

**P101/1740G/21**

**KAHENI PETER**

**Predictive analytics in Business Intelligence**

**Technical Assignment II**

**Predictive Analytics Using Machine Learning**

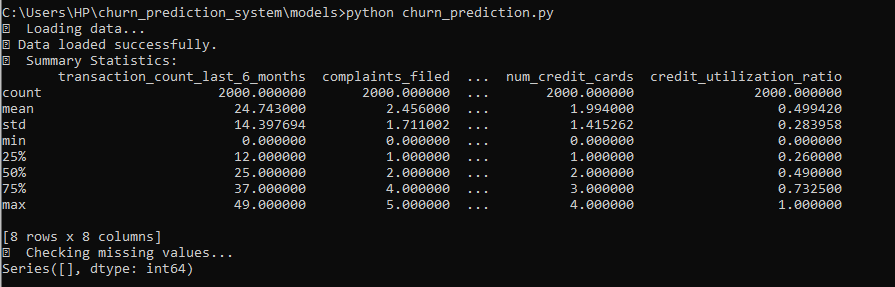
**Dataset: bank\_customer\_churn.csv**

1. **Dataset description**

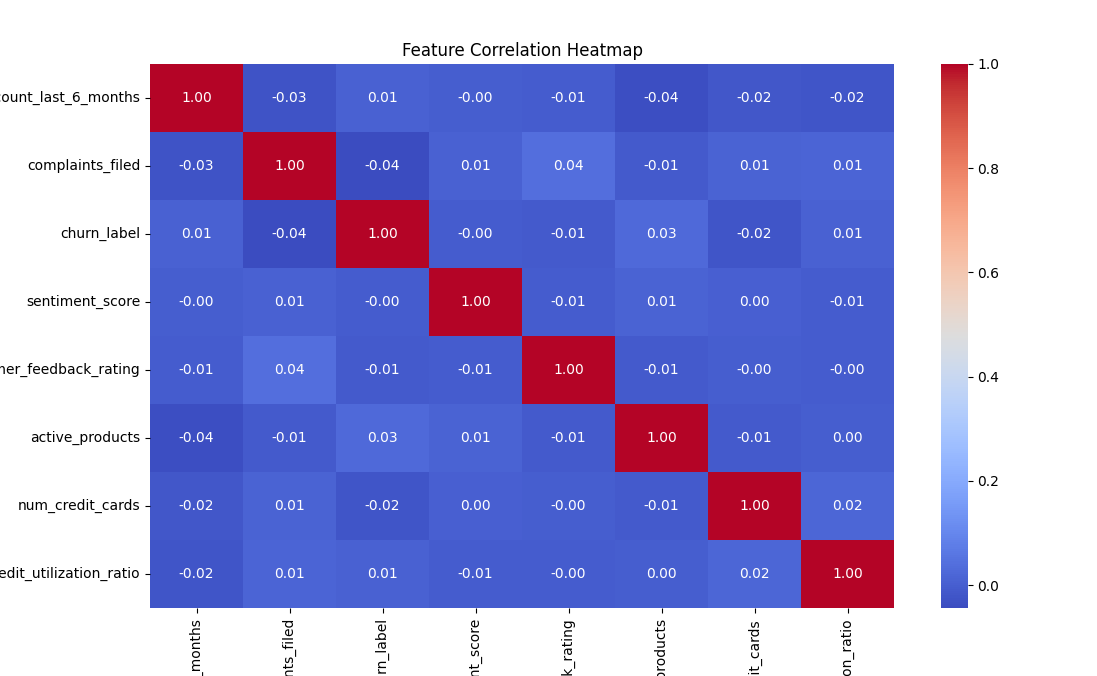
The dataset contain customer-related financial and behavioral data to predict churn (whether a customer leaves or stays with a bank).

|  |  |
| --- | --- |
| 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). |

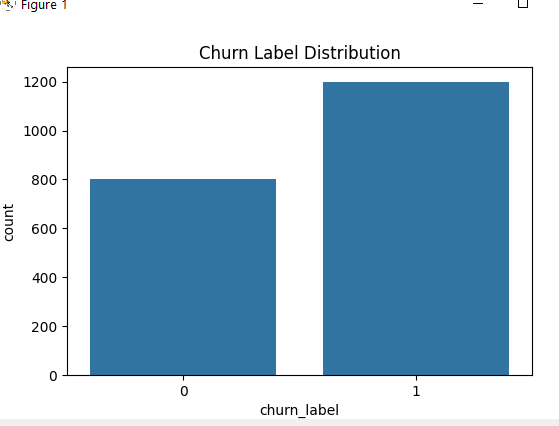
1. **Key findings from EDA**
2. **SUMMARY STATISTICS.**

