**4.2 Tree Model and Feature importance**

How to predict whether customers will default is an important and difficult task for Banks. Each client has a huge amount of past records and information about everything from age, gender, job, family, etc. For the ordinary banking sector, based on these large and complex data, it is difficult to carry out logical and complete analysis and speculation. Above all, there are so many features that bank analysts waste time on irrelevant ones. How to extract the most important feature factor becomes extremely important.

The single model with decision tree as an example and the ensemble model established on this basis can effectively analyze the model after training and integrate the importance of features. This advantage provides an idea for Banks to determine the effective features of their customers. The tree structure firstly looks for the best feature among all features, then looks for all possible classification points, and finds the best classification points for classification, finally obtains an effective decision tree model. Ensemble method can integrate the decision tree model and obtain a more stable model with stronger generalization ability, and the most reliable feature importance data can be obtained by integrating all the decision tree models.

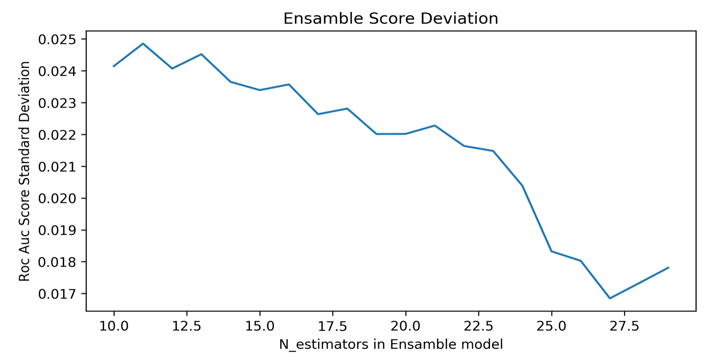
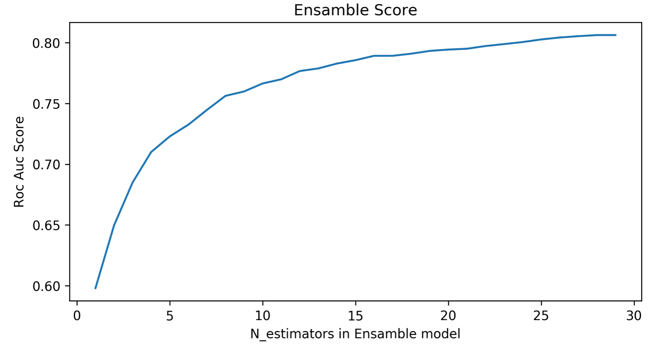
The features of data are not single, but influenced by each other. Through the integration of different data features, many new data features can be obtained, which reflects the non-linear features in the data. Therefore, it is also very important to explore the non-linear features of the data. If a new feature can play a more significant role than the original feature, then for the banking sector, this can be used to simplify the analysis and prediction process and save costs. At the same time, the non-linear feature further illustrates the limitation of human classification, because it is difficult for an analyst to observe or intuit the non-linear feature in the data and make an effective prediction based on it.

**5.2 Model selection and Feature Engineering**

**Model chosen to measure feature importance**

In order to get the insight of the feature importance, a model should be chosen, and it should be precise and robust in prediction to measure the feature importance. Though it is well known that Decision Tree has the ability to measure feature importance, decision tree is note stable and can overfit the data set. It is necessary to build a better model which is more stable than decision tree model. An ensemble method is conducted here. A bagging model can be built based on decision tree model, which means within the bagging model, there are a large quantity of decision trees to train different subsets of the training data. Once all models are trained, the ensemble model can make a prediction for a new instance by simply aggregating the predictions of all predictors. Each individual predictor has a higher bias than if it were trained on the original training set, but aggregation reduces both bias and variance.

Below are Roc Auc score and scores’ standard deviation under 5-fold cross-validation.



It is quite clear that when the number of models in bagging model increases, the bagging model gets better prediction score and low deviation. As a result, the ensemble model will used to measure feature importance.

