Fintech Churn Prediction(Bank)

Application of Data Science in Finance

Presenting by
Manovarma Krishnasamy Thalaivar

FD0003362

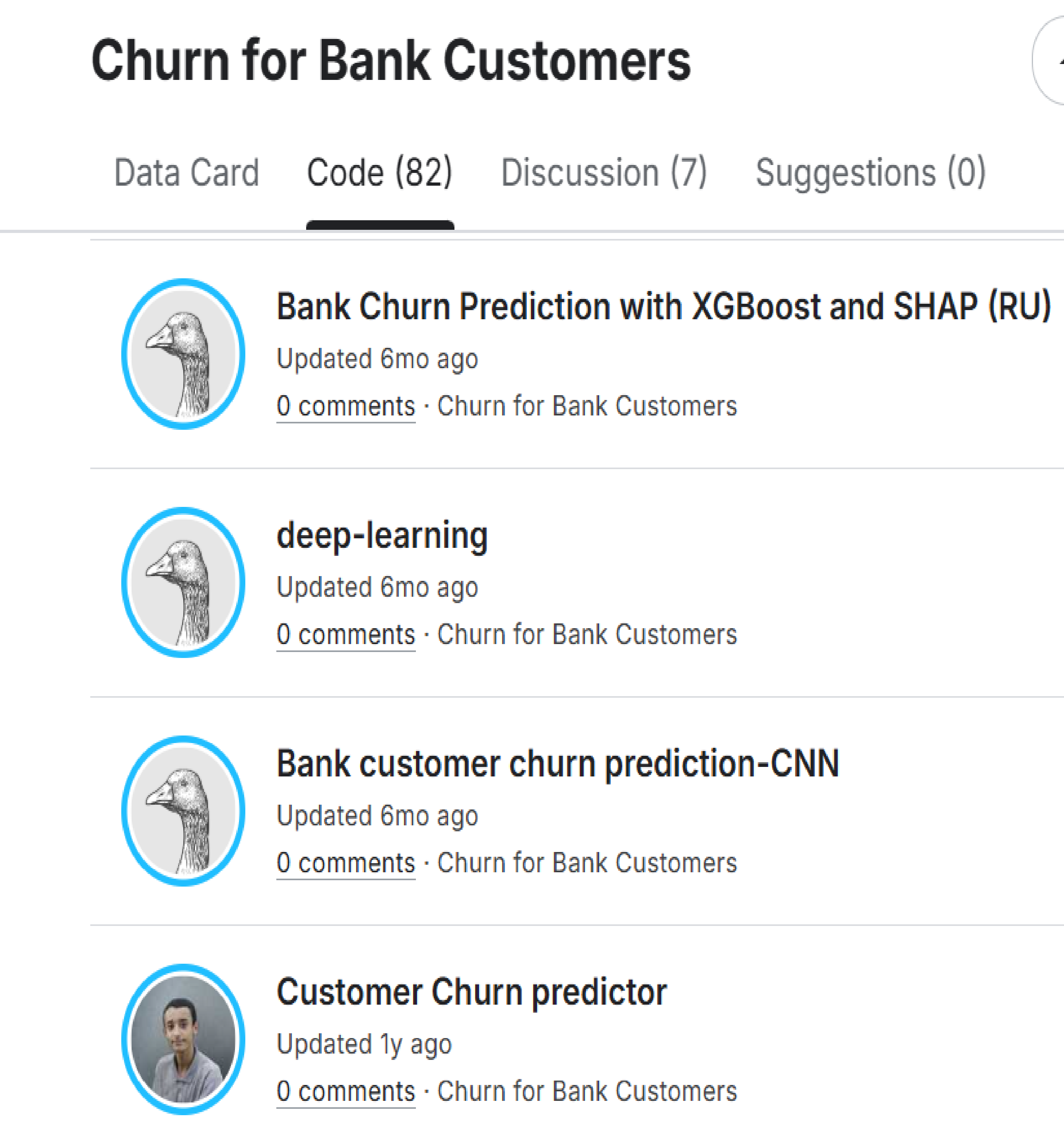
Agenda

1. Introductions
2. Goals
3. Business Understanding
4. Data Pre-processing
5. Models
6. Conclusion

Introduction

- Customer churn = when clients stop using a service.
- Predicting churn helps reduce losses and retain valuable customers.
- Our project compares **three models**:
- **Bayesian Network (BN)** – probabilistic reasoning.
- **Random Forest (RF)** – ensemble tree-based classifier.
- **Logistic Regression (LR)** – baseline linear classifier.

Other Projects



Goals

Using algorithms:

Bayesian Network → probabilistic churn prediction.

Decision Tree / Random Forest → classification model.

Logistic Regression → baseline model.

What are we going to predict?

We will predict whether a customer will exit (churn = 1) or stay (churn = 0).

Business Understanding

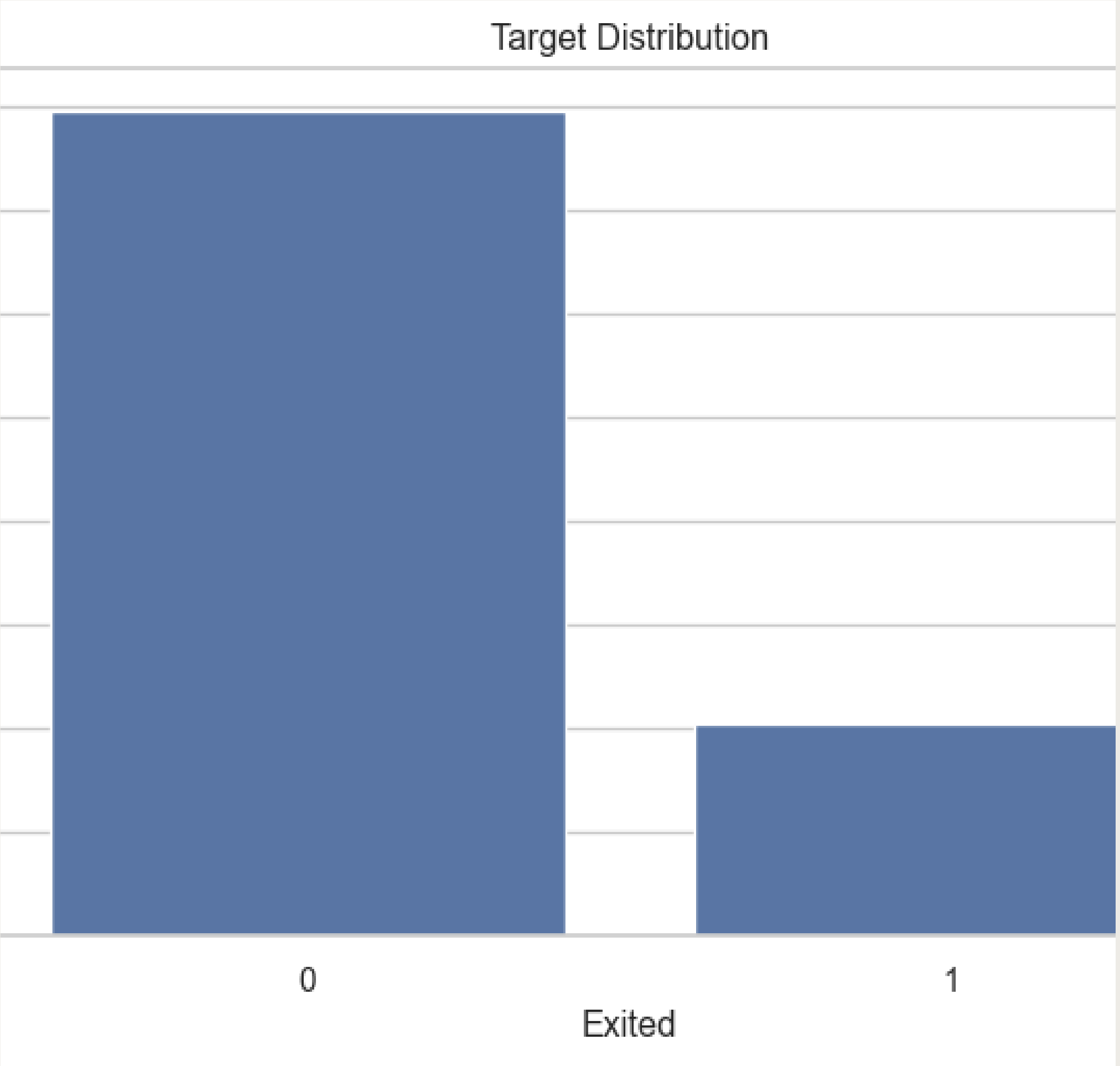
- Source: Bank churn dataset (~10,000 customers). link
- Features: demographic (age, gender, geography), account (balance, tenure, credit score), activity (products, card, active member).

