# Evaluating the Effectiveness of Machine Learning Models for Identifying Superstar Firms

By

**Eric Sesterhenn**

A creative component submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTERS OF SCIENCE

Major: Industrial Engineering

Program of Study Committee:
Dr. Guiping Hu, Major Professor
Dr. Gary Mirka

Iowa State University

Ames, Iowa

2022

## TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF APPENDIX TABLES

# ABSTRACT

The global market competitiveness evaluation within an industry has significantly changed in recent decades. The leading theory that identifies the changes, is the superstar firm phenomenon. This theory demonstrates how highly productive firms, with a decline in labor share have controlled the market. These highly productive companies are defined as superstar firms. However, the factors that identify such firms vary across literature, and are largely defined by subject-matter expertise. It is possible that statistical tools, such as machine learning, are capable of effectively identifying these factors. This paper utilizes a suite of two machine learning algorithms to assess the efficacy of this approach on real world data. The effectiveness of machine learning algorithms is evaluated by (i) identifying high performing factors using a random forest-based algorithm; (ii) imputing the selected factors into an algorithm to predict out-of-sample superstar firms and (iii) assessing the performance of such algorithms with respect to two benchmarks. Various factors are proven to be effective in complex machine learning models, outperforming the naïve solution in most applications. Most of the important features are dependent on the industry. In one industry of interest, textile manufacturing, a random forest model outperforms the naïve and multi-linear regression benchmark models by 15.6% and 12.8% accuracy, respectively. The random forest model is also able to identify the top 200 companies, potential superstar firms, in this industry with an accuracy of 83.4%.

# CHAPTER 1. INTRODUCTION

In recent decades, the phenomenon of superstar firms has risen as the main explanation for company performance in various economic markets (Autor et al., 2020). A superstar firm is a firm that has a significant advantage over other firms in their industry with respect to profits and other measures of the company performance. These highly productive firms can consume the market share of their industry competitors, transforming the industry from a competitive landscape to a “winner take most” scenario. The superstar firm phenomenon is the leading explanation for industry-wide success in the modern economy. The theory was validated to explain company performance in 29 countries, each combining for two-thirds of the worlds GDP, as early as 1991 (Autor et al., 2020). A 2020 study provides a list of several superstar exporting firms in these countries, including Samsung, Intel, and Foxconn (Freund et. al, 2020). Evaluating what factors contribute to the success of such companies will be paramount for discovering future superstar firms.

In a growing global economy, exports have increasingly become more important for firm growth and performance. The Chinese economy has a large reliance on exports with an exports to GDP ratio of 35% compared to 8% and 13% for the United States and India, respectively (Koopman et. al., 2008). China’s growth is highly related to the amount of exports, the country’s exporting is equivalent to a country with a level of income per capita three times larger than that of China’s (Jarreau et. al., 2009). In the global economy a relationship has been identified linking a firms exports, firm performance, and wages (Manasse et. al., 2001). Identification of the firms with the highest exports is significant in identifying the best performing firms. In 32 countries, on average the top firm is responsible for 14% of a countries non-oil exports. On average, the top five firms account for 30% of the non-oil exports (Freund et. al., 2015).

Historically, the factors that identify a superstar firm vary across literature and are discovered through subject matter expert theories and analysis. Autor et al., indicate that a superstar firm is defined as a company that holds above average market price and below average labor share in an industry (Autor et al., 2020). Once discovered, these factors, are proven to be significant through relatively simple, explainable algorithms. Some solutions have proven to be effective, however simple solutions can fail to uncover critical information to assist in algorithm performance.

In recent decades, advancements in computation power have propelled machine learning as an affordable and useful tool in various applications. Machine learning enables researchers to formulate and effectively answer economic research questions, otherwise difficult to interpret from standard techniques. The tool permits result validation, increasing confidence in the machine learning approach (Basuchoudhary et. al., 2017). A study on the application of machine learning in economics demonstrates that machine learning provides value from predictions or empirical insights into economic applications (Gogas et. al., 2021). Machine learning models have potential for discovering superstar firms or providing insight into the characteristics of superstar firms. This paper will evaluate if machine learning algorithms can effectively identify critical factors that contribute to company achieving superstar firm status in their industry. The evaluation will occur in periods of uniform market conditions.

The global economic market is in continual fluctuation between four major market conditions: up, down, normal, and in-financial-crisis (Rashid et al., 2018). To eliminate the bias across market conditions, the evaluation in this paper will only consider one of the four market conditions. Data quality and availability constrains the analysis to occur in normal market conditions. All data is sourced between 2001 and 2007, a period between the 2000 and 2008 recessions.

The data, sourced from the Chinese Industrial Enterprises Database, contains a set of input features and a response variable. The input features represent a company's annual performance along several metrics. Feature engineering, a method used to manipulate or create new features, enhances several of the input features in the provided data. Additionally, the response variable, export market share, is created to represent the percentage of export sales a company has with respect to the industry. To eliminate lookahead bias, the data is split annually. The dataset in 2004 is omitted because it does not contain the response variable, resulting in six distinct datasets.

The six datasets are applied using a variation of K-Fold cross validation where one year is only used to predict the ensuing year. Feature importance is assessed in each training data fold, producing a set of the most important factors to be used to predict the following year. The response variable, export market share, is predicted using two machine learning models (1) random forest and (2) XG boost; and is compared to two benchmarks (1) multi-linear regression and (2) a naïve model. The naïve model assumes that a firms export market share remains constant in the next year. A comparison is made between the machine learning models and the benchmarks to certify the validity of each model.

