DATA SCIENCE TALENT COMPETITION

THE BYTE SQUAD

CHATBOT DEMO

DATA PROCESSING

10%

NULL values

122/124

columns containing NULL

ENQUIRIES

contains negative/ decimal values

DATA PROCESSING

REPLACE

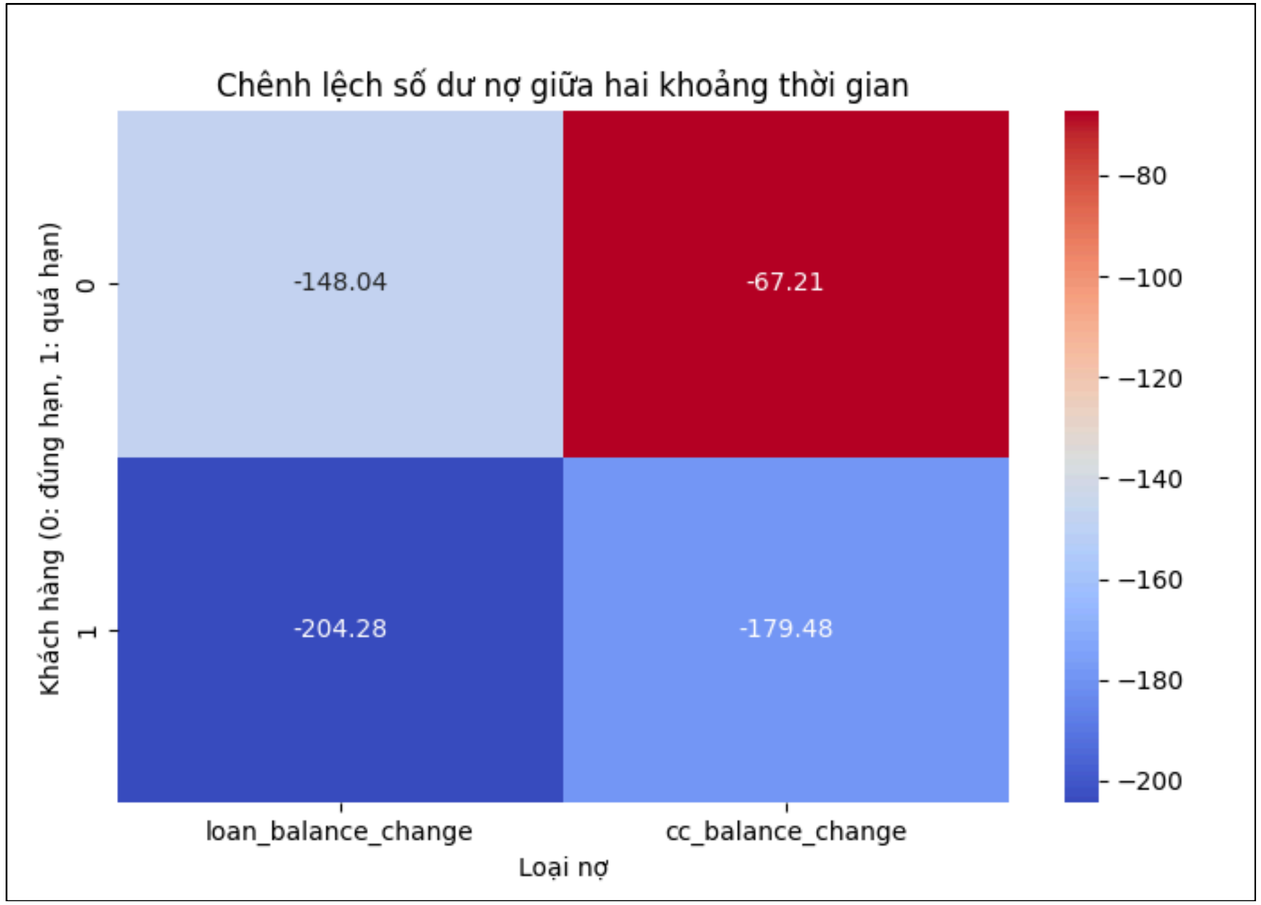
- Linear regression method
- Random Forest Regression using the scikit-learn library
- Replace with median value: INCREASING BAL
- Round positive decimal numbers and change negative numbers to the median.

REMOVE

It was observed that the data cannot be extrapolated, with only about 600 rows, accounting for around 3%, so removing these rows will not significantly affect the data.

