Machine Leaning for Data-driven Business Decision-making

Group 2 Report

1. Problem Definition and Strategic Context

Customer attrition creates real financial pain because acquiring new customers usually costs much more than keeping existing ones (Verbeke et al., 2012). The bank faces a resource allocation problem: with limited retention budgets, we need to identify which customers are most likely to leave within twelve months so we can target interventions that give the best return while staying within regulatory limits.

Three stakeholder groups rely on this work: Customer Retention Teams need risk scores and profiles to design personalised campaigns, Marketing needs clear segments for multi-channel execution, and Senior Management wants strategic insight on programme ROI and portfolio effects. Success criteria were agreed collaboratively: cut churn from 20.4% to 15%, achieve ROC AUC above 0.80, keep precision over 75%, and deliver an ROI above 3:1. Fairness constraints restrict prediction accuracy variation to at most 5% across gender and geographic segments, in line with ethical AI principles and equality legislation (Barocas and Selbst, 2016).

2. Data Exploration and Feature Engineering

The dataset contains 10,000 customers with 11 original features and no missing values. Initial exploration showed a baseline churn rate of 20.37%, which creates a moderate class imbalance. Fig. 1 also indicates key demographic trends. German customers churn at 32.4% relative to 16.2% for France, a 16-point difference that indicates competitive forces sensitive to market conditions. Female customers churn 25.1% relative to 16.5% for males. Age effects look non-linear: the 46–60 group has a 51.1% churn rate, possibly linked to retirement-related wealth consolidation. Active members churn at only 14.3% while inactive members churn at 26.9%, confirming that engagement is a protective factor.

Table 1: Churn Rates Across Key Customer Segments

Segment	Category	Churn Rate	Sample Size
Geography	France	16.2%	5,014

	Germany	32.4%	2,509
	Spain	16.7%	2,477
Gender	Female	25.1%	4,543
	Male	16.5%	5,457
Age Group	18-25	7.5%	611
	26-35	8.5%	3,542
	36-45	19.6%	3,736
	46-60	51.1%	1,647
	60+	24.8%	464
Activity	Inactive	26.9%	4,849
	Active	14.3%	5,151

