P101/1740G/21

KAHENI PETER

Predictive analytics in Business Intelligence

Technical Assignment II

Predictive Analytics Using Machine Learning

Dataset: bank_customer_churn.csv

1. Dataset description

The dataset contain customer-related financial and behavioral data to predict churn (whether a customer leaves or stays with a bank).

Column Name	Description
customer_id	Unique identifier for each customer.
account_type	Type of bank account (e.g., Savings, Checking, Business, Premium).
employment_status	Employment status of the customer (e.g., Employed, Unemployed, Self-Employed).
transaction_count_last_6_months	Number of transactions made in the last six months.

loan_status	Status of any loan associated with the customer (e.g., No Loan, Active Loan, Paid, Defaulted).
complaints_filed	Number of complaints the customer has filed.
churn_label	Target variable indicating churn (e.g., 1 = churned, 0 = stayed).
review_text	Customer's feedback or review text about the bank.
sentiment_score	Numeric sentiment analysis score derived from the review text.
sentiment_label	Sentiment classification of the review (e.g., Positive, Negative, Neutral).
customer_feedback_rating	Rating given by the customer (e.g., 1-5 stars).
service_issue_type	Type of service issue faced by the customer (e.g., Billing, Technical, Account-related).
active_products	Number of active banking products the customer is using.
preferred_transaction_type	Customer's preferred way of making transactions (e.g., Online, ATM, Mobile, In-Branch).
num_credit_cards	Number of credit cards the customer owns.
credit_utilization_ratio	Ratio of used credit to available credit (a financial risk indicator).

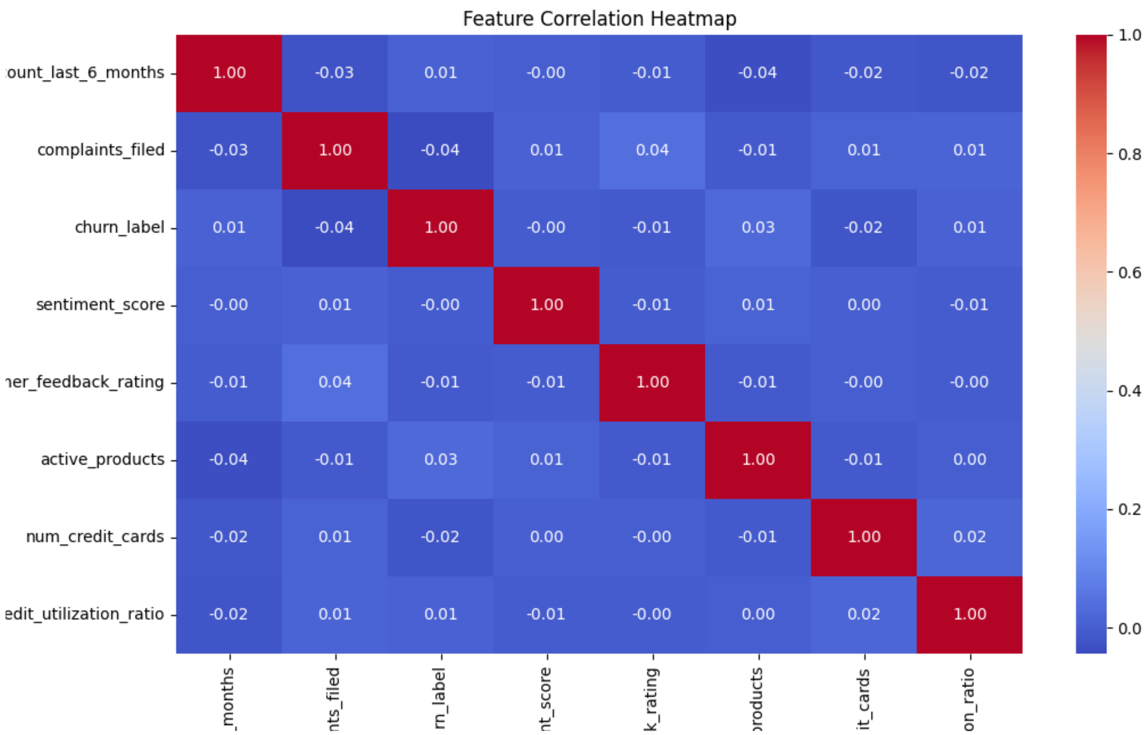
recommended_product	A recommended financial product based on customer behavior (e.g., Credit Card, Personal Loan, Investment Plan).

