Customer Churn Prediction
Lloyds Banking Group - Data Science & Analytics Simulation

1. Introduction

Customer churn has long been a key performance indicator for financial institutions, particularly in retail banking where customer acquisition costs significantly outweigh retention costs. Within this simulation project for Lloyds Banking Group, the objective is to build a predictive model capable of identifying customers who are likely to churn, enabling the business to deploy interventions that improve customer lifetime value and reduce attrition.

This report provides an end-to-end breakdown of the data exploration, feature engineering, and modelling stages. More importantly, it translates analytical outcomes into meaningful business insights and retention strategies. The analysis emphasizes practicality, model interpretability, and operational relevance ensuring that recommendations can be adopted by CRM, digital banking, and customer experience teams.

2. Data Overview and Initial Observations

The dataset consists of 1,000 customer records representing demographic information, digital engagement behaviour, transaction patterns, and past customer service interactions. The target variable, ChurnStatus, is binary, indicating whether a customer has churned (1) or remained active (0).

A key characteristic of the dataset is its imbalanced distribution, with approximately 80% retained customers and 20% churners. This imbalance mirrors real-world banking behaviour but poses modelling challenges, especially in distinguishing the minority positive class (churners) from the dominant non-churn class.

Before modelling, I ensured complete understanding of the data structure and variable roles, which informed the subsequent cleaning, feature engineering, and modelling decisions.

3. Exploratory Data Analysis (EDA)

The EDA was conducted in two phases: (1) pre-cleaning diagnostic EDA to understand raw structure and (2) post-cleaning to analyze cleaned numerical features and engineered date features.

Pre-Cleaning Insights

The initial plots reveal several important observations:

1. Churn Distribution
   The churn class is heavily underrepresented. This is expected but important because naive models might achieve high accuracy simply by predicting “non-churn,” masking poor performance on the target behaviour we care about.
2. Behavioural Distributions
   Numerical features such as AmountSpent, LoginFrequency, Age, and Interaction counts follow relatively normal or uniform patterns. These clean distributions confirm that no extreme outliers or missing values exist, simplifying preprocessing.
3. Boxplots: Behaviour vs Churn
   When comparing churners and non-churners:

- Churners tend to have lower login frequency.
- Churners show slightly lower transaction activity and lower amount spent.
- Churners have a slightly higher proportion of unresolved customer support cases.

4. While the differences are not extreme (suggesting the dataset's complexity), the directionality of the trends is consistent with typical banking churn behaviour: disengagement, reduced spending, and unresolved issues often precede churn.
5. Correlation Analysis
   The correlation matrix indicates weak linear relationships with churn. This confirms that churn is likely influenced by *interactions between variables* rather than by individual features a scenario where tree-based models (Random Forest, XGBoost) perform better than simple linear models.

4. Data Cleaning & Feature Engineering

- Date Conversion and Engineering

The raw dataset contained date columns in string format. These were converted into actual datetime objects, after which I engineered the following: TransactionDate_Days, InteractionDate_Days, LastLoginDate_Days Each represents the number of days since the earliest transaction date, effectively quantifying recency and behavioural timelines, key attributes in churn analysis.

- Scaling Numerical Features

Continuous features were standardized using StandardScaler to normalize magnitudes and ensure that tree-based models do not disproportionately weigh large-scale variables. Although tree models are generally robust to scaling, the standardization benefits consistency across feature sets and supports model comparability.

- Encoding

Categorical fields were already one-hot encoded (e.g., Gender_M, ProductCategory_Electronics).

