5CS037

Concepts and Technologies of AI

Classification Analysis Report

Name: Krishal Maharjan

Group: L5CG19

Canvas ID:2408955

Lecturer: Bibek Khanal

Tutor: Ronit Shrestha

Module Leader: Siman Giri

Submission Date: 11/2/2025

Abstract:

A predictive analysis of Bank Churn data will determine bank customer retention likelihood. Acquiring new customers costs banks more than keeping existing ones exactly why banks face high customer churn rates. Through implementation of a predictive system banks gain the ability to detect customers at risk and respond with retention measures. The project progress moved through five sequential steps which began with data loading and analysis followed by data cleaning and ended in model establishment and training and model evaluation. Logistic Regression proved best because it provides an efficient and straightforward solution for binary classification problems. The model received evaluation through multiple metrics including accuracy alongside precision and recall and F1-score

Contents

Dataset Name: Bank Churn dataset

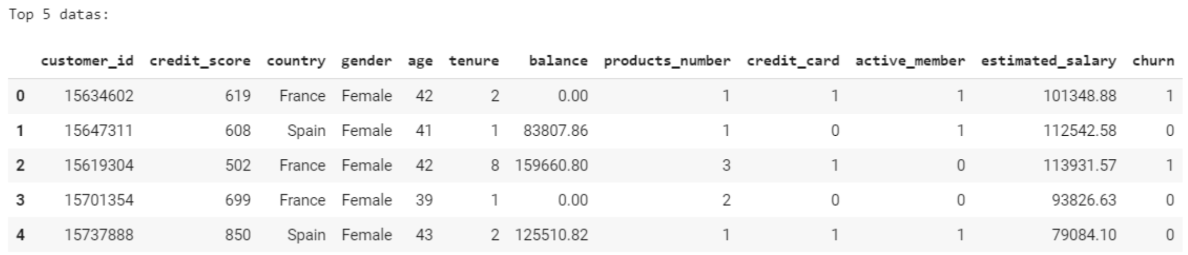
Introduction

Customer churn refers to the situation when bank customers choose to leave their banking relationship behind. Banks face this problem frequently since acquiring new customers costs the institution more than it does to maintain their existing customer base. A smart predictive system receives focus in this project for detecting bank customers who plan to leave. Early identification enables banks to develop strategies which keep their customers from leaving.

The system development process included five vital stages starting with data loading followed by data exploration then moving to data preparation after which model building occurred before conducting model evaluation.

Data Loading and Exploration

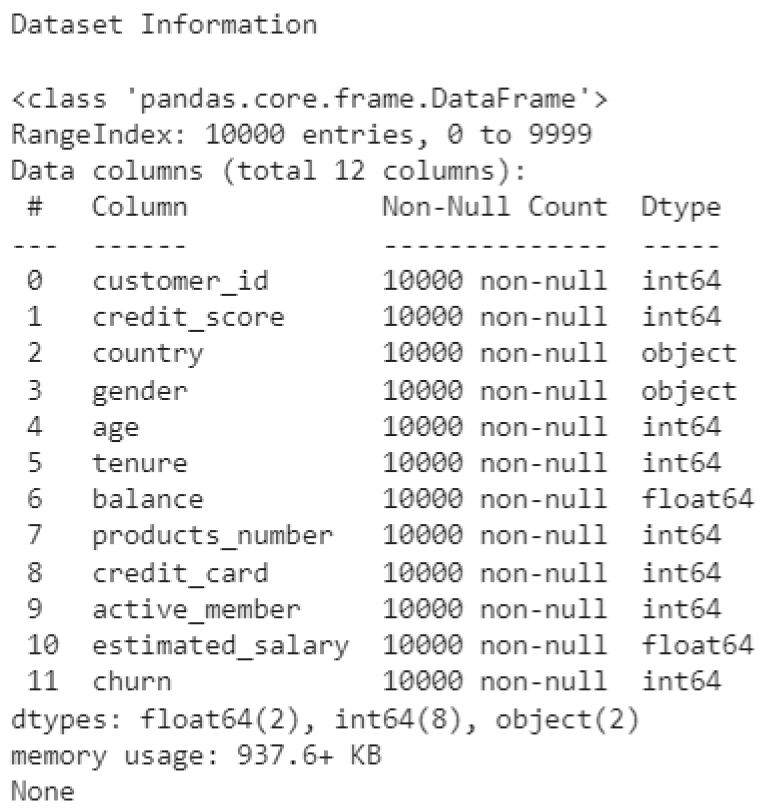
The analysis data was obtained from Kaggle before being moved to a personal drive for further inspection. Bank customer information with their corresponding attributes about age, credit score, and country is found in the available dataset. The initial five records of the dataset show the following information:



