# ECE535 Project Final Report

# Teammate 1: Huan-Rui Zhang

# Teammate 2: Anonymous for privacy reasons

**Section 1: Introduction**

The paper predicted the long-term stock trend by considering sixty features and by applying a nonlinear model instead of linear models to tackle with many factors that may affect the stock price.

This algorithm can be used by individual investors, investment management companies, stockbrokers, or academic researchers. The algorithm can be designed that the brokers have additional information that consumers would not. There are sixty features of each measurement will be considered, which are classified into ten factors: validation, growth, financial quality, leverage, size, momentum, volatility, turnover, liquidity, and technical factors. This paper did not mention which platform the algorithm be implemented, but most of the stock trend predicting models are implemented on Cloud computing platform; the users can access the prediction results through the Internet, as the computational complexity of such model typically very high. Therefore, the model is not suitable for implementing on edge computing device such smartphone since low power consumption is the most important consideration in smartphone.

In terms of special considerations, the paper only considers for the China stock market; the performance of the paper may vary if it applied to other stock markets. Another two limitations mentioned by the authors are the improvement of feature selection algorithms and exploration of new features. For instance, I think the policy of the government toward stock market might implicitly affect the emotion of investors; this effect is currently not modelled by the sixty features yet. Hence, it is a good idea to explore this kind of features. However, there are two ethical considerations in this paper. Firstly, even the proposed model in this paper is great, the model may fail in the future; as stock markets can be affected by many factors. We think that is why the paper mentioned they need to improve the feature selection algorithms and to explore more new features. Secondly, the failure prediction of the proposed model may cause investors to lose money. Hence, the designer of the model should let users know the risk of losing money even if the model has a high prediction accuracy.

In this report, we will describe the data set collection and processing in Section 2. The details of feature selection will be stated in Section 3. Section 4 explains machine learning algorithm used in the paper. The performance evaluation will be discussed in Section 5. Finally, the conclusion of this report will be given in Section 6.

**Section 2: Data Set Collection/Processing**

The sources of data are the stock information from all the Chinese A-share market from January 1, 2010 to January 1, 2018 from the Wind database. To obtain the training data and testing data, a sliding window is applied to split the original dataset into training and testing sets. The author chose 60 features and categorized the features into 10 categories, as listed in Table 1.

To improve the full representation of the distribution of measurements, the paper excluded special treatment stocks and sub-new stocks with less than 3 months trading days as they are riskier to predict. Moreover, the paper used the stock excess returns to be the predict target to avoid the influence of the overall market trend. The stock excess returns are computed by using the stock return minus the return of the Shanghai Stock Index. In terms of the overall stock market trend, it means that when the stock is in a bear market or bull market then the overall stocks will drop or rise, respectively.

There are three data pre-processing schemes, including outlier removal or normalization, used in the paper:

Data preprocessing method 1: Process extreme values of feature data

Data preprocessing method 2: use the average value of the same stocks in the same industry to fill missing feature data

Data preprocessing method 3: normalizes by zero-norm, unit variance

These three data preprocessing methods were applied to mitigate the effect on the model. The first and second data preprocessing methods were based on observations toward the outliers and the inherent defect values like missing feature data. The third data preprocessing method, normalization, is very commonly used in machine learning field. To the best of our knowledge, the authors did not mention how they measured the effectiveness of preprocessing. They removed the extreme values to prevent the model learns those abnormal cases, filled the missing values to solve data defects from the source websites, and normalized the features to mitigate the dominant features. The three data preprocessing methods will not remove any useful features. Firstly, processing extreme values of feature data is just a clipping operation for feature value larger or less than the threshold. Hence, it will remain the overall non-extreme values the same as before. Secondly, using the average value of the same stocks in the same industry to fill missing feature data, which will not remove any useful features; this is a widely used data wrangling scheme. Thirdly, normalization by zero-norm and unit variance adjusts the value range of the features into a better way. Therefore, the useful features remained.

The novel features for the data set collection/processing for this system is that the authors adopted to use 60 features to be the input of the model. The 60 features will be further selected for training and testing the model according to the feature selection algorithms, which will be described in Section 3 in this report.