Key findings include a set of features considered as the most important features and the effectiveness of machine learning models for predicting superstar firms. High performing input features across all industries include main business revenue, main business cost, and the intermediate input. Several other features demonstrate high performance in specific industries, suggesting that some features are industry dependent. Prediction results across each K-Fold in five unique industries indicate that the random forest model outperformed the benchmarks and accompanying XG boost models. The random forest, XG boost, multi-linear regression, and naïve models had an average predication accuracy of 69.6%, 67.0%, 57.2%, and 54.3% in the textile manufacturing industry, respectively. The random forest model predicts the top 200 firms, superstar firms, relative to the export market share, with an accuracy of 83.4%. The random forest model's performance indicates that complex machine learning models are an effective solution for identifying future superstar firms.

# CHAPTER 2. MATERIALS AND METHODS

## DATA SOURCES

The data used throughout the analysis is provided from the Chinese Industrial Enterprises Database and contains annual performance results from a variety of firms located in China. The dataset is applied in several research applications. Some research includes the development of a Chinese economic growth model (Song et. al., 2011) and quantifying resources misallocation leading to marginal productivity between China, India, and the United States (Hsieh et. al., 2009). The dataset is provided in multiple files, each containing financial data for the respective year. The number of features and firms included is highly dependent on the year. Table 1 represents the number of features and firms included in the original dataset for each year considered in the analysis. Prior to the analysis, pre-processing of the data is conducted. After pre-processing, the number of the features and firms change.

**Table 1**
Number of features and firms in the raw dataset each year

| Year | Feature Count | Firm Count |
|---|---|---|
| 2002 | 125 | 181,527 |
| 2003 | 125 | 190,081 |
| 2005 | 124 | 271,789 |
| 2006 | 124 | 301,901 |
| 2007 | 125 | 336,696 |

The global economic market is in continual fluctuation between four major market conditions: up, down, normal, and in-financial-crisis (Rashid et al., 2018). In the United States, a market condition is typically defined by the Dow Industrial Average or other major benchmark indices. When the value of a benchmark increases or decreases over a period time, the market condition as categorized as an up or down market, respectively. Periods of significant value contraction, such as the 2008 financial crisis or 2020 COVID-19 crisis, are considered as in-financial-crisis. In all other conditions, the market is normal, a period with little fluctuation in several of the major indices. Each market condition has leading factors that identify company performance. To eliminate any bias, the evaluation in this paper will only consider one of the four market conditions. Data quality and availability constrains the analysis to occur in normal market conditions. All data is sourced between 2001 and 2007, a period between the 2000 and 2008 recessions.

Data from Chinese Industrial Enterprises Database is available in the period of interests, however, coverage and availability of features fluctuates in certain years. The dataset from 2004 does not contain the feature that creates the response variable, export delivery value, and therefore is not considered in the analysis. The remaining years: 2001, 2002, 2003, 2005, 2006, and 2007 are preprocessed and successively used in the analysis.

## DATA PREPROCESSING

### INPUT PREPROCESSING

Superstar firms, denoted as high performing firms in an industry, must be evaluated with respect to the remainder of the firms in the industry. Respectively, the original data is partitioned and evaluated for a specific industry. Most of the analysis is focused on one industry, textile manufacturing, however the same pre-processing methodology can be applied to the available industries. Four other industries are considered for comparison, a complete list of the industries considered are included in Table 16. The data pre-processing steps is standardized for all industries, and an overview is provided in Figure 1.

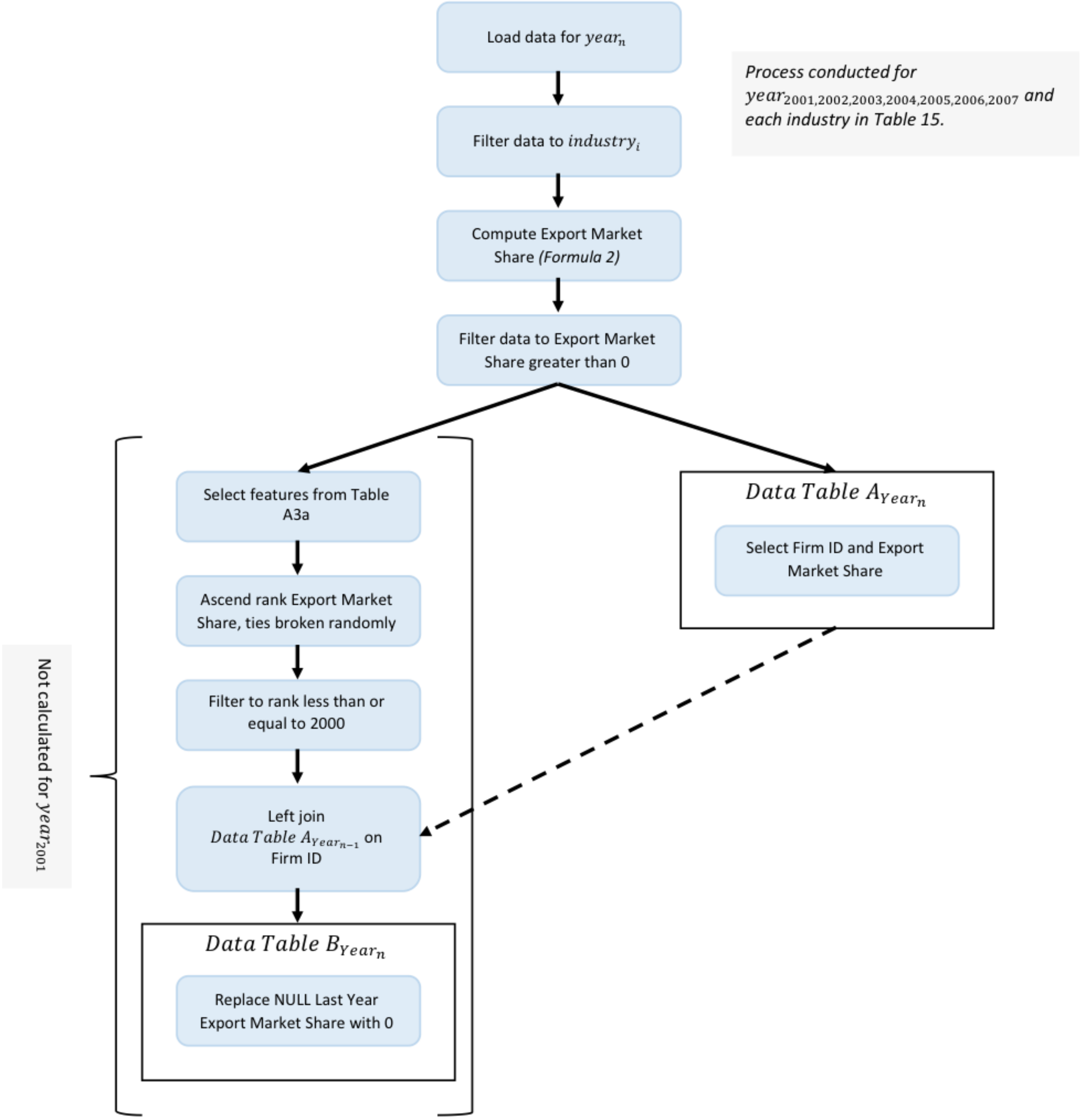
Load data for $year_n$
Filter data to $industry_i$
Compute Export Market Share *(Formula 2)*
Filter data to Export Market Share greater than 0
*Process conducted for $year_{2001,2002,2003,2004,2005,2006,2007}$ and each industry in Table 15.*
Select features from Table A3a
Ascend rank Export Market Share, ties broken randomly
Filter to rank less than or equal to 2000
Left join $Data\ Table\ A_{Year_{n-1}}$ on Firm ID
$Data\ Table\ B_{Year_n}$
Replace NULL Last Year Export Market Share with 0
$Data\ Table\ A_{Year_n}$
Select Firm ID and Export Market Share
Not calculated for $year_{2001}$