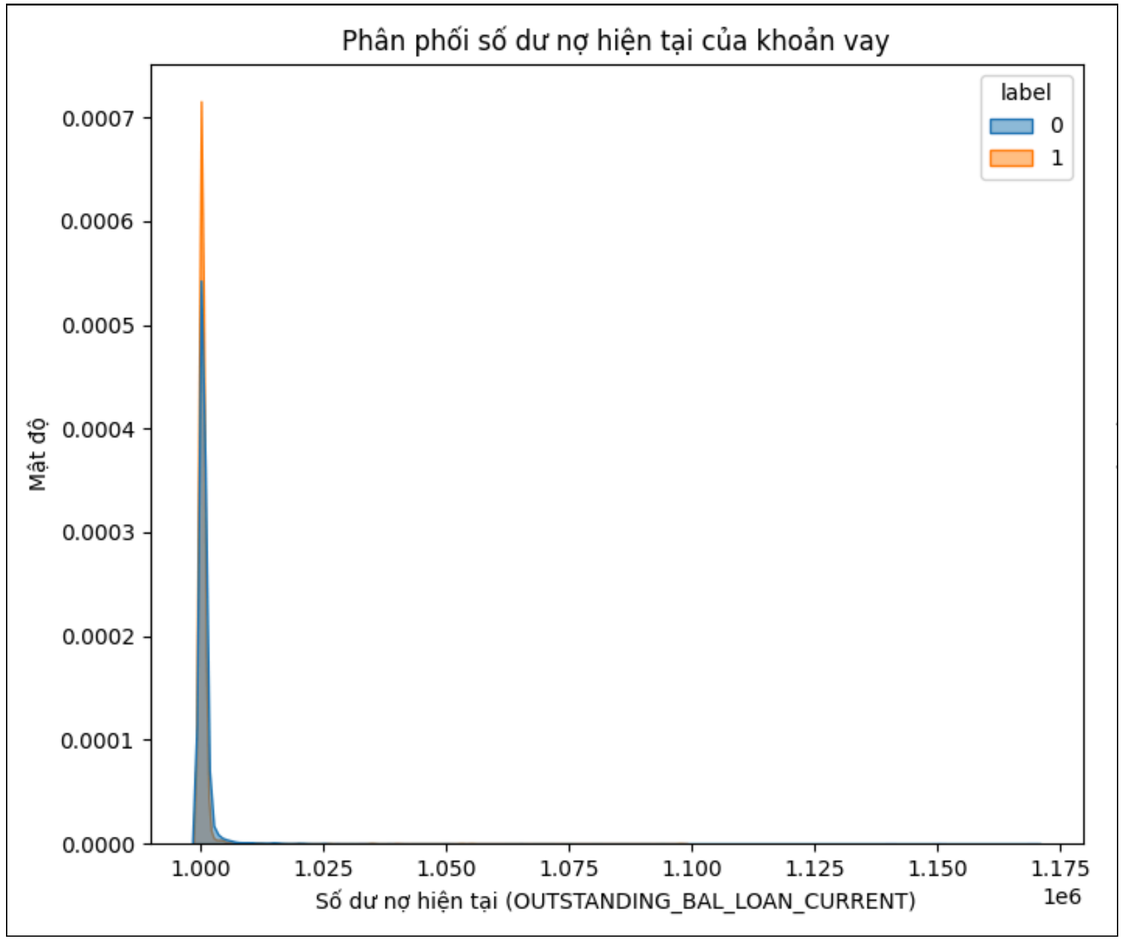
ANALYZE CURRENT OUTSTANDING BALANCE

- Customers who are overdue typically have a higher outstanding balance compared to those who make payments on time.
- This indicates a negative correlation between late payments and outstanding balance, highlighting the need for more effective debt management strategies.



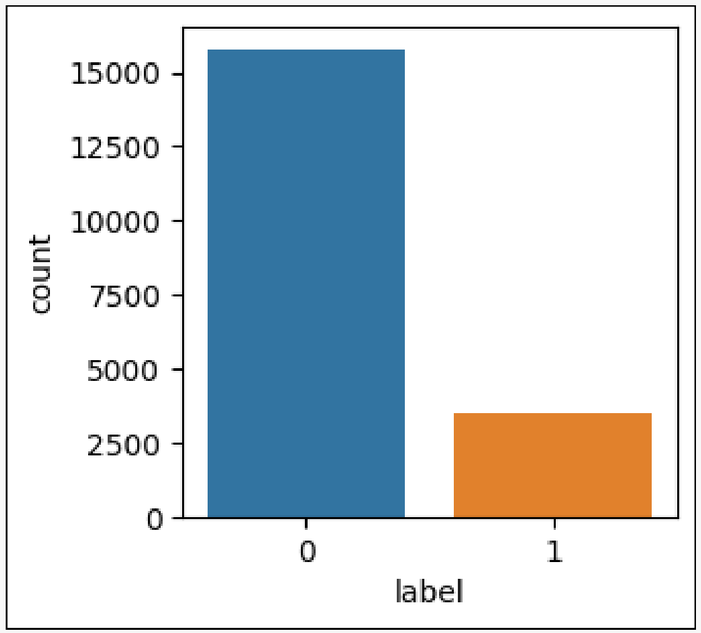
ANALYZE CURRENT OUTSTANDING BALANCE

- Label 1 (late payment) is concentrated at higher outstanding balances compared to label 0.
- This indicates that customers who are late on payments tend to have higher outstanding balances, reflecting that they may be experiencing short-term cash flow stress.



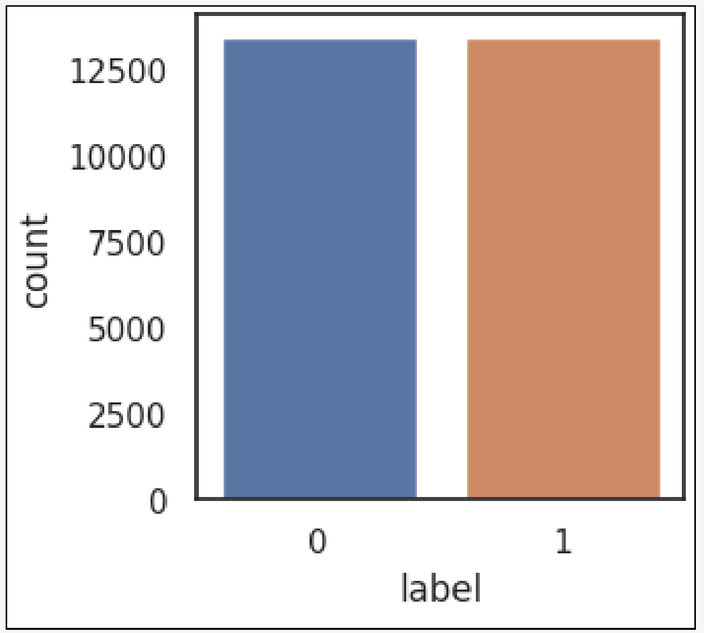
DATA IMBALANCE

- The data distribution is uneven, with **15,277** label 0 and **3,529** label 1.
- Label 1 makes up **only 18.76%** of the dataset.



