Statistical Methods to Predict Industrial Stock Prices

Team members:
Jiahao Zhang/ Pinren Chen/ Hongjie Ren
Peilin Qiu/ Sihuai Yu/ Tangming Li/ Yufei Zhao
Yangyu Niu

Date:
April 29th (Saturday)

Instructor:
Professor Zhiliang Ying

Project Introduction

Why this project?

Stock Market ---- Highly Sophisticated Market

Traditional Scenario:

Fundamental analysis(qualitative analysis)

Basic technical indicators(MACD, Bollinger)

What We Can Do Now:

Statistical Models and Math Methods

Simulation Technology

Better Prediction
More Profits

Why based on industry?

- Most models are designed for the whole market
- Industries are inherently different in many aspects
- Different models are suitable for different situations

A Quick Peek at the 8 Methods

1. C4.5 Decision Trees
2. Random Forest
3. Naive Bayes
4. Logistic Regression
5. SVM
6. Neural Network
7. Linear Discriminant Analysis

Basic idea:

1. Acquire data of **6 industries** from Bloomberg Terminals and initially choose **47 indexes**, including micro indexes and macro indexes
2. Apply **seven classification statistical models**(machine learning algorithm)
3. **Predict the 3-month returns** on 2017.3.31 using the **quarterly data** from 2009.3.31 to 2016.12.31
4. Acquire the **real 3-month returns** from Bloomberg Terminals and compare the predicted value we calculated with these real values

Definition of Accuracy

- If return is greater than or equal to 10%, denote as “Good”
- Otherwise, denote as “Bad”
- If predicted class is the same as true class, denote as “1”
- Otherwise, denote as “0”
- Define “ Accuracy ” as follows:

$$Accuracy = \frac{Total\ Number\ of\ "1"}{(Total\ Number\ of\ "1" + Total\ Number\ of\ "0")}$$

How we choose the data

- Micro Indexes:

Tangible book value per share,
current market capitalization,
change of net income,
dividend yield, earnings per share,
net revenue, sales growth,
price to earnings ratio,
current ratio, book value per share,
asset turnover, net revenue growth,
currency ratio, PE ratio...

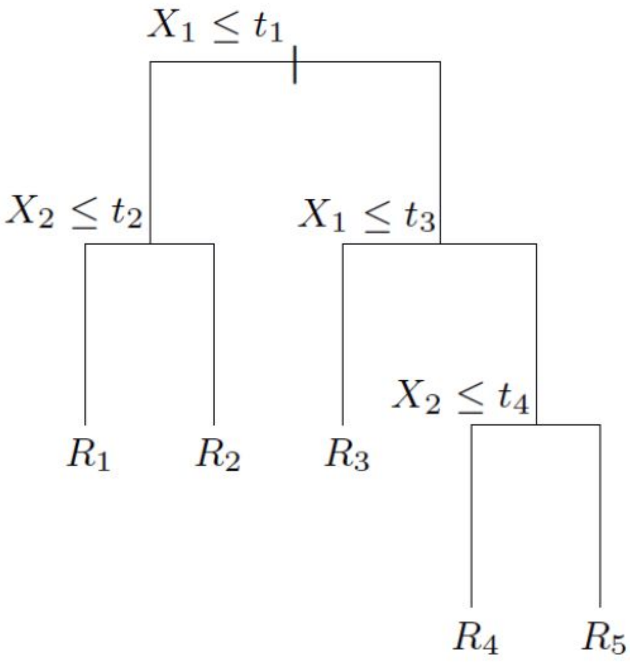
- Macro Indexes:

Velocity of M1,
Velocity of M2,
Gross Domestic Product Index,
Consumer Price Index,
US Trade Balance of Goods and Services,
Amount of import and export,
Unemployment rate
Personal Consumption Expenditure...