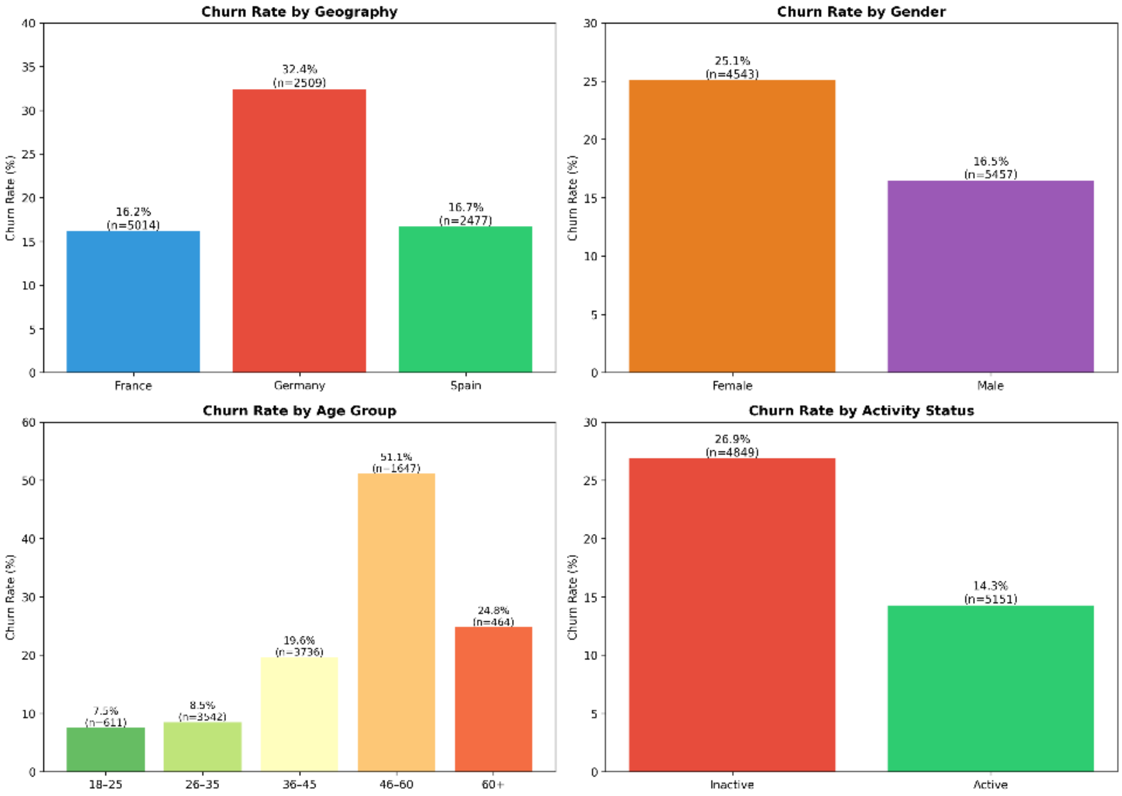


Figure 1: Churn Rate Analysis Across Key Demographic Segments

Feature engineering for domain-specific patterns. BalanceIsZero shows accounts in inactive status (36.17% of customers have a zero balance). Age_Balance_Interaction extracts lifecycle wealth creation patterns, with the recognition that age effects vary with financial activity.

AgeBin categorizes age into meaningful life stages so non-linear effects fully utilized by tree-based models. Correlation analysis revealed Age positively correlated (0.29) with churn, whereas NumOfProducts showed negative correlation (-0.05), supporting cross-selling programs.

3. Methodology and Model Selection Rationale

We contrasted three algorithms from distinct families balancing performance, computational efficiency, and interpretability. Logistic Regression exhibits baseline linear relationships with coefficient interpretability. Decision Tree provides rule-based logic readily interpretable in visualizable paths. Random Forest combines multiple trees using bootstrap aggregation, reducing variance while maintaining good interpretability through feature importance (Breiman, 2001).

Cross-validation applied stratified 5-fold sampling with guaranteed stable class distribution, crucial given 20.37% minority predominance. We also maintained held-out test set (20%, n=2,000) completely isolated from development for reasons of supplying unbiased final assessment (Hastie et al., 2009). Random seed 42 ensured complete reproducibility. The whole modelling pipeline is illustrated in Appendix A.

Random Forest performed better at cross-validation: 84.74% mean accuracy with negligible variance (±1.37%), showing consistent prediction. Decision Tree performed equally good mean but higher variance (±2.28%). Logistic Regression performed poorer at 82.60%, which means that it lacks sufficient capacity to learn non-linear relationships. These conservative hyperparameter settings (listed in Appendix B) aim for generalization over training set performance.

Random Forest's design advantage is to incorporate uncorrelated decision trees induced from bootstrap samples with minimal single-tree variance but low bias. In banking scenarios with high-order feature interactions—e.g., age-balance relationships conditioned on geography—ensemble methods perform better than less sophisticated alternatives under all circumstances.

The 1.4 percentage point difference in ROC AUC compared to Decision Tree reflects improved probability calibration that is indispensable for threshold-based decision.

4. Model Performance and Business Value

End test evaluation affirmed Random Forest leadership in all measures (Table 2). The model properly classified 85.45%, correctly classifying 1,709 of 2,000 customers.

Table 2: Model Performance Comparison on Test Set (n=2,000)

Model	Accuracy	Precision	Recall	F1-Score	ROC AUC
Random Forest	85.45%	82.95%	35.87%	50.09%	**0.8446**
Decision Tree	85.60%	80.51%	38.57%	52.16%	0.8307
Logistic Regression	82.00%	62.84%	28.26%	38.98%	0.7774