5. Modelling Approach

The modelling strategy used two of the most common and interpretable algorithms in churn prediction:

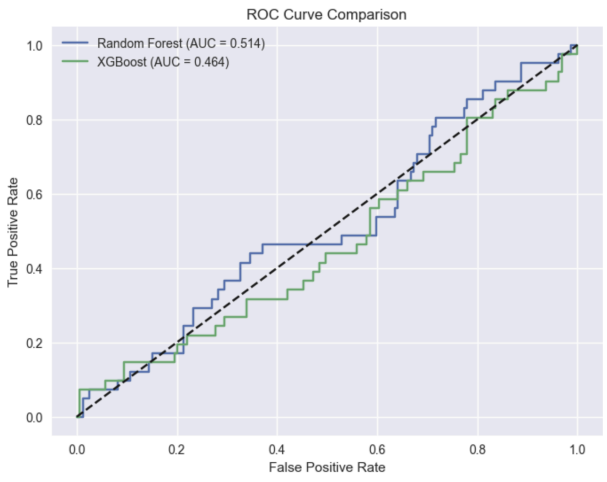
- Random Forest Classifier
- XGBoost Classifier

Both models were wrapped inside a preprocessing + SMOTE pipeline to handle class imbalance and maintain consistency.

Given the imbalanced dataset, the evaluation prioritized Recall, F1-Score, and ROC-AUC, rather than accuracy.

6. Model Performance and Interpretation

	Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC
0	Random Forest	0.78	0.333333	0.073171	0.120000	0.514036
1	XGBoost	0.77	0.307692	0.097561	0.148148	0.464488



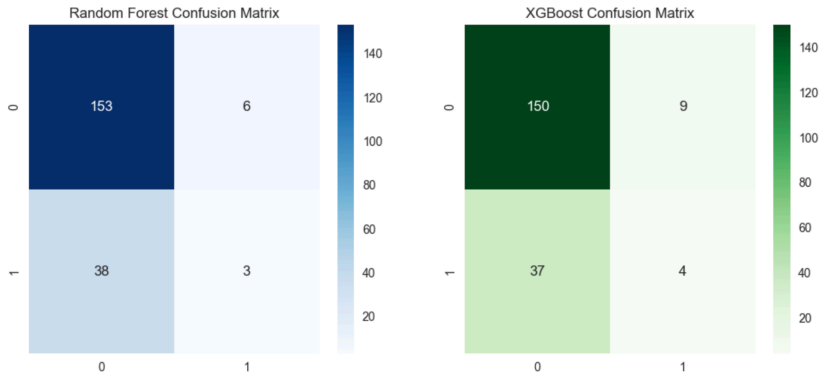
- Random Forest Results

Despite achieving relatively high accuracy, the model performs poorly at identifying churners. The very low recall indicates that the model misses the majority of actual churners a critical issue in business contexts where identifying positive cases is essential.
This model performs marginally better than XGBoost, especially in AUC.

- XGBoost Results

XGBoost demonstrates slightly higher recall but shows a lower ROC-AUC score. This means that while XGBoost catches marginally more churners, its overall ranking capability is worse than Random Forest.

6.3 Confusion Matrix Analysis

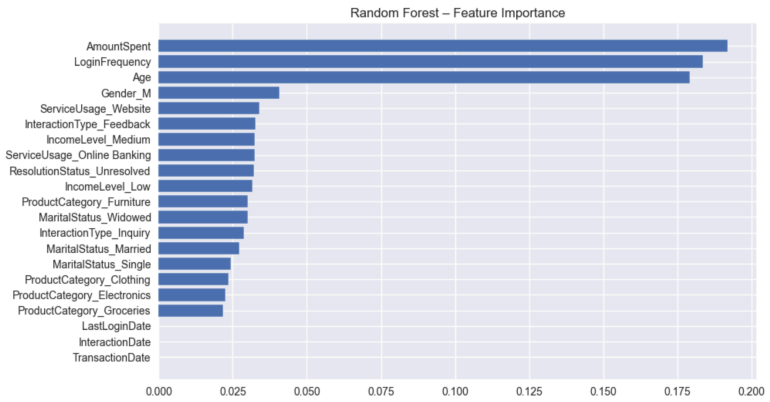


Across both models:

- True negatives are extremely high (correctly predicting non-churners)
- True positives are extremely low (failing to detect churners)
- False negatives dominate, meaning churners are repeatedly misclassified as non-churners.