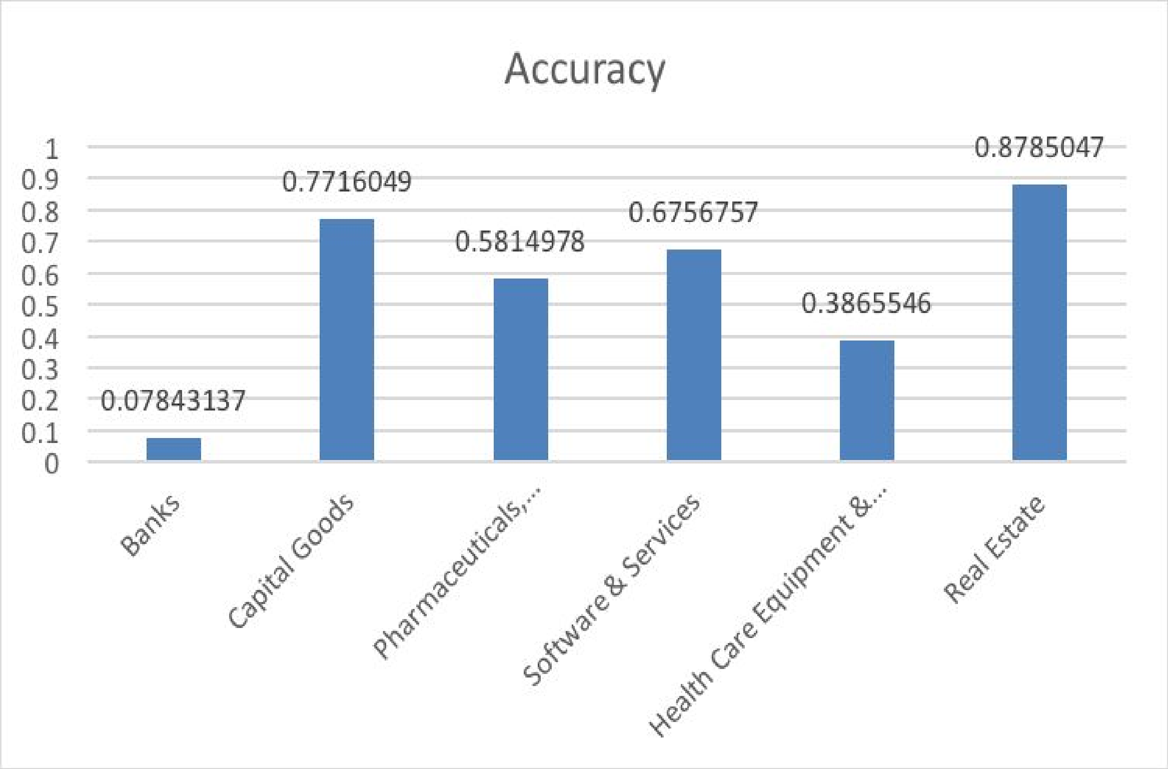
Statistical Model Introduction

C4.5 Decision Trees

- A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility.
- Commonly used in operations research and operations management.



C4.5 Decision Trees



Advantages:

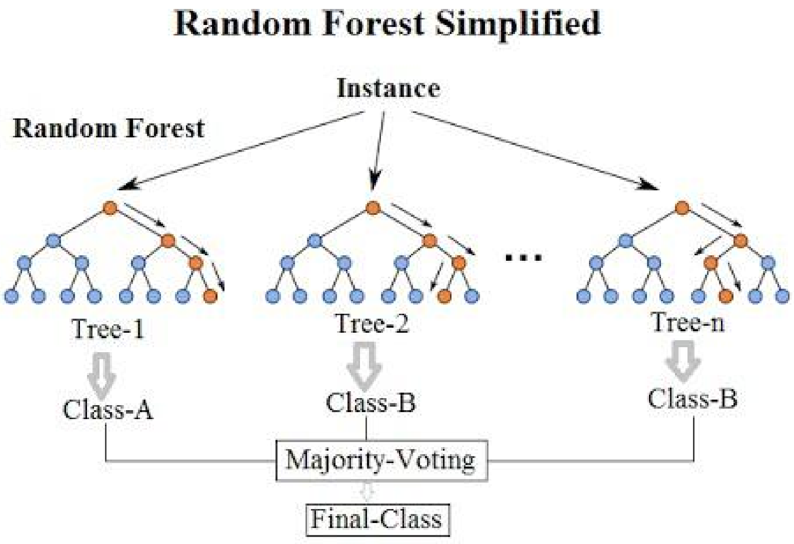
- Able to handle both numerical and categorical data
- Performs well with large datasets

Disadvantages:

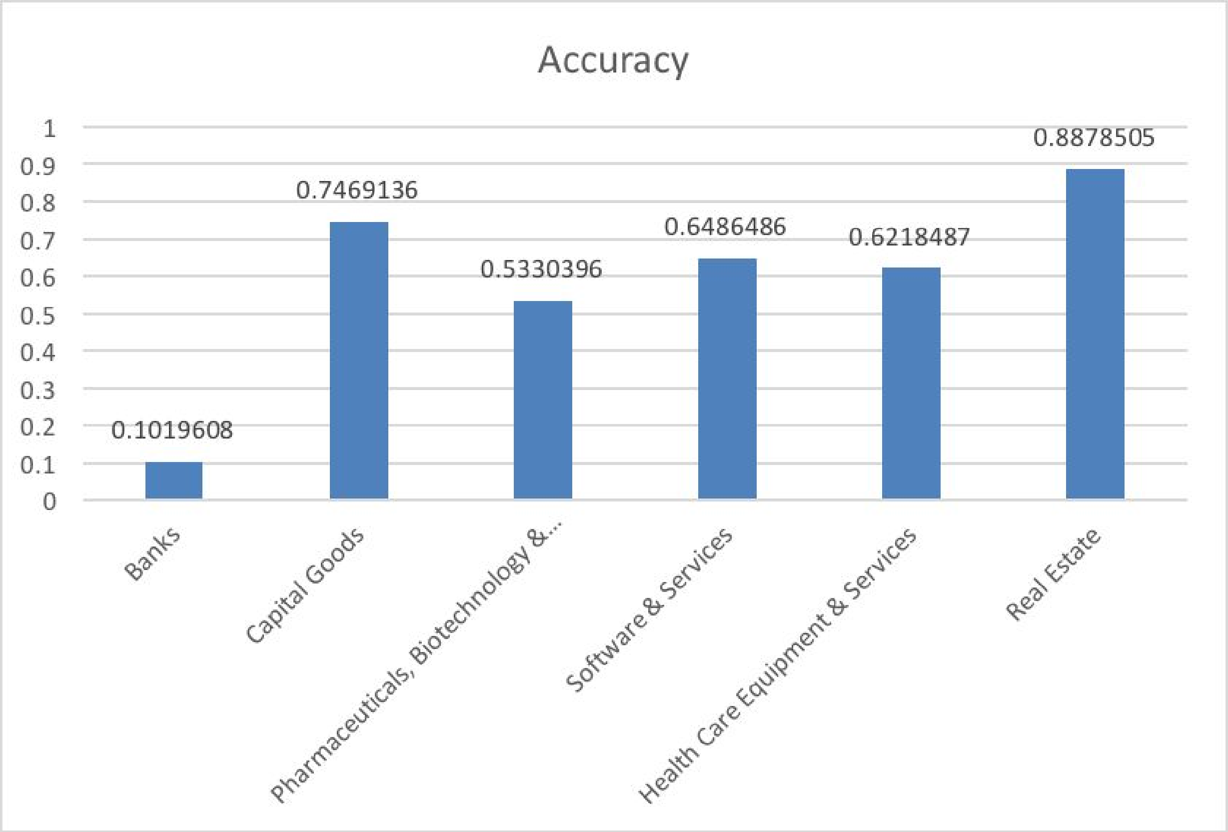
- Trees can be very non-robust
- Decision trees is biased in favor of attributes with more levels