2. ## Key findings from EDA

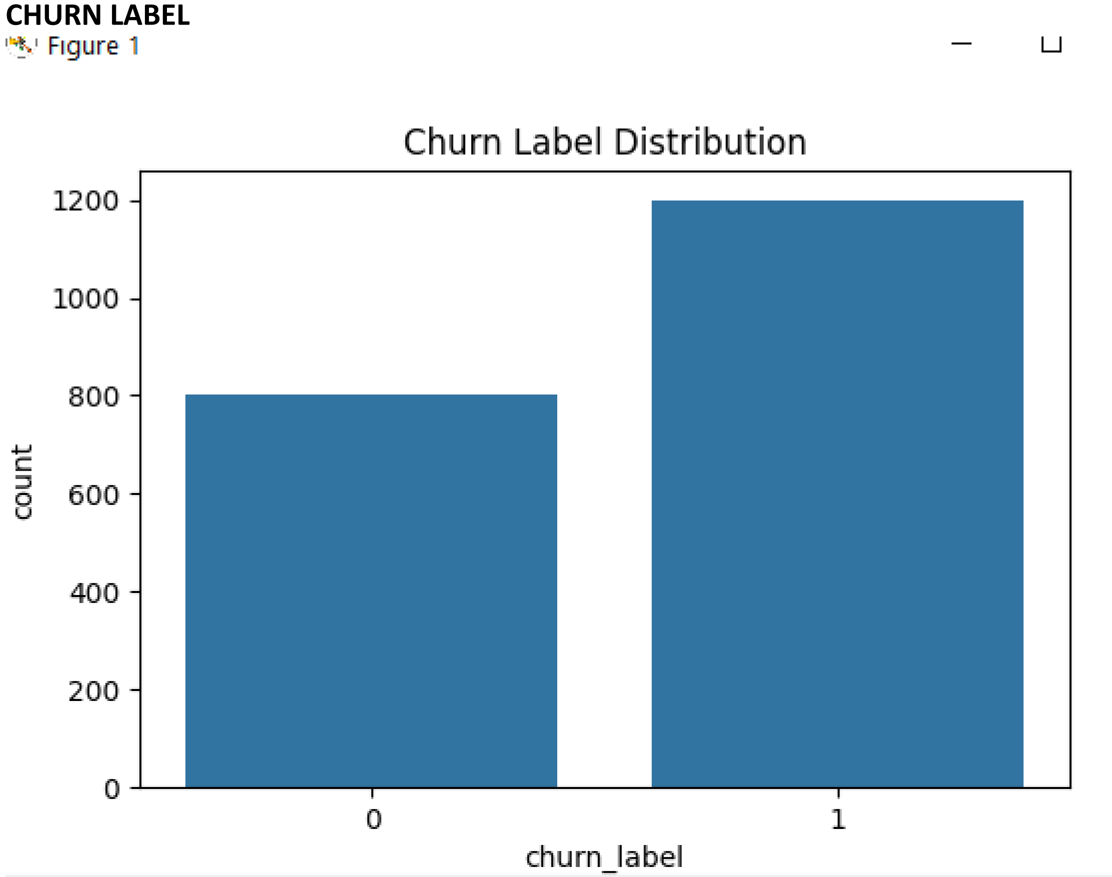
A. SUMMARY STATISTICS.

```
C:\Users\HP\churn_prediction_system\models>python churn_prediction.py
□  Loading data...
□ Data loaded successfully.
□  Summary Statistics:
       transaction_count_last_6_months  complaints_filed  ...  num_credit_cards  credit_utilization_ratio
count                      2000.000000       2000.000000  ...       2000.000000               2000.000000
mean                         24.743000          2.456000  ...          1.994000                  0.499420
std                          14.397694          1.711002  ...          1.415262                  0.283958
min                           0.000000          0.000000  ...          0.000000                  0.000000
25%                          12.000000          1.000000  ...          1.000000                  0.260000
50%                          25.000000          2.000000  ...          2.000000                  0.490000
75%                          37.000000          4.000000  ...          3.000000                  0.732500
max                          49.000000          5.000000  ...          4.000000                  1.000000

[8 rows x 8 columns]
□  Checking missing values...
Series([], dtype: int64)
```