SMOTE

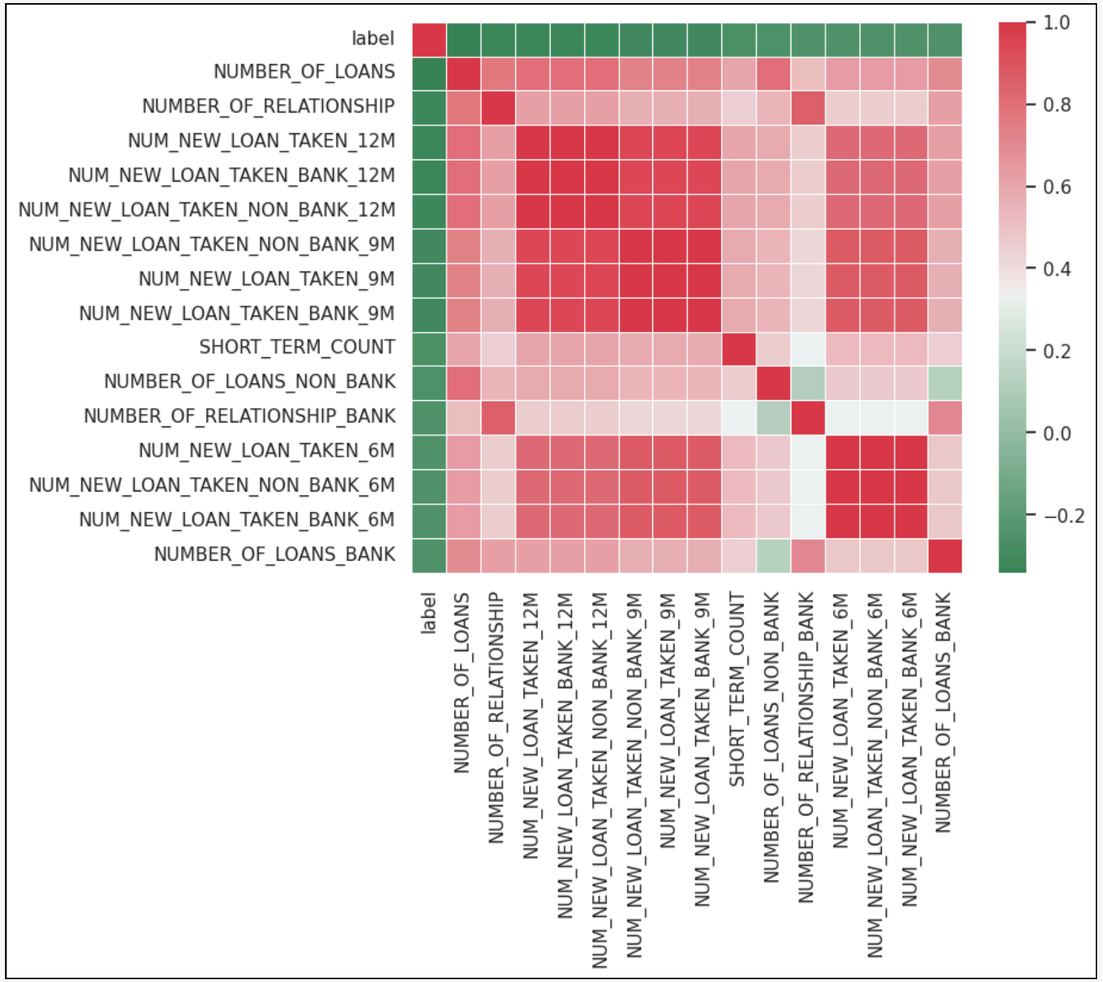
Undersampling



VARIABLE INPUT

- Delete the “**customer_id**“ column as it is not relevant to the customer’s ability and payment term.

- Select the **40 variables** with the highest correlation from the correlation matrix of the variables in the dataset.

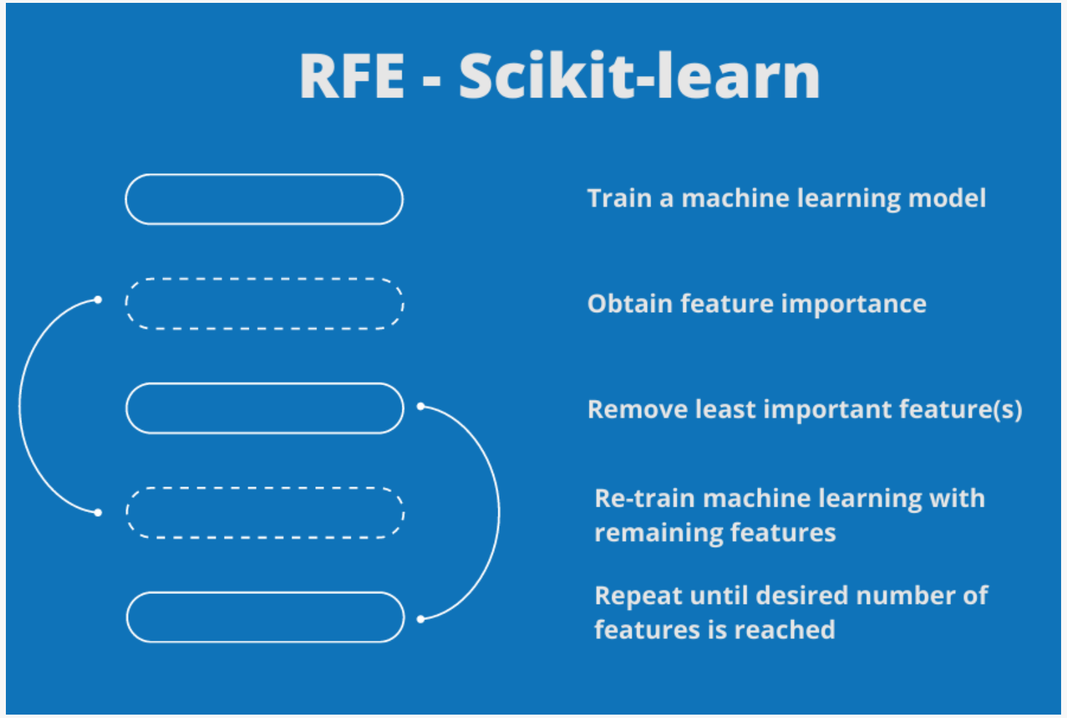


VARIABLE INPUT

- Use **RFE (Recursive Feature Elimination)** in combination with Random Forest.

