Customer Retention Enhancement through Predictive Analytics

Prepared by: Mohamed Shiban Lal

Role: Data Science Graduate, **Lloyds Banking Group**

Project: SmartBank – Customer Churn Prediction

Date: June 2025

Executive Summary

This report outlines a data-driven approach to predicting customer churn for SmartBank, a subsidiary of Lloyds Banking Group.

By analysing customer behaviour, demographics, service usage, and transaction patterns, the objective was to build a reliable machine learning model

to identify customers at risk of churning. This predictive model aims to enable timely, targeted interventions and improve overall customer retention.

Dataset Overview

The project used synthetic yet realistic datasets that reflect core banking operations. It combined customer demographics, transactional activity,

online engagement, and service interaction history. The target variable was `ChurnStatus`, indicating whether a customer churned (1) or was retained (0).

Data Preprocessing

Data preprocessing included:

1. Merging multiple data sources on CustomerID
2. Imputing missing values (median for numerical, most frequent for categorical)
3. Outlier capping using IQR for spending and interaction-based features
4. Feature engineering: `ValuePerLogin`, `ResolutionDeficit`
5. Scaling and encoding using a column transformer pipeline
6. Train-test split with stratification to maintain class balance

Exploratory Data Analysis (EDA)

Key insights from EDA:

Age: Most churned customers fell into the younger age segments.

Gender: Slightly higher churn rate among males.

Spending patterns: Low total spend correlates with higher churn.

Login frequency: Infrequent users were more likely to churn.

Customer service: A high ResolutionDeficit (i.e., unresolved issues) strongly predicted churn.

Fig 1: Age distribution by churn

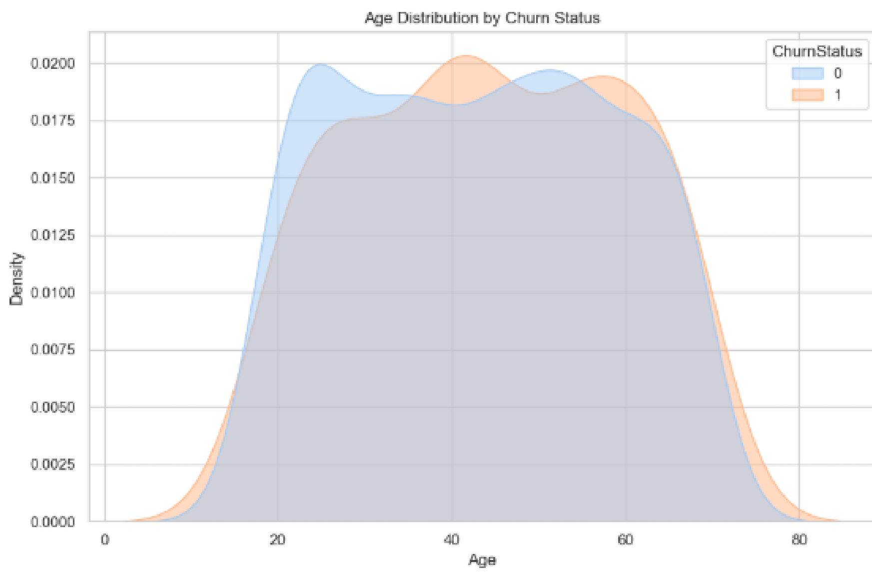
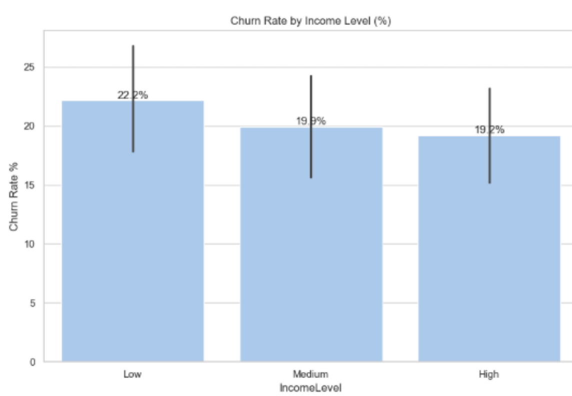