Table 1. Original 60 features for the stock market data set. [1]

| Category | Description of features |
|---|---|
| Validation factors | Net profit(TTM)/Total market value |
| | The net profit after extraordinary gains and losses(TTM)/Total market value. |
| | Net asset(TTM)/Total market value |
| | Operating income(TTM)/Total market value |
| | Net cash flow(TTM)/Total market value |
| | Operating cash flow(TTM)/Total market value |
| | Dividend payable(TTM)/Total market value |
| | Net profit(TTM) compared with the same period of last year/PE_TTM |
| Growth factors | Operating income growth rate compared with the same period of last year. |
| | The net profit after extraordinary gains and losses growth rate compared with the same period of last year. |
| | Operating cash flow growth rate compared with the same period of last year. |
| | ROE compared with the same period of last year. |
| Financial quality factors | ROE(QTD) |
| | ROE(TTM) |
| | ROA(QTD) |
| | ROA(TTM) |
| | Gross profit margin(QTD) |
| | Gross profit margin(TTM) |
| | The net profit margin after extraordinary gains and losses(QTD) |
| | The net profit margin after extraordinary gains and losses(TTM) |
| | Operating cash flow/Net profit(QTD) |
| | Operating cash flow/Net profit(TTM) |
| | Inventory turnover(YTD) |
| | Inventory turnover(TTM) |
| | Total assets turnover(YTD) |
| | Total assets turnover(TTM) |
| Leverage factors | Fixed assets ratio |
| | Total assets/Net assets |
| | Non-current liabilities/Net assets |
| Size factors | *ln*(circulation market value) |
| | Circulation market value/Total market value |
| | *ln*(total market value) |
| Momentum factors | Return rate 1-month |
| | Return rate 3-month |
| | Return rate 6-month |
| | Return rate 12-month |
| | Daily turnover rate* daily return rate 1-month |

| | |
|---|---|
| | Daily turnover rate* daily return rate 3-month |
| | Daily turnover rate* daily return rate 6-month |
| | Daily turnover rate* daily return rate 12-month |
| Volatility factors | Highest price/Lowest price 1 month |
| | Highest price/Lowest price 3 month |
| | Highest price/Lowest price 6 month |
| | Highest price/Lowest price 12 month |
| | Std of daily return rate 1 month |
| | Std of daily return rate 3 month |
| | Std of daily return rate 6 month |
| | Std of daily return rate 12 month |
| Turnover factors | Daily turnover rate 1 month |
| | Daily turnover rate 3 month |
| | Daily turnover rate 6 month |
| | Daily turnover rate 12 month |
| Liquidity factors | Current ratio |
| | Quick ratio |
| Technical factors | MACD(10, 30, 15) |
| | RSI(20) |
| | PSY(20) |
| | BIAS(20) |

The authors used appropriate methods for data set collection and processing. The authors used 60 features as its original data and applied the data preprocessing, including clipping extreme values, filling the missing values, and normalization. The reason is that the stock market is very complicated and is very hard to predict it by using only few information; therefore, it is a good idea to take different factors into account for predicting the stock market. We surveyed the papers in predicting stock market fields and found that many of the papers adopted the market information and technical information as the training data. Although technical information like Moving Average Convergence/Divergence (MACD) and Relative Strength Index (RSI) are very useful technical indicators, they still only based on statistical prospective, and they have delay in reflecting the market situation due to their mathematically definitions. In terms of the data set processing, refined the values of extreme values and the missing values are helpful data wrangling techniques, otherwise the performance of the model will possibly be affected by those incorrect data. Normalization reduces the possibility of some feature values have a higher value range leading to dominant over other features; A simpler way of interpreting the meaning of normalization is that treat all the features as the same importance in training the model.

Exploratory Data Analysis (EDA) can be made to the data processing, which investigate to understand the 60 features using descriptive statistics or a visualization way. EDA helps to understand the features in more statistical way. There are several EDA statistics summary methods such as the mean, medium, mode, frequency, percentile, standard deviation, symmetric distribution, and skewed distribution. The visualization methods in EDA cover 1-D histogram, 2-D histogram, Box Plot-Comparison, and scatter plot. If features are highly similar in their EDA descriptive statistics, then some of them may be redundant. By using EDA techniques, the authors may reduce the 60 features before doing the feature selection methods, rather than use the found all 60 features.

**Section 3: Feature Selection**

The authors adopted two feature selection methods: Support Vector Machine-Recursive Feature Elimination (SVM-REF) and random forest (RF). The authors calculated the importance score of each feature and used only the features with the highest 80% importance score. The importance score of a feature in SVM-REF is the square of the weight of that feature. The higher value of weight of a feature means that feature plays an important role. The RF feature selection method calculates the average out-of-bag error of each feature. The out-of-bag error of one feature is defined as the error between the prediction by using the original features and by shuffling one of the features. Therefore, a feature has zero average out-of-bag error means that feature is not important for the prediction. Suppose we have m sample data and n features. Each feature can have its importance score by computing the out-of-bag error over the n features but shuffle itself. The RF feature selection method will choose the features which have the highest 80% importance score in this paper.

