CREDIT CARD DEFAULT PREDICTION

Submitted by:

Anshul Mehra

OVERVIEW

- Banking/Financial Institutes plays a significant role in providing financial service.
- To maintain the integrity, bank/institute must be careful when investing in customers to avoid financial loss.
- Before giving credit to borrowers, the bank must come to about the potential of customers.
- The term credit scoring, determines the relation between defaulters and loan characteristics.

DATA PREPROCESSING

- Data set is divided in 75:25 ratio for train and test respectively.
- ID column was dropped as its unnecessary for our modeling.
- The attribute name ‘PAY_0’was converted to ‘PAY_1’ and 'default.payment.next.month’ was converts to ‘Default’ for naming convenience.
- Pay_0:No consumption of credit card=-2,Pay duly(paid on time)=-1, payment delay for one mouth=1, payment delay for two months=2, payment delay for nine months and above=-9.
- No Null values in dataset

INSIGHT FROM DATA ANALYSIS

- There are more women than men in our dataset and, apparently, men have a slightly higher chance of default.
- The probability of default was higher for men.
- Most people in our dataset have between 25 and 40 years old. There is also an impression that around that age the chance of default is a little lower.
- Most customers have 200k or less of credit limit. And it seems that we will find a higher concentration of customers in default on that range.

INSIGHT FROM DATA ANALYSIS

- Those who have a negative bill statement have a lower chance of default than the rest. What stands out is that there is a little higher chance of default for those who didn't have a bill in the previous months.
- There is a higher default rate among those who paid nothing in previous months and lower rates among those paid over 25k of NT dollars.

RANDOM FOREST MODEL

The term “Gradient Boosting Classifier” refers to the classification algorithm made up of several decision trees. The algorithm uses randomness to build each individual tree to promote uncorrelated forests, which then uses the forest’s predictive powers to make accurate decisions.

Gradient Boosting Classifier algorithm is used in a lot of different fields, like banking, the stock market, medicine and e-commerce.

INCREASING THE PREDICTIVE POWER

There is the **n_estimators** hyperparameter, which is just the number of trees the algorithm builds before taking the maximum voting or taking the averages of predictions. In general, a higher number of trees increases the performance and makes the predictions more stable, but it also slows down the computation.

max_features, which is the maximum number of features random forest considers to split a node. Sklearn provides several options, all described in the documentation.

The last important hyperparameter is **learning_rate.** This determines the step size at each iteration while moving toward a minimum of a loss function.

INCREASING THE MODEL’S SPEED

The **n_jobs** hyperparameter tells the engine how many processors it is allowed to use. If it has a value of one, it can only use one processor. A value of “-1” means that there is no limit.

The **random_state** hyperparameter makes the model’s output replicable. The model will always produce the same results when it has a definite value of random_state and if it has been given the same hyperparameters and the same training data.

here is the **oob_score** (also called oob sampling), which is a random forest cross-validation method. In this sampling, about one-third of the data is not used to train the model and can be used to evaluate its performance. These samples are called the out-of-bag samples. It's very similar to the leave-one-out-cross-validation method, but almost no additional computational burden goes along with it.

CONCLUSION

- We investigated the data, checking for data unbalancing, visualising the features and understanding the relationship between different features.
- We used train-test split to evaluate the model effectiveness to predict the target value i.e. detecting if a credit card will default next month.
- We started with Logistic Regression, SVC, Decision Tree, Gaussian Naive Bayes, Bernoulli Naive Bayes, Xgboost, Random Forest, Ada boost, and Gradient Boosting. All had different accuracy scores.
- We choose Gradient Boosting model base on the accuracy score which were lower for the other models.
- This would also inform the issuer’s decisions on who to give a credit card and what credit limit to provide.

THANK YOU