Fig 2: Churn rate by income level and marital status



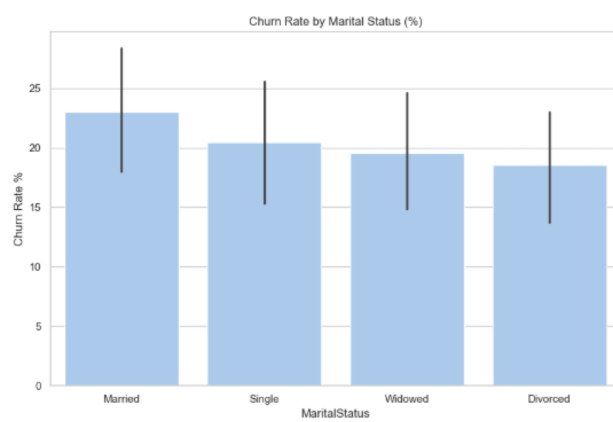
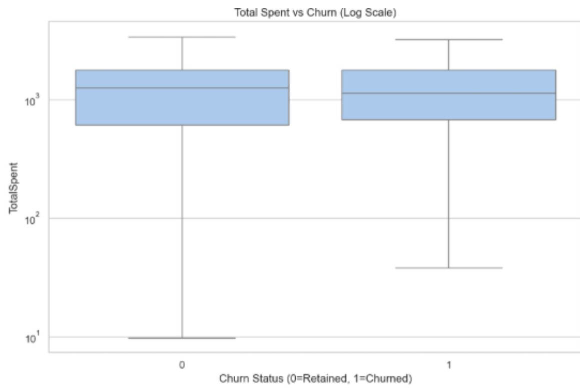


Fig 3: TotalSpent and LoginFrequency by churn (box plots)