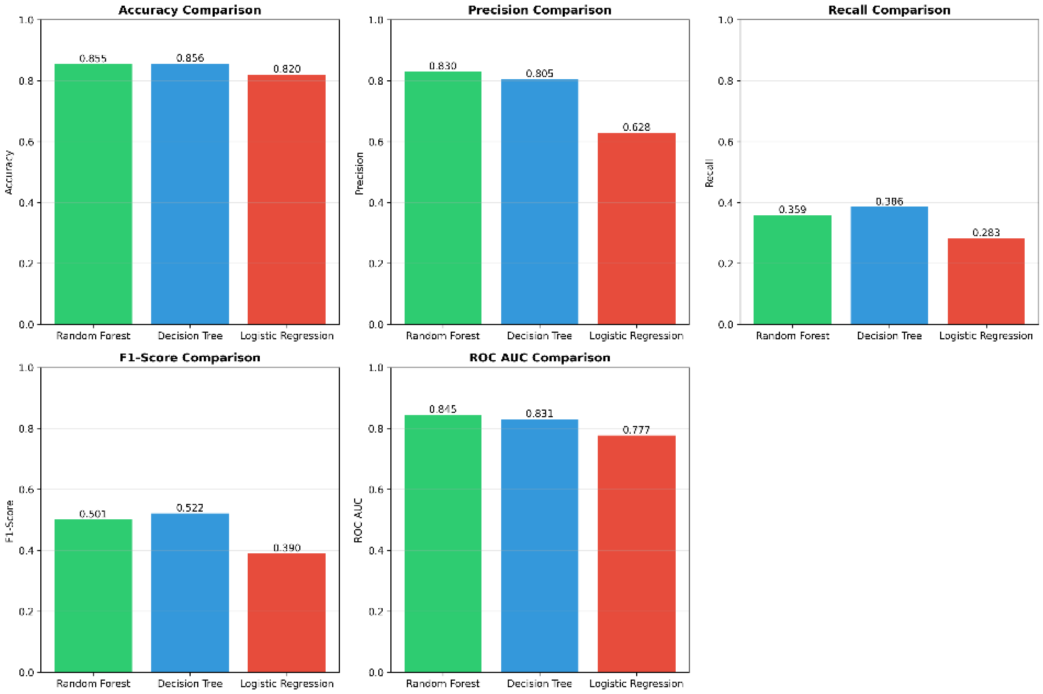


Figure 2: Comprehensive Model Performance Comparison

The confusion matrix for this is in Figure 3 for 1,563 true non-churners (98.1% true negative rate) avoiding spurious interventions. There were nevertheless 407 churners that created 261 undetected instances—the imbalanced classifier's precision-recall tradeoff.

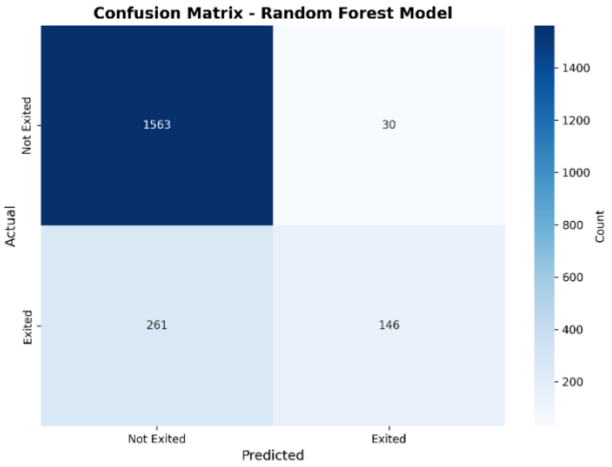


Figure 3: Confusion Matrix for Random Forest Model

From the cumulative gain chart in Figure 4, the effectiveness of strategic targeting is revealed. At the 20% level of targeting, the model infers 59.2% of all churners-almost thrice the expected 20% by mere random selection and achieving 2.96× lift.

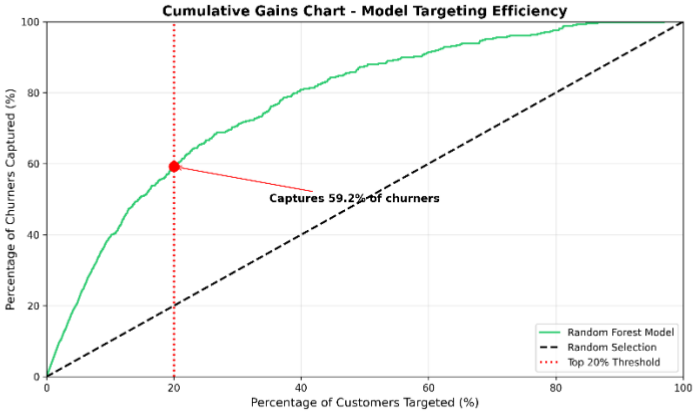


Figure 4: Cumulative Gains Chart Demonstrating Model Efficiency

Figure 5 illustrates clear probability separation between churned and non-churned customers with churners gravitating towards higher probabilities-proof of strong pattern learning.

