My project is on vehicles loan default,
this is a classification problem and the goal is to predict whether a vehicle loanee is going to default in the first EMI on due date.
Since Financial institutions can incur significant losses due to the default of vehicle loans, it is necessary for them to predict the potential defaulter.
The data is from Kaggle provided by an Indian financial institution.

From EDA, I found that the risk level of a loanee' s credit has a clear correlation with the target variable. Also, most of the features related to the loanee' s secondary account are zeros, so I just combine the primary account information and secondary account together to get a total features.

Since the dataset is imbalanced, I used the stratified splitting method to take 20% of the data as test set. The other 80% as training and validation set. After dropping some uninformative columns and combine some columns together, there are total 26 features left. 4 scalers are used to transform the features.
L2(improve the generalization of the model)
I tried logistic regression with l2 regularization, random forest, k nearest neighbor and xgboost classifier. In order to get he best performance of each model, I used gridserchcv with stritifed 4 fold cross validation to tune the hyperparameters. Since the dataset is imbalanced, I picked f1 score as the metric to evaluate the performance so that both false negative and false positive can be taken into account. Then five different random states are used to measure uncertainties that caused by splitting and non-deterministic methods

Here is the f1 scores of these 4 models, xgboost classifier has the highest score.
The baseline f1 score which is calculated by taking everything as positive is around 0.4. The trained xgboost classifier has a mean score of 0.84 with std of 0.04. The baseline score is 43 std lower than the trained score.

I used the xgboost built in feature importance method with 3 different importance type. According to the three graph, the most important features are the risk level of the loanee' s credit, the cost of the car, the amount of their loan, the loanees' total overdue account and their current balance.

This is a force plot for a single point based on xgboost. We can see that the total current balance push the prediction lower and the disbursed amount push the prediction higher.

Finally, in order to improve the performance, the possible things we can do are like solving the problem of imbalanced data like using resampling method SMOTE or assign more weights to class 1. Or we do more feature reduction to reduce the uninformative features and collect more meaningful real-world feature like their loan history of other property.