Top 5 datas:

	customer_id	credit_score	country	gender	age	tenure	balance	products_number	credit_card	active_member	estimated_salary	churn
0	15634602	619	France	Female	42	2	0.00	1	1	1	101348.88	1
1	15647311	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
2	15619304	502	France	Female	42	8	159660.80	3	1	0	113931.57	1
3	15701354	699	France	Female	39	1	0.00	2	0	0	93826.63	0
4	15737888	850	Spain	Female	43	2	125510.82	1	1	1	79084.10	0

The data collection contains three types of data features including numeric values, categorical values and booleans that provide various exploration options.



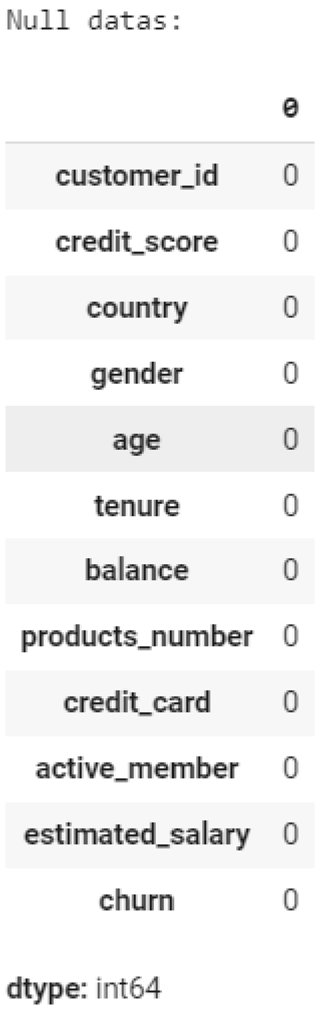
```
Dataset Information

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 12 columns):
 #   Column            Non-Null Count  Dtype  
---  ------            --------------  -----  
 0   customer_id       10000 non-null  int64  
 1   credit_score      10000 non-null  int64  
 2   country           10000 non-null  object 
 3   gender            10000 non-null  object 
 4   age               10000 non-null  int64  
 5   tenure            10000 non-null  int64  
 6   balance           10000 non-null  float64
 7   products_number   10000 non-null  int64  
 8   credit_card       10000 non-null  int64  
 9   active_member     10000 non-null  int64  
 10  estimated_salary  10000 non-null  float64
 11  churn             10000 non-null  int64  
dtypes: float64(2), int64(8), object(2)
memory usage: 937.6+ KB
None
```

Data Preprocessing

The dataset included no missing nor duplicate data points.

A model construction required selection of its feature variables along with target variables.

Standard scaling occurred through the `StandardScaler` function to normalize numberical values across different ranges.



Null datas:

	0
customer_id	0
credit_score	0
country	0
gender	0
age	0
tenure	0
balance	0
products_number	0
credit_card	0
active_member	0
estimated_salary	0
churn	0