B. CORRELATION ANALYSIS.



C. CHURN LABEL



3. Model selection and evaluation results

I have trained my model using four models namely: `Logistic Regression, Decision Tree, Random Forest and Gradient Boosting.`

For Logistic

```
Logistic Regression Metrics:
Accuracy: 0.5925, Precision: 0.6005, Recall: 0.9583, F1-score: 0.7384
Confusion Matrix:
[[  7 153]
 [ 10 230]]
              precision    recall  f1-score   support

           0       0.41      0.04      0.08       160
           1       0.60      0.96      0.74       240

    accuracy                           0.59       400
   macro avg       0.51      0.50      0.41       400
weighted avg       0.53      0.59      0.47       400
```

For Decision Tree

```
Decision Tree Metrics:
Accuracy: 0.5475, Precision: 0.6166, Recall: 0.6500, F1-score: 0.6329
Confusion Matrix:
[[ 63  97]
 [ 84 156]]
              precision    recall  f1-score   support

           0       0.43      0.39      0.41       160
           1       0.62      0.65      0.63       240

    accuracy                           0.55       400
   macro avg       0.52      0.52      0.52       400
weighted avg       0.54      0.55      0.54       400
```

For Random Forest

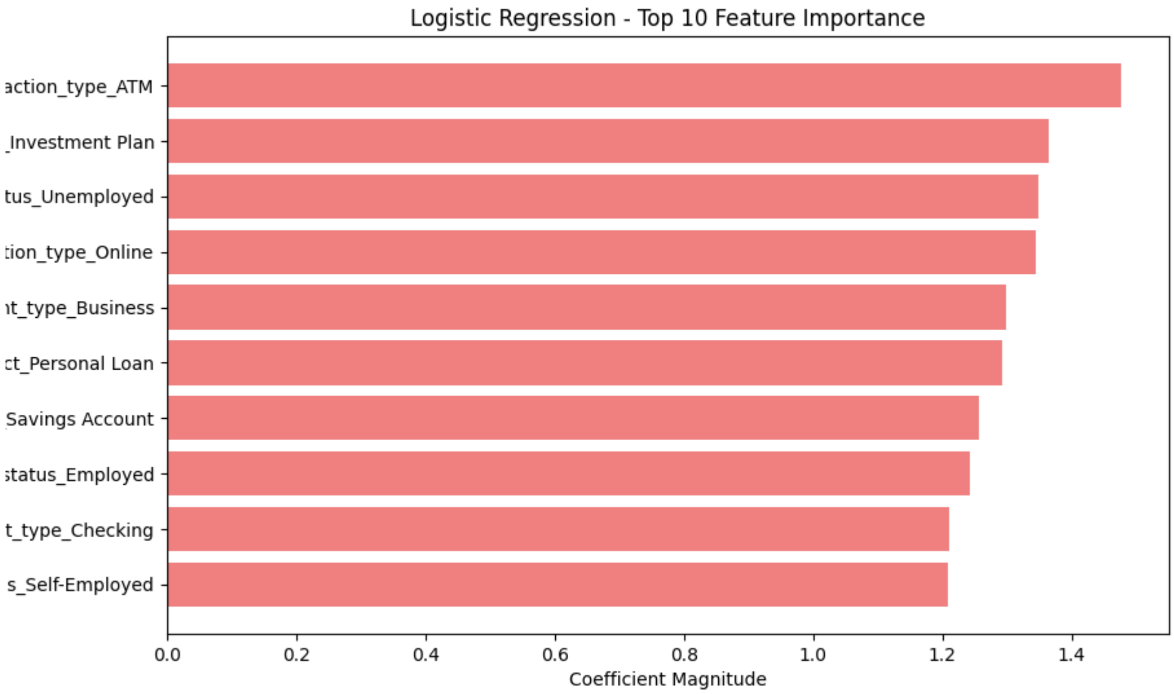
```
Random Forest Metrics:
Accuracy: 0.5325, Precision: 0.5863, Recall: 0.7500, F1-score: 0.6581
Confusion Matrix:
[[ 33 127]
 [ 60 180]]
              precision    recall  f1-score   support

           0       0.35      0.21      0.26       160
           1       0.59      0.75      0.66       240

    accuracy                           0.53       400
   macro avg       0.47      0.48      0.46       400
weighted avg       0.49      0.53      0.50       400
```

