To select the bank features that regulation institutions need to keep an eye on to prevent bank failure, we selected the most important features.

Specifically, we selected features based on their importance in predicting whether a bank will fail within 6 quarters by classification. For all the banks that have failed, we used data from the last 6 quarters that are available. For the banks that have never failed in history, we randomly pick data in 6 continuous quarters. We balanced the training dataset after sampling and splitting.

The algorithm is to combine the elastic net and penalized logistic regression with standardized features and the target variable as input. We took an average of the features' estimated coefficients yielded by two regressions and got a list with features sorted by their estimated coefficients.

we then trained an ensemble learning model including the bagging tree, the Gaussian Naïve Bayesian, and the KNN models with K selected features as input, with K ranging from 5 to 10 inclusively. The model's accuracy on the testing dataset is approximately 97%, which proves the success of our feature selection.

Top 5 features: ['ILNDOMQR' 'LNLSNETR' 'ASSTLTR' 'ROA' 'total\_loans\_equity']

Chart

Description automatically generated

For k = 5 the Ensemble accuracy: 0.9445809177557237

For k = 5 the probability: [[0.97787021 0.02212979]

[0.89426518 0.10573482]

[0.93804997 0.06195003]

...

[0.97262073 0.02737927]

[0.94266025 0.05733975]

[0.67787129 0.32212871]]

For k = 5 the probability: [0 0 0 ... 0 0 0]

Top 6 features: ['RBC1AAJ' 'ILNDOMQR' 'LNLSNETR' 'ASSTLTR' 'ROA' 'total\_loans\_equity']

A picture containing chart

Description automatically generated

For k = 6 the Ensemble accuracy: 0.960499164783335

For k = 6 the probability: [[0.99456277 0.00543723]

[0.93512551 0.06487449]

[0.96772416 0.03227584]

...

[0.9916891 0.0083109 ]

[0.97294648 0.02705352]

[0.8603654 0.1396346 ]]

For k = 6 the probability: [0 0 0 ... 0 0 0]

Top 7 features: ['RB2LNRESR' 'RBC1AAJ' 'ILNDOMQR' 'LNLSNETR' 'ASSTLTR' 'ROA'

'total\_loans\_equity']

A picture containing table

Description automatically generated

For k = 7 the Ensemble accuracy: 0.9646261177164194

For k = 7 the probability: [[0.99596388 0.00403612]

[0.8624328 0.1375672 ]

[0.96493965 0.03506035]

...

[0.98649081 0.01350919]

[0.98000326 0.01999674]

[0.87230799 0.12769201]]

For k = 7 the probability: [0 0 0 ... 0 0 0]

Top 8 features: ['LNATRES' 'RB2LNRESR' 'RBC1AAJ' 'ILNDOMQR' 'LNLSNETR' 'ASSTLTR' 'ROA'

'total\_loans\_equity']

Graphical user interface, application

Description automatically generated

For k = 8 the Ensemble accuracy: 0.9714061118207723

For k = 8 the probability: [[0.99594447 0.00405553]

[0.87314198 0.12685802]

[0.96812866 0.03187134]

...

[0.98556621 0.01443379]

[0.98490013 0.01509987]

[0.87810001 0.12189999]]

For k = 8 the probability: [0 0 0 ... 0 0 0]

Top 9 features: ['SCMTGBKR' 'LNATRES' 'RB2LNRESR' 'RBC1AAJ' 'ILNDOMQR' 'LNLSNETR'

'ASSTLTR' 'ROA' 'total\_loans\_equity']

Graphical user interface, application, Teams

Description automatically generated

For k = 9 the Ensemble accuracy: 0.972585241230225

For k = 9 the probability: [[0.99480213 0.00519787]

[0.95000432 0.04999568]

[0.97304252 0.02695748]

...

[0.98605134 0.01394866]

[0.98798195 0.01201805]

[0.95186536 0.04813464]]

For k = 9 the probability: [0 0 0 ... 0 0 0]

Top 10 features: ['SCSNHAFR' 'SCMTGBKR' 'LNATRES' 'RB2LNRESR' 'RBC1AAJ' 'ILNDOMQR'

'LNLSNETR' 'ASSTLTR' 'ROA' 'total\_loans\_equity']

Graphical user interface, application, Teams

Description automatically generated

For k = 10 the Ensemble accuracy: 0.9707182863319249

For k = 10 the probability: [[0.98970708 0.01029292]

[0.91073845 0.08926155]

[0.95841592 0.04158408]

...

[0.97316841 0.02683159]

[0.98103013 0.01896987]

[0.93815099 0.06184901]]

For k = 10 the probability: [0 0 0 ... 0 0 0]