Data Preprocessing

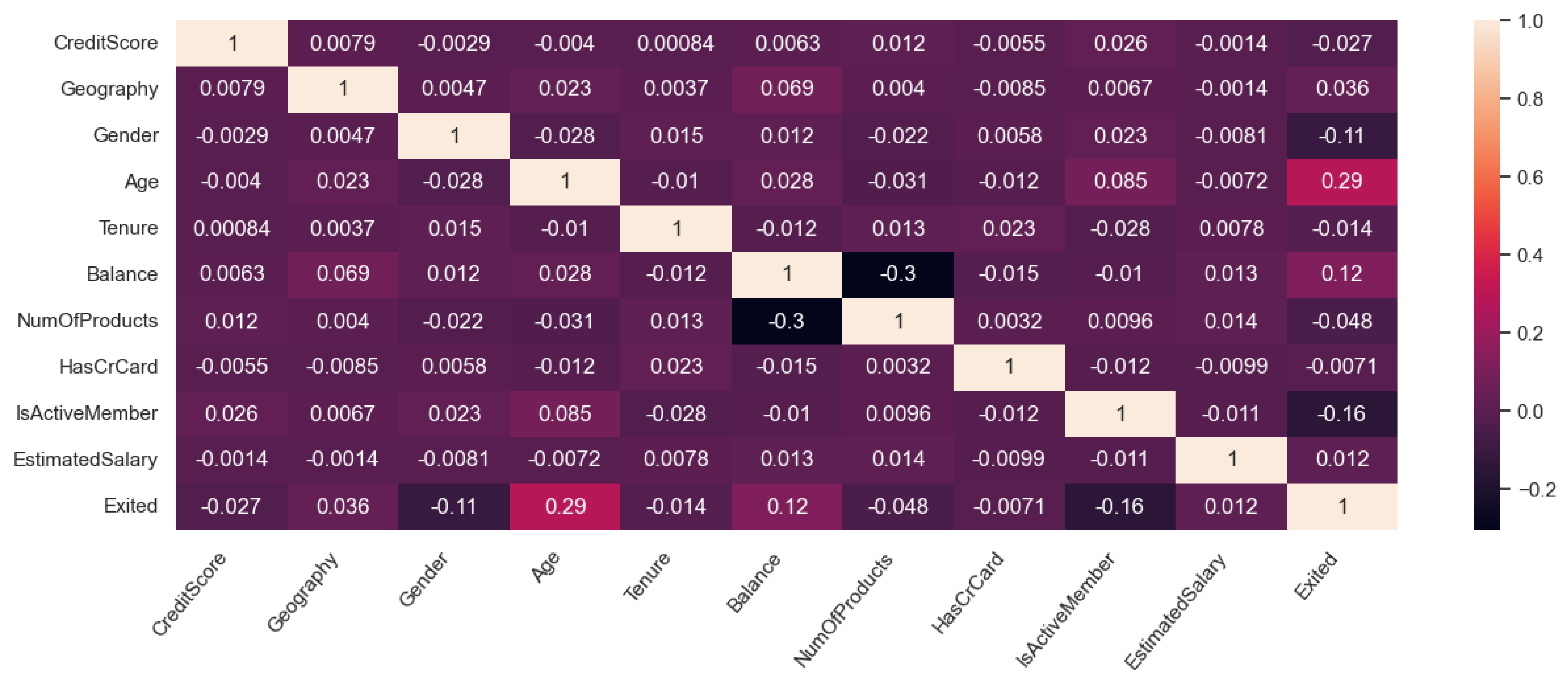
- Target: Exited (0 = stayed, 1 = churned).

Preprocessing steps:

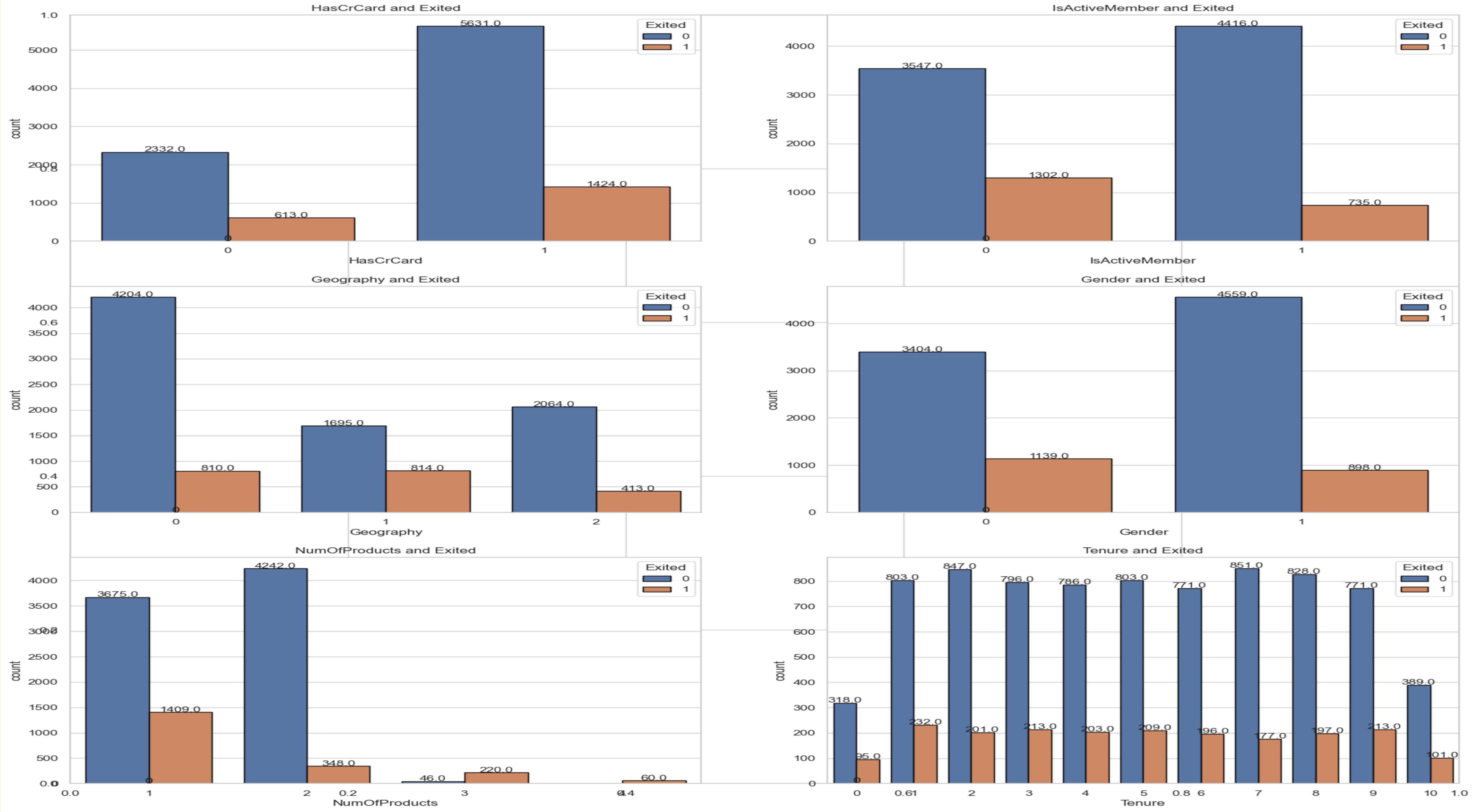
- Removed irrelevant IDs (RowNumber, CustomerId, Surname).
- Discretization (KBins) for BN.
- One-hot encoding for categorical variables in RF/LR.
- Balanced data using SMOTE to handle class imbalance.

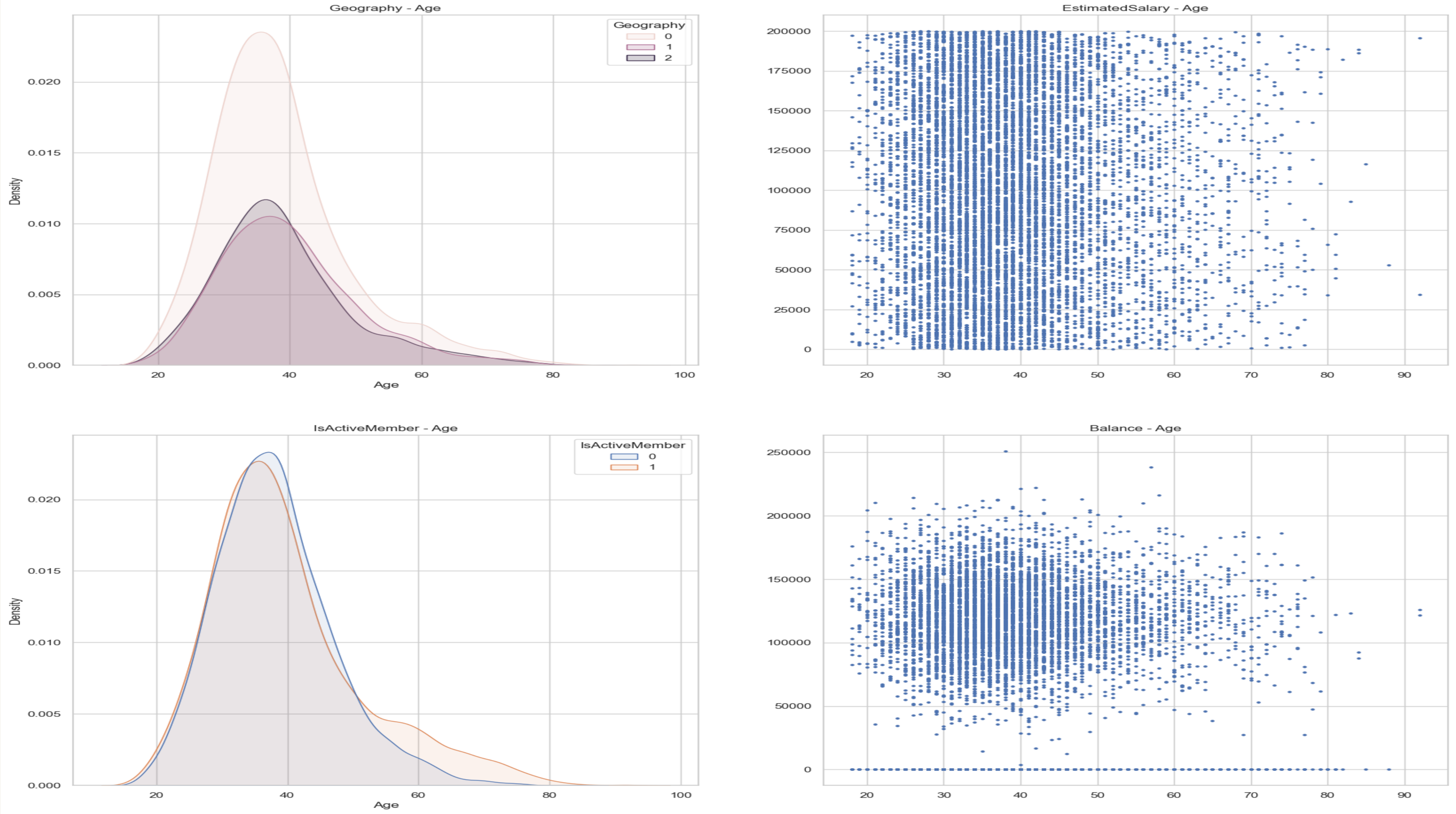



```
Duplicates: 0
Missing values per column:
 CreditScore        0
Geography          0
Gender             0
Age                0
Tenure             0
Balance            0
NumOfProducts      0
HasCrCard          0
IsActiveMember     0
EstimatedSalary    0
Exited             0
dtype: int64
```