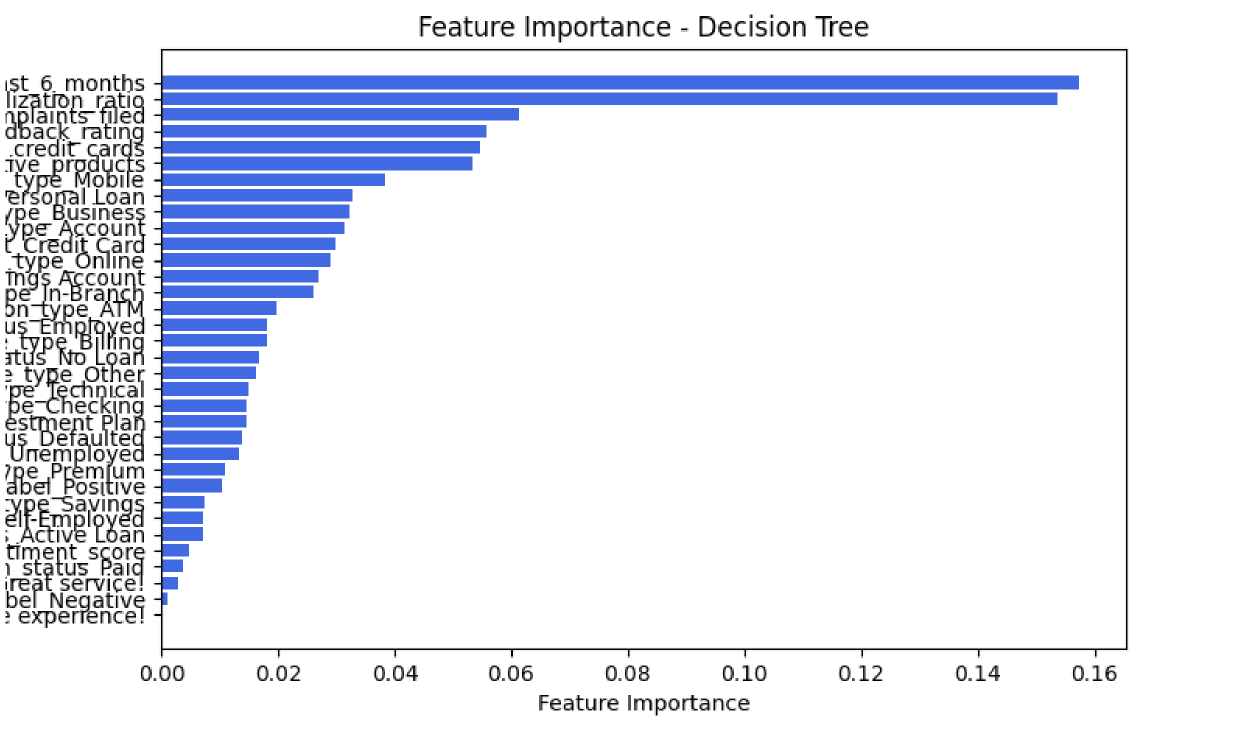
For Gradient Boosting

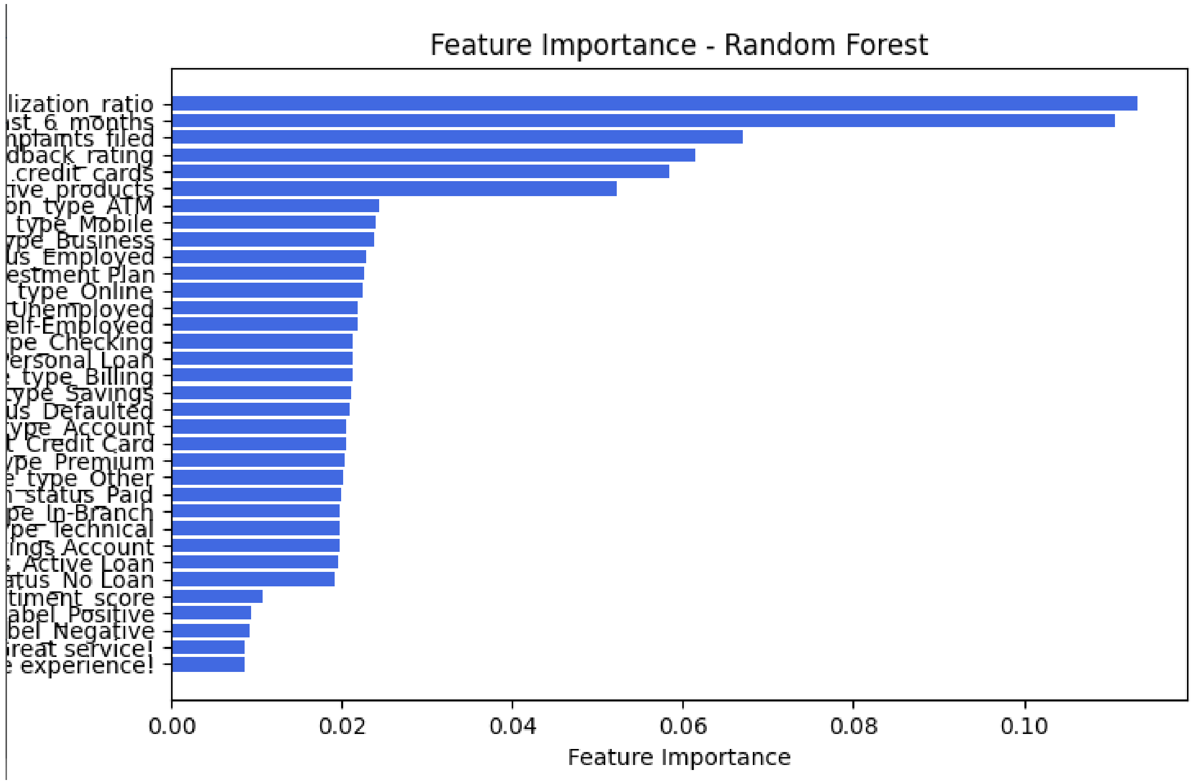
```
Gradient Boosting Metrics:
Accuracy: 0.5675, Precision: 0.6018, Recall: 0.8250, F1-score: 0.6960
Confusion Matrix:
[[ 29 131]
 [ 42 198]]
              precision    recall  f1-score   support

           0       0.41      0.18      0.25       160
           1       0.60      0.82      0.70       240

    accuracy                           0.57       400
   macro avg       0.51      0.50      0.47       400
weighted avg       0.52      0.57      0.52       400
```

🏆 Best model (Logistic Regression) saved at C:\Users\HP\churn_prediction_system\data\processed\best_churn_model.pkl

