Individual Final Report – Aihan Liu

Financial institutions incur significant losses due to the default of Vehicle Loans. Our project uses profile information to predict whether the customer tends to default or not. Our dataset is from Kaggle. We shared the work for each part and cooperated with each group member.

Contributions

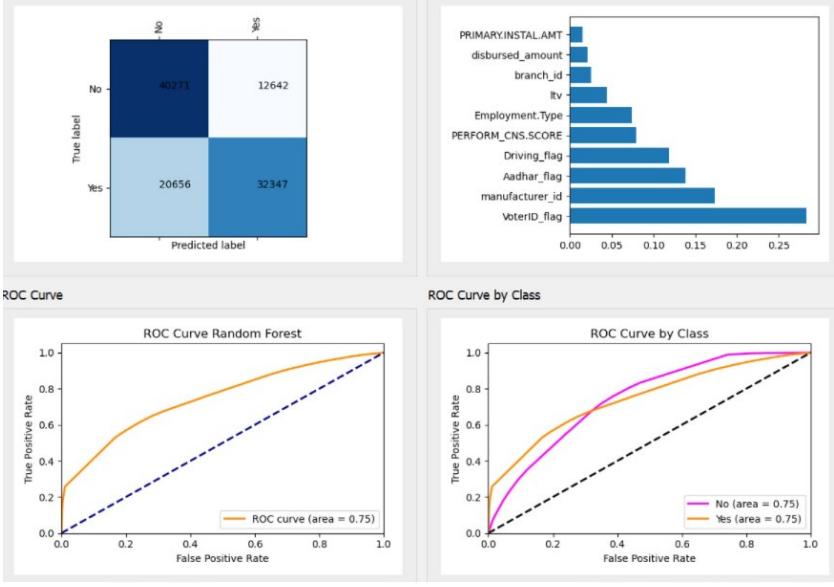
My work for this project is based on code tidying up and problem-solving. Once we finished a task, I would extract the helpful part and manage them into a code file, including data downloading, preprocessing, modeling, and GUI. I first planned to create classes for each portion of the code, but since we spent too much time on GUI, it was hard to make the arrangement after. The only part I finished was preprocessing. I developed the GUI layouts and the EDA part in the GUI. One of the questions I didn't solve is how to use other visualization packages such as seaborn in pyqt5. It is regrated that I could not put the woebin plots that Sara created into the GUI because they were generated by scorecardpy package. I also developed the final model based on other members' previous work. (The question I asked after the class was a tiny misunderstanding.) I appended the code into the Code file.

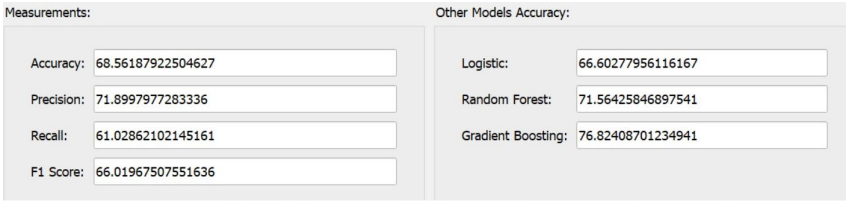
Results

Classification is the process of assigning data points to predefined classes or categories. In this project, we have implemented four classification Algorithms.

Decision tree:

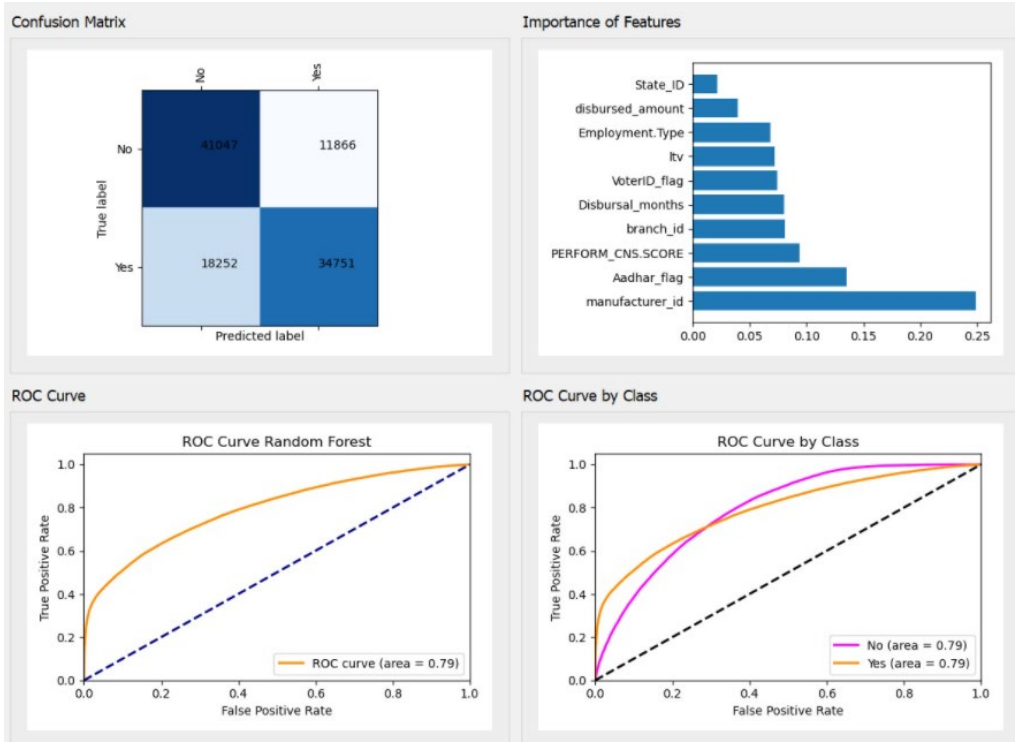
Decision Tree is a tree flowchart-like structure that divides the data into different subgroups based on conditions to classify the data. A condition is selected such that the classification is as pure as possible. At each node of the tree, a decision is made about splitting the data and getting the purest nodes. We can use different measures like Gini, entropy or misclassification error to calculate what attribute to split on. When you travel down the tree, finally, at leaf nodes, we find the labels of the data of a particular sample.

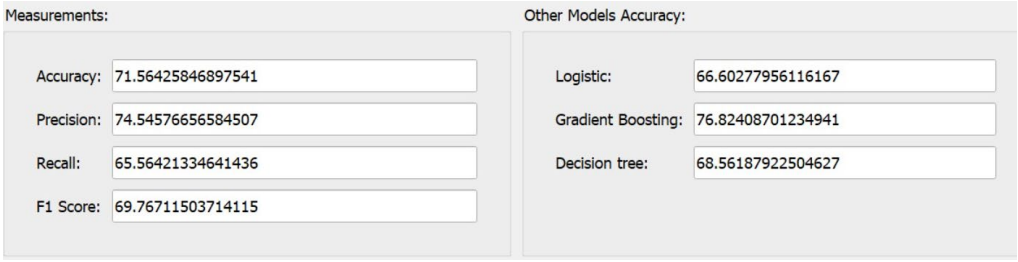




Random Forest:

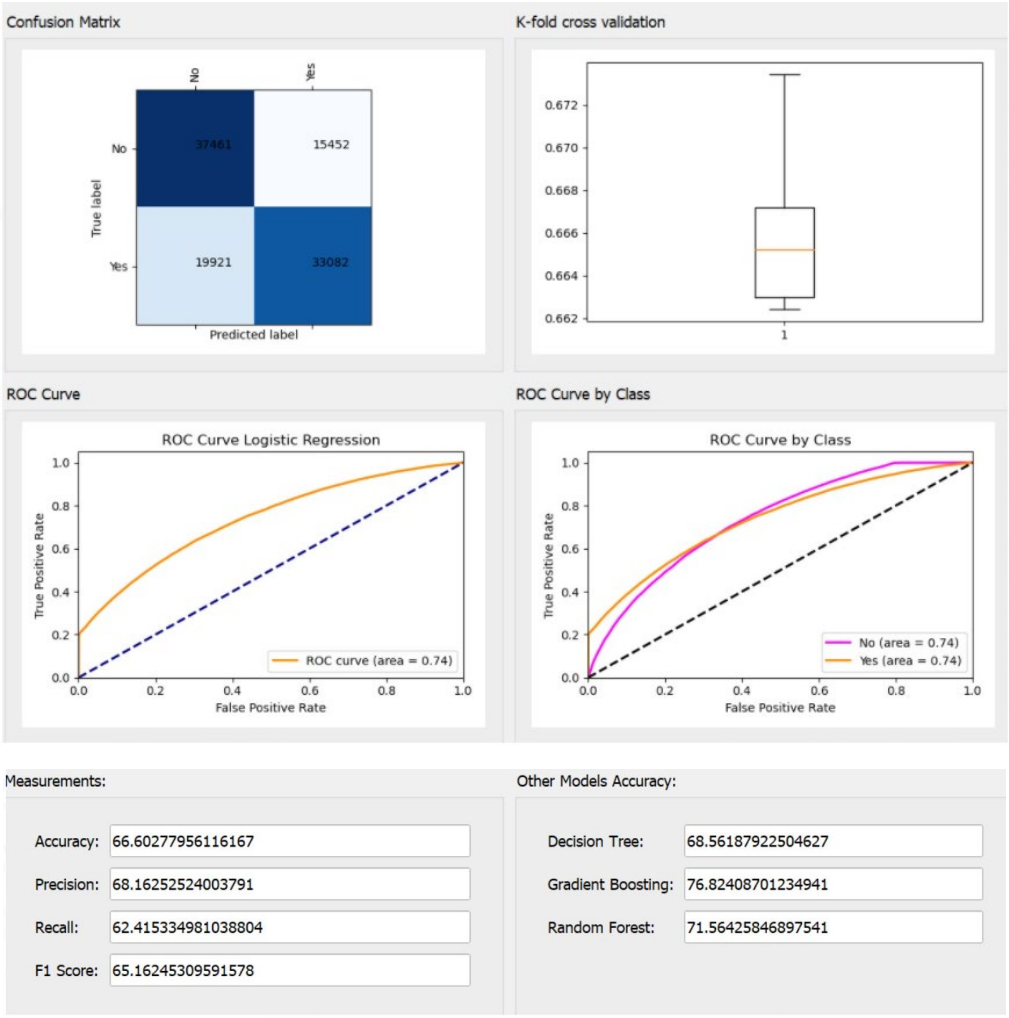
For random forest, we select random features to check for the best split attribute. And we use the max voting classifier to classify the data.





Logistic Regression:

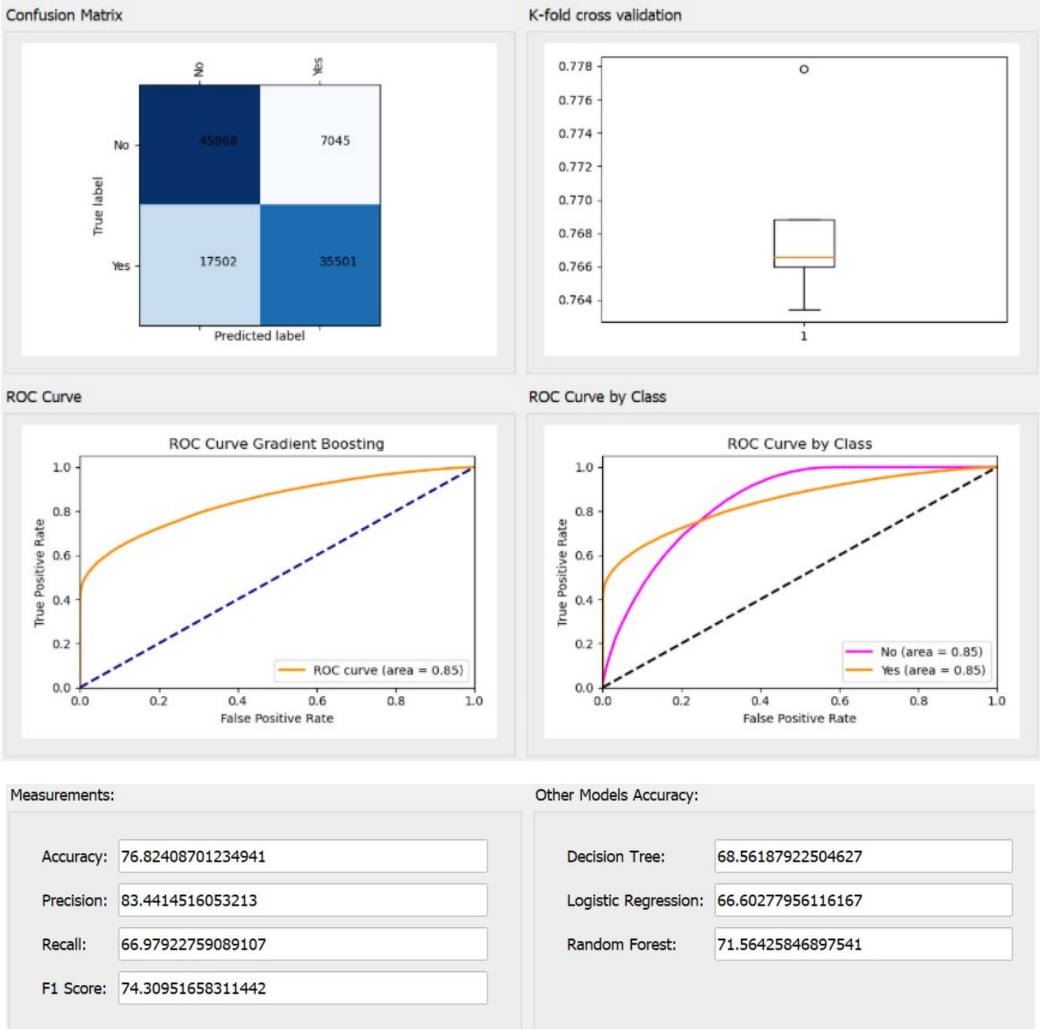
Logistic Regression is a Statistical Learning technique. It is one of the Supervised Machine Learning methods used in Classification tasks. We used K-fold Cross-validation as one metric for this classification.



Gradient Boosting:

Gradient boosting is a machine learning technique used in classification and Regression. It relies on the intuition that the best possible next model minimizes the overall prediction error when compared with previous models. This extraordinary ensemble learning technique combines several weak learners into strong learners. This works by each model paying attention to its predecessor's mistakes.

The following shows the results of all the Algorithms run so far. We can see that Gradient boosting gives the best prediction with a high precision of 83.4% and an accuracy of 76.8%.



METRICS	RANDOM FOREST	LOGISTIC REGRESSION	GRADIENT BOOSTING	DECISION TREE
Accuracy	71.6%	66.6%	76.8%	68.6%
F1_score	69.8%	65.2%	74.3%	66.0%
Precision score	74.5%	68.2%	83.4%	71.9%
Recall	65.6%	62.4%	67.0%	61.0%

Summary and conclusions.

The features that most affect the loan default are: Adahar_flag, voterID_flag, perform_cns.score, driving_flag, ltv, employer type and state_id.
The model we trained with the highest accuracy is Gradient boosting. However, it has the lowest recall.

Another work that may improve this dataset's accuracy is applying PCA or other feature selection techniques. We can also use other ensemble methods to get better results.

I have learned a lot in this project, and it's the first time for me to create a GUI. I did a lot of research to generate the result. The codes I copied from the internet or the lecture codes were about 70% for GUI layout, 50% for modeling, and 30% for preprocessing.