**Figure 1.** Flow diagram of the data-preprocessing steps for identifying superstar firms. Source: Authors’ data engineering process for preparing models.

One study suggest that the current market share of a firm can be an effective solution for predicting future market share (Lemoine, 2003). Feature engineering is conducted to generate an input factor that represents the firms previous export market share. This feature is constructed using Formula 1, a calculation related to Formula 2.

$$P_f(t) = \frac{V_f(t-1)}{\sum_{1}^{f_i(t-1)} V_f(t-1)} * 100 \qquad (1)$$

where:

$P_f(t)$ = *Previous years export market share of a firm (f) in a year (t)*

$V_f(t-1)$= *Export value of a firm (f) in the previous year (t-1)*

$f_i(t-1)$= *Number of firms in an industry (i) in the previous year (t-1)*

The previous year's export market share is merged onto the current years data. The 2001 data is only used to compute the previous year's exports for the final dataset in 2002. To ensure firms are not penalized for not being available in the previous year, it is assumed that the firm did not have any exporting sales and are imputed as zero.

All input features, including the previous year's export market share, are candidates for the multi-linear regression, random forest, and XG boost models used in the analysis. The naïve model assumes that the previous market share remains consistent and therefore only the previous year's export market share is considered.

## MODEL RESPONSE VARIABLE

The response variable in the dataset, the annual export market share of a firm, is computed to understand the firms position within each industry. This annual export market share is computed using the exporting sales for the firm. Firms in an industry with the largest market share are considered the superstar firms of the industry. The export market share is created using Formula 2.

$$S_f(t) = \frac{V_f(t)}{\sum_1^{f_i(t)} V_f(t)} * 100 \qquad (2)$$

where:

$S_f(t)$ = Export market share of a firm (f) in a year (t)

$V_f(t)$= Export value of a firm (f) in a year (t)

$f_i(t)$= Number of firms in an industry (i) in a year (t)

For each industry considered, the sum of the export market share is 100%. It is anticipated that the variance of the response variable will change depending on the amount of competition in an industry. Firms with higher market share are considered superstar firms. The export market share is the response variable for training and testing each model, however it is not considered in model performance assessment. Research suggests that continuous input features can predict continuous responses more effectively than discrete responses.

## PERFORMANCE RESPONSE VARIABLE

Due to the nature of the problem, the absolute value of the export market share is less valuable than the order of the firms relative to their export market share. After training the models using the selected features, a prediction of the export market share is generated for the respective testing dataset. The export market share for a firm is ranked relative to its peers in the industry, breaking ties at random. Ties in export market share are uncommon in all industries. The ranking ranges from one to the number of firms in the industry, where one is the highest export market share. Firms with an export market share ranking greater than 2000 are removed from the dataset to eliminate outlier firms. These outlier firms consume a significantly small amount of market share and skew the performance metrics. Removing these firms ensures only the models ability of identifying the potential superstar firms is considered.

The predicted and actual rankings in the testing dataset are separated into three unique group classifications. The classifications include extended, standard, and condensed, where the number of firms contained in a group is maximized, respectively. The formulation of the groups is defined in Table 2, where group sub-sets are highlighted. Accuracy of each model is measured based on if the prediction matches the actual grouping value. These group classifications are evaluated separately using a confusion matrix, evaluating if the actual and predicted classifications are equivalent for a firm.