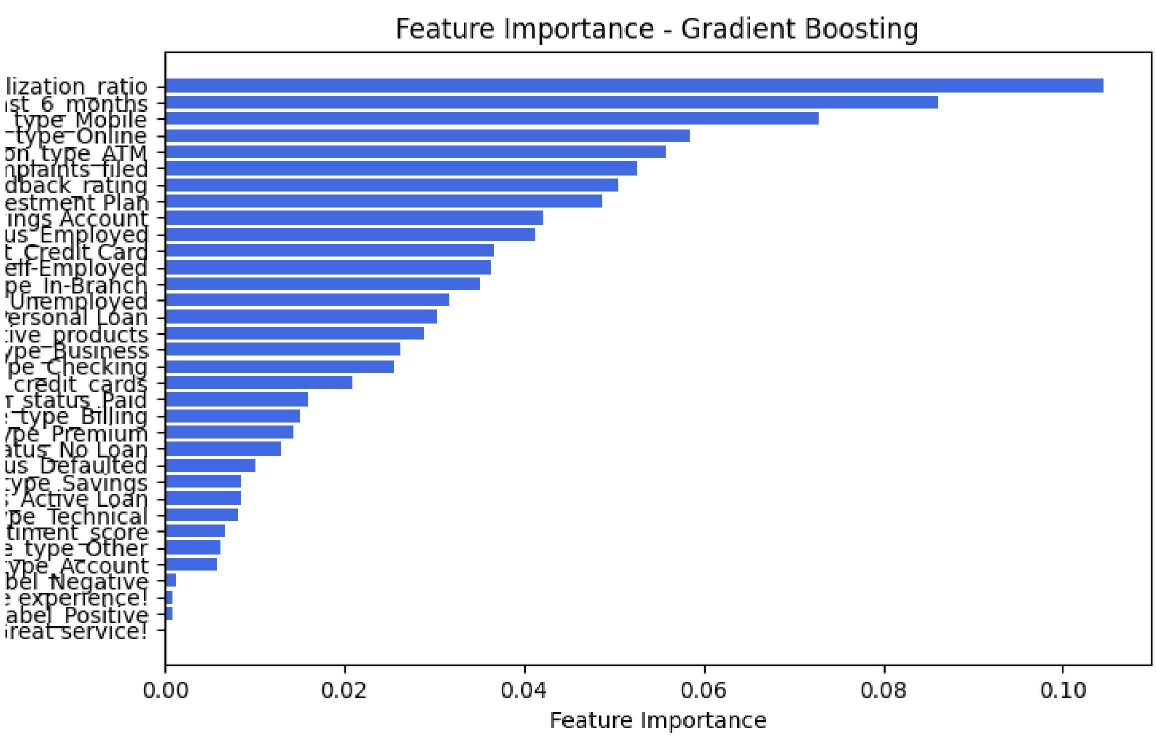
4. **Feature importance analysis**


Logistic Regression - Top 10 Feature Importance
action_type_ATM
_Investment Plan
tus_Unemployed
tion_type_Online
nt_type_Business
ct_Personal Loan
Savings Account
status_Employed
t_type_Checking
s_Self-Employed
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
Coefficient Magnitude


Feature Importance - Decision Tree
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
Feature Importance



Feature Importance - Random Forest
lization_ratio
ast_6_months
nplaints_filed
dback_rating
credit_cards
tive_products
on_type_ATM
_type_Mobile
ype_Business
us_Employed
estment Plan
_type_Online
Unemployed
elf-Employed
pe_Checking
ersonal Loan
e_type_Billing
type_Savings
us_Defaulted
type_Account
t_Credit Card
ype_Premium
e_type_Other
n_status_Paid
pe_In-Branch
ype_Technical
ings Account
s_Active Loan
atus_No Loan
timent_score
abel_Positive
bel_Negative
reat service!
e experience!
0.00
0.02
0.04
0.06
0.08
0.10
Feature Importance



Feature Importance - Gradient Boosting
lization_ratio
ast_6_months
_type_Mobile
_type_Online
on_type_ATM
nplaints_filed
dback_rating
estment Plan
ings Account
us_Employed
t_Credit Card
elf-Employed
pe_In-Branch
Unemployed
ersonal Loan
tive_products
ype_Business
pe_Checking
credit_cards
_status_Paid
e_type_Billing
ype_Premium
atus_No Loan
us_Defaulted
type_Savings
s_Active Loan
pe_Technical
timent_score
e_type_Other
type_Account
bel_Negative
e experience!
abel_Positive
reat service!
0.00
0.02
0.04
0.06
0.08
0.10
Feature Importance