CreditScore 1 0.0079 -0.0029 -0.004 0.00084 0.0063 0.012 -0.0055 0.026 -0.0014 -0.027
Geography 0.0079 1 0.0047 0.023 0.0037 0.069 0.004 -0.0085 0.0067 -0.0014 0.036
Gender -0.0029 0.0047 1 -0.028 0.015 0.012 -0.022 0.0058 0.023 -0.0081 -0.11
Age -0.004 0.023 -0.028 1 -0.01 0.028 -0.031 -0.012 0.085 -0.0072 0.29
Tenure 0.00084 0.0037 0.015 -0.01 1 -0.012 0.013 0.023 -0.028 0.0078 -0.014
Balance 0.0063 0.069 0.012 0.028 -0.012 1 -0.3 -0.015 -0.01 0.013 0.12
NumOfProducts 0.012 0.004 -0.022 -0.031 0.013 -0.3 1 0.0032 0.0096 0.014 -0.048
HasCrCard -0.0055 -0.0085 0.0058 -0.012 0.023 -0.015 0.0032 1 -0.012 -0.0099 -0.0071
IsActiveMember 0.026 0.0067 0.023 0.085 -0.028 -0.01 0.0096 -0.012 1 -0.011 -0.16
EstimatedSalary -0.0014 -0.0014 -0.0081 -0.0072 0.0078 0.013 0.014 -0.0099 -0.011 1 0.012
Exited -0.027 0.036 -0.11 0.29 -0.014 0.12 -0.048 -0.0071 -0.16 0.012 1
CreditScore
Geography
Gender
Age
Tenure
Balance
NumOfProducts
HasCrCard
IsActiveMember
EstimatedSalary
Exited
1.0
0.8
0.6
0.4
0.2
0.0
−0.2


HasCrCard and Exited
IsActiveMember and Exited
Geography and Exited
Gender and Exited
NumOfProducts and Exited
Tenure and Exited
Exited
count
HasCrCard
IsActiveMember
Geography
Gender
NumOfProducts
Tenure



Geography - Age
Geography
0
1
2
Density
Age
EstimatedSalary - Age
IsActiveMember - Age
IsActiveMember
0
1
Density
Age
Balance - Age