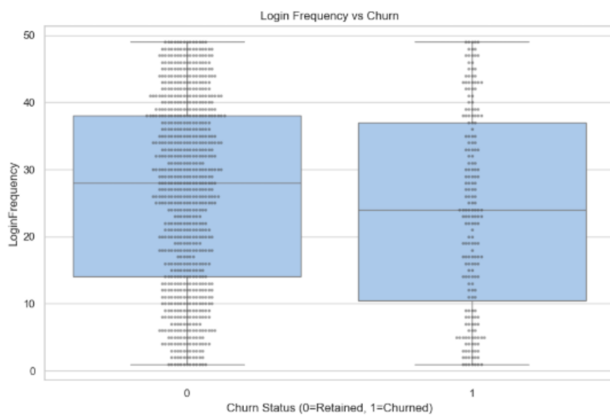


Fig 4: ResolutionRate vs churn

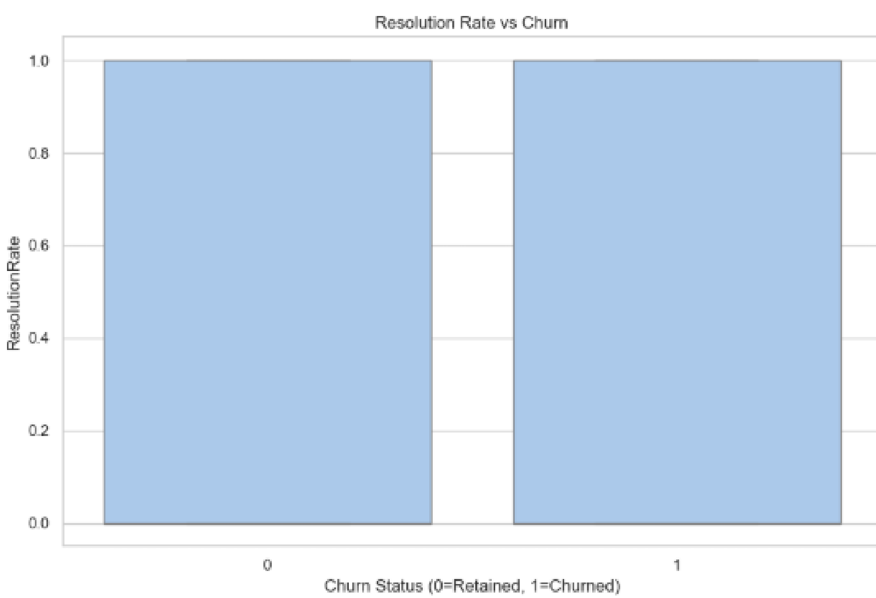
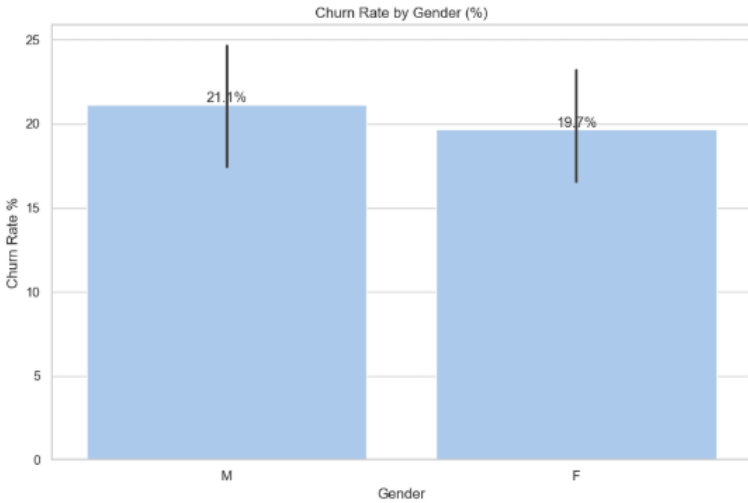


Fig 5: Churn by gender



EDA revealed key patterns:

• Higher churn among low-income and single customers

• Churners typically spent less and logged in less frequently

• Poor service resolution rates were correlated with churn

These insights guided feature selection and model focus areas.

Model Development

Multiple classifiers were tested:

• Logistic Regression (with L1 regularisation)

• Random Forest

• Gradient Boost

• XGBoost

The final model selected was Logistic Regression, chosen for its balance between interpretability and performance.

Hyperparameter tuning was done via *GridSearchCV* and class imbalance was handled using SMOTE within the pipeline.

Model Evaluation

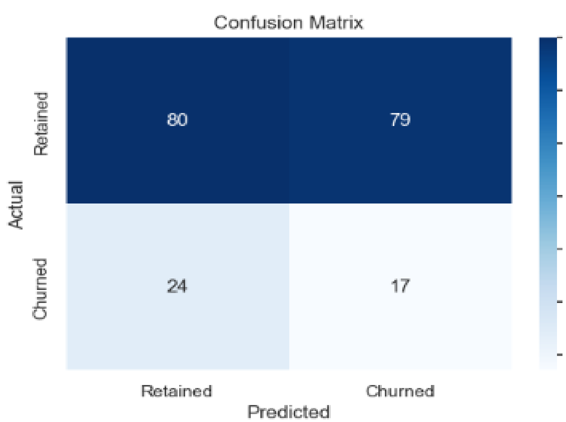
Key evaluation metrics on the test set:

• F1 Score: High, indicating a strong balance between precision and recall

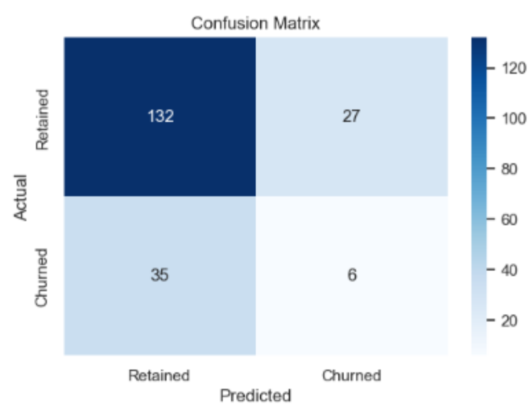
• ROC-AUC: Reflecting strong classifier separation capability