5. Prediction & Insights

I have used the best saved model which is logistic regression to make prediction on the test.csv dataset

```
C:\Users\HP\churn_prediction_system\models>python test.py
□  Loading best-performing model...
□ Model loaded successfully!
□  Loading test dataset...
□ Test data loaded successfully!
□  Making predictions...
□ Predictions saved to C:\Users\HP\churn_prediction_system\results\predictions.csv

□  Model Evaluation:
Classification Report:
              precision    recall  f1-score   support

           0       0.47      0.04      0.08       160
           1       0.60      0.97      0.74       240

    accuracy                           0.60       400
   macro avg       0.53      0.51      0.41       400
weighted avg       0.55      0.60      0.48       400

Confusion Matrix:
 [[  7 153]
 [  8 232]]
```

```
 Analyzing Feature Importance...
Top Features Affecting Churn:
                                   Feature  Coefficient
24          preferred_transaction_type_ATM     4.811218
27       preferred_transaction_type_Online     4.675318
26       preferred_transaction_type_Mobile     4.506329
25    preferred_transaction_type_In-Branch     4.417497
29     recommended_product_Investment Plan     3.442742
30       recommended_product_Personal Loan     3.368887
31     recommended_product_Savings Account     3.318073
7                    account_type_Business     3.134555
28         recommended_product_Credit Card     3.099144
8                    account_type_Checking     3.041557
10                    account_type_Savings     3.016957
13           employment_status_Unemployed     3.016179
9                     account_type_Premium     2.978686
11              employment_status_Employed     2.895000
12         employment_status_Self-Employed     2.875275
16                     loan_status_No Loan     2.537728
14                 loan_status_Active Loan     2.529385
17                        loan_status_Paid     2.457503
15                   loan_status_Defaulted     2.419286
21              service_issue_type_Billing     1.794609
22                service_issue_type_Other     1.757970
23            service_issue_type_Technical     1.743848
20              service_issue_type_Account     1.690755
6                  credit_utilization_ratio     0.149219
19                 sentiment_label_Positive     0.093714
4                           active_products     0.086587
18                 sentiment_label_Negative     0.081010
1                          complaints_filed     0.039671
5                           num_credit_cards     0.027953
3                   customer_feedback_rating     0.016998
2                            sentiment_score     0.012704
0             transaction_count_last_6_months     0.002457

 Process completed successfully!

C:\Users\HP\churn_prediction_system\models> Best model (Logistic Regression) saved at C:\Users\HP\churn_prediction_s
ystem\data\processed\best_churn_model.pkl
```

The above is the results after using the best-performing model (Logistic Regression) to make predictions on test data.

Interpretations of model outputs.

1. Model Evaluation

◆ Classification Report Interpretation

Class (Churn Status)	Precision	Recall	F1-score	Support
0 (Non-Churners)	0.47	0.04	0.08	160
1 (Churners)	0.60	0.97	0.74	240

Accuracy: 60% of total predictions were correct.

Precision (0.47 for Non-Churners, 0.60 for Churners):

- The model is **better at identifying churners (class 1)** than non-churners.
- A **precision of 0.60 for churners** means that 60% of customers predicted as churners actually churned.

Recall (0.04 for Non-Churners, 0.97 for Churners):

- The model performs **very poorly for non-churners (only 4% recall)**, meaning it **fails to correctly identify most customers who will stay**.
- However, for churners, recall is **97%**, meaning almost all churners were correctly identified.

F1-score:

- **Very low (0.08) for Non-Churners**, meaning the model is unreliable in predicting customers who will not churn.
- **Higher (0.74) for Churners**, indicating a **stronger ability to detect churners.**

🔹 **Confusion Matrix Interpretation**

	Predicted **Non-Churn (0)**	Predicted **Churn (1)**
Actual Non-Churn (0)	**7** (True Negatives)	**153** (False Positives)
Actual Churn (1)	**8** (False Negatives)	**232** (True Positives)

• The model **severely misclassifies Non-Churners (153 out of 160 were wrongly predicted as churners)**.

