Proposal

Our group is interested in lending. Lending has always been a very important part of finance. The functioning of lending system largely depends on people making their payments on time, that is non-default. This raises our question, "How to make accurate predictions about default?"

The data set 'train_public' we focus on is part of "Regular season: Individual loan default forecast" set on https://aistudio.baidu.com/datasetdetail/130186?lang=en. Due to the website limit, it has to be downloaded before opening. There are 10000 samples and 38 variables, with missing values less than 10 percentages. These variables generally cover the personal information factors, loan information factors and credit history factors described as their background introduction. IsDefault is served as the outcome, which is 0 or 1, representing whether the customer has defaulted.

Although the data comes from professional bank records, some preprocessing is still required, imputation of missing values and variables selection for instance. After preprocessing, our group decide to use KNN, logistic regression, Adaboost and Xgboost respectively to build the model and make predictions. KNN method just use the train set itself to fit instead of building a model. The prediction of one point relies on its k neighbors. Logistic regression fit a linear model with $P(Y=1)=\frac{1}{1+exp\,(-w^Tx+b)}$ and is useful with binary classification problem. Both Adaboost and Xgboost method are both one of the integrated learning, applying decision tree as base learner. Adaboost proposed by Freund & Schapire (1997)[1] use adaptive re-weighting to continuously improve the performance of the base learner (decision tree) and do weighted integration according to the performance of these base learners and make predictions. XGBoost proposed by Chen & Guestrin (2016)[2] weights between loss and complexity. It sequentially adds base learners (decision tree), each focusing on reducing the residuals from the previous step, thereby enhancing prediction accuracy.

In our previous work, decision tree itself may be overcomplicated low-accurate to deal with high dimensional data, since high correlation exists between different features. Then we found Ensemble Learning Algorithms and Boost family, which is built on decision tree. We want to apply what we find to the real data and compare their accuracy to other algorithms like KNN and logistic regression.

In our paper, we decide to consider AUC and error rate as the indicators to judge the effect of our models. Area Under the Curve (AUC) is a metric to evaluate the performance of a classification model. It refers to the area under the Receiver Operating Characteristic (ROC) curve. This curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings. Higher AUC represents higher accuracy.

---

[1] Freund Y., Schapire R., A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting, *the Journal of Computer and System Sciences*, 1997, 55(1), 119-139.

[2] Chen T., Guestrin,C., XGBoost: A Scalable Tree Boosting System, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16, August 13-17, 2016, San Francisco, CA, USA)*, 2016.