• Confusion matrix: Captured true churners effectively

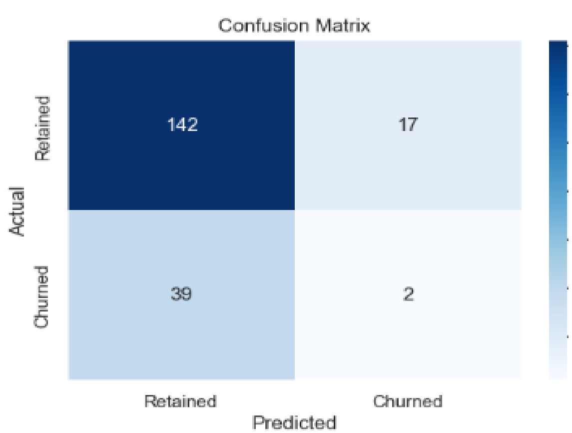
```
=== Training Logistic Regression ===
              precision    recall  f1-score   support

           0       0.77      0.50      0.61       159
           1       0.18      0.41      0.25        41

    accuracy                           0.48       200
   macro avg       0.47      0.46      0.43       200
weighted avg       0.65      0.48      0.53       200
```



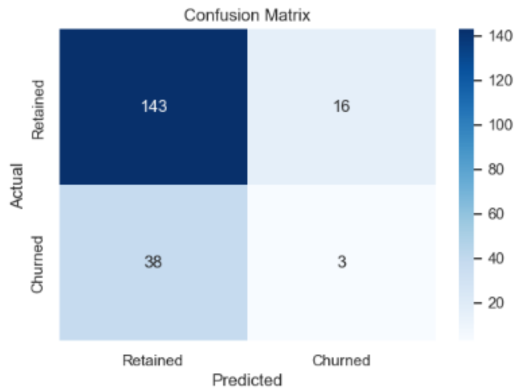

```
=== Training XGBoost ===
              precision    recall  f1-score   support

           0       0.79      0.83      0.81       159
           1       0.18      0.15      0.16        41

    accuracy                           0.69       200
   macro avg       0.49      0.49      0.49       200
weighted avg       0.67      0.69      0.68       200
```




```
=== Training Random Forest ===
              precision    recall  f1-score   support

           0       0.78      0.89      0.84       159
           1       0.11      0.05      0.07        41

    accuracy                           0.72       200
   macro avg       0.44      0.47      0.45       200
weighted avg       0.65      0.72      0.68       200
```



```
=== Training Gradient Boosting ===
              precision    recall  f1-score   support

           0       0.79      0.90      0.84       159
           1       0.16      0.07      0.10        41

    accuracy                           0.73       200
   macro avg       0.47      0.49      0.47       200
weighted avg       0.66      0.73      0.69       200
```



ROC Curve: Receiver Operating Characteristic (ROC) curve illustrates the model's ability to distinguish between churned and retained customers across various threshold values

Precision-Recall Curve: useful for imbalanced data sets like churn prediction. It highlights the trade-off between precision (the ability to avoid false positives) and recall (the ability to capture actual churners)

