Keisuke Yokota

ID: 12174207

HW3 Report

Let true positive (TP) be 1 and mean that someone will have financial trouble in the next two years and true negative (TN) be 0 otherwise. We try seven classifiers; Linear Regression, Decision Tree, Bagging, Adaboost, Random Forest, K-nearest neighbor and SVM.

To go straight to the bottom line, Bagging model is the best among the seven classifiers.

From the only view point of accuracy, or the correct identification rate of those who will face financial distress or not, Adaboost model is recommended because it got the highest accuracy score.

But we will need to care about only either of them. For example, credit card companies and bank want to know identify who is bad customers to avoid losses brought by their default. In this case it is important to reduce the probability to predict actual bad customers as good customers. Here, we can use recall score. Recall score is the ratio of how well the model predicts persons as those who at financial risk when the persons are actual risk holders. Even though Random Forest won the highest recall score among the classifiers, Decision Tree (DT) is better in this case instead because we care about only TP in this case and the highest score is provided by DT.

Still, we also might have to take precision into consideration. Precision is the ratio of actual risk holders among predicted risk holders. If financial institutions counted on only recall score, lots of people who are actually good customers would be refused to make a credit card or loan. This mean financial companies would lose potential good customers. Among the classifiers, Bagging model got the highest precision score.

Besides, the Bagging has the third highest recall score and accuracy score and they are relatively high compared to other classifiers.

From accuracy, recall and precision score, Bagging model will be the most useful.

Table: Evaluation of classifiers by accuracy, precision and recall score

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Classifier  Metrics | LR | DT | BAG | AB | RF | KNN | SVM |
| Accuracy | 0.851 | 0.876 | 0.877 | **0.880** | 0.878 | 0.841 | 0.840 |
| Precision (avg) | 0.85 | 0.86 | 0.85 | 0.87 | 0.87 | 0.81 | 0.74 |
| (about 1 only) | 0.86 | 0.69 | **0.88** | 0.71 | 0.73 | 0.64 | 0.23 |
| (about 0 only) | 0.84 | 0.89 | 0.85 | 0.90 | 0.89 | 0.84 | 0.84 |
| Recall (avg) | 0.85 | 0.82 | 0.88 | 0.88 | 0.88 | 0.84 | 0.83 |
| (about 1 only) | 0.10 | **0.45** | 0.43 | 0.43 | 0.44 | 0.00 | 0.02 |
| (about 0 only) | 0.99 | 0.89 | 0.96 | 0.96 | 0.96 | 1.00 | 0.98 |

(LR: Linear Regression, DT: Decision Tree, BAG: Bagging, AB: Adaboost, RF: Random Forest, KNN: K-nearest neighbor, SVM: Support Vector Machine and attributes in red: top 3 high score in the metrics)