**Table 2**
Final prediction group definitions by ranking received relative to the export market share.

| Group | Extended | Standard | Condensed |
|---|---|---|---|
| 1 | 1 | 1 | 1 |
| 2 | 2-5 | 2-5 | 2-5 |
| 3 | 6-10 | 6-10 | 6-10 |
| 4 | 11-20 | 11-20 | 11-20 |
| 5 | 21-30 | 21-30 | 21-50 |
| 6 | 31-40 | 31-40 | 51-100 |
| 7 | 41-50 | 41-50 | 101-150 |
| 8 | 51-60 | 51-75 | 151-200 |
| 9 | 61-70 | 76-100 | 201-500 |
| 10 | 71-80 | 101-150 | 501-1000 |
| 11 | 81-90 | 151-200 | 1001-2000 |
| 12 | 91-100 | 201-250 | - |
| 13 | 101-150 | 251-500 | - |
| 14 | 151-200 | 501-1000 | - |
| 15 | 201-250 | 1001-2000 | - |
| 16 | 251-300 | - | - |
| 17 | 301-350 | - | - |
| 18 | 351-400 | - | - |
| 19 | 401-450 | - | - |
| 20 | 451-500 | - | - |
| 21 | 501-1000 | - | - |
| 22 | 1001-2000 | - | - |

## COMPUTATIONAL SETTING

Several models support the evidence of the efficacy of machine learning in prediction of superstar firms. Each model used is a supervised learning algorithm, which uses one or more inputs to predict a known output. The random forest algorithm, an iteration of the Decision Tree algorithm, is predominately used in the analysis.

### FEATURE SELECTION

For each training data set, the Boruta algorithm is selected to filter each of the input features to the most important features. The Boruta algorithm returns a list of features that are confirmed to be important, inconclusive, and confirmed unimportant in evaluating the response. To generate the resultant list, Boruta copies the training data set and shuffles the values in each feature randomly. The reconstructed features are known as shadow features. The algorithm fits the original features and shadow features to the response variables in the respective dataset using a random forest model. The importance of the original and shadow features are evaluated based on their percent increase in mean square error. This measurement represents the likely increase in prediction error if the feature was not included in the model. When a features percent increase in mean square error is higher than the maximum of all the shadow features percent increase in mean square error, it is recorded as a hit.

This process is iterated fifteen times, recording the hits for the features in each iteration. Features that have a number of hits, out of fifteen, in the top and bottom 0.5% of a binomial distribution are considered as important and unimportant, respectively. The remaining features are considered to be inconclusive or tentative. Only the important and tentative features are used in the models as the number of important features is minimal in most cases.

# RANDOM FOREST

A Decision Tree generates a hierarchical structure by splitting the input features repetitively to form a prediction for the response variable. The random forest supervised learning algorithm is a bagging algorithm that improves the variance and accuracy of predictions relative to single predictions, such as a single decision tree (Sarkar et. al., 2019). A bagging algorithm initially creates multiple unique copies of the training data set. The training datasets are used to produce multiple models which predict the target response variable. Each of the predictions from the models are averaged together, resulting in one prediction for each entity in the initial dataset. Bagging produces the largest benefit in unstable environments, where slight changes in the training data lead to larger changes in the response (Sarkar et. al., 2019).

The general bagging logic is applied to the decision tree algorithm, creating multiple decision trees that are averaged to form one random forest prediction. The number of decision trees included in the bagging for the random forest model can be constrained using a hyper parameter.

The random forest algorithm is utilized in a variety of applications. One benefit of the algorithm is its versatility. The algorithm can employ binary, categorical, and numerical input features to predict a regression or classification response. The random forest algorithm performs well in high dimensional data. In high dimensional data, the number of features is close to or larger than the number of observations. The algorithm also demonstrates strong performance in medium to large sample sizes. Due to the algorithm’s nonlinear nature, discovering non-linear relationships to the response variable is trivial. Random forest algorithms are used in a wide variety of applications; however, some common use cases include e-commerce, stock market prediction, and fraud call detection.

The primary model of interest, random forest, contains several hyper parameters that can be tuned to increase the efficacy of the model. Two principal hyper parameters in the R programming language include the number of variables randomly selected as variables at each split (mtry) and number of trees produced (ntree). Each hyper parameters is tuned to an optimal combination on the training dataset.

Adjusting the hyper parameters in each direction assist in the prevention of under fitting or over fitting. The hyper parameters are tuned by searching a range of values for the optimal combination. For regression problems, the default mtry value is defined in Formula 3 (Breiman et. al., 2018).

$$d = \frac{p}{3} \tag{3}$$

where:

*$d$ = The default mtry value*

*$p$= The number of features considered for a model*

In a random forest model, the default ntree hyper parameter is 500, however values between 200 and 800 were evaluated for major increases in performance. An optimal mtry parameter was automatically selected by the model, however models across several industries indicated that the ntree hyper parameter did not improve performance after 300 trees. For all random forest models 300 trees were produced to also assist in computation speed.

## NAÏVE MODEL

The random forests' model results were compared to various other models, including: a naive model, multiple linear regression, and XG boost. The naive model and multiple linear regression model were selected as simple benchmarks. Both simple models are highly interpretable and transparent compared to the random forest and XG boost models.

The naive model assumes a firms' last year export market share is the same as the testing market share. The naïve methodology can perform well in industries with little fluctuation in firm export market share. The previous year market share and testing market share are converted into the extended, standard, and condensed group classifications and compared.

## MULTIPLE LINEAR REGRESSION

The multiple linear regression model is trained with the same input features selected from the random forest based feature selection method, Boruta. Like the random forest model, a multiple linear regression model is trained on the training set and tested with the corresponding testing set. Each model is constructed with R's base linear model function, lm(), with no hyper parameters defined. All selected factors in the model are equally weighted. A linear regression model is expected to perform with input features that have a strong linear relationship with the response. Several correlations in Table 10 indicate a potential for high performance in a linear space.

## XG BOOST

The XG boost and random forest models, considered as complex machine learning models, typically sacrifice interpretability and transparency for performance. Each model requires considerable appraisal to understand the relationship between the input features and the final prediction. The tradeoff in prediction accuracy must be considered when selecting a final model.

An XG boost (Extreme Gradient Boosting) model is generated using a tree boosting technique. The boosting logic is different the bagging methodology used in a random forest model. XG boost models create an initial decision tree, but methodically improve the initial tree by iteratively refining it. Each iteration of the initial tree enhances the prediction accuracy. The final prediction is an ensemble of all the trees created. The ensemble of the trees is prone to overfitting if not avoided by adjustments to the hyper parameters. Hyper parameters used in the model include nrounds, lambda, alpha, and eta which tune the number of trees included in the final model, L2 regularization on leaf weights, L1 regularization on leaf weights, and the learning rate, respectively. L2 and L1 regularization and the learning rate, controlled by lambda, alpha and eta, control the over or under fitting of the model. A grid search, which enables the model to determine the optimal combination of hyper parameter values, was defined using the range of values defined in Table 3.