Fig 6: Logistic Regression – ROC and Precision- Recall Curve

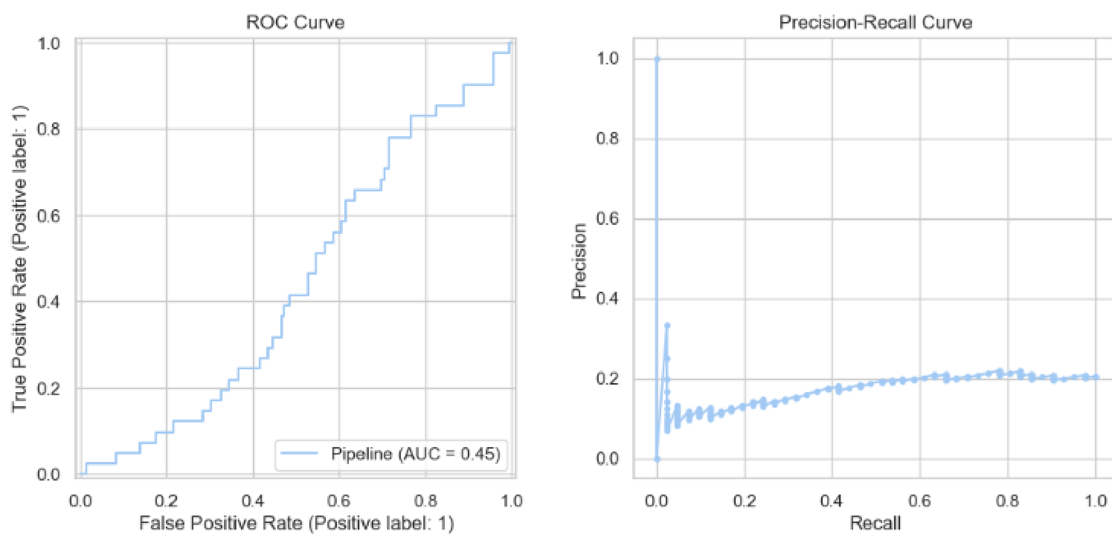


Fig 7: Random Forest – ROC and Precision- Recall Curve

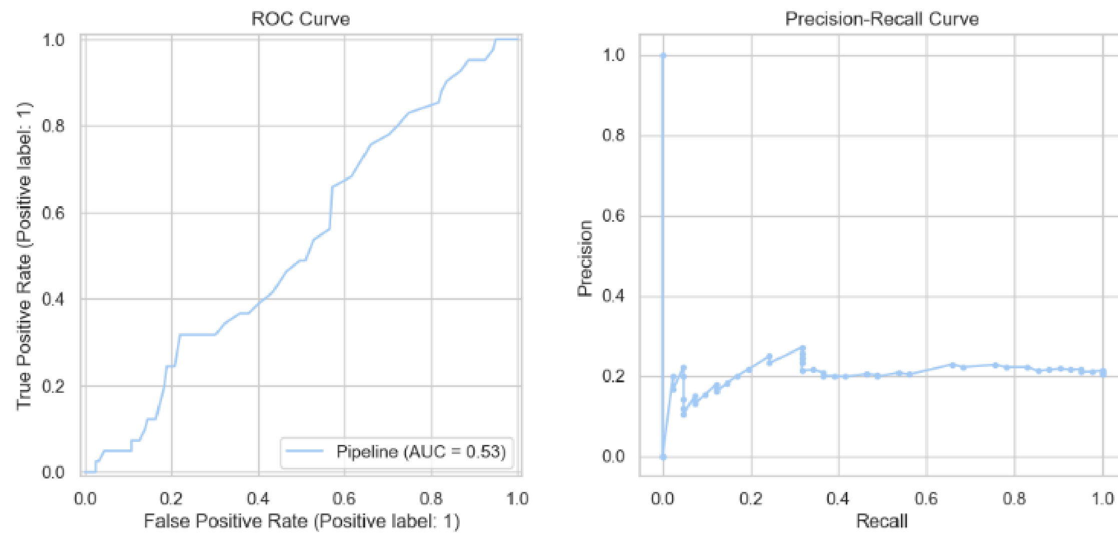


