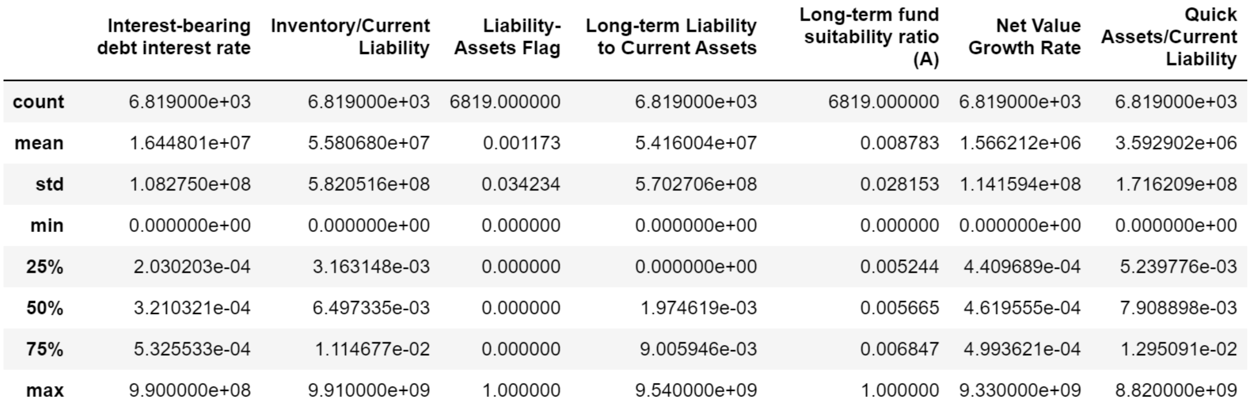
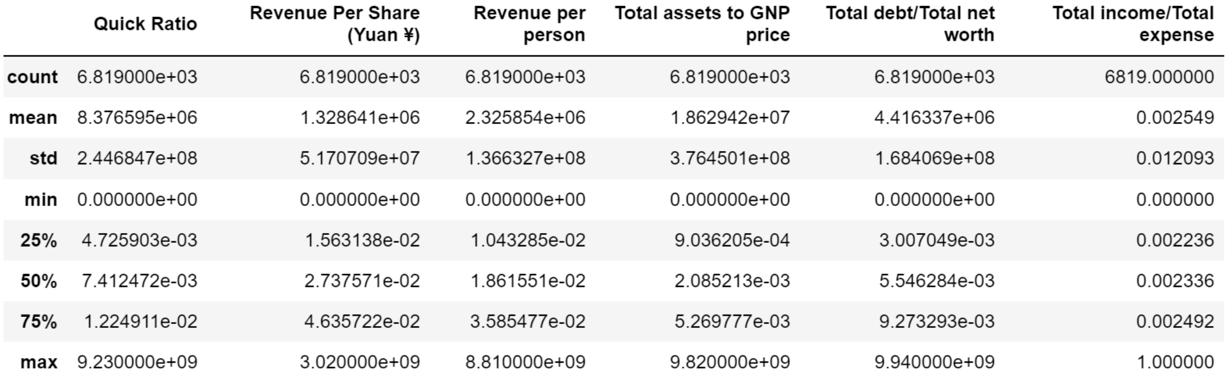
1. What is the statistics of the data? E.g., the mean, max., min. stdev. of the data in each feature?

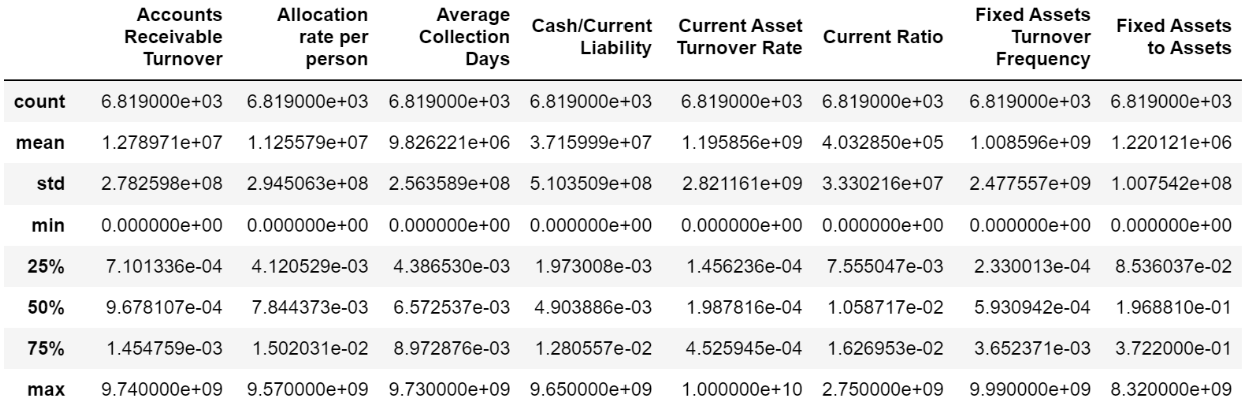
The followings are the statistics of our modified data:



	Interest-bearing debt interest rate	Inventory/Current Liability	Liability-Assets Flag	Long-term Liability to Current Assets	Long-term fund suitability ratio (A)	Net Value Growth Rate	Quick Assets/Current Liability
count	6.819000e+03	6.819000e+03	6819.000000	6.819000e+03	6819.000000	6.819000e+03	6.819000e+03
mean	1.644801e+07	5.580680e+07	0.001173	5.416004e+07	0.008783	1.566212e+06	3.592902e+06
std	1.082750e+08	5.820516e+08	0.034234	5.702706e+08	0.028153	1.141594e+08	1.716209e+08
min	0.000000e+00	0.000000e+00	0.000000	0.000000e+00	0.000000	0.000000e+00	0.000000e+00
25%	2.030203e-04	3.163148e-03	0.000000	0.000000e+00	0.005244	4.409689e-04	5.239776e-03
50%	3.210321e-04	6.497335e-03	0.000000	1.974619e-03	0.005665	4.619555e-04	7.908898e-03
75%	5.325533e-04	1.114677e-02	0.000000	9.005946e-03	0.006847	4.993621e-04	1.295091e-02
max	9.900000e+08	9.910000e+09	1.000000	9.540000e+09	1.000000	9.330000e+09	8.820000e+09



	Quick Ratio	Revenue Per Share (Yuan ¥)	Revenue per person	Total assets to GNP price	Total debt/Total net worth	Total income/Total expense
count	6.819000e+03	6.819000e+03	6.819000e+03	6.819000e+03	6.819000e+03	6819.000000
mean	8.376595e+06	1.328641e+06	2.325854e+06	1.862942e+07	4.416337e+06	0.002549
std	2.446847e+08	5.170709e+07	1.366327e+08	3.764501e+08	1.684069e+08	0.012093
min	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000
25%	4.725903e-03	1.563138e-02	1.043285e-02	9.036205e-04	3.007049e-03	0.002236
50%	7.412472e-03	2.737571e-02	1.861551e-02	2.085213e-03	5.546284e-03	0.002336
75%	1.224911e-02	4.635722e-02	3.585477e-02	5.269777e-03	9.273293e-03	0.002492
max	9.230000e+09	3.020000e+09	8.810000e+09	9.820000e+09	9.940000e+09	1.000000



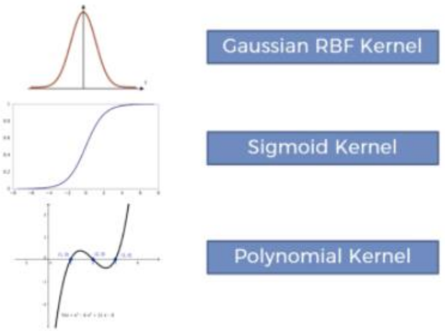
	Accounts Receivable Turnover	Allocation rate per person	Average Collection Days	Cash/Current Liability	Current Asset Turnover Rate	Current Ratio	Fixed Assets Turnover Frequency	Fixed Assets to Assets
count	6.819000e+03	6.819000e+03	6.819000e+03	6.819000e+03	6.819000e+03	6.819000e+03	6.819000e+03	6.819000e+03
mean	1.278971e+07	1.125579e+07	9.826221e+06	3.715999e+07	1.195856e+09	4.032850e+05	1.008596e+09	1.220121e+06
std	2.782598e+08	2.945063e+08	2.563589e+08	5.103509e+08	2.821161e+09	3.330216e+07	2.477557e+09	1.007542e+08
min	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
25%	7.101336e-04	4.120529e-03	4.386530e-03	1.973008e-03	1.456236e-04	7.555047e-03	2.330013e-04	8.536037e-02
50%	9.678107e-04	7.844373e-03	6.572537e-03	4.903886e-03	1.987816e-04	1.058717e-02	5.930942e-04	1.968810e-01
75%	1.454759e-03	1.502031e-02	8.972876e-03	1.280557e-02	4.525945e-04	1.626953e-02	3.652371e-03	3.722000e-01
max	9.740000e+09	9.570000e+09	9.730000e+09	9.650000e+09	1.000000e+10	2.750000e+09	9.990000e+09	8.320000e+09

