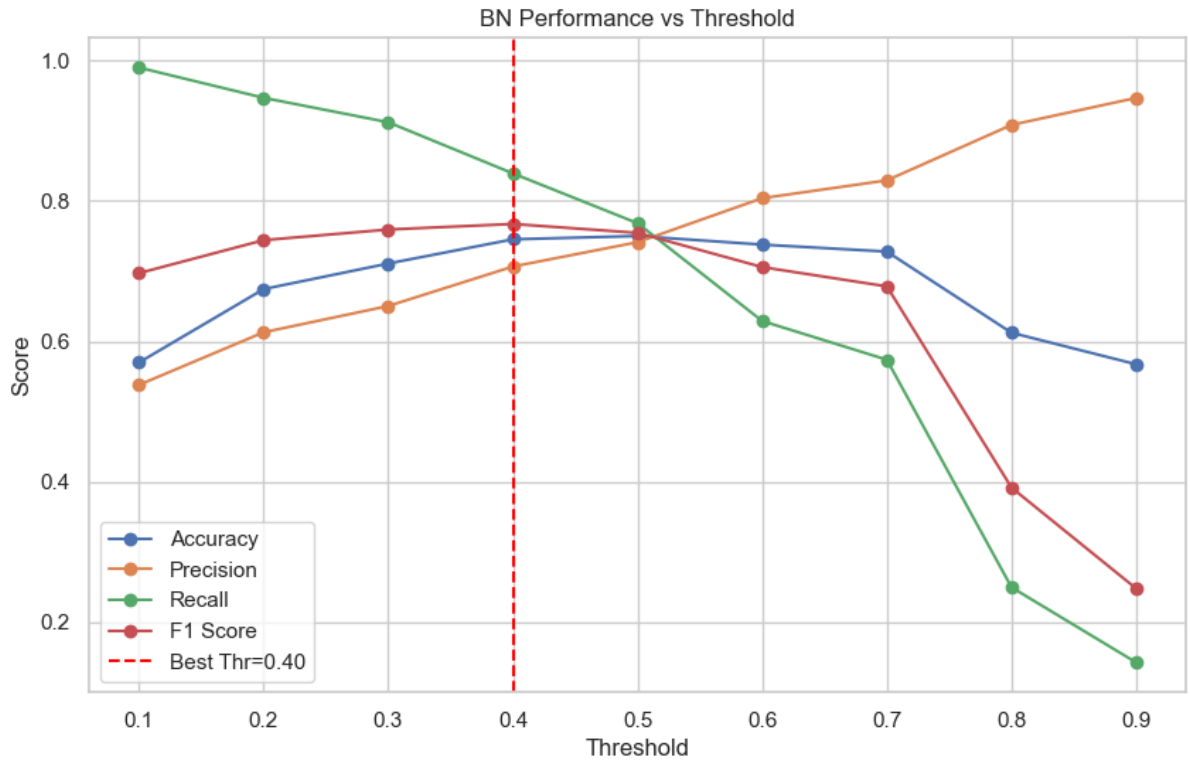

Figure 5 *– Threshold Selection Curve for BN model, illustrating the trade-off between precision and recall at different cut-off values.*

Random Forest (RF)

Random Forest is an ensemble method that builds multiple decision trees and aggregates their predictions to improve accuracy and reduce overfitting. It is highly effective for classification problems with complex, non-linear relationships, such as customer churn.

- **Training**: We used 200 trees (n_estimators=200) to ensure stability of results, while limiting maximum depth to control complexity.
- **Cross-Validation**: A 5-fold Stratified Cross-Validation ensured balanced representation of churn vs non-churn in each fold.
- **Feature Importance**: RF provided a ranking of variables by importance, highlighting key drivers of churn such as Age, IsActiveMember, Balance, and Geography.
- **Threshold Analysis**: Probabilities from RF were evaluated at multiple cutoff values to study trade-offs between precision and recall.