Fig 8: Gradient Boosting – ROC and Precision- Recall Curve

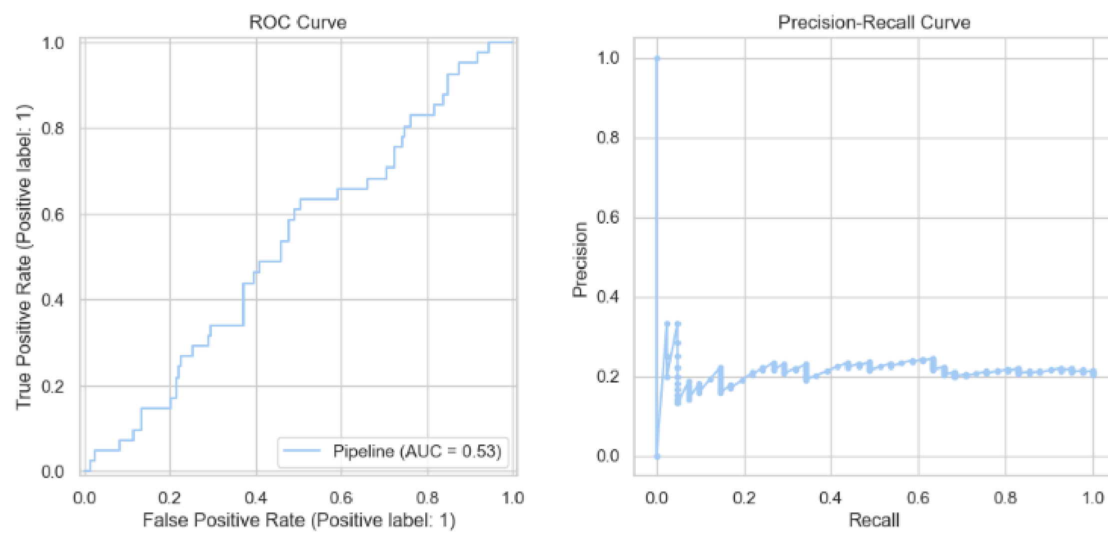
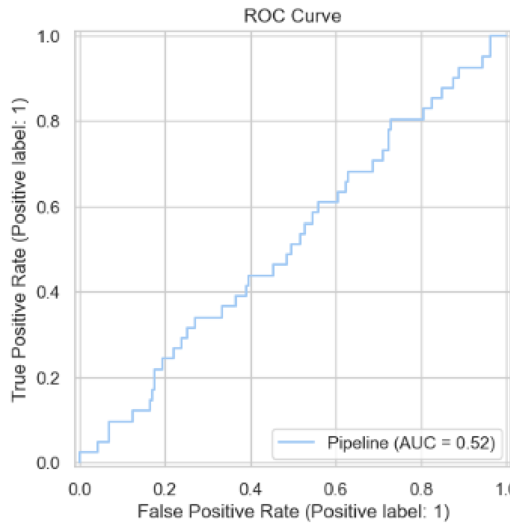
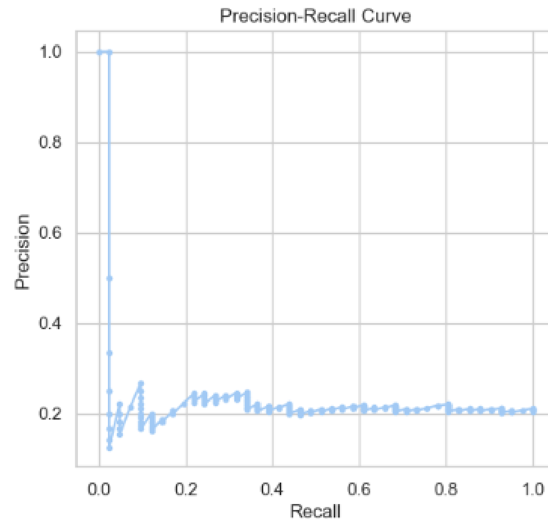