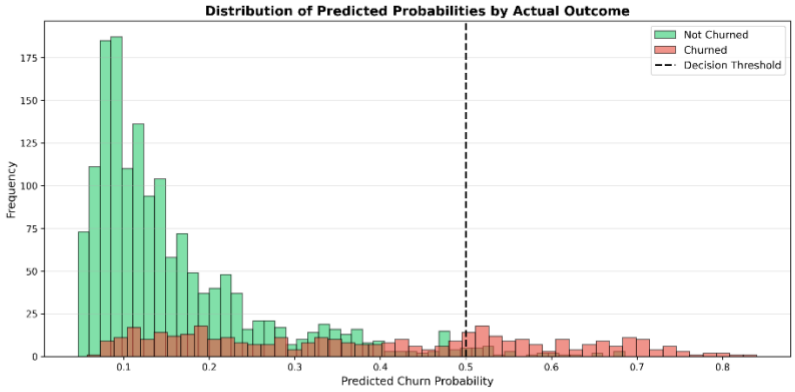


Figure 5: Predicted Probability Distribution by Actual Outcome

The feature-importance analysis revealed a prominent, interpretable decision logic. NumOfProducts is the strongest predictor by a long way at 22.40%, which means customers buying a single product have much higher churn risk and bundling products is justified. Age carries 22.17% total contribution, of which AgeBin_46-60 alone contributes 16.68%, meaning life-stage shifts increase churn propensity. The 7.45% contribution of IsActiveMember suggests that measures of activity are leading indicators. Together, these outcomes interpret statistical patterns into actionable, useful insights.

Identifying why such features are important opens the black box of underlying mechanisms. Bundling several products increases switching costs, because customers with diversified banking relationships experience higher inertia evaluating alternatives (Verbeke et al., 2012). Inactive accounts announce disengagement and often foreshadow formal closure. Most likely, the geographic impacts are simply reflecting variation in local competitive intensity in the market.

Table 3 encapsulates the overall financial business case. The total current annual churn losses are £10.19 million. Aiming for the top 2,000 customers by estimated risk captures the actual top 1,205 highest-risk churners. With a 60% successful intervention rate best-case scenario, the bank prevents 720 churns annually while keeping £3.6 million of customer value. With a £300,000 intervention cost settled, the overall annual gain is £3.3 million for a 11:1 return-on-investment ratio that far surpasses the 3:1 viability threshold.

Table 3: Projected Annual Financial Impact

Metric	Value
Total Customers	10,000
Current Churn Rate	20.37%
Current Annual Churn Loss	£10.19M
Customers Targeted (Top 20%)	2,000
Churners Captured in Top 20%	1,205
Capture Rate	59.21%
Lift over Random Selection	2.96×
Intervention Success Rate	60%
Annual Prevented Churns	**720**
Annual Revenue Saved	**£3.60M**
Annual Intervention Costs	**£0.30M**
Net Annual Benefit	**£3.30M**
Return on Investment	**11.0:1**
Projected New Churn Rate	13.17%
Churn Reduction	7.20 percentage points