****

1. **CORRELATION ANALYSIS.**

****

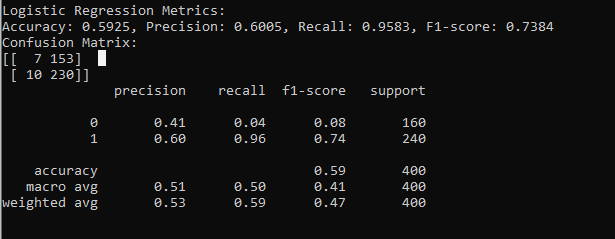
1. **CHURN LABEL**

****

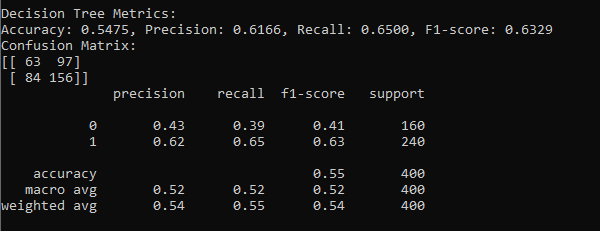
1. **Model selection and evaluation results**

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

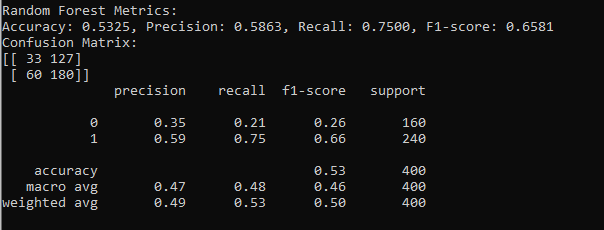
**For Logistic**



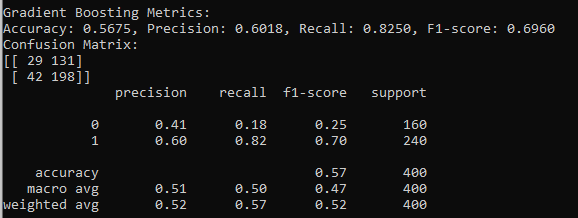
**For Decision Tree**



**For Random Forest**

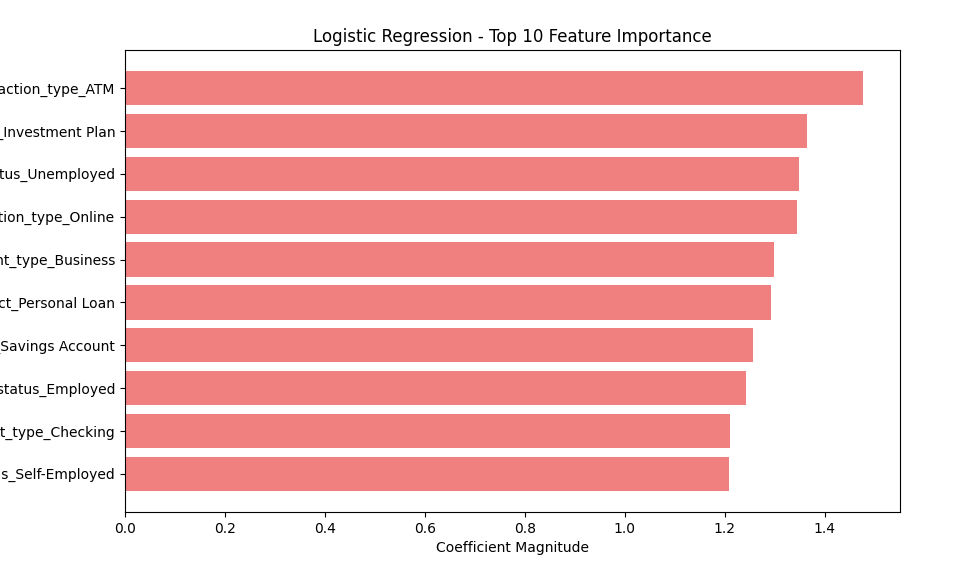


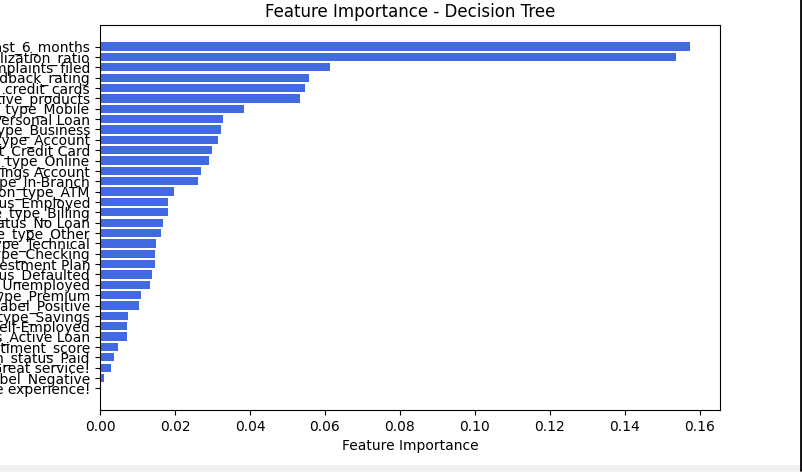
**For Gradient Boosting**

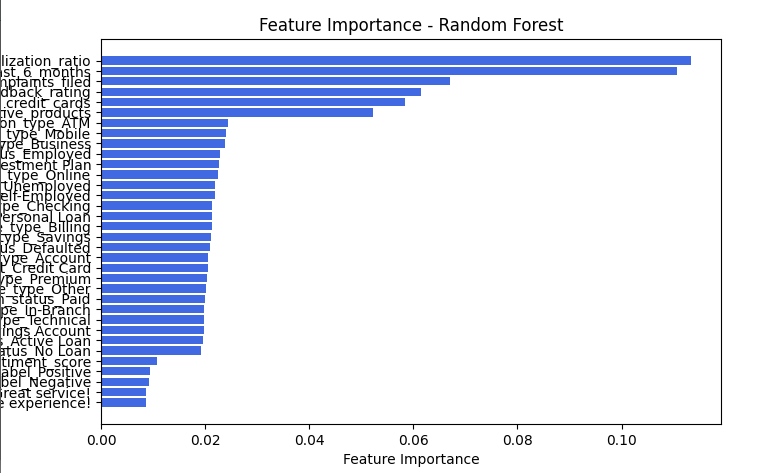


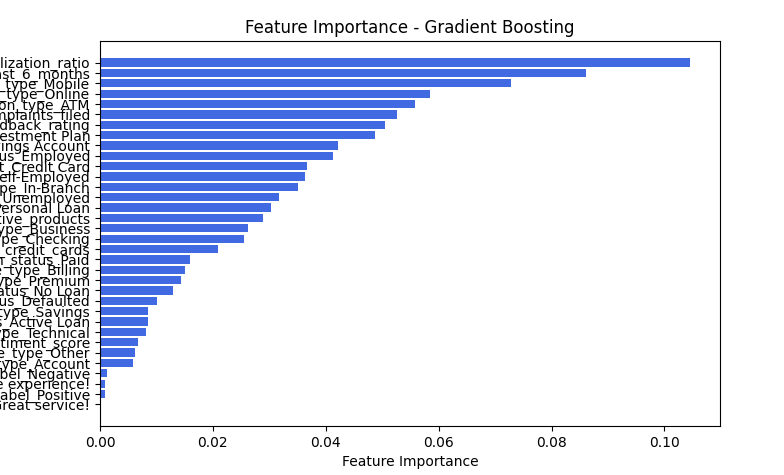
🏆 **Best model (Logistic Regression) saved at C:\Users\HP\churn\_prediction\_system\data\processed\best\_churn\_model.pkl**

1. **Feature importance analysis**

****

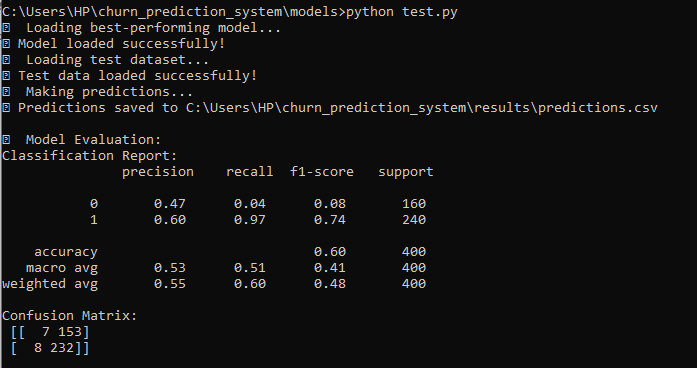
