**Feature Scaling**

Feature scaling could be defined as the method used to standardize the range of variables. It is one of the key task which should be applied. When the variables have very different ranges, algorithms could not predict well. Scaling the variables and making the variables in the same range is important in the step data preprocessing.

There are two main method to feature scaling: min-max scaling and standardization.

The min-max scaling method is to scale the data to [0,1] or [-1, 1]. Scaling to what extent depends on the nature of the data. The formula for this method is as follows:

.

In this formula, is the initial value, and is the scaled value.

According to the formula, the values are shifted and retuned to make their values from 0 to 1 by subtracting the min value and dividing by the max minus the min. In this project, a transformer called **MinMaxScaler** provided by Scikit-Learn is applied. The code is as follows:

图片包含 屏幕截图

已生成极高可信度的说明

Standardization is different from the min-max scaling. It makes the variables have the zero mean and unit variance. The formula for this method is as follows:

According to the formula, first it subtracts the mean value (so standardized

values always have a zero mean), and then it divides by the variance so that the resulting distribution has unit variance. Standardization is much less affected by outliers than Min-max scaling (Géron,2017). In this project, a transformer called **StandardScaler** provided by Scikit-Learn is applied. The code is as follows:

图片包含 屏幕截图

已生成极高可信度的说明

**Feature selection**

Feature selection could be also called variable selection or attribute selection which is an important data reprocessing task. It is the selection of the features which are most relevant to the target value. Suppose dealing with a supervised learning problem, the number of features is very large, but there may be only a few characteristics that will affect the results. Even dealing with a simple linear category, if the number of sample features exceeds n, but the VC dimension of the function is still n, then unless the number of training sets is greatly expanded, there could lead to the problem of overfitting. In such cases, feature selection could be used to reduce the number of features.

In general, considering the selection of features from some aspects: On the one hand, consider whether the features diverge. If a feature does not diverge, for example, the variance is close to 0, which means that there are no differences between the values of this feature, this feature is not useful for the prediction. On the other hand, consider the correlation between features and target values. It is obvious that the features with high relevance should be priority selected. The features with low relevance should be dropped.

Feature selection has lots of benefits: promoting data understanding, reducing measurement and storage requirements, reducing training and test time, avoiding overfitting and improving prediction (Guyon,2003). Feature selection could help build more effective prediction models. It can choose features that make prediction better. Feature selection could help build much easier model. Fewer features make it spend less time on running the program.

There are differences between feature selection and dimensionality reduction. Both approaches try to reduce the number of features. Dimensionality reduction do this through the relation between features, such as combining different features to new features, which changes the original feature space. The method of feature selection is to select a subset from the original feature data set, which is a kind of included relationship, without changing the original feature space.

According to the form of feature selection, the feature selection method can be divided into three types:

1. Filter：Filter feature selection methods score each feature according to its divergence or correlation and rank the features by the scores. The methods often set threshold value or number of selected features then select features by the threshold value and the number. Filters evaluate the features as a pre-processing step, which are independent of the model (Guyon,2003).
2. Wrapper: Wrapper feature selection methods consider the selection of a subset as a search optimization problem which generate different combinations, evaluate combinations, and compare them with other combinations. By this way, wrapper feature selections find the best subset of features which makes the prediction best. This approach is closely related to predictor model and evaluating systems.
3. Embedded: Embedded feature selection integrates feature selection process and machine training process into one. Both of them are completed in the same optimization process, i.e. the feature selection is automatically carried out during the learning process.

In this project, filter methods and wrapper methods are applied to the feature selection.

1. **Filter method**

Wendong’part

1. **Wrapper method**

Wrapper method is to build lots of subset of features and then evaluate which feature subset make the prediction best. By this way, the methods drop the un redundant features and find the best features for model. Generally, the wrapper methods could be more effective than filter method because this method choose the features according to the predictor model.

Guyon (2003) argues that there are three things should be defined in this method: (i) how to search all possible subset of the features; (ii) which model to use for prediction; (iii) how to evaluate the feature with the given model. In this project, all of these are defined as follows: (i): this project searches the subset of features by the backward selection method. (ii): this project uses Linear Regression model, DecisionTree model and RandomForest model to find the best features for each three models. (iii): this project uses CrossValidation to evaluate the features, according to calculate the MSE (mean-square error).

Suppose dataset have n features, so there are possible feature subsets. If n is very large, it is hard to search all possible feature subsets because it could cost too much time. Therefore, some heuristic algorithms such as forward selection or backward selection could be used to realize the research of the subset of features. This project uses backward selection method to search features for three data sets (NewFinal30, NewFinal50, NewFinal97).

The example wrapper method feature selection code for NewFinal30 data set using Linear Regreesion model is as follows:

图片包含 屏幕截图

已生成极高可信度的说明

图片包含 屏幕截图

已生成极高可信度的说明

As the code shows, the first step is to build the subset of all features, which named F. Use Linear Regression to fit the variables and target values. Get the prediction of target value. Then evaluate the subset of all features by calculating the MSE (mean-square error) according to the prediction target value and real target value. The second step is to drop one of feature from this subset. The new subset is named F\_1. Then calculate the MSE. After that, drop another feature from F and also calculate the MSE. Loop this step for every feature in F. If there are n features in F, there are n Mse score. The feature which corresponds to the smallest MSE score is the feature need to drop. Just drop it. The third step is to loot all above steps. There is a subset of features whose score is smallest. When drop any feature, the MSE could not decline. This subset is the best subset of features.

The best features for NewFinal30 data set using Linear Regreesion model are ['calculatedbathnbr', 'finishedsquarefeet12', 'fips', 'fullbathcnt', 'longitude', 'lotsizesquarefeet', 'propertycountylandusecode', 'rawcensustractandblock', 'structuretaxvaluedollarcnt', 'taxvaluedollarcnt', 'assessmentyear', 'landtaxvaluedollarcnt', 'taxamount', 'year0', '02', '03', '04', '05', '06', '07', '08'].

The MSE for these best features is 0.00803076188201, while the MSE for all features is 0.00932025465749. After this this feature selection method, there is a much smaller MSE, which means the method is effective.

This project applied wrapper feature selection for three data sets (NewFinal30, NewFinal50, NewFinal97) by Linear Regression model, DecisionTree model and RandomForest model. All the MES have a significant drop after wrapper feature selection.