- Train the model to select the **20** most relevant variables to the label.

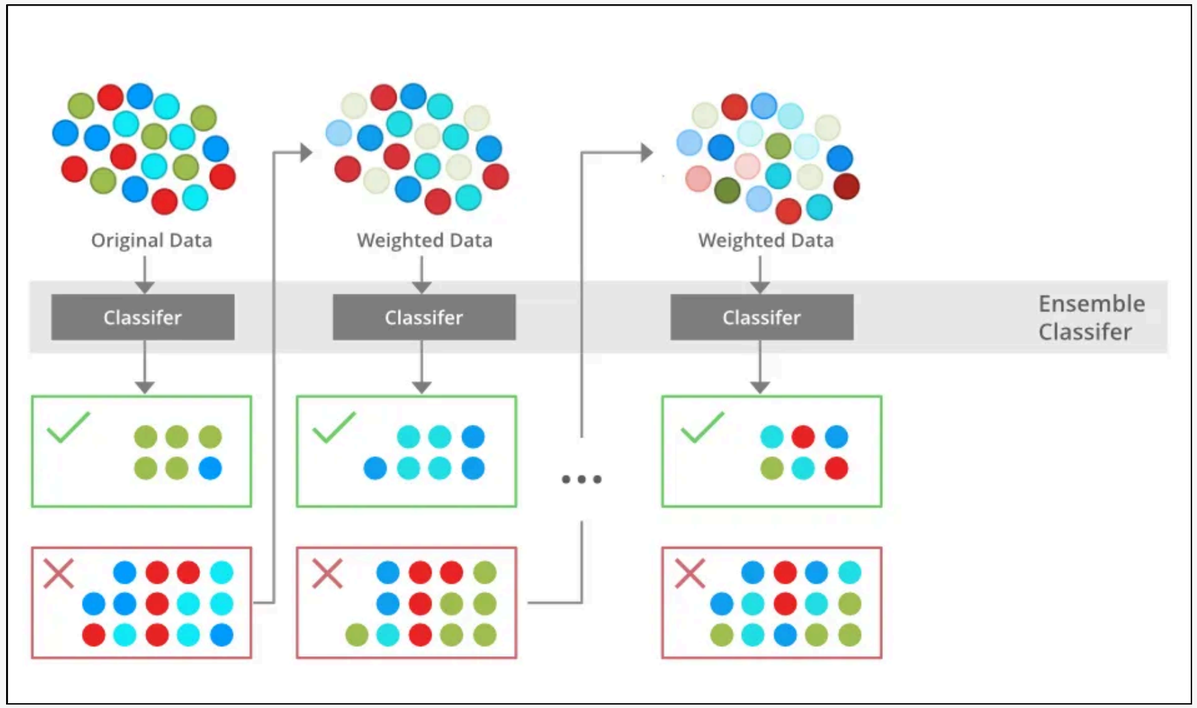
- The most influential variables on the label are:
  - Number of new loans
  - Financial relationship & number of credit cards
  - Number of short-term loans



MODEL RECOMMENDATION

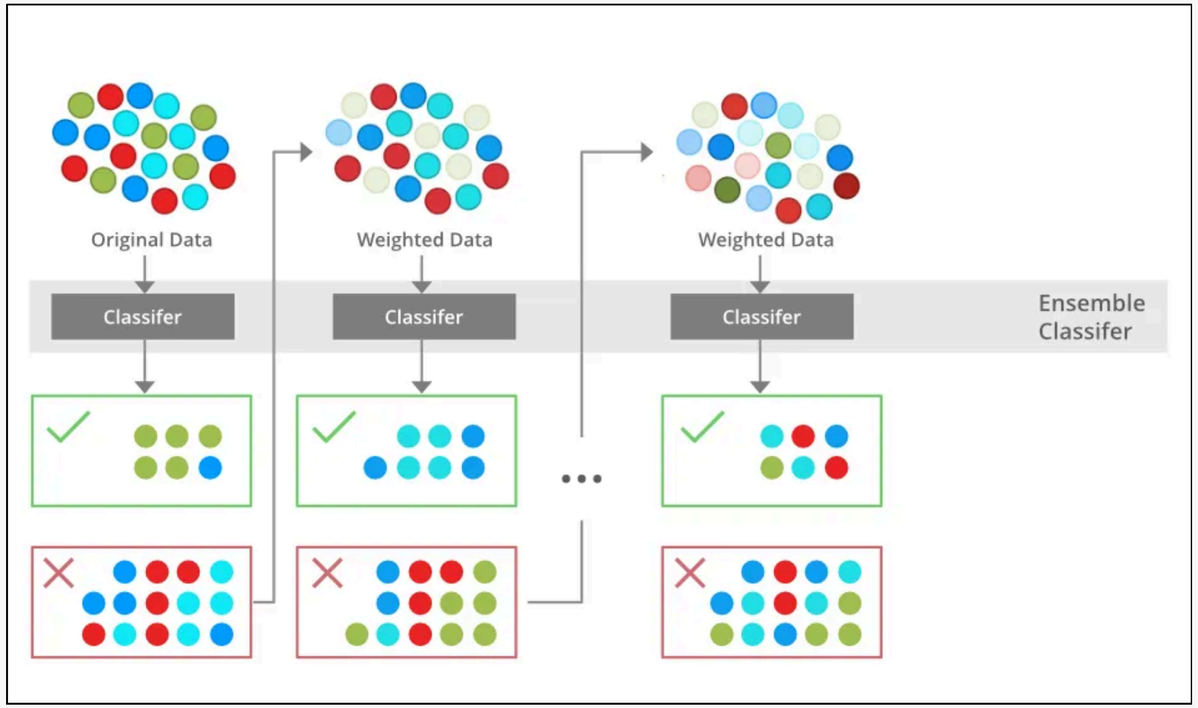
- Use **Decision Tree** combined with Gradient **Boosting techniques** to improve accuracy compared to traditional machine learning models (Random Forest, Decision Tree, Logistic Regression, etc.).