Random Forest

- Fit a decision tree to different Bootstrap samples
- Select a random sample of m<p predictors to consider in each step when growing each tree
- This will lead to **uncorrelated** trees for each sample
- Grow 500 trees in total
- Finally, average the prediction of each tree



Random Forest



Advantages:

- Improve accuracy if 50% of the data are classified correctly in each tree

Disadvantages:

- Computationally intensive
- Overfitting Risk

Naive Bayes

In machine learning, Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes's theorem.

<u>Main Assumption:</u>

Features are independent with each other

Simplest form

- Random variables $X \in \mathbf{X}$ and $Y \in \mathbf{Y}$, where $\mathbf{X}, \mathbf{Y}$ are finite sets.
- Each possible value of X and Y has positive probability.

Then

$$P(X = x, Y = y) = P(y|x)P(x) = P(x|y)P(y)$$

and we obtain

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)} = \frac{P(x|y)P(y)}{\sum_{y \in \mathcal{Y}} P(x|y)P(y)}$$

It is customary to name the components,

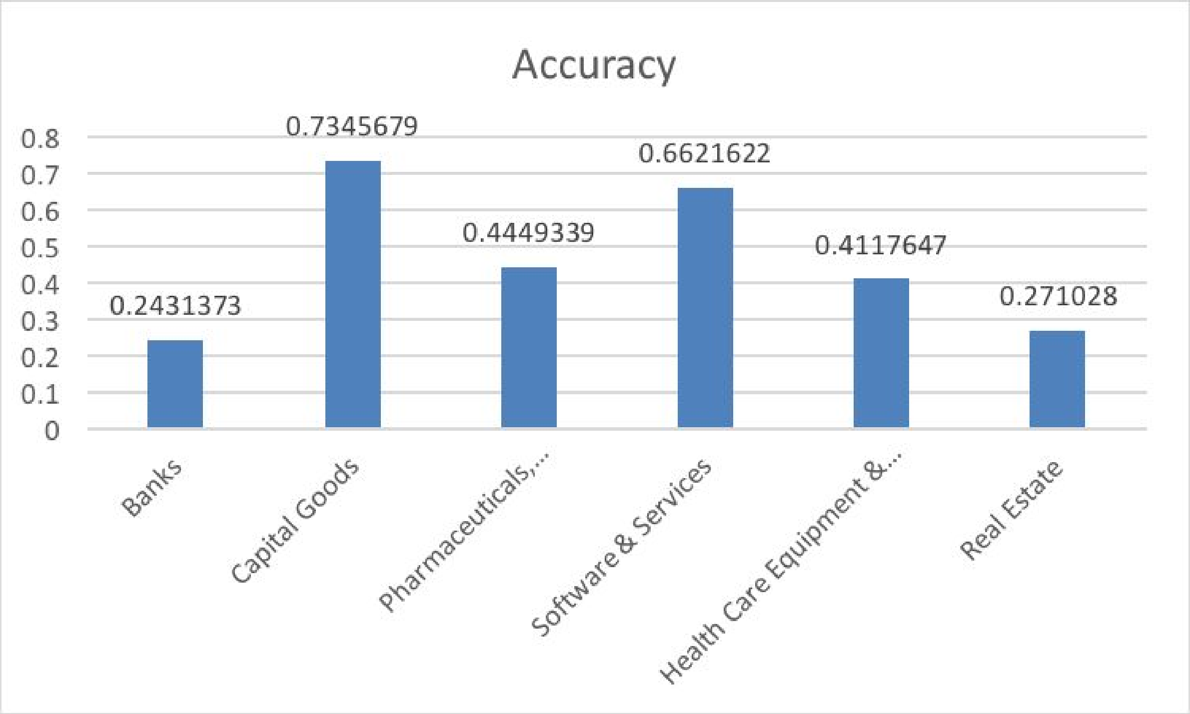
$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$

In terms of densities

For continuous sets $\mathbf{X}$ and $\mathbf{Y}$,

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} = \frac{p(x|y)p(y)}{\int_{\mathbf{Y}} p(x|y)p(y)dy}$$