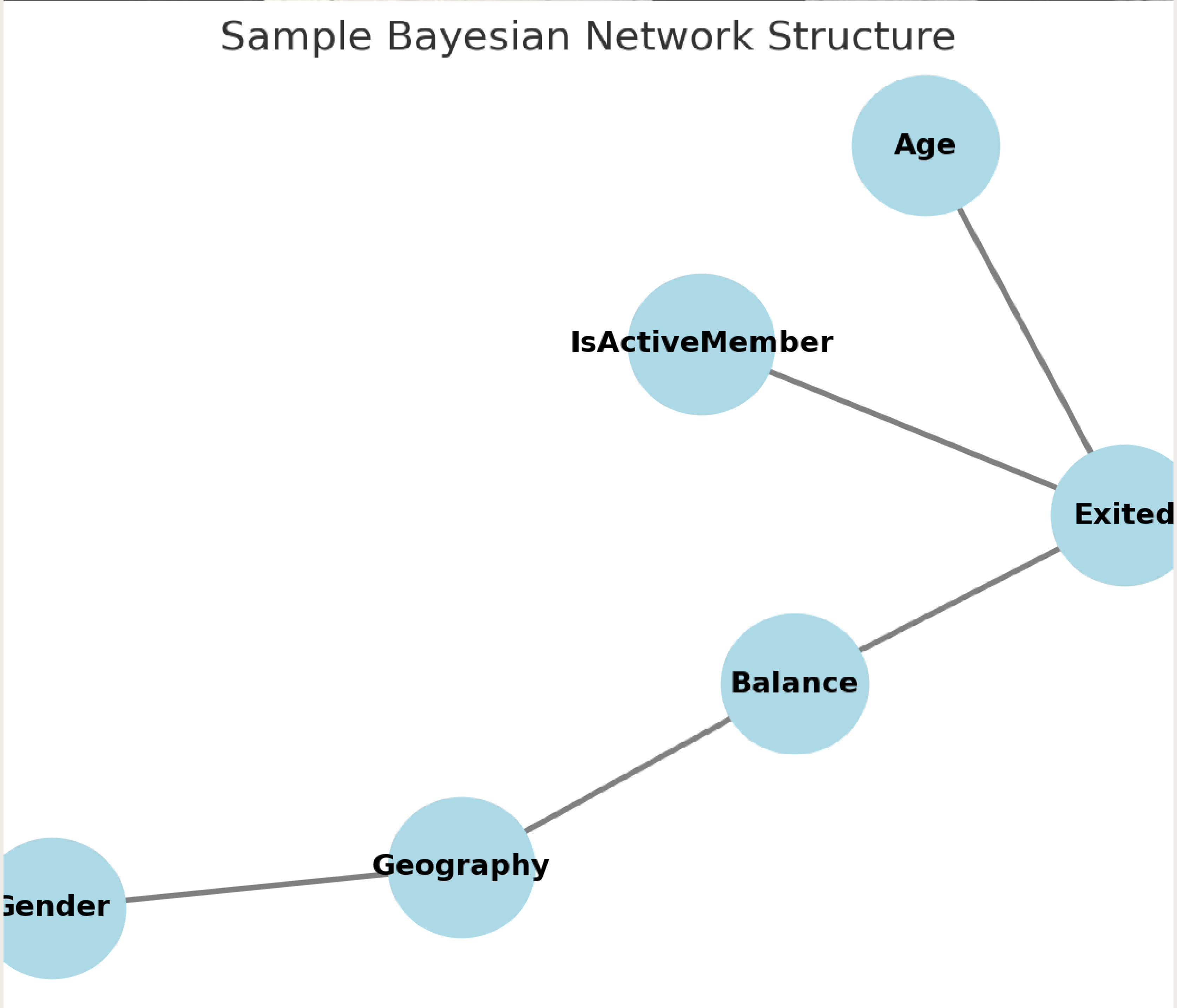
Models

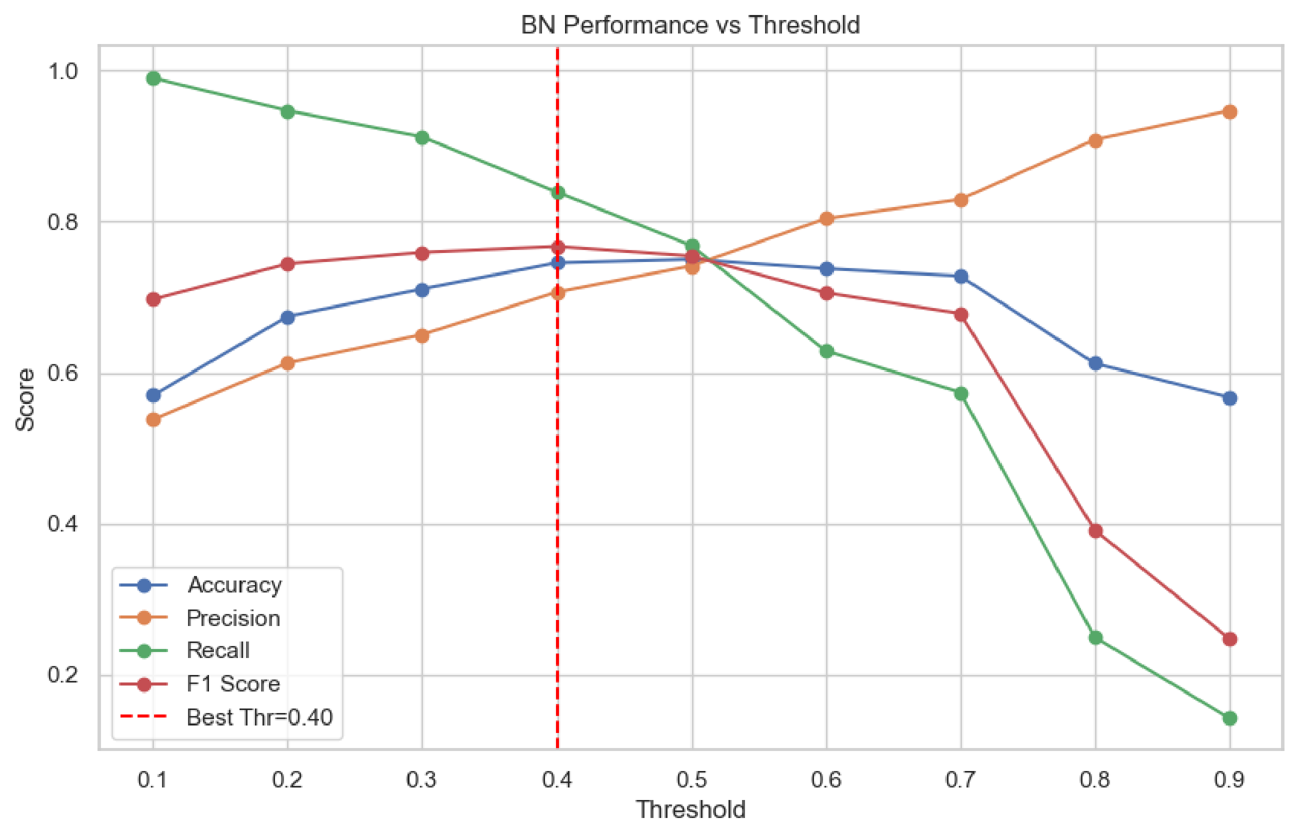
Bayesian Network

- Learned structure using Hill Climb Search + BIC score.
- Fitted CPDs with Maximum Likelihood Estimation.
- Inference via Variable Elimination.
- Output: probabilistic prediction P(Exited=1 | evidence).

	Threshold	Accuracy	Precision	Recall	F1	AUC	Type I Error	Type II Error
0	0.1	0.569886	0.537984	0.989828	0.697090	0.827694	0.850057	0.010172
1	0.2	0.674369	0.612877	0.946754	0.744078	0.827694	0.598016	0.053246
2	0.3	0.710787	0.650309	0.911968	0.759226	0.827694	0.490393	0.088032
3	0.4	0.745259	0.706492	0.839131	0.767120	0.827694	0.348612	0.160869
4	0.5	0.749969	0.741185	0.768178	0.754440	0.827694	0.268241	0.231822
5	0.6	0.737724	0.803852	0.628909	0.705700	0.827694	0.153460	0.371091
6	0.7	0.727804	0.829220	0.573779	0.678245	0.827694	0.118172	0.426221
7	0.8	0.612206	0.908177	0.249655	0.391647	0.827694	0.025242	0.750345
8	0.9	0.567311	0.946667	0.142660	0.247954	0.827694	0.008037	0.857340

BN STRUCTURE





	Actual	RF_Pred	LR_Pred
0	0	0	0
1	0	0	1
2	0	0	0
3	0	0	0
4	0	0	0
5	0	0	0
6	0	0	0
7	0	0	1

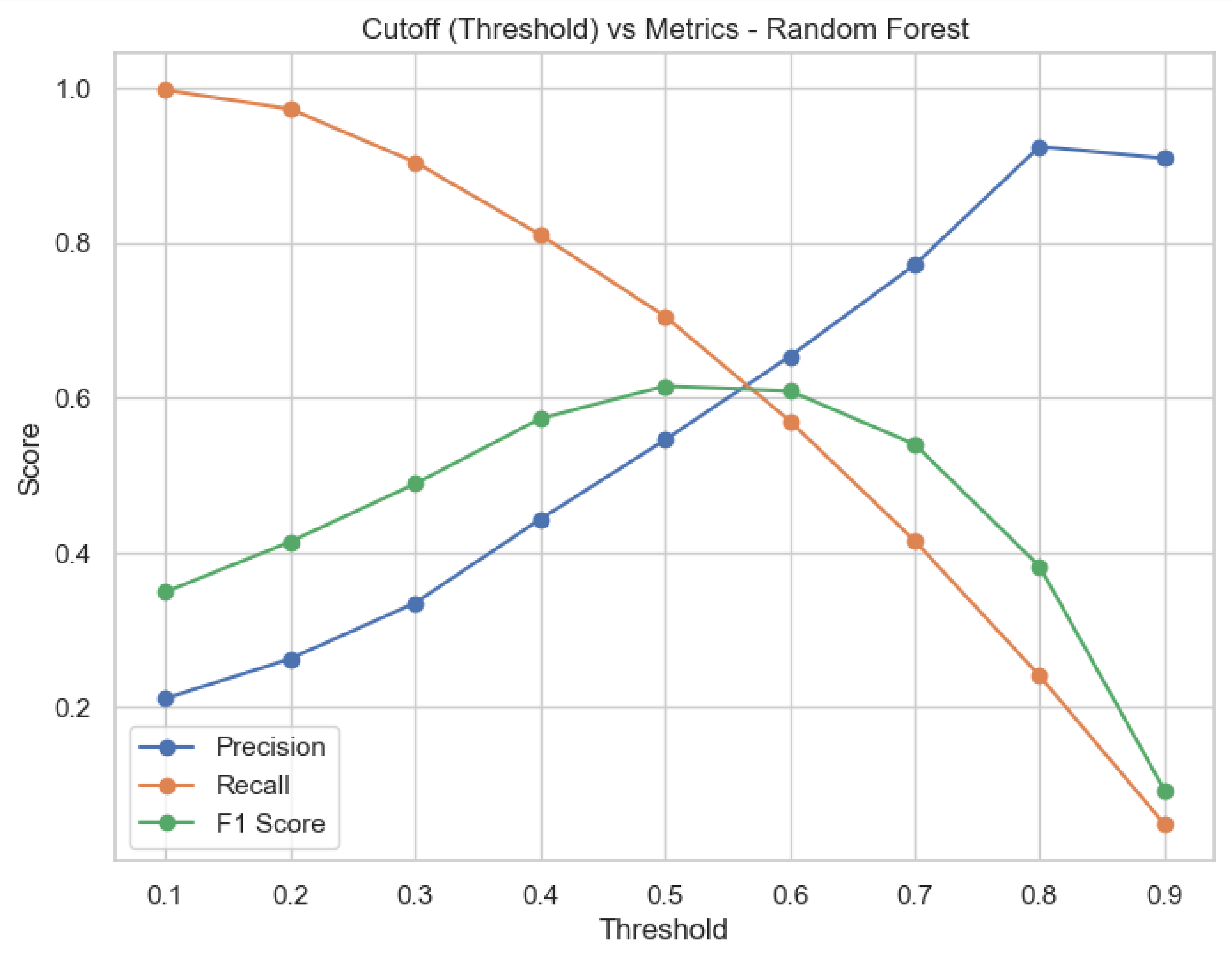
The Actual and Model predcited values for RF and LR.