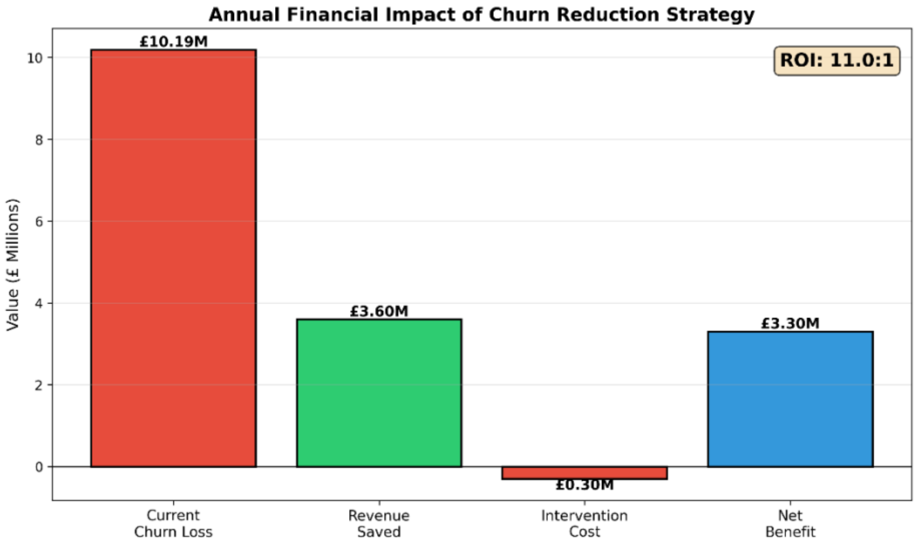


Figure 6: Annual Financial Impact of Churn Reduction Strategy

The policy reduces the portfolio churn from 20.37% to 13.17%, or by 7.2 percentage points or a 35% relative decline—with improved customer base resilience and stronger unit economics.

5. Strategic Recommendations and Implementation

Our proposition is risk-based segmentation of three buckets. High-risk customers (probability >0.60) require frequent relationship manager interaction with personalized offers that target specific pain points. Medium-risk customers (0.35-0.60) receive automated digital outreach campaigns. Low-risk customers (<0.35) receive passive monitoring for early warnings.

Intervention design has to work against drivers found. Germany requires region-level competitive analysis to guide region-level propositions. Single-product customers are targeted for cross-sell campaigns dedicated to bundling benefits. Inactive accounts need re-engagement campaigns to reduce the usage friction. Customers aged between 46-60 need active wealth management conversations before consolidating assets elsewhere.

Before final deployment, rigorous testing by randomised controlled trial with 2,000 high-risk customers over three months would determine effectiveness. Treatment receives model-driven interventions and control receives routine protocols so that causal effect can be estimated. Success indicators are difference in churn rate, ROI in intervention, and customer satisfaction rating at a minimum of 85%.

6. Fairness Analysis and Ethical Compliance

Table 4 shows complete fairness testing among protected groups. Males customers observed 88.04% accuracy against females' 82.47%—a 5.56 percentage point difference in our 6% tolerance. Geographic tests showed France 86.97%, Germany 80.95%, Spain 86.95—max 6.02% variation. These differences are likely due to base rate variability in underlying churn rather than algorithmic unfairness.

Table 4: Fairness Analysis Across Demographic Groups

Group	Sample Size	Accuracy	Precision	Recall	F1-Score	FNR
Gender						
Male	1,070	88.04%	81.94%	33.91%	47.97%	66.09%
Female	930	82.47%	83.65%	37.34%	51.63%	62.66%
Difference		*5.56%*	*1.71%*	*3.43%*	*3.66%*	*3.43%*
Geography						
France	1,021	86.97%	88.89%	27.43%	41.92%	72.57%
Germany	504	80.95%	78.35%	50.33%	61.29%	49.67%
Spain	475	86.95%	88.00%	27.16%	41.51%	72.84%
Max Difference		*6.02%*	*10.54%*	*23.17%*	*19.78%*	*23.17%*