Naive Bayes



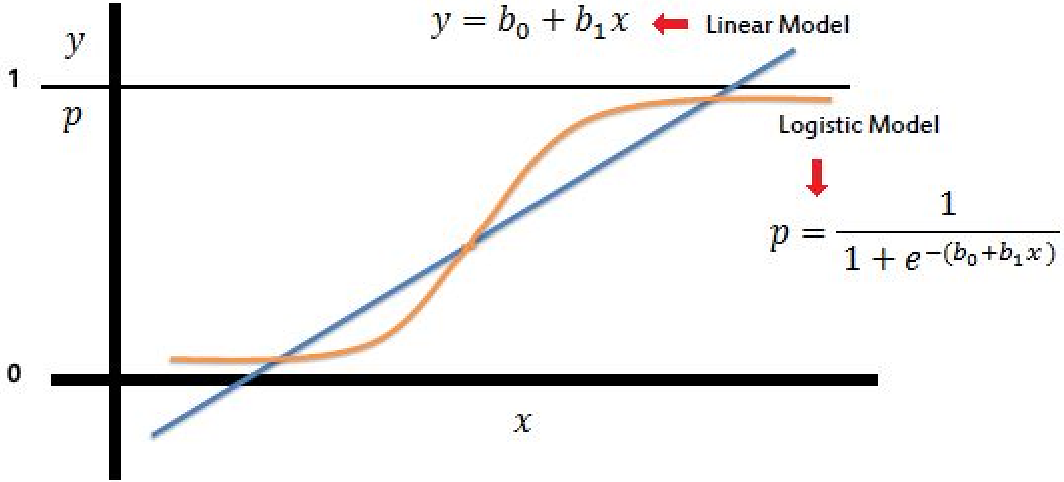
Advantages:

- Computing very fast

Disadvantage:

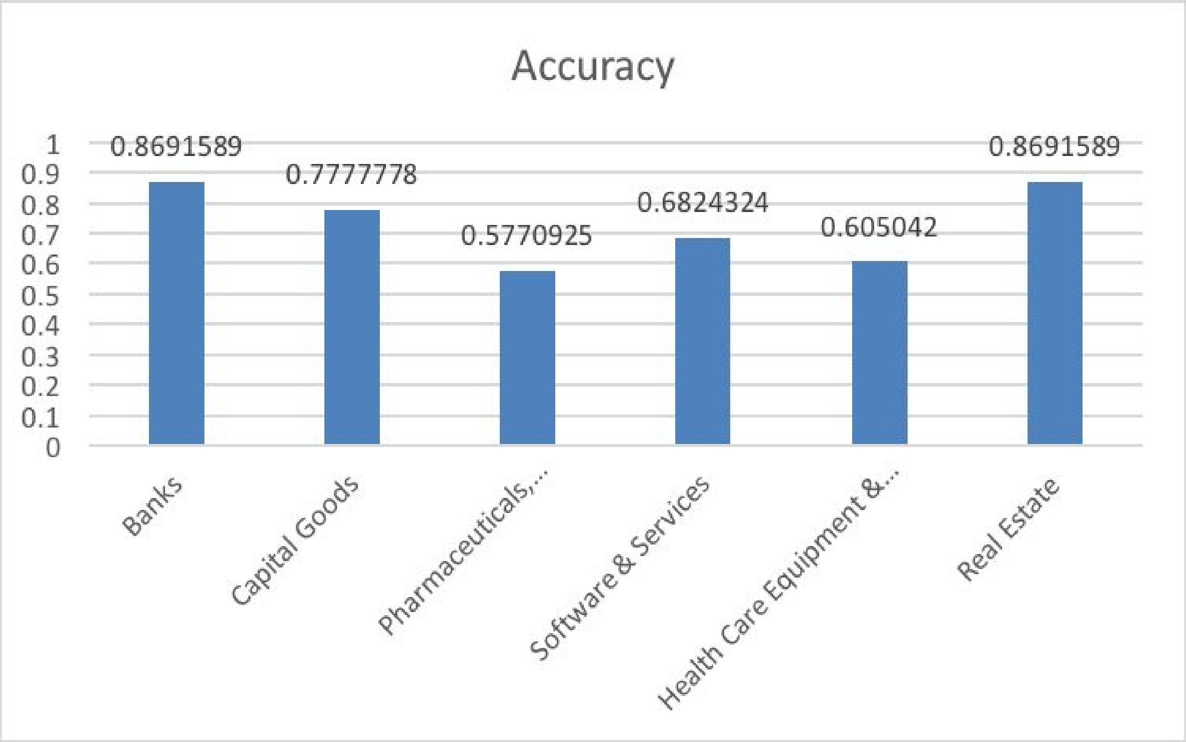
- Response has to be categorical
- Independent Features Required

Logistic Regression



- Logistic regression is a statistical method for analyzing a dataset in which there are one or more **independent variables** that determine an outcome. The outcome is measured with a dichotomous variable.
- Mainly used in medical and pharmaceutical area

Logistic Regression



Advantages:

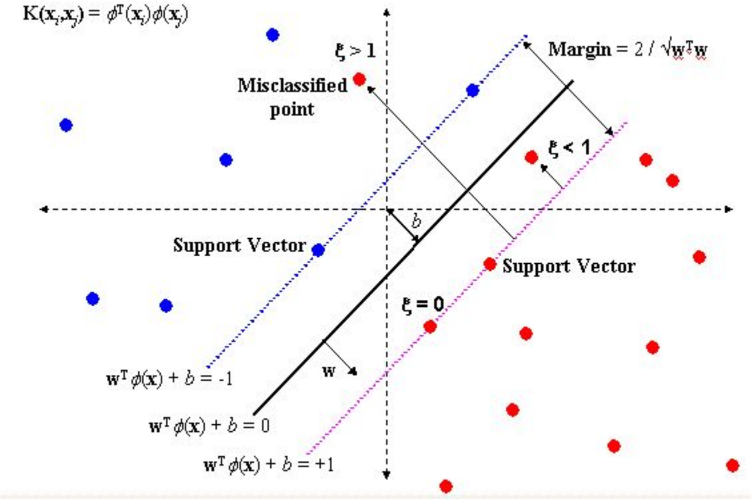
- Easy to implement
- Variables can be either categorical or continuous

Disadvantages:

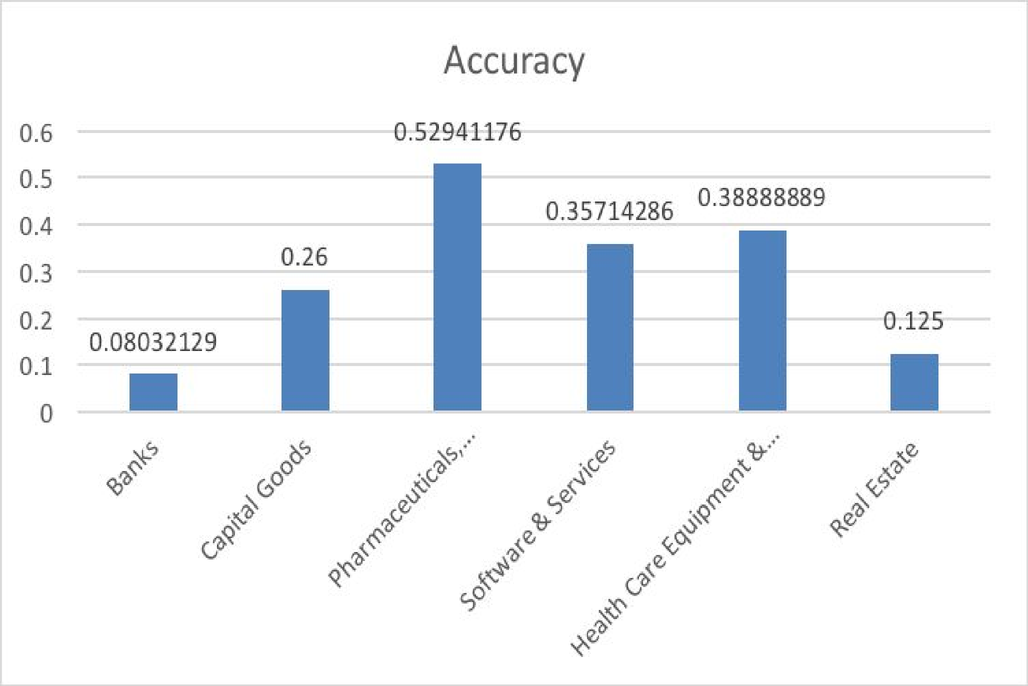
- Multicollinearity risk
- Limited Outcome Variables
- Demand Independent variables

Support Vector machine

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples.



Support vector machine



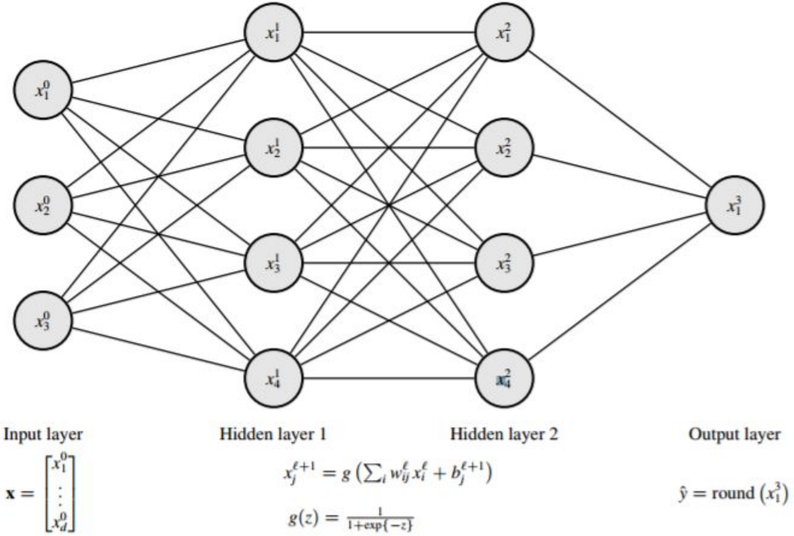
Advantage:

- provide a good out-of-sample generalization
- With the choice of an appropriate kernel, one can put more stress on the similarity between companies

Disadvantage:

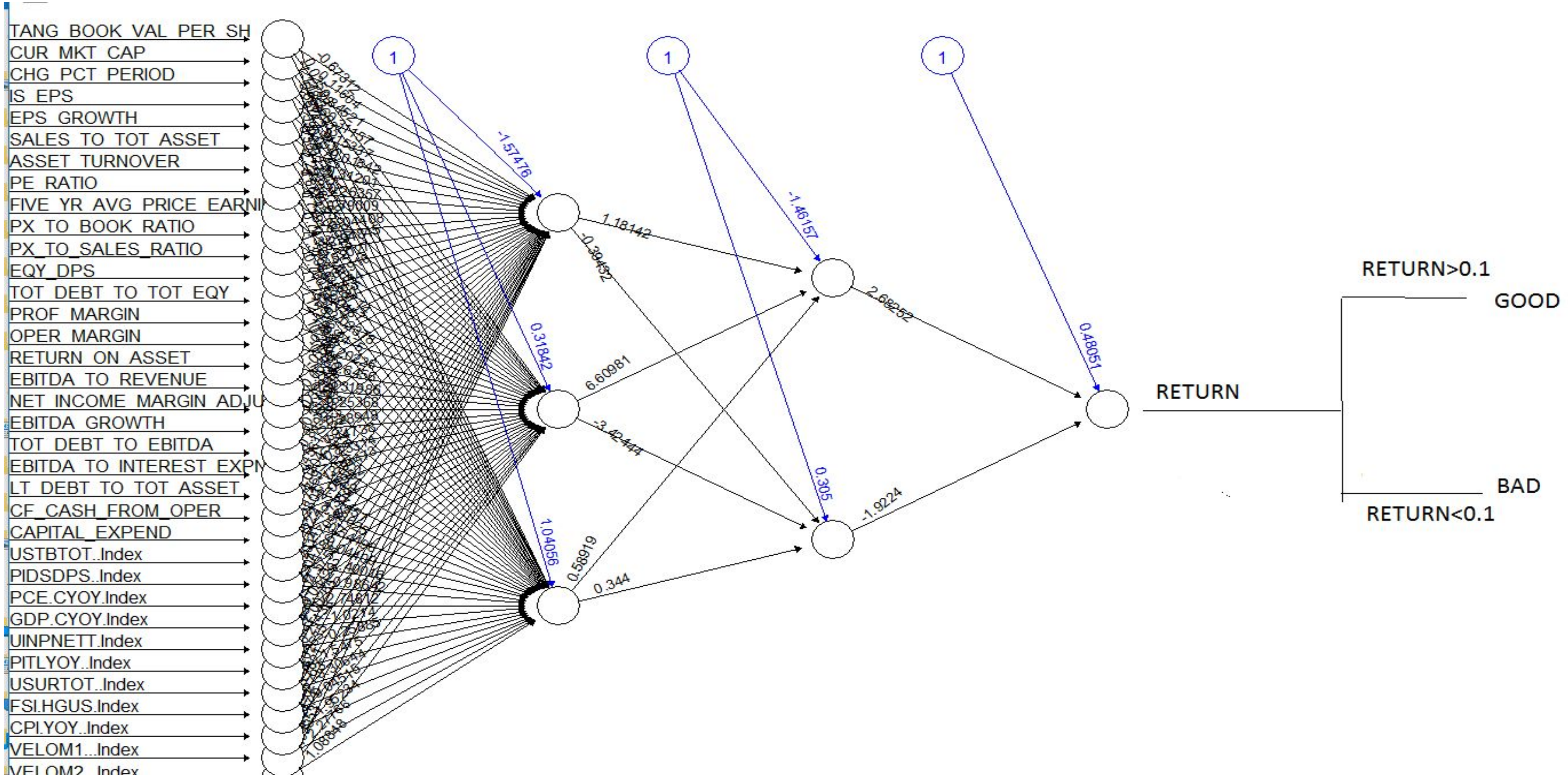
- SVMs cannot represent the score of all companies as a simple parametric function

Neural Network

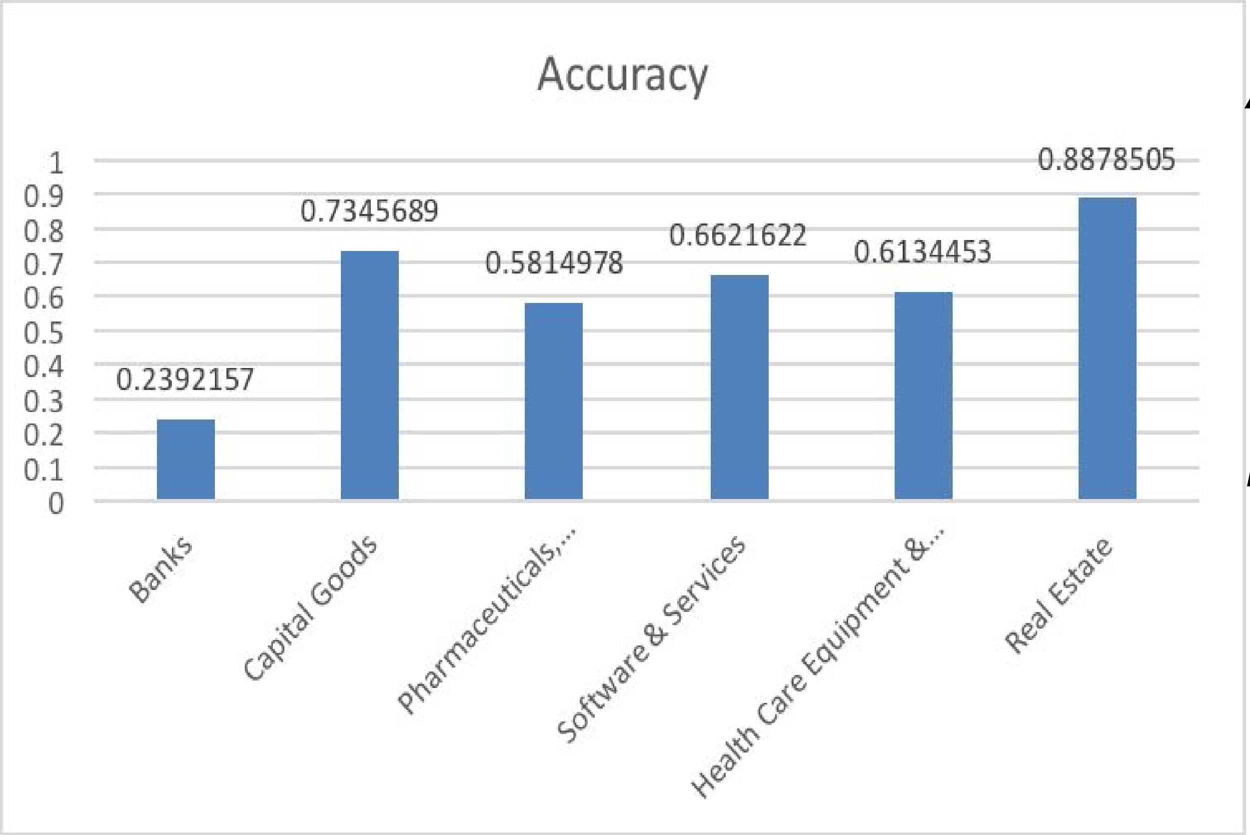


- Given weights $W = \left\{w_{ij}^\ell, b_j^\ell\right\}_{i,j,\ell}$, this is just a classifier $f_W : \mathbb{R}^d \to \{0, 1\}$.
- $\hat{y} = x_1^3 \to$ regression network (or $\hat{y} = \left[x_1^L \dots x_D^L\right]^\top$).
- As with every other method, the work is to optimize the w_{ij}^ℓ and b_j^ℓ.

