Classification steps:

1 Data Exploration & Statistics analysis

2 Data Preparation

1) Normalization

features\_raw[skewed] = data[skewed].apply(lambda x: np.log(x + 1))

2) Feature scaling

[sklearn.preprocessing.MinMaxScaler](http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html)

3) Translate category features to numbers

a) One hot encoding独热编码： change each category to new features with value 0 or 1

[pandas.get\_dummies()](http://pandas.pydata.org/pandas-docs/stable/generated/pandas.get_dummies.html?highlight=get_dummies#pandas.get_dummies)

b) EncodeLabel: Encode labels with value from 0 to class-1

class sklearn.preprocessing.LabelEncoder

c) train/validation/test data split

3 Modeling, scoring and parameter optimization

Scoring: accuracy and F1-score

4 Feature importance, weight, reduction

5 Perform test on test data