- **Strength:**
  - Works effectively with tabular data.
  - Has the ability to generalize imbalanced data.
  - Fast training and prediction speed.



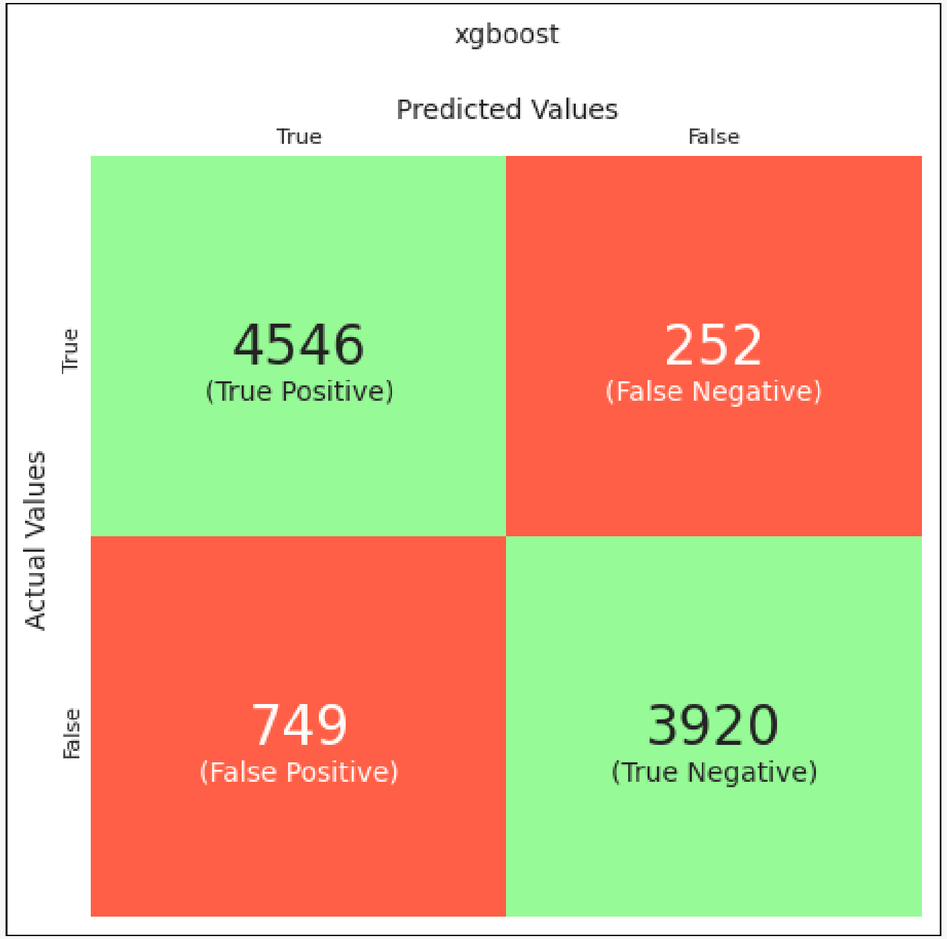
MODEL RECOMMENDATION

- **Weakness**:
  - The model can still be sensitive to imbalanced data.
  - It is computationally expensive.
  - Low visualization capability.
  - Lacks extrapolation ability.



SETUP & DEPLOY MODEL

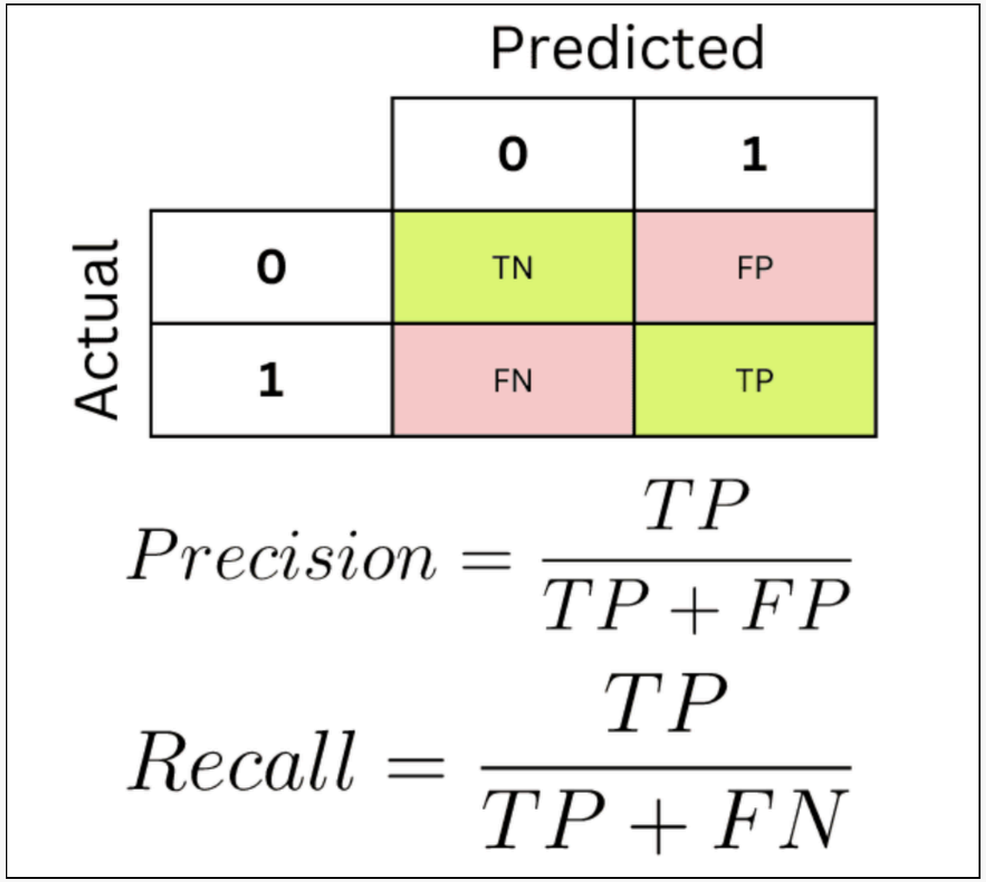
- Use the XGBoost library for training and prediction
  - Use the **XGBClassifier** class to initialize the prediction model.
  - Combine it with the **DMatrix** data structure to reduce training time.