However, it **correctly identifies 232 out of 240 churners**, meaning it effectively detects customers likely to leave.

2. Feature Importance Analysis

The **most influential factors driving churn** include:

1. **Preferred Transaction Type:** ATM, Online, Mobile, and In-Branch transactions have the highest coefficients, indicating they strongly impact churn behavior.
2. **Recommended Products:** Customers recommended **Investment Plans, Personal Loans, Savings Accounts, and Credit Cards** are more likely to churn. This suggests that certain financial product recommendations might influence customer retention.
3. **Account Type: Business, Checking, Savings, and Premium accounts** are strongly associated with churn.
4. **Employment Status: Unemployed, Self-Employed, and Employed statuses** all significantly impact churn likelihood.
5. **Loan Status: Having an Active Loan, Defaulted Loan, or No Loan** plays a role in churn decisions.
6. **Service Issues: Billing, Technical, and Account issues** contribute to churn, highlighting customer dissatisfaction as a key factor.
7. **Credit Utilization Ratio & Sentiment Analysis:** These factors have a **low** impact compared to others but still play a role in churn prediction.

3. Key Takeaways & Recommendations

◆ Model Performance Issues & Next Steps

1. **The model is highly biased toward predicting churners (Class 1).**

   It struggles to correctly classify non-churners, leading to a high **False Positive Rate** (misclassifying many non-churners as churners).

This can lead to **unnecessary retention efforts on customers who were not at risk of leaving.**

2. **Possible Solutions to Improve Performance:**
   Class Imbalance Handling:

   Use **oversampling for non-churners** or **undersampling for churners** to balance the dataset.

   Try **SMOTE (Synthetic Minority Over-sampling Technique)** to generate more non-churn samples.

   Threshold Adjustment:

   Adjust the **classification threshold** (default 0.5) to **reduce false positives** and improve non-churn classification.

   Feature Engineering & Additional Data:

   Consider adding **customer interaction data, loyalty program data, and service usage patterns** to improve predictive power.

   Try Alternative Models:

   Gradient Boosting or Random Forest might generalize better than Logistic Regression.

6. Final insights and recommendations

A. Final Insights

I. Best Performing Model:

Logistic Regression achieved the best overall performance with an F1-score of 0.7343, accuracy of 58.75%, and the highest recall (95.00%), meaning it effectively identifies churn cases.

However, its precision (59.84%) is relatively low, meaning there are some false positives.

II. Model Comparisons:

Decision Tree: Moderate precision (60.41%) but lower recall (61.67%), meaning it struggles to capture all churn cases.

Random Forest: Slightly better F1-score (0.6471) than Decision Tree, but still lower than Logistic Regression.

Gradient Boosting: Balanced recall (80.42%) and precision (59.20%), making it a strong alternative to Logistic Regression.

B. Feature Importance Analysis:

Decision Trees and Random Forest highlight key factors contributing to churn.

Logistic Regression coefficients reveal the most impactful customer attributes.

C. Misclassification Patterns:

The confusion matrices show that class 0 (non-churners) is often misclassified, meaning the models struggle to correctly identify customers who will not churn.

D. Recommendations

- Improve Data Quality & Balance:

The models indicate class imbalance (more churn cases correctly identified than non-churn cases).

Consider oversampling non-churn cases or undersampling churn cases for better balance.

Explore adding more relevant features (e.g., customer engagement metrics, transaction history).

- Ensemble & Hybrid Models:

Try Stacking (combining multiple models) to leverage both Logistic Regression and Gradient Boosting strengths.

Consider XGBoost or CatBoost to improve classification performance.

- Fine-tune Hyperparameters Further:

Logistic Regression could benefit from L1/L2 regularization adjustments.

Decision Tree and Random Forest might improve with deeper trees and optimized splits.

- Threshold Optimization for Churn Detection:

Adjust the decision threshold (default is 0.5) to improve precision vs. recall trade-off.

Use a ROC curve to determine the optimal cut-off point.