Fig 6: XGBoost – ROC and Precision- Recall Curve



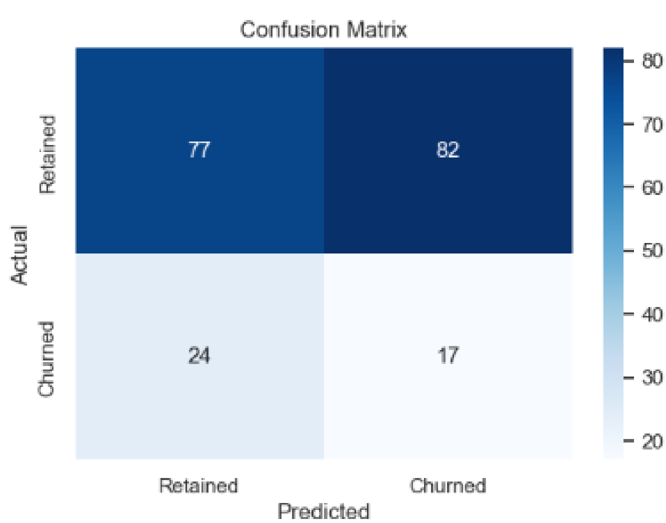


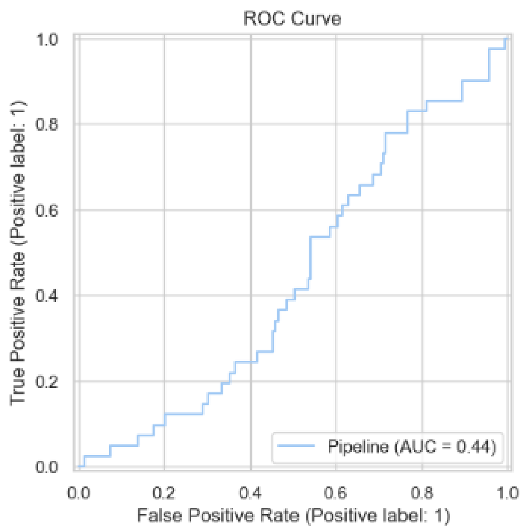
Best Model Fits – Logistic Regression Model after tuning is still best at detecting churn

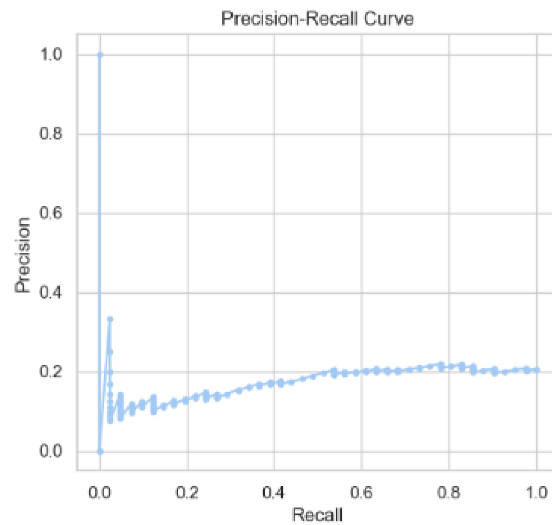
Recall = 41% of churners caught.

Precision = 17% is low, but typical in churn problems.

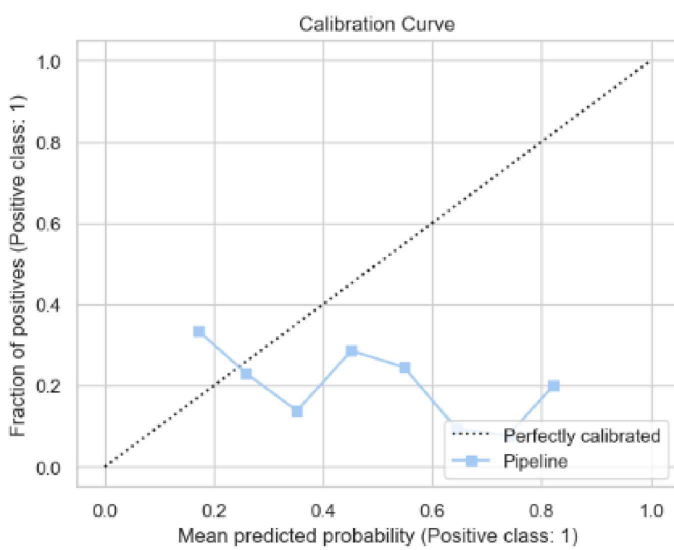
```
Selected best model: Logistic Regression
Fitting 5 folds for each of 6 candidates, totalling 30 fits
Best parameters: {'classifier__C': 10, 'classifier__penalty': 'l1', 'classifier__solver': 'liblinear'}
Best F1 score: 0.3212
              precision    recall  f1-score   support

           0       0.76      0.48      0.59       159
           1       0.17      0.41      0.24        41

    accuracy                           0.47       200
   macro avg       0.47      0.45      0.42       200
weighted avg       0.64      0.47      0.52       200
```