**Table 3**
Hyper parameter grid search minimum and maximum values for the XG boost model.

| **Hyperparameter** | **Minimum Value** | **Maximum Value** |
|---|---|---|
| nrounds | 500 | 500 |
| lambda | 0 | 1 |
| alpha | 0 | 1 |
| eta | 0 | 1 |

# CHAPTER 3. CASE STUDY

An industry in China, textile manufacturing, is studied and validated against the methods due to its size and dependence on exports. In terms of employment, the textile manufacturing industry is the largest industry among all manufacturing industries in China. The number of firms reflects the size. According to a 2016 study, China is the world’s largest textile producer (Huang et. al., 2016). The industry is also export dependent, according to Table 15 the number of firms that export exceeds 58%. The Chinese textile production accounts for 56.3% of global production (Huang et. al., 2016). Due to size, data availability, and the dependence on exports the textile manufacturing industry is used as the primary industry in the analysis.

## VERIFICATION METHODS

Before models are constructed, the data is split into a testing and training datasets. Two methods for splitting the data were considered for the model. The options contain advantages and disadvantages.

> Option A: Subset the data by year, such that the firms from one year will predict the firms market share the following year.
>
> Option B: Randomly split the collection of years, such that 80% of the data is used for training and 20% is used for testing.

In the textile manufacturing industry, the number of firms deviates each year. The number of firms available in the dataset for the industry is provided in Table 4. The variation in the number of firms demonstrates concerns for underperformance in years data is less available.

**Table 4**
Number of firms in the textile manufacturing industry each year.

| Year | Firm Count |
|---|---|
| 2002 | 5052 |
| 2003 | 5429 |
| 2005 | 6487 |
| 2006 | 6826 |
| 2007 | 7172 |

Splitting the complete dataset at random introduces selection bias, a phenomenon where the observations used in the training set differ greatly from the intended sample. Each firm's values vary over time, a primary concern when firms could be counted multiple times in a training or testing dataset during random selection. The change in the most important feature's values across the time horizon is demonstrated in the sample statistics in Tables 5 and 6.

**Table 5**
The annual mean for the selected factors and the trend of their values.

| Feature | Mean | | | | | Slope of Best Fit Line |
|---|---|---|---|---|---|---|
| | 2002 | 2003 | 2005 | 2006 | 2007 | |
| Other Business Profits | 221.1 | 281.1 | 238.6 | 285.5 | 460.7 | 48.4 |
| Net Receivables | 6805.6 | 9066.0 | 9493.4 | 11511.9 | 14337.8 | 1751.0 |
| Current Liabilities | 23716.4 | 27660.2 | 36435.8 | 42222.6 | 47701.9 | 6253.3 |
| Total Profit | 3020.2 | 3920.4 | 5190.6 | 6137.9 | 7870.5 | 1191.8 |
| Total Liabilities | 26211.3 | 31520.4 | 40066.7 | 46643.3 | 52517.9 | 6773.6 |
| Accumulated Depreciation | 6760.0 | 7606.9 | 9842.9 | 11183.9 | 12813.9 | 1568.5 |
| Management Costs | 2854.9 | 3272.1 | 4356.3 | 5061.7 | 6189.9 | 846.0 |
| Employment | 548.1 | 585.1 | 632.2 | 672.1 | 702.3 | 39.5 |
| Input Tax | 5733.5 | 7547.7 | 9669.4 | 11569.8 | 13184.7 | 1892.4 |
| Owner Equity | 21445.7 | 25481.6 | 31999.3 | 37439.9 | 44104.0 | 5727.5 |
| Operating Profit | 2737.0 | 3979.9 | 5291.0 | 6162.2 | 8459.4 | 1362.7 |
| Payroll Payable | 5704.2 | 6687.5 | 8892.1 | 10696.3 | 14741.9 | 2208.4 |
| Main Business Wages | 5349.9 | 6360.7 | 8559.7 | 10291.8 | 13690.7 | 2061.3 |
| Balance of Current Assets | 26082.1 | 31268.00 | 40903.7 | 48252.9 | 55762.8 | 7634.6 |
| Total Assets | 47657.0 | 57001.96 | 72065.7 | 84083.2 | 96622.0 | 12501.1 |
| Last Year Export Market Share | 0.03 | 0.03 | 0.02 | 0.03 | 0.03 | -0.00 |
| Main Business Revenue | 65537.7 | 80280.7 | 108291.0 | 126702.9 | 150254.3 | 21585.5 |
| Main Business Cost | 57388.5 | 69922.2 | 94363.7 | 110341.4 | 128657.2 | 18295.7 |
| Intermediate Input | 54063.5 | 64332.5 | 84430.8 | 97647.9 | 113951.8 | 15309.2 |
| Output Tax | 3204.2 | 3587.2 | 4304.3 | 4859.5 | 5187.0 | 523.8 |
| Net Total Fixed Assets | 7155.7 | 7694.2 | 8140.2 | 8638.6 | 9136.3 | 490.6 |
| For Production | 7950.0 | 7950.0 | 8959.2 | 9810.0 | | 608.8 |