Proabability of BN.

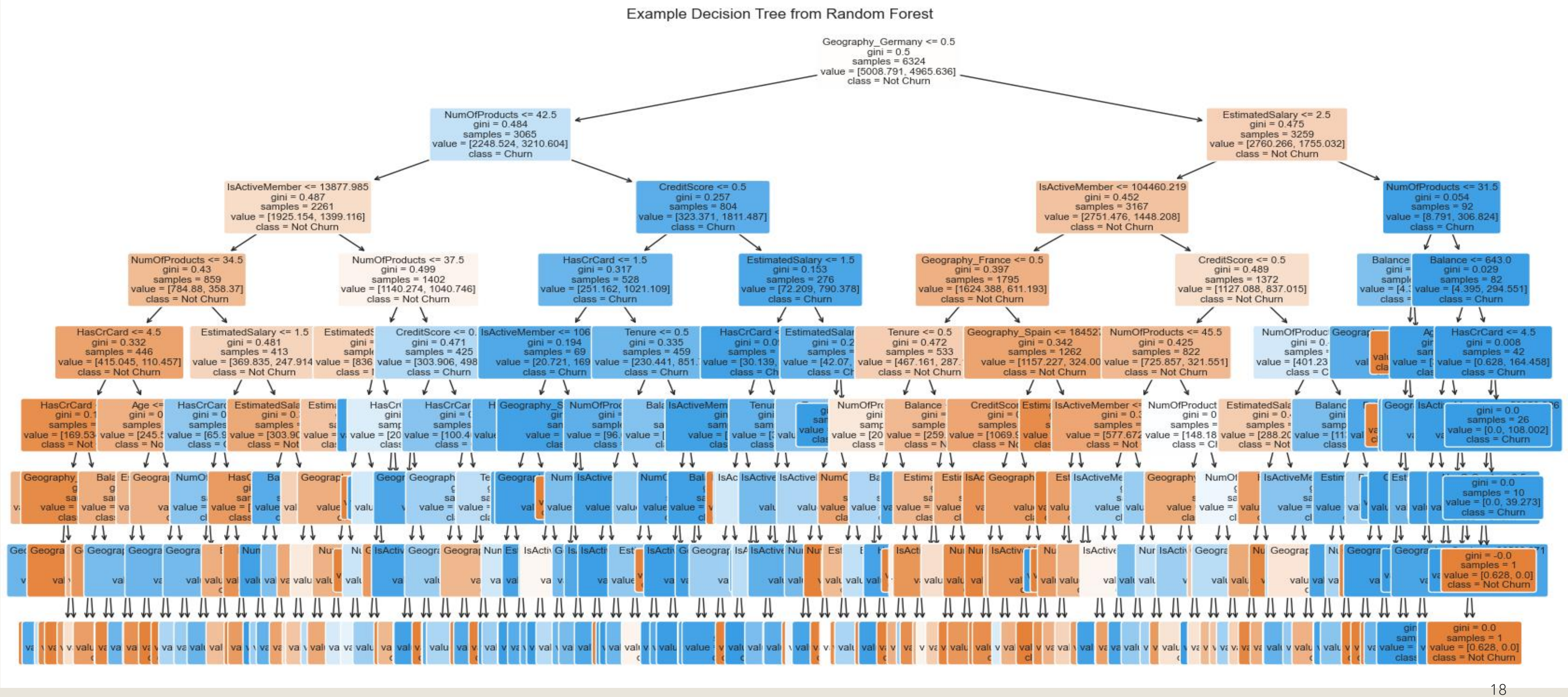
Actual	P(Exited=0)	P(Exited=1)	Predicted
1	55.91%	44.09%	1
0	49.58%	50.42%	1
1	3.14%	96.86%	1
0	62.01%	37.99%	0
0	62.86%	37.14%	0
1	86.74%	13.26%	0
0	95.21%	4.79%	0
1	0.00%	100.00%	1
0	92.02%	7.98%	0
0	76.00%	24.00%	0

Random Forest

- Used Pipeline with preprocessing + RandomForestClassifier.
- Hyperparameters: 200 trees, max depth = 8, class_weight = balanced.
- Evaluated with 5-fold Stratified CV.
- Feature importance analyzed (e.g., Age, Balance, Geography top drivers).



RF TREE - 200 Trees and 8- depth



Logistic Regression

- Simple linear baseline.
- Used with class weighting to handle imbalance.
- Evaluated with same 5-fold CV.

Logit(P(Exited=1)) = -0.1234 + (-0.0667 * CreditScore) + (0.9006 * Age) + (-0.0281 * Tenure) + (0.1537 * Balance) + (-0.0669 * NumOfProducts) + (-0.0143 * HasCrCard) + (-0.4759 * IsActiveMember) + (0.0307 *EstimatedSalary) + (-0.3320 * Geography_France) + (0.5328 *Geography_Germany) + (-0.3242 * Geography_Spain) + (0.2362 *Gender_Female) + (-0.3597 * Gender_Male)

$$\text{Logit}(P) = \beta_0 + \beta_1 \cdot \text{CreditScore} + \beta_2 \cdot \text{Age} + \beta_3 \cdot \text{Balance} + \cdots + \beta_k \cdot \text{IsActiveMember}$$