dtype: int64

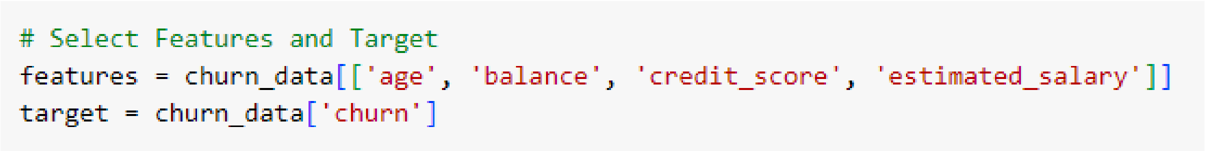
Model Building and Training

1. To evaluate model learning an 80/20 split was applied for data training and data testing purposes.

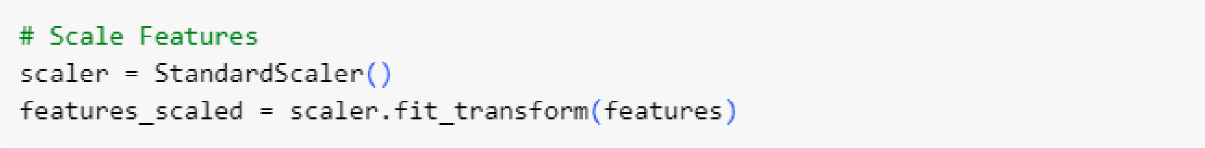
2. Logistic Regression received selection as the best model because it maintains simplicity and high efficiency when dealing with binary classification tasks.

3. The RFE algorithm automatically selected 10 most influential features from the initial pool through recursive elimination.

4. The Logistic Regression model received its hyperparameter values through GridSearchCV optimization methods.



```
# Select Features and Target
features = churn_data[['age', 'balance', 'credit_score', 'estimated_salary']]
target = churn_data['churn']
```



```
# Scale Features
scaler = StandardScaler()
features_scaled = scaler.fit_transform(features)
```

Model Evaluation

This model achieved evaluation through analysis based on accuracy and precision and recall and the F1-Score and confusion matrix metrics.

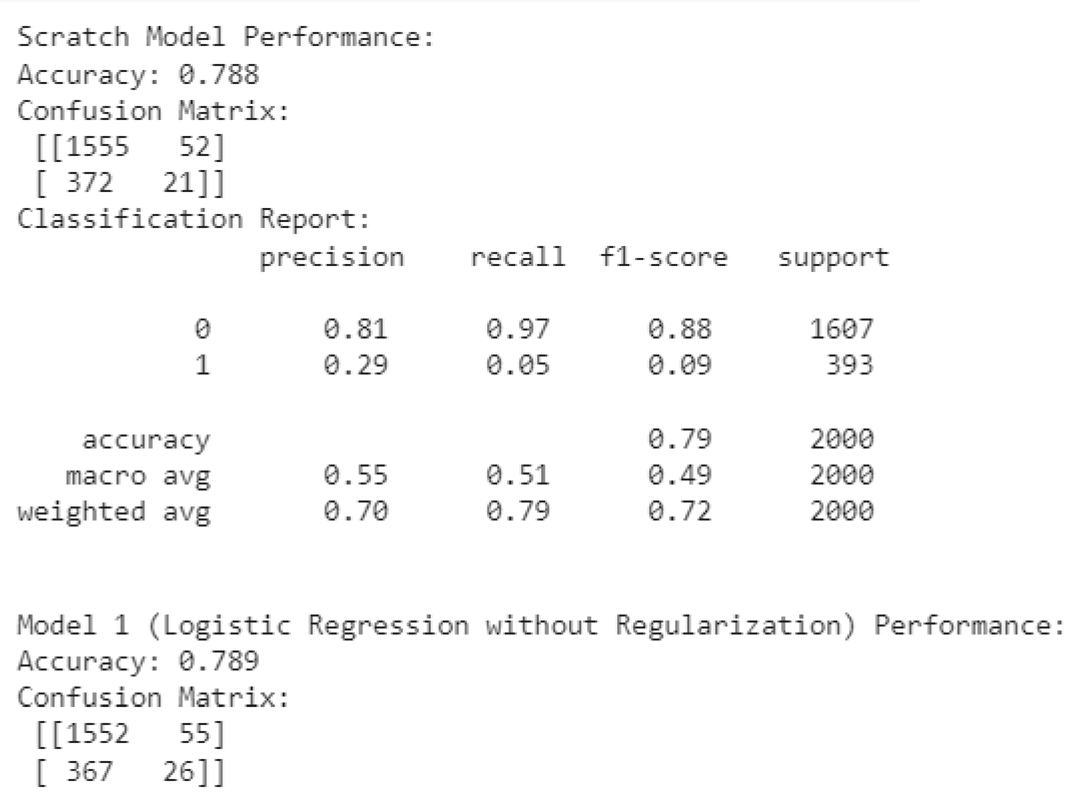
- Accuracy: Proportion of correct predictions out of all predictions.