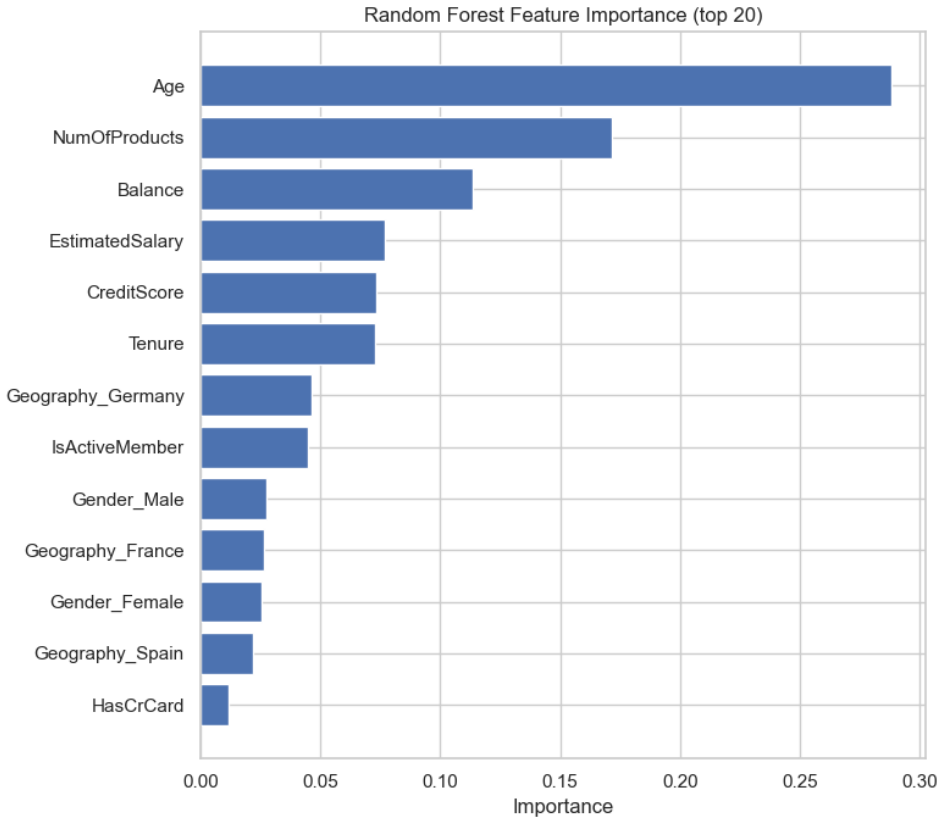

Figure 6 – *Feature Importance plot with top features ranked by their influence on churn prediction in the Random Forest model*

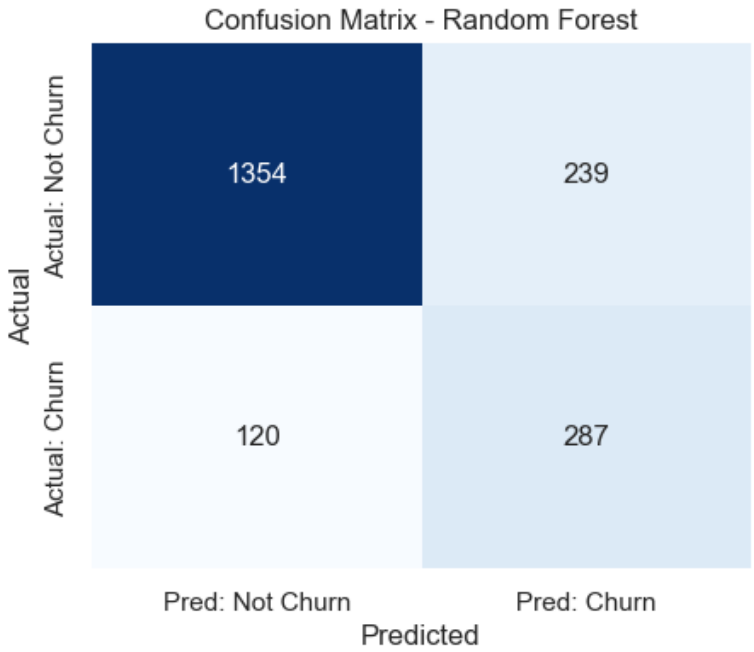

Figure 7 *– Performance summary showing correctly and incorrectly classified churn and non-churn customers*