EVALUATION METRICS

- Use the accuracy score (the percentage of correct predictions compared to the actual values).
- However, due to the uneven distribution of the data, using only the accuracy score is **not sufficient.**

-> Use the F1-Score (the harmonic mean of **Recall** and **Precision**) to more accurately evaluate the model's performance.

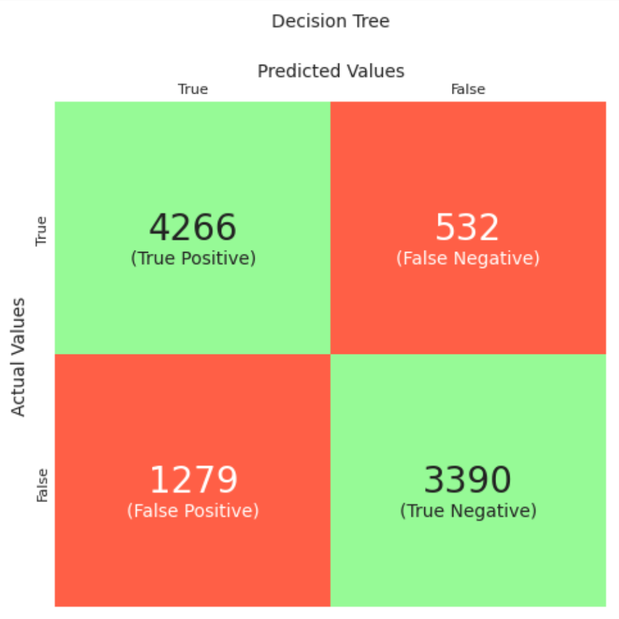


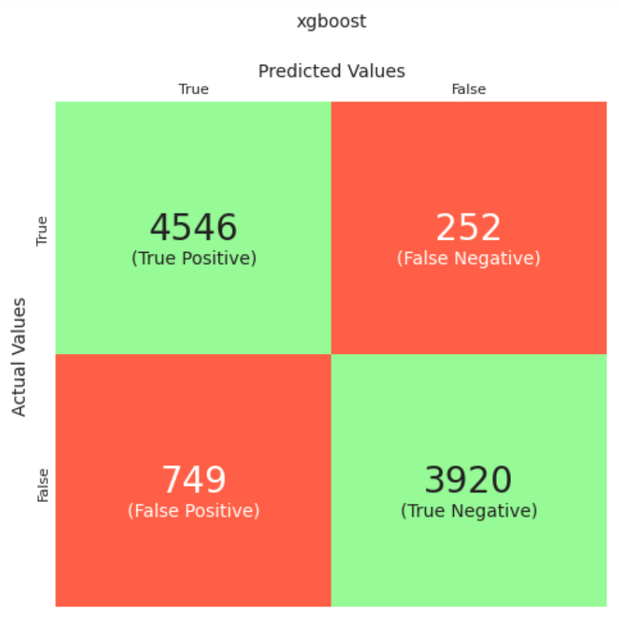
RESULTS

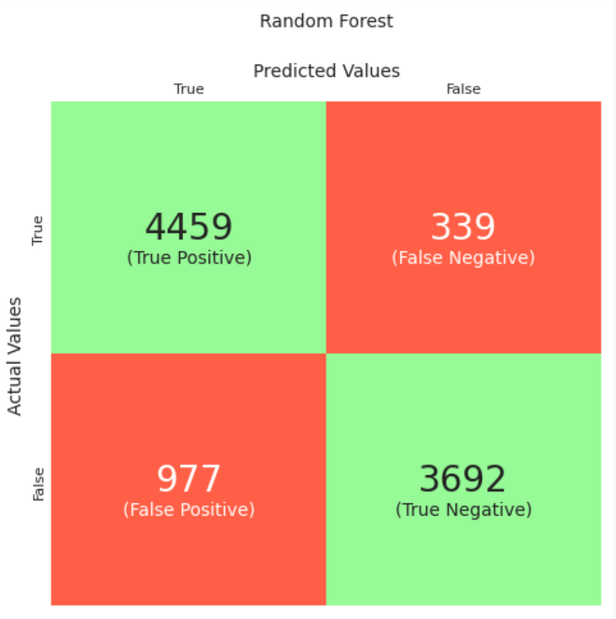
- **Decision Tree**
  - Accuracy: 80.87%
  - F1 Score: 0.789