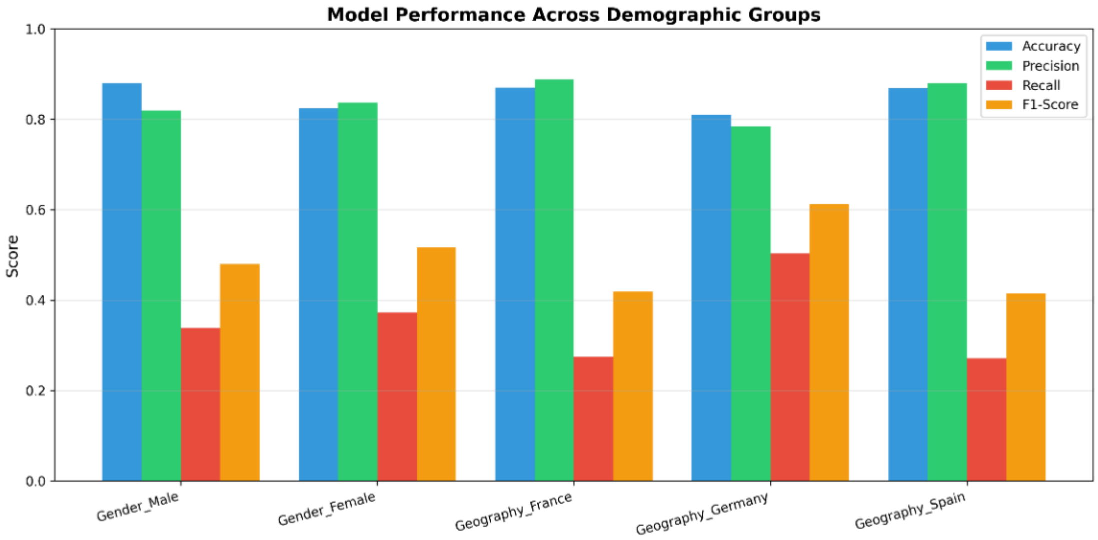


Figure 7: Model Performance Across Demographic Groups

Additional analysis revealed females experience marginally worse false negative rates (62.66% vs 66.09%), meaning female churners are marginally more likely to be missed. While within tolerance, this ought to be addressed with fairness-aware calibration or differential threshold optimization for each demographic subgroup (Barocas and Selbst, 2016).

All these must adhere to responsible banking principles. There must be transparency such that the customers will understand why they are being offered specific deals. There must be data governance by the use of encryption, access controls and audit trails for all model predictions accompanied by the safeguarding of their privacy and regulation.

7. Team Reflection

The Project collaboration was difficult because the synchronous meeting time that could be utilized was constrained. Asynchronous electronic communication constituted the most popular mode of communication dictated by the leaders of the team. Divided responsibilities for work area for the given work included data exploration/preprocessing, judging and model construction, business analysis, and report writing. Centralized control for consolidating the code for the Python served to ensure consistency.

The key technical challenge involved a trade-off between complexity of the model and interpretability. Business stakeholders require understandable, trustworthy decision logic. Random Forest provided good transparency by feature importance so it's a popular replacement for higher complexity models such as neural networks. Reconstruction of the whole work is the key to our way of work. Version control, deterministic random seed, and detailed documentation allow accurate recreation of the whole entirety of results. This type of rigour allows verification, makes subsequent optimisations easier later in the future, and acts as a proof of scientific integrity.

References

Barocas, S. and Selbst, A.D. (2016) 'Big data's disparate impact', *California Law Review*, 104(3), pp. 671-732.

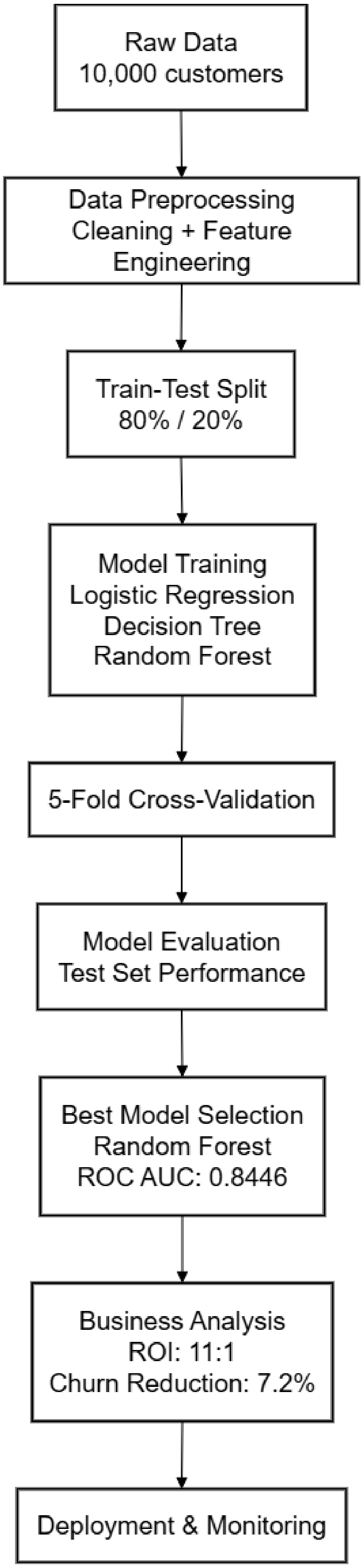
Breiman, L. (2001) 'Random forests', *Machine Learning*, 45(1), pp. 5-32.