To the best of my knowledge, the authors did not explain why they choose the features have the highest 80% importance score. But we think this number can be determined by a better way rather than just a fixed threshold; this problem can be evidenced at the end of the paper that the authors pointed out the decision of the number of features selected still needs to be optimized.

The authors did not use any special steps to deal with invariant problem. The authors showed a novel feature of the feature selection that by choosing the feature with the highest 80% importance score to train the model can achieve higher model performance as people might think the original 60 features are useful for training a model. This method makes me reflect that not each feature is as useful as we think.

Currently, the authors only focus on the stock trend predicting in China stock market, but they can apply the same data set collection/processing method to global stock market such as U.S. Furthermore, it is reasonable to take the economic indicators as a part of the data set for a long-term stock trend prediction. The economic indicators cover from macroeconomic factors to microeconomic factors. The microeconomic factors are information for a company. While the macroeconomic factors are related to a country or the global economy. There are many macroeconomic factors can be used in the stock market analysis applications, such as the Year-of-Year in Consumer Price Index (CPI YoY), unemployment rates, labor force participation rates, United States (U.S) dollar index, federal funds interest rate, U.S. 2-year Treasury Bond Yields, and U.S. 10-year Treasury Bond Yields. Some researchers have explored the relationship between macroeconomic features and the stock market return in other countries. The CPI YoY is an indicator that shows the fluctuations of the living cost by year, which is highly related to the inflation situation. The difference between the 2-year U.S. treasury bond yield and 10-year U.S. treasury bond yield is usually referred the "yield curve spread" or simply the "yield spread". The yield of the 10-year U.S. treasury bond is usually higher than the 2-year U.S. treasury bond. When the yield of the 2-year U.S. treasury bond higher than the 10-year U.S. treasury bond, this is called an inverted yield curve. The inverted yield curve implies that the market is becoming more pessimistic about the economic prospects for the near future, which will possibly affect the stock market.

**Section 4: Machine learning algorithms**

In this article, the author applies several machine learning algorithms for stock price trend prediction, such as Support Vector Machines (SVM), Random Forest (RF), and Artificial Neural Networks (ANN). SVM is a cheap technique for classification, RF is for its high prediction accuracy, and ANN can capture complex nonlinear relationships. All these algorithms can dive into complex financial market environments to predict the long-term trends for the Chinese stock market.

- Support Vector Machines

The idea of SVM, from what we learned in class, is to maximize the margin hyperplane, which can effectively classify the sample data into two different categories. The author states that the use of the SVM is because of its strong ability to manage high-dimensional data [1]. The stock data is very complex and stochastic, which is common in the financial sector. By leveraging the power of the SVM classifier, it is capable of handling non-linear decision boundaries through kernel functions. Thus, the non-linear SVM can align well with the complex nature of the financial markets. I agree with the author about using SVM for prediction because of its low cost, efficiency, and effectiveness.

The SVM's parameters are mainly the regularization parameter C and the kernel function's parameter γ. The C is tested from 0.0001 to 1, and the gamma is tested from 0.001 to 10, and every time, it is increased by ten times. All two parameters are tested in pairs. In order to get the optimal parameters, the time window slicing cross-validation is deployed to help the SVM get the result for each model and determine the best-performing model. However, the setup of the SVM model with optimal regularization parameter C and the kernel function's parameter γ can be computationally intense. Focusing on the high dimensional financial data with non-linear kernels, this makes SVM less suitable for the embedded system with limited computational capacity. Thus, simpler kernels or feature reduction techniques are used to reduce computational complexity.

- Random Forest

RF is composed of multiple decision trees. There is no correlation between each decision tree because random sampling is used to obtain the sub-datasets. The final result will be the decision tree with the most classification. Due to this step, the RF can avoid overfitting problems, and it is suited for stock trend prediction. The model can generalize well from historical data for future data and avoid the algorithm sticking to the greedy local maximum. I agree with the author that choosing RF for trend prediction is a reasonable option because of its stability and accuracy.

For RF, parameters such as the number of decision trees, the depth of the trees, and the maximum number of features for each tree can be considered at each iteration. In this paper, the author reduces the complexity of the model and sets the number of decision trees to 100. In order to determine the maximum number of features in each decision tree, the author tests different values by using 2, 5, 10, 50, and 100 in pairs to find the optimal minimum number of samples for the internal node and a minimum sample of the leaf node. RF can be computationally intensive due to storing numerous decision trees and managing multiple datasets simultaneously during training. For deployment on embedded systems, reducing the number of trees and depth of each tree might mitigate memory, and developing a parallel structure might reduce computational cost and maintain accuracy.