**Table 6**
The annual standard deviation for the selected factors and the trend of their values.

| Feature | Standard Deviation | | | | | Slope of Best Fit Line |
|---|---|---|---|---|---|---|
| | 2002 | 2003 | 2005 | 2006 | 2007 | |
| Other Business Profits | 1688.4 | 2190.1 | 1646.0 | 1693.4 | 3468.0 | 306.3 |
| Net Receivables | 18160.6 | 25550.4 | 26324.2 | 32603.6 | 47305.4 | 6534.3 |
| Current Liabilities | 74644.9 | 81960.9 | 126661.8 | 148199.8 | 157519.5 | 23198.8 |
| Total Profit | 15751.3 | 17456.6 | 26453.5 | 33408.7 | 38004.2 | 6045.8 |
| Total Liabilities | 80845.8 | 106947.8 | 144657.5 | 171522.9 | 180938.5 | 26476.1 |
| Management Cost | 6329.4 | 7075.8 | 10978.9 | 14064.8 | 16113.1 | 2655.6 |
| Input Tax | 12604.0 | 19256.1 | 32513.4 | 41854.7 | 42687.4 | 8276.5 |
| Owner Equity | 87127.5 | 101897.3 | 125793.8 | 156070.4 | 187376.1 | 25467.0 |
| Operating Profit | 13567.7 | 16580.6 | 25540.3 | 32087.7 | 43077.8 | 7452.7 |
| Payroll Payable | 9813.2 | 11048.2 | 17205.9 | 19803.5 | 44154.0 | 7743.7 |
| Main Business Wages | 9470.7 | 10725.4 | 16936.6 | 19342.3 | 30396.9 | 5046.9 |
| Balance of Current Assets | 73394.3 | 88234.1 | 144087.9 | 176117.0 | 212366.3 | 36582.7 |
| Total Assets | 162855.7 | 202391.2 | 261842.0 | 315697.0 | 355765.1 | 49912.5 |
| Last Year Export Market Share | 0.07 | 0.06 | 0.06 | 0.06 | 0.06 | -0.00 |
| Main Business Revenue | 141425.6 | 169520.8 | 282351.6 | 372852.9 | 425711.7 | 77190.4 |
| Intermediate Input | 115121.6 | 140437.7 | 223772.2 | 291949.5 | 335381.8 | 59203.2 |
| Output Tax | 9758.3 | 10392.0 | 18364.3 | 22979.4 | 21266.5 | 3560.4 |
| Net Total Fixed Assets | 37172.2 | 41718.6 | 40375.2 | 42845.7 | 41928.1 | 963.9 |
| For Production | 34804.4 | 40565.8 | 48609.8 | 54173.1 | | 6615.0 |

Each of the values average and standard deviations increase across the years of interest. If the values are mixed and randomly selected, the importance of several features are likely to be reduced by noise in the data. This can be attributed to a firm’s beta; a measure of a firms changes in performance caused by movements unrelated to the individual firm’s behavior. Beta can include inflation, market growth or decay, and other related measurements. To mitigate bias and the influence of beta in the results, annual partitions of the data will be exerted to train and validate the models.

Due to the exploratory analysis of the textile manufacturing industry, models are trained and tested using a form of K-Fold cross validation. One year is used to predict the next year, producing K models. The performance metrics from K models are averaged, producing one aggregated metric for the available years. Table 7 demonstrates how the model is trained and tested, producing four folds between 2002 and 2007.

**Table 7**
The formulation of K-Fold training and testing datasets.

| Fold Number | Training Data | Testing Data |
|---|---|---|
| 1 | 2002 | 2003 |
| 2 | 2003 | 2005 |
| 3 | 2005 | 2006 |
| 4 | 2006 | 2007 |

## INPUT FEATURES

Sixteen features were selected in the textile manufacturing industry, deriving a default mtry value of five. The importance of each feature included in the textile manufacturing industry is dependent on the training year. The important features selected for each training period aim to enhance the models performance in prediction of the next. The features identified distinguish the features that are critical in identifying superstar firms in a collection or a specific industry. Feature selection for the models are administered through a random forest based wrapper feature selection algorithm, Boruta. The selected k-folds' training data features and response, export market share, are parameters for the feature selection algorithm.

Using the standard Boruta method, a list of the important and inconclusive features for the textile manufacturing industry are outlined in Table 8 and 9, respectively. These are the features that are used in the models for each training year.

**Table 8**

Boruta characterized confirmed features for the textile manufacturing industry in each training year.

| Training Year | Confirmed Features |
|---|---|
| 2002 | Last Year Export Market Share<br>Main Business Revenue<br>Main Business Cost<br>Main Business Profit<br>Intermediate Input |
| 2003 | Last Year Export Market Share<br>Main Business Revenue<br>Main Business Cost<br>Intermediate Input |
| 2005 | Last Year Export Market Share<br>Main Business Revenue<br>Main Business Cost<br>Intermediate Input |
| 2006 | Last Year Export Market Share<br>Total Assets<br>Owner Equity<br>Main Business Revenue<br>Main Business Cost<br>Intermediate Input |

**Table 9**
Boruta characterized tentative features for the textile manufacturing industry in each training year.

| Training Year | Confirmed Features |
|---|---|
| 2002 | Net Receivables<br>Balance of Current Assets<br>Accumulated Depreciation<br>Total Assets<br>Owner Equity<br>Total Profit |
| 2003 | Net Receivables<br>Balance of Current Assets<br>Total Assets<br>Current Liabilities<br>Total Liabilities<br>Operating Profit<br>Input Tax |
| 2005 | Balance of Current Assets<br>Net Total Fixed Assets<br>For Production<br>Total Assets<br>Other Business Profits<br>Operating Profit<br>Total Profit<br>Main Business Wages |
| 2006 | Balance of Current Assets<br>Management Costs<br>Total Profit<br>Payroll Payable<br>Input Tax<br>Output Tax |

# CHAPTER 4. RESULTS AND ANALYSIS

## TEXTILE MANUFACTUING INPUT ANALYSIS

Using the methods and case study analysis, various importance metrics were produced for the input features in the confirmed and tentative feature selection pool for the industry. This initial analysis provides transparency into the most critical features for identifying superstar firms in the industry. The percentage of training years the feature was confirmed or tentative, correlation with the response, and percent increase in mean square error for each feature are detailed in Table 10.