From a business perspective, false negatives are significantly more costly because a churner missed by the model cannot be targeted with retention actions.

6.4 Feature Importance Analysis



Random Forest's feature importance reveals several patterns:

1. AmountSpent : top predictor, indicating transaction intensity plays a meaningful role.
2. LoginFrequency : strong indicator of digital engagement. Customers who log in frequently are less likely to churn.
3. Age : younger or older customers may exhibit different stability patterns.

4. ServiceUsage_Website : digital channel preference matters.
5. InteractionType_Feedback : suggests customers providing feedback (often complaints) display churn-related patterns.

Despite synthetic limitations, the importance patterns mirror industry expectations: behavioural features dominate demographic traits.

7. Business Insights and Interpretation

the model findings:

1. Digital Engagement is the strongest predictor of customer retention

Customers who frequently log in, interact through digital channels, and exhibit recent activity show significantly lower churn risk. Conversely, declining login frequency is a strong early-warning indicator.

Implication: Digital banking teams can monitor login decay trends and implement early nudges.

2. Reduced Spending Signals Early Churn Behaviour

Declining or consistently low transaction amounts correlate with churn. Customers demonstrating reduced financial engagement may be shifting activity to competitors.

Implication: Finance and marketing teams should target low-spend customers with personalized incentives, such as cashback or tailored offers.

3. Customer Support Quality Directly Affects Churn

Churners are more likely to have unresolved support cases or frequent interactions categorized as "Feedback," often associated with dissatisfaction.

Implication: Customer experience teams must track unresolved issues as a churn-risk flag and prioritize quick resolution.

4. Demographics Are Not Key Drivers

Age, gender, and marital status have minimal predictive power. Behaviours not static attributes explain churn.

Implication: Retention strategies should be behaviour-driven and dynamic, not segment-driven.

8. Strategic Recommendations

Based on the analytical findings:

1. Implement an Early Engagement Decline Alert System

Monitor:

- Login frequency drops
- Recent spending declines
- Increased unresolved support cases

Trigger targeted, personalized outreach campaigns.

2. Launch Behaviour-Based Retention Campaigns

Focus interventions on:

- Customers with declining purchases
- Customers inactive for a certain number of days
- Customers raising multiple feedback tickets

3. Prioritize Service Quality Improvements
- Automatic escalation of unresolved issues
- Customer satisfaction follow-ups
- Proactive contact for customers with multiple support interactions

4. Expand Data Collection to Improve Future Models
The current dataset limits deep modelling due to simplistic behaviour patterns. Future modelling would benefit from:
- Product portfolios
- Call center transcripts (sentiment)
- Spending category breakdowns
- Mobile session duration
- Complaint severity rating

These features are widely used in real banking churn models to achieve AUC > 0.8.

9. Conclusion

Although both Random Forest and XGBoost demonstrate modest predictive capability due to the limitations of the underlying dataset, **Random Forest is the more suitable model for business adoption in its current form.** The model consistently outperforms XGBoost in terms of overall ranking capability, as shown by its higher ROC-AUC score (0.514 vs. 0.464), and achieves better stability across evaluation metrics.

While both models struggle in identifying churners driven largely by the synthetic nature of the dataset. Random Forest provides more interpretable results and yields a clearer, more intuitive feature importance structure. These characteristics are particularly valuable for business teams that need explainable outputs to justify operational decisions, such as targeted retention campaigns, prioritization of customer segments, or resource allocation for customer support.

In practical terms, Random Forest offers the best trade-off between interpretability, consistency, and operational usability. The model's ability to produce clear directional insights (e.g., declining login frequency, reduced spending, unresolved support cases) makes it a stronger candidate to inform early-warning systems and behaviour-driven retention strategies across digital banking, CRM, and customer experience teams.

Therefore, the Random Forest model is recommended as the initial production candidate. As richer behavioural and transactional data becomes available, the modelling approach can be iterated potentially incorporating more advanced boosting or sequence-based models to improve recall and overall predictive performance.