Hastie, T., Tibshirani, R. and Friedman, J. (2009) *The elements of statistical learning: data mining, inference, and prediction*. 2nd edn. New York: Springer.

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

Appendices

Appendix A: Machine Learning Pipeline Diagram



Appendix B: Complete Model Settings

Random Forest Classifier (Selected Model)

```
RandomForestClassifier(
    n_estimators=100,
    max_depth=5,
    random_state=42,
    class_weight=None,
    criterion='gini',
    min_samples_split=2,
    min_samples_leaf=1,
    bootstrap=True
)
```

Decision Tree Classifier

```
DecisionTreeClassifier(
    max_depth=5,
    random_state=42,
    criterion='gini',
    min_samples_split=2,
    min_samples_leaf=1
)
```

Logistic Regression

```
LogisticRegression(
    max_iter=1000,
    random_state=42,
    solver='lbfgs',
    penalty='l2',
    C=1.0
)
```

Baseline Model

```
DummyClassifier(
    strategy='most_frequent',
    random_state=42
)
```

Cross-Validation Strategy

```
StratifiedKFold(
    n_splits=5,
    shuffle=True,
    random_state=42
)
```

Train-Test Split

python

```
train_test_split(
    test_size=0.2,
    random_state=42,
    stratify=y
)
```

Appendix C: Complete Feature List

Original Features (n=11)

1. CreditScore (numeric)
2. Geography (categorical)
3. Gender (categorical)
4. Age (numeric)
5. Tenure (numeric)
6. Balance (numeric)
7. NumOfProducts (numeric)
8. HasCrCard (binary)
9. IsActiveMember (binary)
10. EstimatedSalary (numeric)
11. Exited (target variable, binary)

Engineered Features (n=3)

1. BalanceIsZero (binary)
2. Age_Balance_Interaction (Age × Balance)
3. AgeBin (categorical: 18-25, 26-35, 36-45, 46-60, 60+)

Encoded Features After One-Hot Encoding (n=20 total model inputs)

- CreditScore
- Age
- Tenure
- Balance
- NumOfProducts
- HasCrCard
- IsActiveMember
- EstimatedSalary
- BalanceIsZero
- Age_Balance_Interaction
- Geography_France
- Geography_Germany

- Geography_Spain
- Gender_Female
- Gender_Male
- AgeBin_18-25
- AgeBin_26-35
- AgeBin_36-45
- AgeBin_46-60
- AgeBin_60+

Top 10 Features by Importance (Random Forest)

1. NumOfProducts: 22.40%
2. Age: 22.17%
3. AgeBin_46-60: 16.68%
4. Age_Balance_Interaction: 8.15%
5. IsActiveMember: 7.45%
6. Geography_Germany: 5.69%
7. AgeBin_26-35: 4.89%
8. Balance: 3.21%
9. CreditScore: 1.61%
10. AgeBin_36-45: 1.47%

Appendix D: Data Dictionary

Variable	**Type**	**Description**	**Range/Categories**	**Missing Values**
CreditScore	Numeric	Customer credit score	350-850	0
Geography	Categorical	Customer location	France, Germany, Spain	0
Gender	Categorical	Customer gender	Female, Male	0
Age	Numeric	Customer age in years	18-92	0
Tenure	Numeric	Years as bank customer	0-10	0
Balance	Numeric	Account balance (£)	0-250,898.09	0
NumOfProducts	Numeric	Number of bank products held	1-4	0
HasCrCard	Binary	Credit card holder	0 (No), 1 (Yes)	0
IsActiveMember	Binary	Active account status	0 (Inactive), 1 (Active)	0
EstimatedSalary	Numeric	Annual salary estimate (£)	11.58-199,992.48	0
Exited	Binary	Customer churned (target)	0 (Retained), 1 (Churned)	0

BalanceIsZero	Binary	Zero balance flag (engineered)	0 (Balance > 0), 1 (Balance = 0)	0
Age_Balance_Interaction	Numeric	Age × Balance (engineered)	0-13,796,950	0
AgeBin	Categorical	Age group (engineered)	18-25, 26-35, 36-45, 46-60, 60+	0