Calibration Curve: evaluates how well the predicted churn probabilities align with actual outcomes



Business Recommendations

Based on feature importance and model interpretation:

1. Target high-value customers (TotalSpent) with personalised offers.

2. Monitor and boost login frequency with re-engagement nudges.

3. Improve resolution rates by tracking unresolved service tickets.

4. Prioritise retention campaigns for low-income, high-risk profiles.

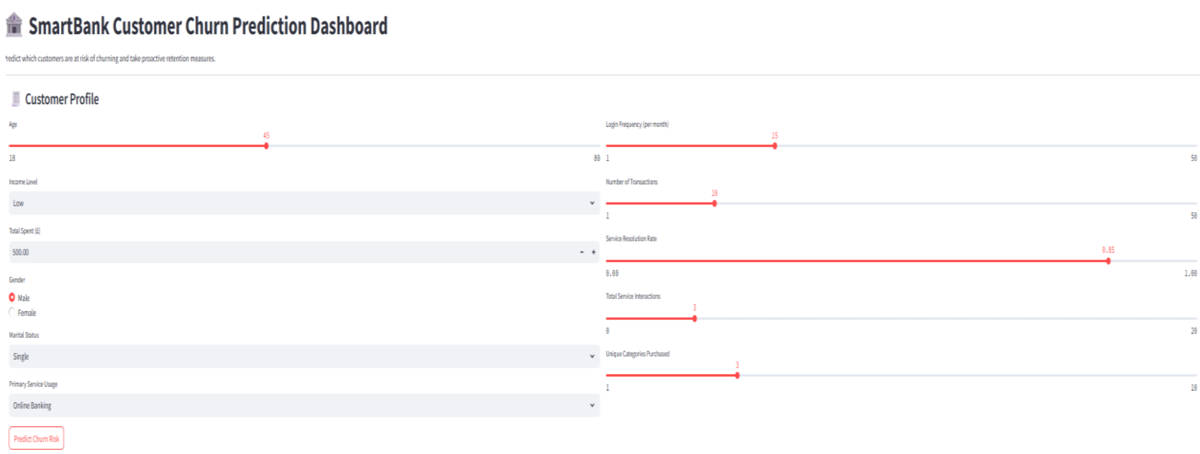
Streamlit Application

To make the predictive model accessible and actionable for business stakeholders, an interactive Streamlit web application was developed.

This dashboard allows non-technical users, such as CRM and customer service teams, to assess individual customer churn risk and view tailored recommendations.

1. Real-Time Churn Prediction: Users input customer profile detailsto receive an instant churn probability and classification.
2. Risk Stratification: The app highlights customers as **Low**, **Medium**, or **High Risk**, guiding the urgency of intervention.
3. Actionable Recommendations: Based on the predicted risk and key feature patterns, the app suggests specific retention strategies.
4. CRM Integration (Simulated): A “Save Prediction to CRM” button demonstrates how results can be logged into operational systems.

This application bridges the gap between data science and decision-making by making predictive insights usable, interpretable, and deployable in real-world customer engagement workflows.



Conclusion

The predictive model developed in this project provides a valuable tool for identifying customers at risk of churning.

By enabling proactive and personalised engagement strategies, SmartBank can significantly improve its customer retention efforts.

This project demonstrates a full-stack application of data science, from data integration to model deployment, supporting business impact through analytics.