****

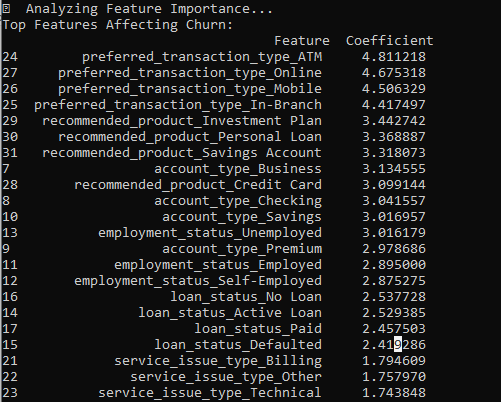
****

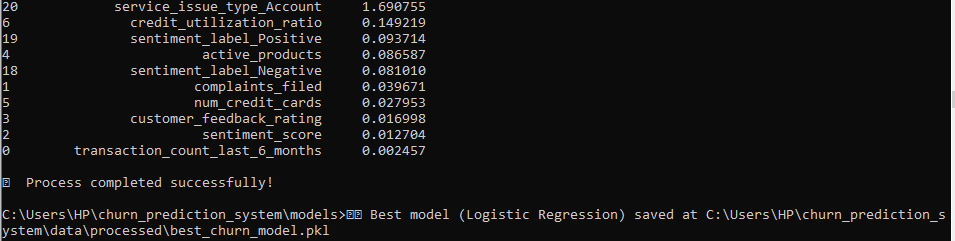
****

1. **Prediction & Insights**

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







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.**

1. **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.

1. **Final insights and recommendations**
2. **Final Insights**
3. **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.

1. **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.

1. **Feature Importance Analysis:**

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

Logistic Regression coefficients reveal the most impactful customer attributes.

1. **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.

1. **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.