Neural Network



Neural Network



Advantage:

- Dynamic Non-Linear System
- A good model for the dynamic stock market

Disadvantage:

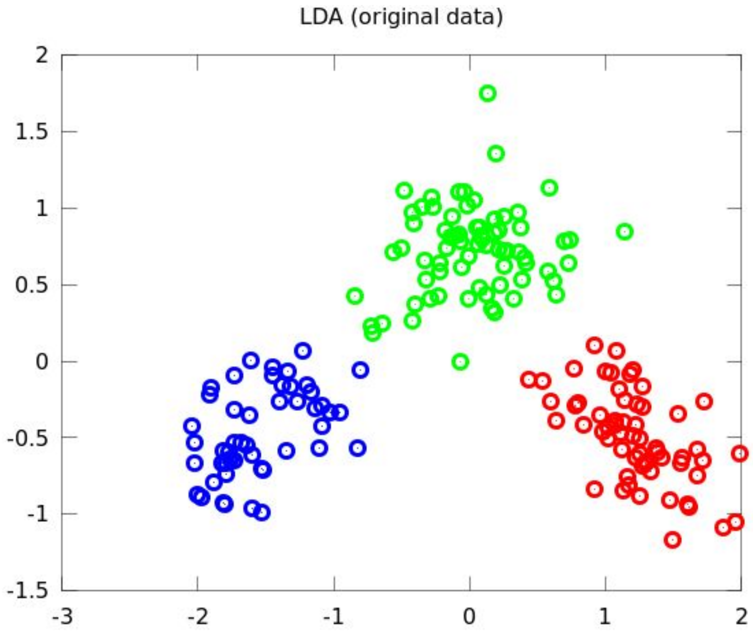
- Computational Complex
- Low Efficient

Linear Discriminant Analysis

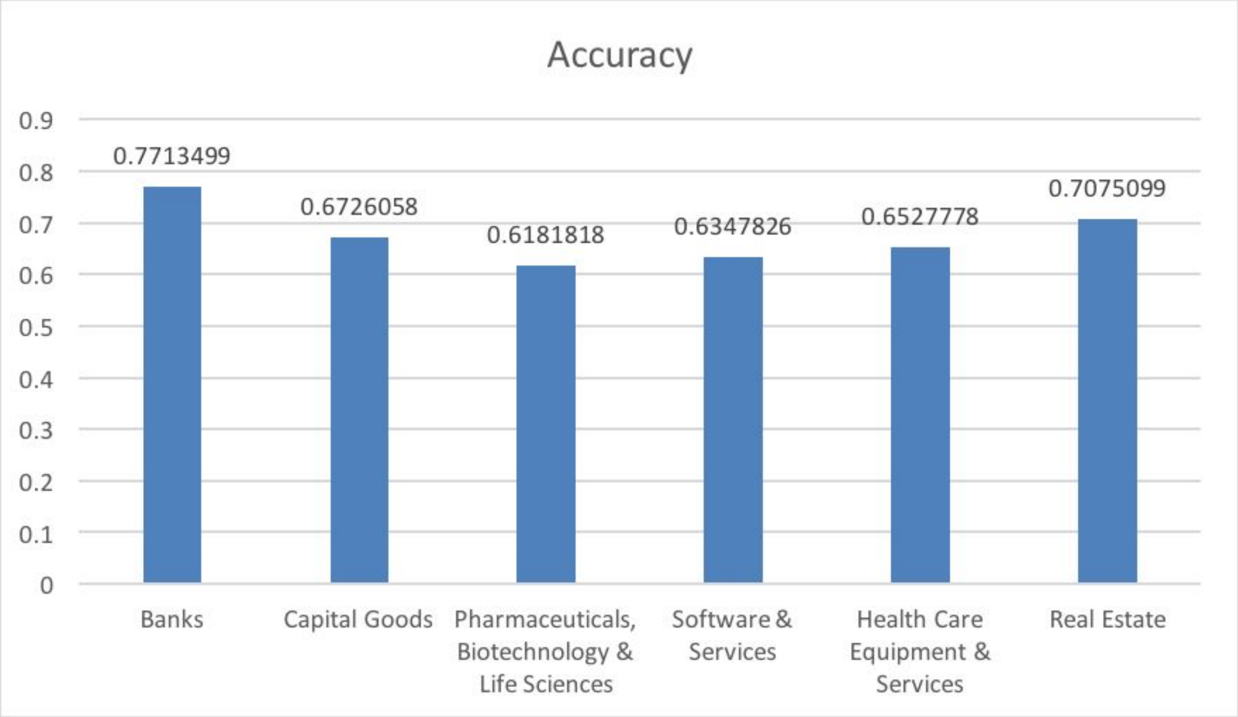
- Linear discriminant analysis(LDA) is a generalization of Fisher’s Linear Discriminant, a method used in Statistics Pattern Recognition and Machine Learning.

Two assumptions:

- Gaussian distributed classes
- Equal class covariance.