- Precision: Proportion of correct positive predictions.

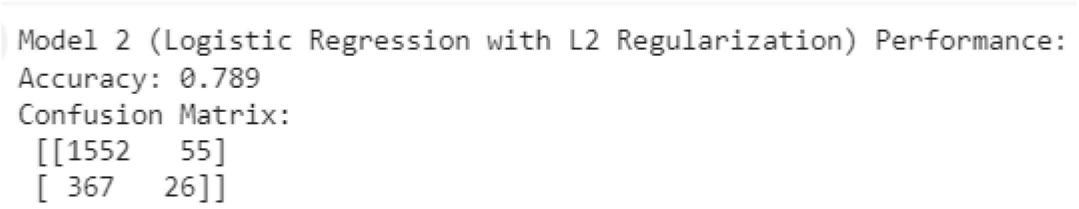
Recall stands for the ratio of correctly detected actual positive results.

F1-Score represents the harmonic average between recall and precision values.

- Confusion Matrix: A matrix representation of true positives, true negatives, false positives, and false negatives.

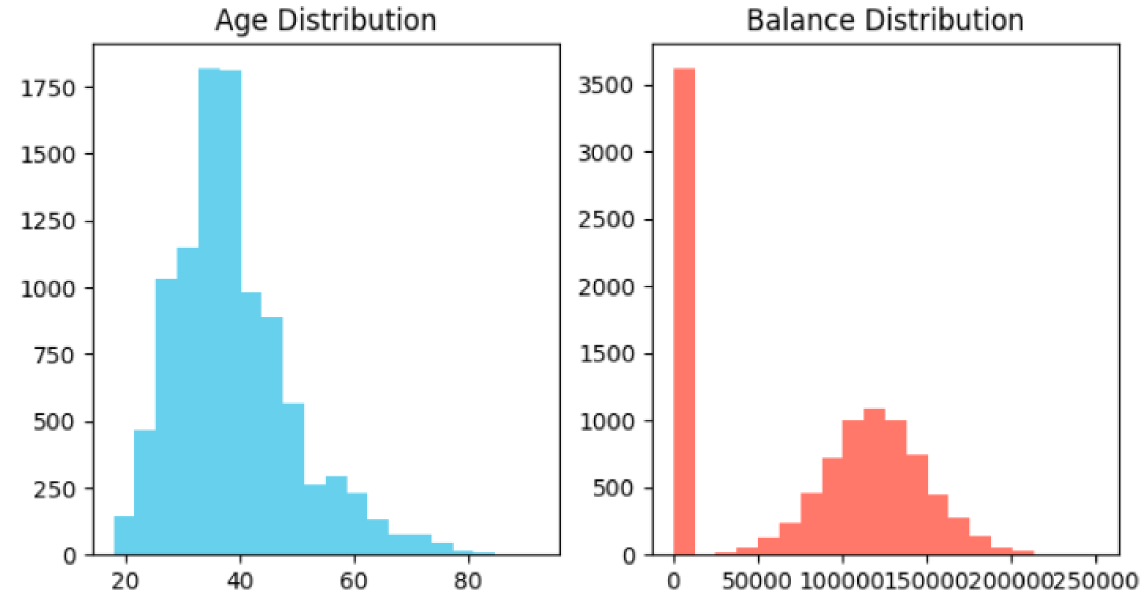


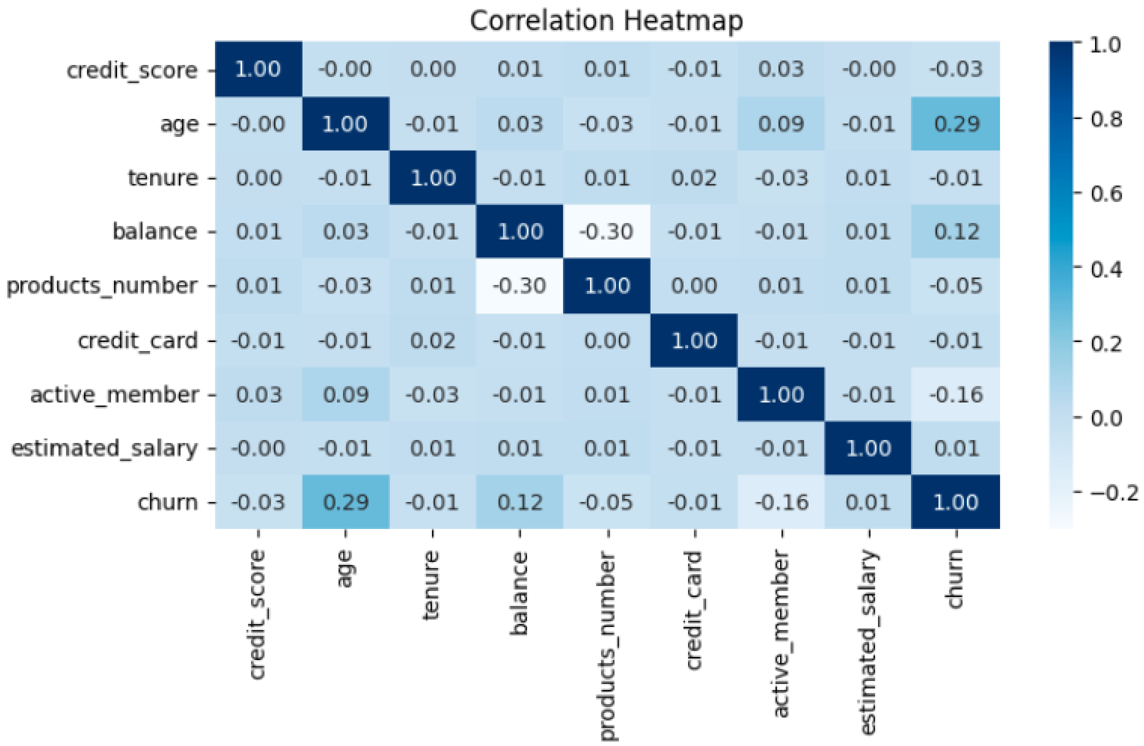
```
Scratch Model Performance:
Accuracy: 0.788
Confusion Matrix:
 [[1555   52]
 [ 372   21]]
Classification Report:
               precision    recall  f1-score   support

           0       0.81      0.97      0.88      1607
           1       0.29      0.05      0.09       393

    accuracy                           0.79      2000
   macro avg       0.55      0.51      0.49      2000
weighted avg       0.70      0.79      0.72      2000


Model 1 (Logistic Regression without Regularization) Performance:
Accuracy: 0.789
Confusion Matrix:
 [[1552   55]
 [ 367   26]]
```



```
Model 2 (Logistic Regression with L2 Regularization) Performance:
Accuracy: 0.789
Confusion Matrix:
 [[1552   55]
 [ 367   26]]
```

Visualization:





	credit_score	age	tenure	balance	products_number	credit_card	active_member	estimated_salary	churn
credit_score	1.00	-0.00	0.00	0.01	0.01	-0.01	0.03	-0.00	-0.03
age	-0.00	1.00	-0.01	0.03	-0.03	-0.01	0.09	-0.01	0.29
tenure	0.00	-0.01	1.00	-0.01	0.01	0.02	-0.03	0.01	-0.01
balance	0.01	0.03	-0.01	1.00	-0.30	-0.01	-0.01	0.01	0.12
products_number	0.01	-0.03	0.01	-0.30	1.00	0.00	0.01	0.01	-0.05
credit_card	-0.01	-0.01	0.02	-0.01	0.00	1.00	-0.01	-0.01	-0.01
active_member	0.03	0.09	-0.03	-0.01	0.01	-0.01	1.00	-0.01	-0.16
estimated_salary	-0.00	-0.01	0.01	0.01	0.01	-0.01	-0.01	1.00	0.01
churn	-0.03	0.29	-0.01	0.12	-0.05	-0.01	-0.16	0.01	1.00

Conclusion:

The predictive model for bank customer churn prediction reached satisfactory accuracy levels throughout development. Through this model banks obtain the capability to detect customers at risk so they can employ preventive actions that include special promotions or improved service quality to maintain customer loyalty.