**Table 10**
Average importance metrics across training years for selected features in the textile manufacturing industry.

| Feature | Percentage Years Selected | Average Correlation with Response | Percent Increase in Mean Square Error |
|---|---|---|---|
| Output Tax | 25% | 0.402 | 0.8% |
| Current Liabilities | 25% | 0.455 | 0.9% |
| Input Tax | 50% | 0.507 | 1.7% |
| Net Receivables | 50% | 0.561 | 0.2% |
| Total Assets | 100% | 0.604 | 2.1% |
| Total Liabilities | 25% | 0.588 | 0.9% |
| Net Total Fixed Assets | 25% | 0.620 | -0.2% |
| Owner Equity | 50% | 0.626 | 1.9% |
| Balance of Current Assets | 100% | 0.634 | 2.2% |
| Total Profit | 75% | 0.636 | 2.0% |
| Other Business Profits | 25% | 0.666 | 2.0% |
| Main Business Wages | 25% | 0.669 | 1.3% |
| Operating Profit | 50% | 0.687 | 1.0% |
| Payroll Payable | 25% | 0.702 | 1.6% |
| For Production | 25% | 0.724 | 0.1% |
| Main Business Revenue | 100% | 0.740 | 9.9% |
| Main Business Costs | 100% | 0.742 | 9.2% |
| Intermediate Input | 100% | 0.758 | 8.9% |
| Last Year Export Market Share | 100% | 0.908 | 17.8% |

The metrics in Table 10 are distinct measurements of the success of the feature in the future model. However, features that rank highly in one metric are reasonably considered as the most important. As detailed in Table 10, the Last Year Export Market Share feature is selected in each of the training years and has the highest correlation and percent increase in mean square error. It is anticipated that this feature will add the most value to each of the models for the industry. Several other input features screen highly and can be expected to be key drivers in performance for the machine learning models. Features that are only selected in a subset of years could have similar relationship in the final model, but only in the training folds they were selected in.

Even lower correlated input features can have an impact on the response. Total Assets, a feature only selected in one of the training years with a correlation of 0.604, has a high percent increase in mean square error. Non-linear machine learning models, random forest and XG boost, are likely to extrapolate more information gain from Total Assets than the multi-linear regression model may.

## TEXTILE MANFUACTURING RANDOM FOREST PERFORMANCE

The accuracies for each testing year of the random forests model for the textile manufacturing industry is generated and recorded. Three results are generated for each ranking group classification using a confusion matrix. Due to the design of the group classifications, it is expected that the accuracy of the model increases across the extended, standard, and condensed responses, respectively. The condensed groups are highly concentrated and are therefore easier to predict. Our hypothesis is confirmed by the average accuracy across each testing year of the random forest Model shown in Table 11.

**Table 11**
Average accuracy across testing years for the random forest model in the textile manufacturing industry.

| Group | Accuracy |
|---|---|
| Extended | 62.2% |
| Standard | 69.6% |
| Condensed | 73.6% |

The results indicate that each firms' group can be predicted to a degree of accuracy dependent on the group composition. The provided metrics, however, do not indicate if there was deviation in the accuracy across the testing years. We can expect slight deviations in the accuracy for each year depending on the performance of the input features. Table 12 below discloses the performance of each training year for each group composition.

**Table 12**
Accuracy for each year for the random forest model in the textile manufacturing industry.

| Testing Year | Extended | Standard | Condensed |
|---|---|---|---|
| 2003 | 62.1% | 69.5% | 73.1% |
| 2005 | 57.8% | 67.4% | 71.2% |
| 2006 | 64.6% | 71.5% | 76.2% |
| 2007 | 64.4% | 70.1% | 74.0% |

According to the table, the accuracy was consistent in years 2003, 2006, and 2007. In 2005, the accuracy decreased by about 5% across each group classification. This is likely due to the 2005 being trained on the 2003 dataset. Therefore, it can be hypothesized that ensuring the years training and testing years are sequential is important to the model's success. Modification of the order of the folds produces decreased performance.

The confusion matrix that produces the final accuracies can be partitioned amongst the groups or a collection of the groups. This partition provides a more precise evaluation of the model's performance in the best and worst performing firms. To gauge the prediction accuracy in the superstar firms of the industry, the top ten firms were separated from the dataset. Two methods assess the model's accuracy of these firms (i) standard evaluation using the condensed grouping and (ii) group exclusion, only evaluating if a firms is accurately predicted to a top ten firm. The results are shown in Table 13, where each assessment is reported as group and top ten only, respectively.

**Table 13**
The accuracy of predicting the top ten firms in the textile manufacturing industry using the condensed grouping and top ten only.

| Testing Year | Group Accuracy | Top Ten Only Accuracy |
|---|---|---|
| 2003 | 50% | 80% |
| 2005 | 50% | 70% |
| 2006 | 80% | 80% |
| 2007 | 20% | 80% |
| **Average** | 50% | 77.5% |

The results indicate that the model struggled to predict the top ten firms with respect to the remainder of the industry in the grouped analysis. However, the model was able to identify if a firm is a top ten firm with an accuracy of 77.5%. It is expected that the accuracy would increase in the top ten only analysis. The results are an indication that the model performs highly in identifying superstar firms but is not always able to distinguish the order of these firms. The testing data in 2005, trained from 2003 data, is not subject to the same decline in performance as we saw in the industry wide outcome. Instead, a significant decline in performance in the grouped analysis is realized in the 2007 dataset.

A similar analysis is reproduced for the top 10% of firms, the top 200 in the industry. These firms, although not as prominent as the top ten, still bear a substantial percentage of the market share with respect to their peers. The results in Table 14 indicate the condensed group performance in comparison to the prediction of the top 200 firms only.