Linear Discriminant Analysis



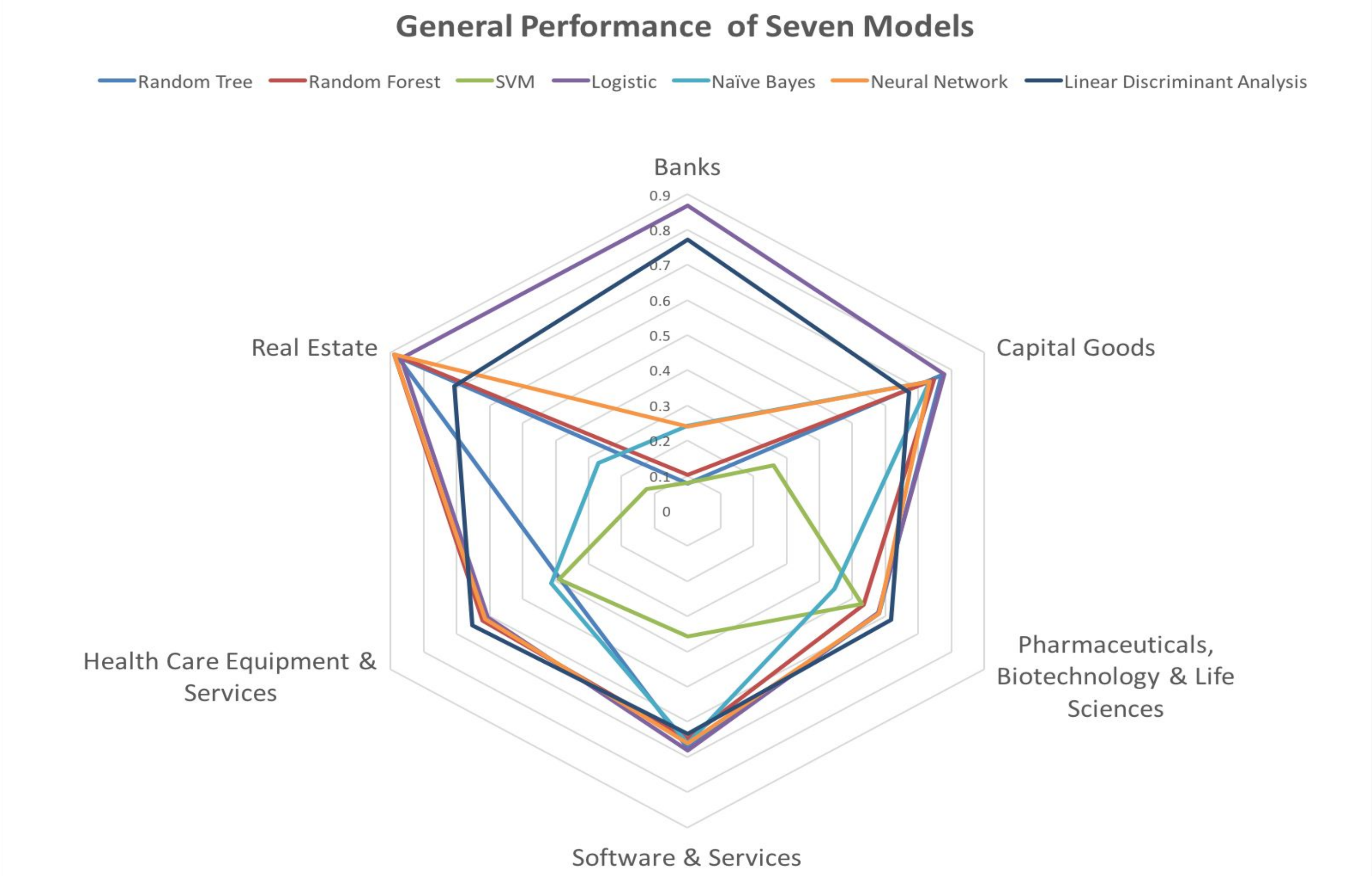
Advantage:

- Performs better with small samples with many variables

Disadvantage

- Not flexible (Linear boundaries only)

Results Analysis



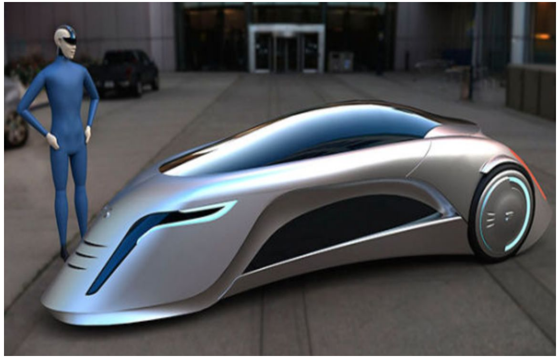
General Performance of Seven Models
Random Tree
Random Forest
SVM
Logistic
Naïve Bayes
Neural Network
Linear Discriminant Analysis
Banks
Capital Goods
Pharmaceuticals, Biotechnology & Life Sciences
Software & Services
Health Care Equipment & Services
Real Estate
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0

Conclusion

Industry	Best Model
Banks	Logistic
Capital Goods	Linear Discriminant Analysis
Pharmaceuticals, Biotechnology & Life Sciences	Linear Discriminant Analysis
Software & Services	Logistic
Health Care Equipment & Services	Linear Discriminant Analysis
Real Estate	Neural Network

Moving On

- Mixture model analysis-- Adaboost
- Consider inter-industrial impacts
- Consider international market impacts
- Consider political factors (text mining)
- More and more!!



Thank you!

Q & A