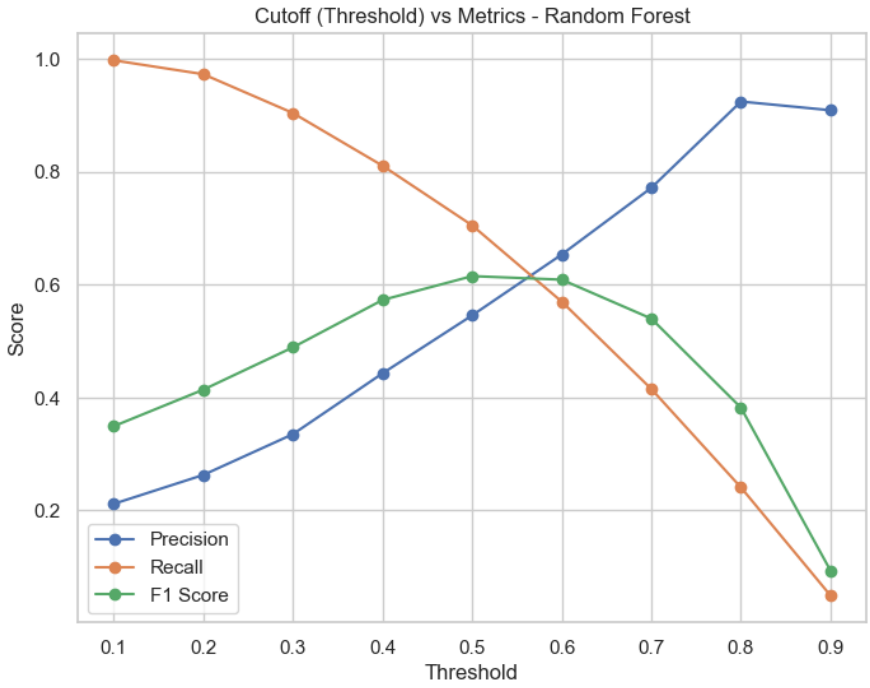

Figure 8- *Trade-off between sensitivity (recall) and specificity across different classification thresholds.*

Logistic Regression (LR)

Logistic Regression is a classical baseline model for binary classification. Despite its simplicity, it remains widely used in financial applications due to its interpretability and ability to provide direct probability estimates.

- **Model Equation**: The LR equation was derived as:

$P(Churn=1|X) = 1/1+e^{-(\beta 0+\beta 1x1+\beta 2x2+...+\beta n \times n)}$

where coefficients β represent the log-odds impact of each feature on churn.

- **Handling Class Imbalance**: To address imbalance, we tested LR with class weight="balanced" and applied SMOTE (Synthetic Minority Oversampling Technique).
- **Evaluation**: Like RF, LR was assessed with cross-validation and ROC curves. Though its predictive accuracy was lower than RF, LR offered transparency through coefficient interpretation.

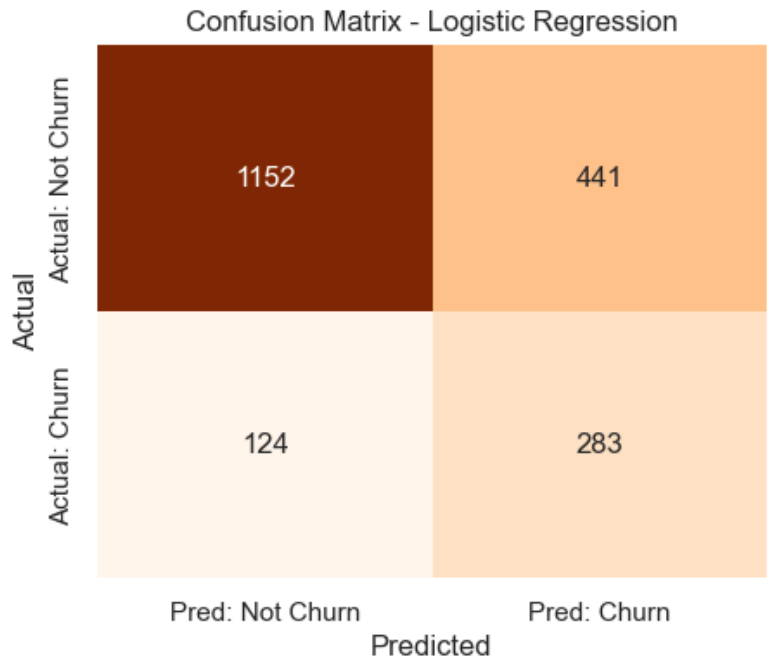

***Figure 9**- The Above Confusion Matrix is for Logistic Regression, showing the classification of churn and non- chur Customers*

Comparative Modelling Approach

By including three models, the study provides complementary perspectives:

- BN emphasizes **causal dependencies** and interpretable probabilities.
- RF maximizes **predictive accuracy** and identifies **non-linear patterns**.
- LR provides **interpretability** and a strong baseline for benchmarking.