- Artificial Neural Networks

ANN is widely used for stock price trend prediction for its strong ability to develop a nonlinear model. In this paper, the author constructs a three-layer, fully connected neural network model to capture the random walk trend in the stock market. The use of ANNs in financial prediction through deep learning architectures allows the algorithm to learn detailed features from large volumes of data. I agree that using ANN is one of the trend prediction tools because of its powerful ability to learn from complicated environments. Meanwhile, traditional methods might be unable to capture all relevant patterns to make the correct prediction.

ANN used in the paper involves more parameters than other methods, such as the number of hidden layer neurons, optimization function, and regularization term. To maximize learning the features, the author set neurons for the first hidden layer to 20 and the second hidden layer to 10. After that, the regularization penalty term runs from 0.00001 to 0.1, and the max iteration is set from 100 to 1000. These parameters are determined through cross-validation to ensure the network architecture is sufficiently complex to model the nonlinear relationships in the data without being overly overfitting. Because ANN involves more tuning parameters like the number of layers, neurons, activation function, and penalty, all these parameters present the complexity of the model. Deep learning networks require vast computational resources and more time to process, making ANN a challenge to implement on embedded platforms.

Although the authors did not mention which computational platform would be implemented, in practice, this kind of system is likely to be implemented on the cloud. Utilizing cloud services, we can simplify workflows by automating parameter tuning, model training, and deployment and help us manage computational costs. During the finding of the optimal parameters, the time window slicing cross-validation is applied to all three algorithms. This approach ensures that the method can operate effectively in predicting stock trends by avoiding the pitfalls of overfitting while maintaining robust generalization capabilities. I agree with the selection of these algorithms due to their strong capabilities in handling diverse and complex datasets in stock markets. SVM is able to deal with high-dimensional spaces, RF can make predictions accurately without overfitting, and ANN is capable of modeling non-linear relationships, making them suitable choices for predicting stock price trends. These features ensure that the model can not only learn from the historical data but also be able to generalize well for future data.

**Section 5: Performance evaluation**

The performance evaluation for the models is divided into two parts. Firstly, the authors tested the stock price trend prediction models using classification metrics like accuracy and AUC derived from a confusion matrix. These two metrics provide a comprehensive measurement independent of thresholds. Furthermore, the authors back-tested historical data to evaluate how the model performs under real-world stock market data. They applied financial indicators such as annualized returns, Sharpe ratio, and maximum drawdown, which are critical for evaluating the model's practical application in financial markets. The evaluation sets contained historical stock data from January 1, 2010, to January 1, 2018. The authors organized this data into training and testing datasets using a sliding window for one month, which used one year of data for training and a consequent month for testing. Additionally, to maintain the independence of the evaluation sets from the training data, the sliding window strategy dynamically and sequentially adjusts the data windows for each data set without any overlap, as shown in Figure 1. This method maintains the model's integrity and ensures evaluation sets that are continuously updated and provide accurate predictions because the latest stock information is more critical when making decisions.