**Feature Importance and Feature engineering**

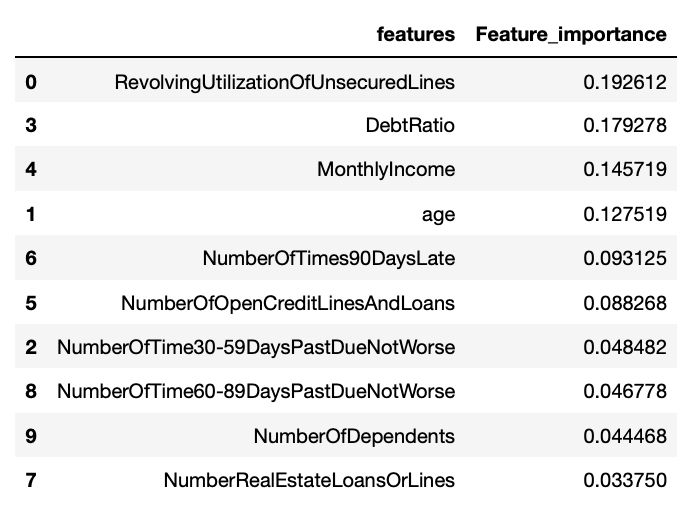
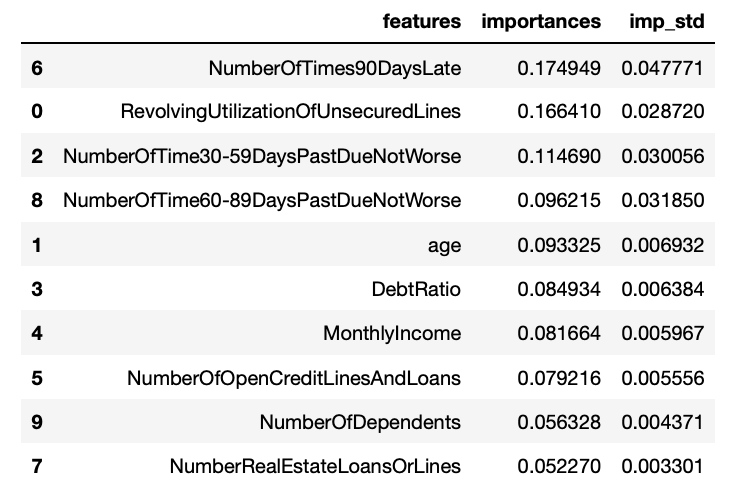
**Feature importance at first glimpse**

It is one of the greatest advantages of Decision Trees that they can automatically calculate feature importance and help people do feature selection and interpreter the whole model. The importance of a feature is computed as the (normalized) total reduction of the criterion brought by that feature. It is also known as the Gini importance. The extreme tree model is used to determine the feature importance and the variation.

The extreme tree model is an ensemble model that contains large amounts of trees. For each tree in the classifier, a model has been trained and the feature importance is determined. However, the feature importance for one Tree model may easily become biased and vary largely. by aggregating these different trees, an ensemble model will result more robust output of feature importance. What is more, the variation of the importance in every tree model can be calculated to measure whether the feature is robust enough to make predictions.

The classifier Random Forest can measure feature importance as well. A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

Below is the result shown by these two models, the first is extreme tree model and the second is Random Forest:



It can be seen that both classifier methods choose the same set of features that contribute to the model training from Gini index. However, the features are arranged in different orders. That is because of the algorithms employed under extreme tree model and Random Forest are different. The random forest uses many subsets of the original features to train the model, but the extreme tree model uses different subsets of features and decide the threshold of each feature randomly to train the model.

**Feature Engineering**

A feature engineering method is conducted to add more polynomial combinations of the existed features to better measure the importance of the features and the combinations of the features. A degree of three is used to see the nonlinear combinations of features existed.

Below is part of the new features after feature engineering.



**Recalculate Feature Importance**

After feature engineering, new features are used to reconduct the feature importance. The result is shown below:



It can be seen that the importance level is diversified. Different from the results before feature engineering, new features have higher feature importance than their original feature, which means combinations of features in this training set have higher contributions in prediction.

**6.2 Feature Discussion and Application in banks**

First of all, from the feature importance, it is quite clear that the past record of whether the client has negative record matters. In real life, the bank can first check the past record of the client and have a first prediction of whether the client will default this time.

The differences between the importance before feature engineering and after can be understood in feature selection step in the random forest model. The tree model finds the best variable and the split point in every node in the tree structure, so in the previous, the important model is being chosen many times in different nodes of the tree. However, the engineered features show different aspects of the original data, so the feature is less chosen in the nodes of the tree structure. As a result, the feature importance is diversified. It can also be concluded that features should be combined together to interpreter the model and make predictions.

In a business level, it means it is not reasonable to predict in daily life just base on simple math manipulation. The classification should consider every variable and also the combinations of them. It is necessary to build a classification system in banks to make classifications if the banks want to have a better result.