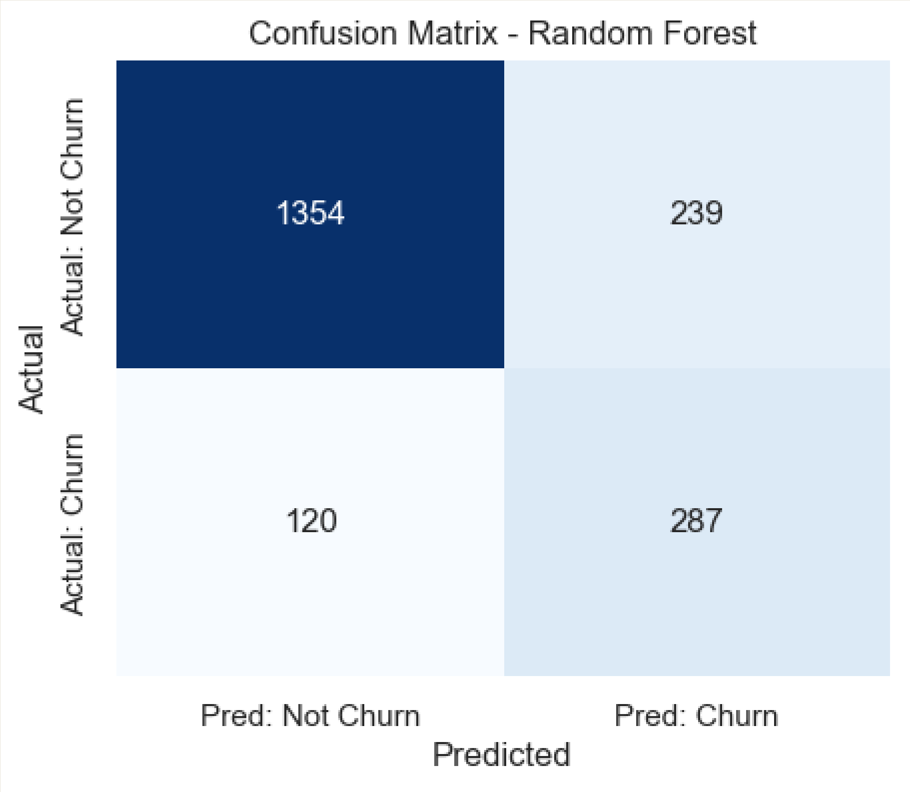
So:

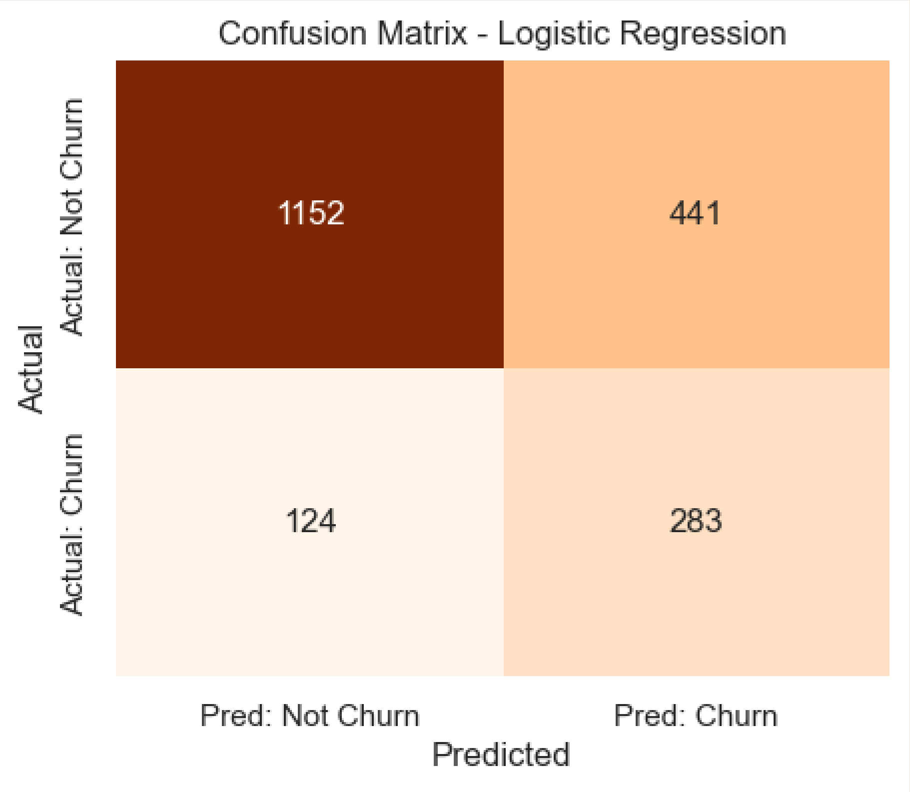
$$P(\text{Exited} = 1) = \frac{1}{1 + e^{-\text{Logit}(P)}}$$

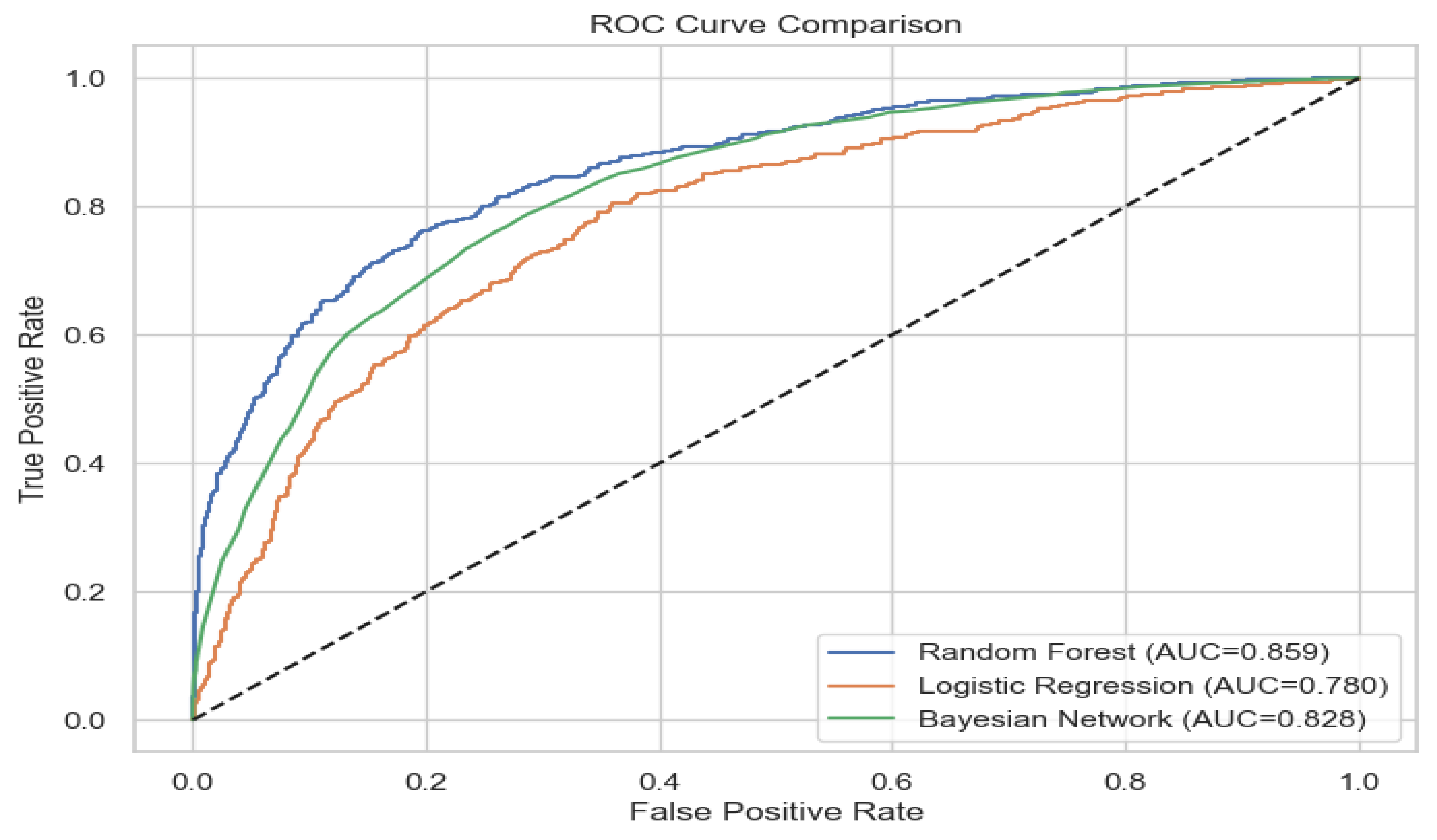
	Model	CA	Precision	Recall	F1	AUC	Type I Error	Type II Error	Threshold
0	BN (Balanced)	0.7453	0.7065	0.8391	0.7671	0.8277	0.3486	0.1609	0.4
1	Random Forest (CV)	0.8230	0.5534	0.6839	0.6115	0.8581	0.1414	0.3161	0.5
2	Logistic Regression (CV)	0.7164	0.3895	0.6917	0.4983	0.7689	0.2773	0.3083	0.5