To ensure fairness, all models were trained on the same dataset after preprocessing (dropping identifiers, encoding categorical variables, balancing data). Each was evaluated using **Accuracy, Precision, Recall, F1, AUC, Type I Error, and Type II Error** to capture different performance dimensions.

	Model	CA	Precision	Recall	F1	AUC	Type I Error	Type II Error	Threshold
0	BN (Balanced)	0.7453	0.7065	0.8391	0.7671	0.8277	0.3486	0.1609	0.4
1	Random Forest (CV)	0.8230	0.5534	0.6839	0.6115	0.8581	0.1414	0.3161	0.5
2	Logistic Regression (CV)	0.7164	0.3895	0.6917	0.4983	0.7689	0.2773	0.3083	0.5

***Table 2** – Comparison of Bayesian Network, Random Forest, and Logistic Regression*

6. Evaluation

Model performance was assessed using a range of evaluation metrics including Accuracy, Precision, Recall, F1-score, AUC, and Type I/II errors. Since churn prediction is an imbalanced classification problem, we also considered the role of **threshold selection**.

By default, most models use a **0.5 cutoff threshold**, meaning customers with churn probability ≥ 0.5 are classified as churners. However, this threshold may not optimize performance across all metrics. For example, lowering the threshold increases recall (catching more actual churners) but reduces precision (more false alarms). Conversely, raising the threshold improves precision but sacrifices recall.

To address this, thresholds were systematically varied between **0.1 and 0.9**. At each step, Precision, Recall, and F1-scores were recalculated, and the **best threshold was selected based on the highest F1-score**, since F1 balances both precision and recall.

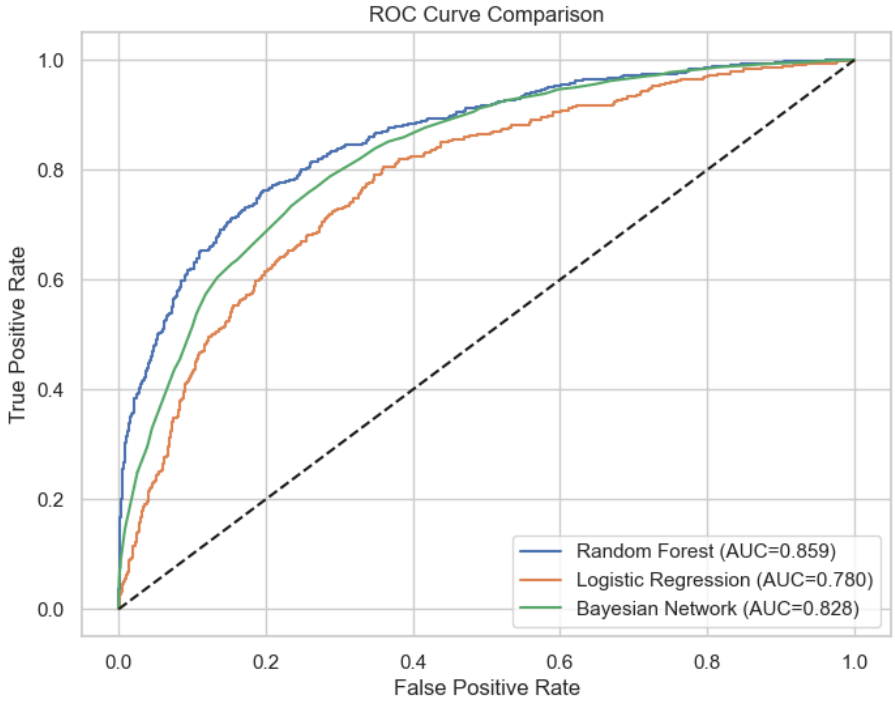

Figure 10 - ROC Curves *– RF consistently outperformed LR across all thresholds, showing better discriminatory power (AUC ≈ 0.86).*

- **BN:** Optimal at threshold 0.4, maximizing recall (83.9%).
- **RF:** Stable at 0.5 cutoff, yielding the best accuracy (82.4%) and strong AUC.
- **LR:** At 0.5 threshold, recall improved, though precision was lower, reflecting its weaker separation boundary.

