2.3 Suitable learning algorithms

After we analyzed the dataset and choose different feature selection methods, now we decide to find some appropriate algorithms for our dataset. We find that the target values of our dataset are 1 or 0 (unsatisfied or satisfied), which means they are discrete. Thus, we narrow down the possible options of algorithms to classifier models instead of regression models since it is hard to use regression models to fit a dataset which only have two labels. Considering the size of features (370 different features), we decide not to use KNN since too large dimensions in our dataset will result in meaningless distance calculation in KNN model. Also, for the learning algorithm SVM (support vector machine), we decide not to choose it because the number of our data are too large, the efficiency of SVM will be very low. Based on the lecture slides of COMP9417, we choose Logistic Regression as our first model since it is the relatively simple algorithm in classifiers models. Then, we decide to choose decision tree, the random forest and Naïve Bayes as our training models. Due to the advanced implementation of these models, we trained all of them by using the python library SKLEARN. The detailed descriptions of the models are shown in Appendix.

|  |  |
| --- | --- |
| Machine learning algorithm | Description |
| Logistic Regression | To ﬁnd the best ﬁtting and most parsimonious, clinically interpretable model to describe the relationship between an outcome (dependent or response) variable and a set of independent (predictor or explanatory) variables. The independent variables are often called covariates. The most common example of modelling, and one assumed to be familiar to the readers of this text, is the usual linear regression model where the outcome variable is assumed to be continuous (Hosmer, David W.). |
| Decision tree | The main characteristic of decision trees is a recursive subsetting of a target field of data according to the values of associated input fields or predictors to create partitions, and associated descendent data subsets (called leaves or nodes), that contain progressively similar intra-leaf (or intra-node) target values and progressively dissimilar inter-leaf (or inter-node values) at any given level of the tree(De Ville, Barry Hoboken). |
| Random forest | Random forest is a commonly-used machine learning algorithm trademarked by Leo Breiman and Adele Cutler, which combines the output of multiple decision trees to reach a single result. Its ease of use and flexibility have fueled its adoption, as it handles both classification and regression problems. |
| Naïve Bayes | A Naive Bayes classifier is a probabilistic machine learning model that’s used for classification task. The crux of the classifier is based on the Bayes theorem. |

Hosmer, David W., Jr., et al. *Applied Logistic Regression*, John Wiley & Sons, Incorporated, 2013.*ProQuest Ebook Central*, http://ebookcentral.proquest.com/lib/unsw/detail.action?docID=1138225.

De Ville, Barry Hoboken, USA: John Wiley & Sons, Inc, peer\_reviewed Wiley interdisciplinary reviews. Computational statistics, 2013-11, Vol.5 (6), p.448-455

2.2.3   Data set after removing features which are highly correlated

Correlation is a measure of the linear relationship of 2 or more variables. Through correlation, we can predict one variable from the other.

Good variables are highly correlated with the target.

Correlated predictor variables provide redundant information.

Correlation is a statistical technique that can show whether and how strongly pairs of variables are Linearly related. There are several different correlation techniques. Some of the most important correlation is Pearson Correlation. Pearson correlation technique works best with linear relationships: as one variable gets larger, the other gets larger (or smaller) in direct proportion. It does not work well with curvilinear relationships (in which the relationship does not follow a straight line).

Correlation is a statistical method that can measure the relationship of 2 or more variables and show how strongly pairs of variables are linearly related. There are several different correlation techniques. Some of the most important correlation is Pearson Correlation. We will use Pearson Correlation in our implementation. The technique works well when distinguish linear relationships: when one variable increases, the other one changes in the same trend. The value of 1 means positive correlation and -1 means negative correlation. If the value is close to 0, it means the variables does not have too much relationship with each other. We will remove those variables which have strong connections with others from our dataset.

|  |  |  |
| --- | --- | --- |
| Models | Cross Validation Strategy | Hyperparameters to adjust |
| Random forest | Non-nested | n\_estimators: The number of trees in the forest  max\_depth: The maximum depth of the tree. If None, then nodes are expanded until all leaves are pure or until all leaves contain less than min\_samples\_split samples.  min\_samples\_split: The minimum number of samples  required to split an internal node |
| Naïve bayes | nested | var\_smoothing: Portion of the largest variance of all features that is added to variances for calculation stability.  priors: Prior probabilities of the classes. If specified the priors are not adjusted according to the data. |
| Logistic regression | nested | Penalty: The regularization expression.  C: Inverse of regularization strength;  Must be a positive float. Like in support vector machines, smaller values specify stronger regularization. |
| Decision Tree | nested | max\_depth: The maximum depth of the tree. If None, then nodes are expanded until all leaves are pure or until all leaves contain less than min\_samples\_split samples.  min\_samples\_split: The minimum number of samples required to split an internal node.  max\_features: The number of features to consider when looking for the best split. |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| rank | mean test score | std test score | max depth | min samples split | N estimators | Feature |
| 1 | 0.83462522 | 0.00865463 | 24 | 150 | 155 | Corr |
| 2 | 0.83353412 | 0.00865153 | 20 | 170 | 130 | Ridge |
| 3 | 0.83159144 | 0.00841704 | 23 | 150 | 100 | Selected |
| 4 | 0.83059448 | 0.0089132 | 24 | 160 | 150 | Original |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| rank | mean test score | std test score | max depth | min samples split | n\_estimators | Feature | Train\_AUC | Test\_AUC |
| 1 | 0.83462522 | 0.00865463 | 24 | 150 | 155 | Corr | 0.916606 | 0.831029 |
| 2 | 0.83353412 | 0.00865153 | 20 | 170 | 130 | Ridge | 0.912587 | 0.830551 |
| 3 | 0.83159144 | 0.00841704 | 23 | 150 | 100 | Selected | 0.913363 | 0.828318 |
| 4 | 0.83059448 | 0.0089132 | 24 | 160 | 150 | Original | 0.911112 | 0.826416 |