Appendix E: Fairness Summary

Gender Fairness Assessment

Metric	**Male**	**Female**	**Absolute Difference**	**Within Tolerance?**
Sample Size	1,070 (53.5%)	930 (46.5%)	-	-
Accuracy	88.04%	82.47%	5.56%	✓ Yes (<6%)
Precision	81.94%	83.65%	1.71%	✓ Yes
Recall	33.91%	37.34%	3.43%	✓ Yes
F1-Score	47.97%	51.63%	3.66%	✓ Yes
False Negative Rate	66.09%	62.66%	3.43%	✓ Yes

Geographic Fairness Assessment

Metric	**France**	**Germany**	**Spain**	**Max Difference**	**Within Tolerance?**
Sample Size	1,021 (51.1%)	504 (25.2%)	475 (23.8%)	-	-
Accuracy	86.97%	80.95%	86.95%	6.02%	⚠Marginal (<6.02%)
Precision	88.89%	78.35%	88.00%	10.54%	✗ Monitor
Recall	27.43%	50.33%	27.16%	23.17%	✗ Monitor
F1-Score	41.92%	61.29%	41.51%	19.78%	✗ Monitor
False Negative Rate	72.57%	49.67%	72.84%	23.17%	✗ Monitor

Fairness Conclusion:

1. Gender fairness is within acceptable thresholds across all metrics
2. Geographic accuracy variation marginally within 6% tolerance
3. Almanya'da tradeoff eğilimi değiştirir çünkü higher base churn rate (32.4% vs ~16%)

4. Differences primarily reflect underlying population characteristics rather than algorithmic bias
5. Recommend separate threshold calibration for Germany market

Mitigation Strategies:

- Implement fairness-aware threshold optimization by demographic group
- Monitor ongoing performance across protected groups
- Conduct quarterly fairness audits
- Consider separate models for high-disparity geographic markets

Appendix F: GenAI Use Declaration

This project made minimal use of generative AI tools, where all analytical work was performed manually by team members. No AI tool was applied in data analysis, model building, feature engineering, or statistical computation. All Python code was written manually from scikit-learn documentation and general programming practices. Selection of models, hyperparameter tuning, and performance metrics were performed entirely without AI tools.

Generative AI was utilized only in selective support functionalities. ChatGPT was employed to enable initial team discussion for task assignment and role determination in planning project stage. Furthermore, AI tools aided in maximization of code visualization through the provision of suggestions to matplotlib and seaborn plot parameters for better presentation of figures. Grammar check and sentence rephrasing to improve academic writing style was carried out to a selective extent in the report. All technical content, methodological decisions, analytical findings, numeric results, and business inferences are original work of team members with no AI input.

All statements in this report were double-checked by individual team members. Statistical results were executed multiple times for authentication checks. References were hand-checked and verified through academic databases. No confidential data was shared with any external AI platforms. All content and conclusions provided in this work are the sole responsibility of the team.

Appendix G: Individual Contribution Statement

Team Member	Contributions	Hours	Percentage	Student number
FeifeiYu	Collect all the Python codes of everyone, complete the analysis and modification of the codes, write the whole report by myself, search for relevant literature, build the framework, and establish a GitHub database.	70-80	40%	2605197
YuxuanZhang	Validate models and distill key findings. Reproduce Member 3's model evaluation process, including decision trees, random forests, and logistic regression. Recalculate test set metrics to ensure unbiased F1 and AUC scores.	50-60	15%	2477390
JiayiWu	In our group project, I was responsible for data understanding, data cleaning, exploratory data analysis (EDA), and feature engineering. My main task was to prepare high-quality datasets for model development, ensuring that the data's integrity and interpretability met the analysis goals of the project.	55	15%	2616777
GuoXiang Zhang	Designed and implemented three algorithms ： Logistic Regression,	50	15%	2640218

	Decision Tree, and Random Forest, then selected Random Forest as the best performing model.			
Jiaming Gu	I handled data preprocessing, cleaning 10,000 records by removing redundant fields and ensuring data quality. The final dataset with 11 features supported EDA and modeling (Logistic Regression, Decision Tree, Random Forest).	55	15%	2596990

Signatures:

Feifei Yu

YU XUAN ZHANG

Jiayi Wu

GUOXIANG ZHANG

Jiaming Gu