- **XGBoost**
  - Accuracy: **89.42%**
  - F1 Score: **0.887**

- **Random Forest**
  - Accuracy: 86.1%
  - F1 Score: 0.848





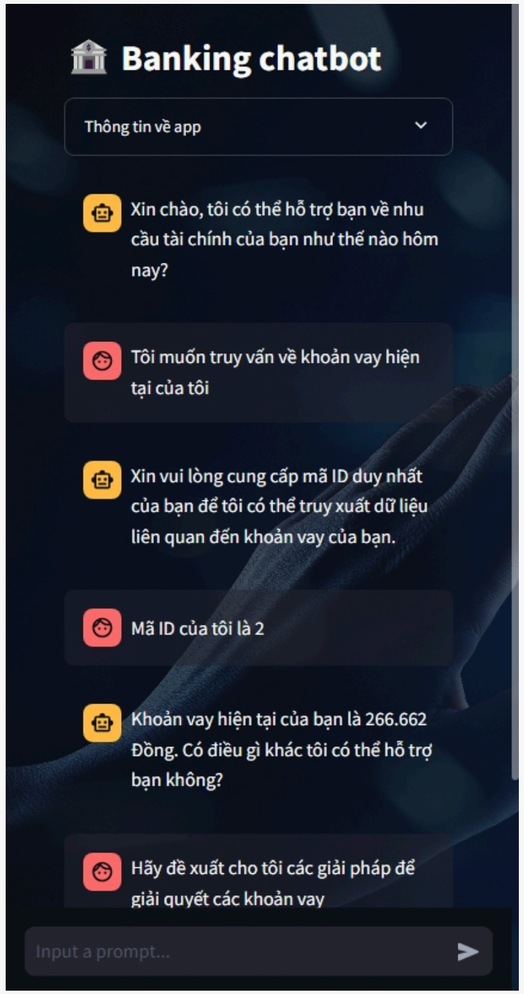


SOLUTION

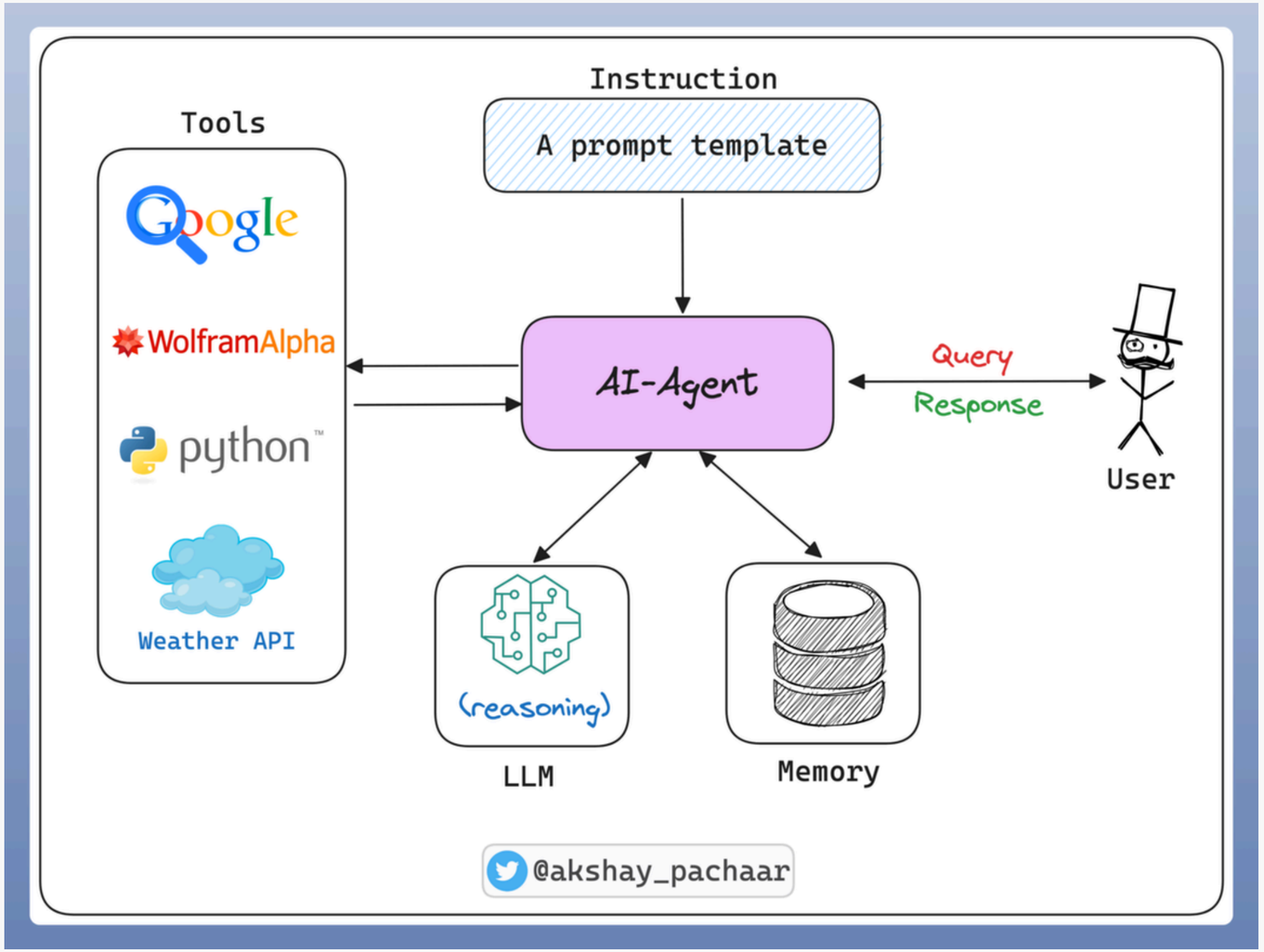
FOR BUSINESSES

BANKING CHATBOT

- Use the OpenAI API (GPT-4, TTS-1) in combination with RAG (Retrieval Augmented Generation).
- Help users retrieve financial information, financial solutions, or relevant legal documents.
- Support text-to-speech (TTS) to make querying easier.
- Combine System Instruction to personalize the responses.