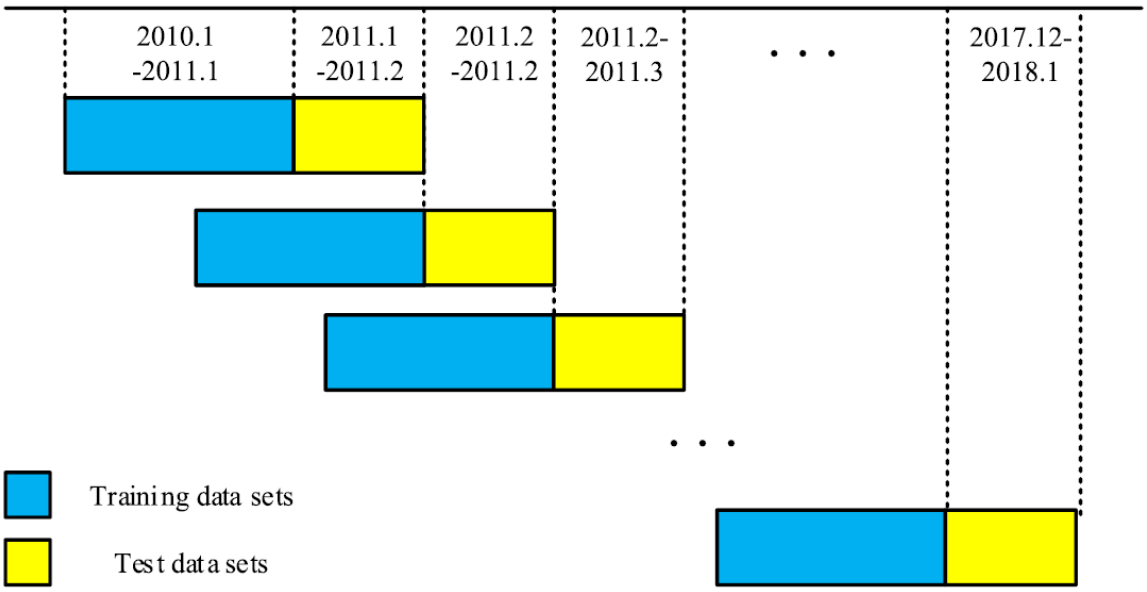


Figure 1. Illustration of sliding window of each case. [1]

The authors primarily used the training and testing data sets, which were organized by the sliding window by one month from January 1, 2010, to January 1, 2018. They did not require other sets for final model evaluation. Because the financial data is time series data, they applied time window slicing cross-validation to generate the sequential segments of the dataset for training and immediate subsequent segments for testing to learn model parameters. This kind of cross-validation is particularly appropriate for financial data, as the temporal sequence and latest data availability are critical. The authors determine the validity of their evaluation by applying several metrics, such as accuracy, AUC, and financial indicators, such as annualized returns and maximum drawdowns. These metrics not only evaluated the predictive accuracy of the model but also showed its utility and risk-adjusted performance in real trading scenarios. In addition, they integrated a long-short trading strategy based on the selected model and showed the overall performance. Thus, in terms of accuracy rate, AUC, annualized return, win rate, and profit-loss ratio, the RF-RF model, which is a random forest model integrated with random forest feature selection, performed the best [1].

The authors pointed out that further evaluations are necessary before their algorithm can be generally deployed. They would conduct more testing on overseas stock markets, optimize the feature selection algorithm, and explore new features to make the model more predictable. During the performance evaluation, the authors use the time window slicing cross-validation to process the time series data. Additionally, integrating financial indicators into the evaluation process and taking back-testing provides a holistic view of the model's performance. In my opinion, the authors properly evaluated the performance of their models through statistical measurements like accuracy, AUC, and long-short strategy performance based on the integrated model. Because of the fluctuating nature of the stock market, sometimes, the model only learns the data patterns of the uptrend or downtrend. So, using a longer period of data could give more accurate measures.

**Section 6: Conclusion**

Overall, the system design within the paper is well organized. The authors deployed feature selection and integrated it with the Support Vector Machines, Random Forest, and Artificial Neural Networks to predict stock price trends. In order to handle the financial time series data, they utilize time window slicing cross-validation to guarantee the framework is assessed under real-world scenarios. To analyze the execution results of different models, they utilized statistical metrics and financial indicators to assess the empirical experiment and found that RF-RF, which is a random forest model with the random forest feature selection, got the best performance. In this way, I concur with the conclusions drawn by the authors. This article clearly interpreted the integration of feature selection with machine learning, which is tailored for the Chinese stock market. They have detailed data preprocessing, parameter tuning, and assessment methodologies. The measurements robustly assessed the system's adequacy and generated the portfolio with the chosen algorithm to show how it works in real-world scenarios.

Furthermore, a few adjustments must be made to proceed with this project in the future. First, the model needs to have more evaluation sets. Extracting information from different worldwide stock markets over long-term periods can improve our model. We will enhance the model’s robustness and generalization ability by applying more data over different financial conditions. Secondly, the authors simplified the state issue, and I would develop the algorithm to incorporate day trading abilities. The system must be adjusted to handle high-frequency stock data efficiently and near real-time data analysis ability. This change makes the model more fit for the real-world situation. Lastly, applying the most recent machine learning methods, such as deep learning or reinforcement learning, may reveal more complex data patterns. The system design appears reasonable for long-term stock selection and trend prediction in the China Stock Market. The random forest with random forest feature selection model has demonstrated its ability to deal with the complex nature of stock market information. Before it is widely deployed, some questions still need to be resolved. How successfully does the model perform over diverse market conditions, such as sudden downtrends or uptrends? Does the model follow international financial regulations, particularly concerning automated trading systems?

## Reference

[1] X. Yuan, J. Yuan, T. Jiang, and Q. U. Ain, “Integrated Long-Term Stock Selection Models Based on Feature Selection and Machine Learning Algorithms for China Stock Market,” IEEE Access, vol. 8, pp. 22672–22685, 2020, doi: https://doi.org/10.1109/access.2020.2969293.