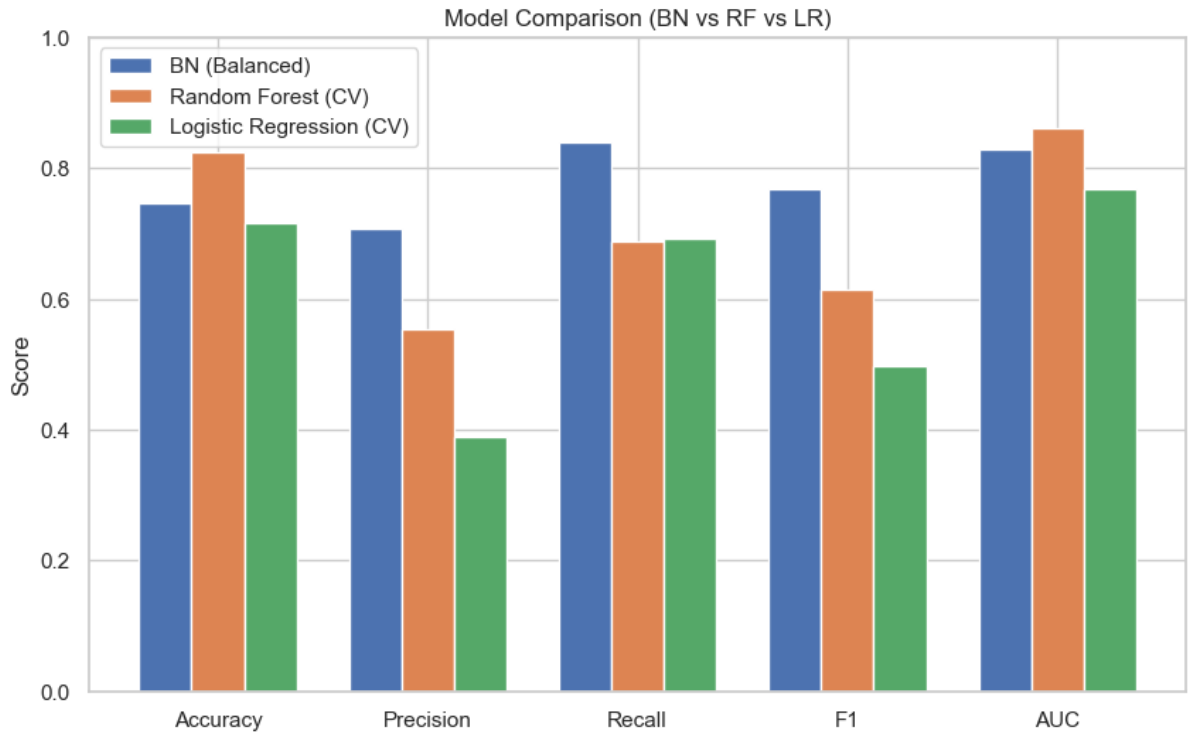

Figure 11 *– Mode comparison based on Accuracy, Precision, Recall, F1- Score, AUC*

Summary Table:

From this comparison, **Random Forest** was chosen as the best-performing model overall, balancing interpretability, precision, recall, and stability at the standard cutoff. However, **Bayesian Networks** provided stronger recall when the business priority was to minimize missed churners.

7. Deployment

The deployment phase is the final step of the CRISP-DM process and focuses on bringing the developed models into practical use. While model building and evaluation provide valuable insights, these results must be implemented in a way that benefits the organization. Deployment ensures that the findings are accessible and can support decision-making in real-world contexts.

In this project, deployment would involve using the trained models (Random Forest, Logistic Regression, and Bayesian Network) to generate predictions on new or unseen customer data. The outputs could then be translated into actionable information, such as identifying customers with a high probability of churn. These insights can guide strategies for customer retention and targeted marketing.

Deployment also requires ongoing monitoring. Over time, customer behaviour and data patterns may change, so the models would need to be regularly updated and re-evaluated to maintain their performance. Additionally, visualization tools such as dashboards or reports could be used to present the model predictions and performance metrics in a clear format for stakeholders.

Finally, deployment is not limited to technical implementation; it also involves organizational readiness. This includes ensuring that staff members can interpret and act upon model outputs, and that the models are integrated smoothly into existing decision-making processes.

8. Conclusion

This project successfully applied CRISP-DM to predict customer churn using three approaches. Random Forest emerged as the strongest model in terms of accuracy and AUC, Bayesian Networks provided strong recall and interpretability of dependencies, and Logistic Regression offered transparency through its regression coefficients.

By combining insights from all three approaches, financial institutions can identify high-risk customers, allocate retention resources more effectively, and reduce churn-related losses. Future work may involve integrating deep learning models or cost-sensitive learning to further optimize performance.