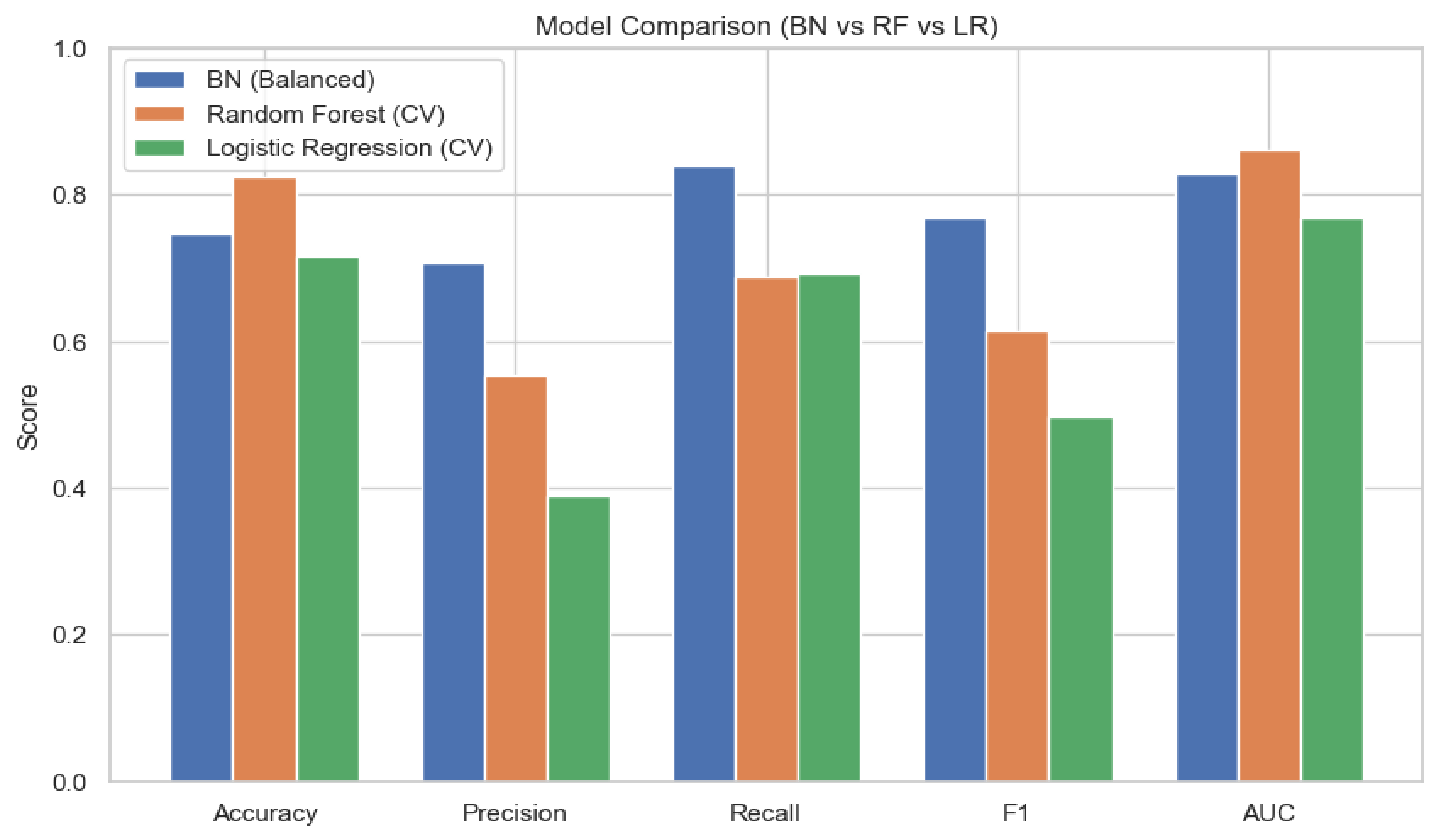
Overall Model Comparision

Confusion matrix






ROC Curve Comparison
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Random Forest (AUC=0.859)
Logistic Regression (AUC=0.780)
Bayesian Network (AUC=0.828)



Model Comparison (BN vs RF vs LR)
BN (Balanced)
Random Forest (CV)
Logistic Regression (CV)
Score
1.0
0.8
0.6
0.4
0.2
0.0
Accuracy
Precision
Recall
F1
AUC

- Random Forest is the most balanced model with the highest overall performance.
- Bayesian Network gives stronger recall, making it useful when identifying churners is the priority.
- Logistic Regression is less accurate but provides interpretability with clear coefficients.
- Age, activity status, and credit score emerged as important features in churn prediction.
- No single personal detail alone decides churn; rather, it is the interaction of multiple factors.

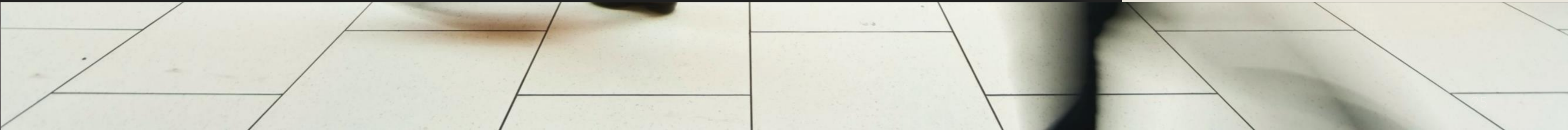
Conclusion

Thank you

Presented by

Manovarma Krishnasamy Thalaivar

FD0003362



Back-up slides