2. What are the hyperparameters in SVM? How to tune them?

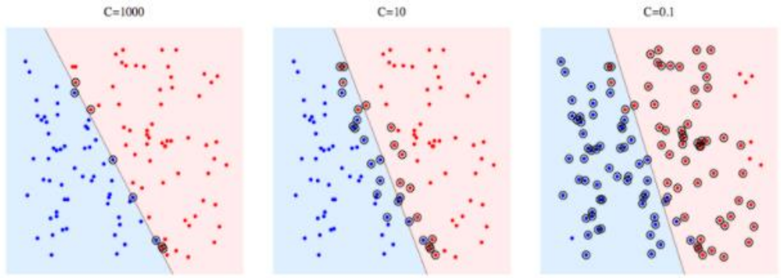
Some key hyperparameters in SVM are kernels, C(Regularization), Gamma.

Among them, kernel is the main hyperparameter of the SVM which maps the observations into some feature space.

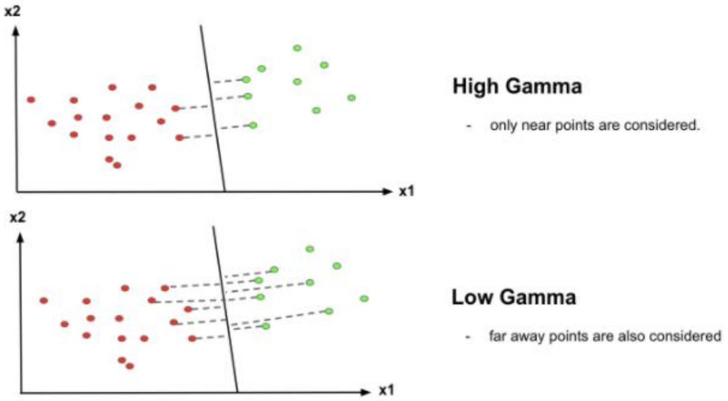
Kernel: To take low dimensional input space and transform it into a higher-dimensional space. It is mostly useful in non-linear separation problem.



C: This is a penalty parameter. The misclassification or error term tells the SVM optimization how much error is bearable.



Gamma: It defines how far influences the calculation of plausible line of separation. When gamma is higher, nearby points will have high influence; low gamma means far away points also be considered to get the decision boundary.


x2
x1
High Gamma
- only near points are considered.
x2
x1
Low Gamma
- far away points are also considered

[1]

Use GridSearch to tune the hyperparameters of an estimator.

We used Grid Search to tune hyperparameters as an approach which methodically builds and evaluates a model for each combination of algorithm parameters specified in a grid. GridSearchCV helps us combine an estimator with a grid search preamble to tune hyperparameters.

[1] https://www.vebuso.com/2020/03/svm-hyperparameter-tuning-using-gridsearchcv/

3. What is main way to handle the imbalanced issue?

We know there are at least 5 different methods for dealing with imbalanced datasets:

1.Change the performance metric 2.Change the algorithm 3.Oversample minority class 4.Undersample majority class 5.Generate synthetic samples

We believe oversample minority class, SMOTE are among the best of the options.

As shown in our report on page 4, “since the data ratio of bankruptcy and non-bankruptcy is severely unbalance, we need to do some preprocessing work to balance the data. We try SMOTE to balance our data set. After SMOTE, we have 13,198 samples in total and the sample ratio of bankruptcy and non-bankruptcy is 1:1.”