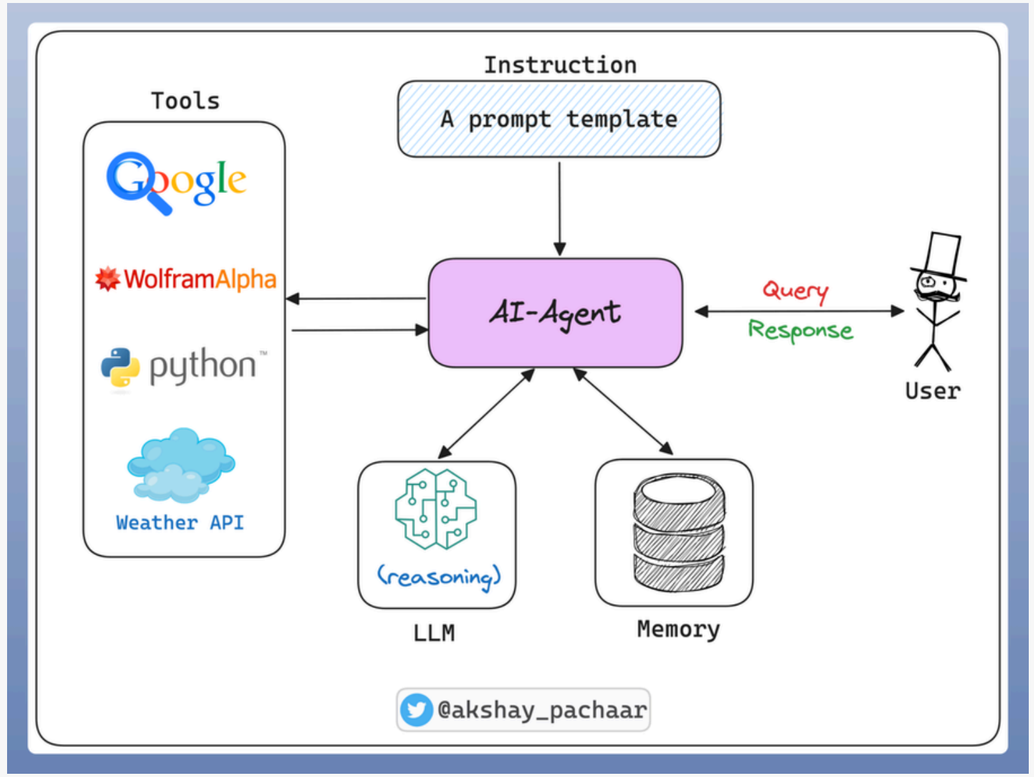


AI AGENTS


Instruction
A prompt template
Tools
Google
WolframAlpha
python
Weather API
AI-Agent
Query
Response
User
(reasoning)
LLM
Memory
@akshay_pachaar

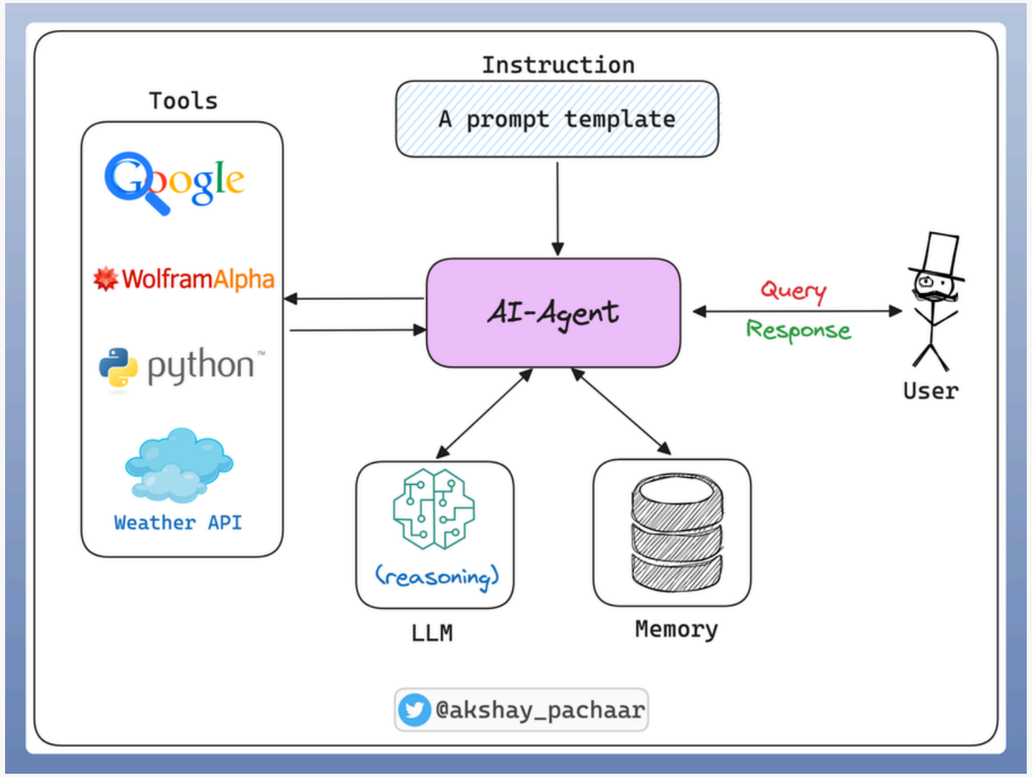
STRENGTH & WEAKNESS

- **Strength:**
  - Reduce labor costs
  - Increase processing efficiency
  - Personalize customer experience
  - Enhance scalability
  - Operate 24/7



STRENGTH & WEAKNESS

- **Weakness:**
  - Cannot handle complex situations
  - Lacks personal interaction and trust
  - Dependent on data quality



THANK YOU