**Table 14**
The accuracy of predicting the top 200 firms in the textile manufacturing industry using the condensed grouping and top 200 only.

| Testing Year | Group Accuracy | Top 200 Only Accuracy |
|---|---|---|
| 2003 | 49% | 85% |
| 2005 | 45% | 82% |
| 2006 | 51% | 86% |
| 2007 | 40% | 81% |
| **Average** | 46% | 83% |

The trends identified in the top ten predictions were reproduced for the top 200. Grouped accuracy consistently underperformed the top 200 only evaluation as anticipated. The degree of underperformance also remained consistent. The grouped accuracies also underachieve the industry-wide results, with an average prediction accuracy of 46%. The random forest model, however, did demonstrate high accuracy in identifying the top 200 firms outright, with an average accuracy of 83%. The notion of the model straining to distinguish the specific position of the firm in the industry is validated. This occurrence is expected, the lower ranking firms tend to have lower variance in export market share. The model underperformed in the 2007 testing data, but unexpectedly performs well in 2005.

## TEXTILE MANUFACTURING MODEL COMPARISON

It is unclear if another model is capable of outperforming the random forest model. Accuracies of each model type in the textile manufacturing industry is produced and demonstrated in Figure 2.

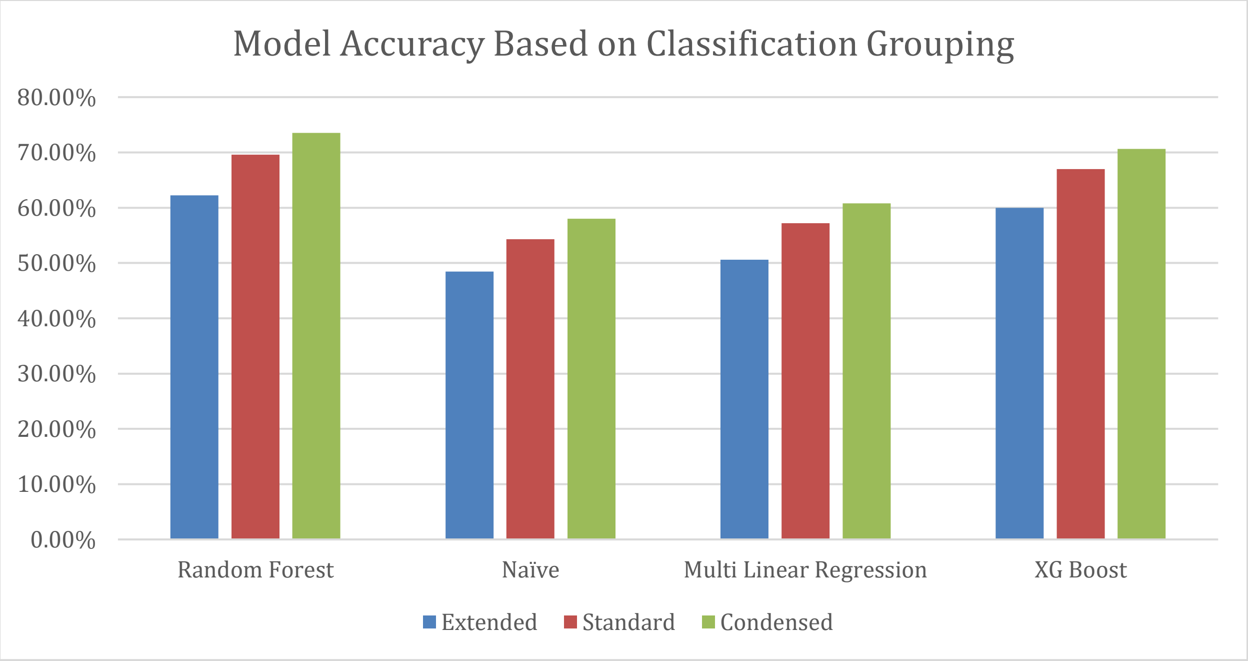

**Figure 2.** Comparison of the selected model performance for each group classification in the textile manufacturing industry.
Source: Authors' computations based on Chinese Industrial Enterprises financial data.

The results indicate that the random forest model outperforms the benchmark and XG boost models across all group classifications. Each of the benchmark models, Naïve and multi-linear regression, underperformed the complex machine learning models with an accuracy of 58% and 61% in the condensed ranking, respectively.

The single feature model was not able to accurately predict future superstar firms. The naïve model underperformed each of the models, including the multi-linear regression. In the textile manufacturing industry, the high correlations of the input features did not translate to performance in the multi-linear regression model. The inputs and response appear to not be explained best in a linear relationship. Due to concerns of overfitting in the XG boost the model did exceed the accuracy of the bagging random forest model.

## INDUSTRY MODEL PERFORMANCE COMPARISON

Textile manufacturing results substantiate evidence for outperformance in the complex models. Additional industries, of various structure, are studied with the same methodology. Alternative characteristics of an industry are anticipated to differentiate results in the models (i) the number of firms in the industry, (ii) the number of years the response value is available, (iii) the percentage of firms who export, (iv) the percentage of total sales from exporting, and (v) the average Herfindahl-Hishman Index (HHI) value. The HHI value is an indicator of the competition or concentration of the industry. The value ranges from 0 to 10,000, where an industry where one firm controls all of the market share (un-concentrated) has an HHI value of 10,000. The industries and the corresponding characteristics are listed in Table 15.

**Table 15**
Characteristics of industries included in the analysis.

| Industry Name | Industry Size | Years Present | Percent of Exporting Firms | Ratio Between Exports and Total Sales | Average HHI |
|---|---|---|---|---|---|
| Textile manufacturing | 10,782 firms | 4 | 58.38% | 32.53% | 10.779 |
| Electronic parts manufacturing | 2,514 firms | 2 | 50.14% | 31.23% | 502.667 |
| Manufacturing of cotton and chemical fiber knitwear and woven products | 3,783 firms | 2 | 50.41% | 30.42% | 345.35 |
| Leather shoes manufacturing | 2,717 firms | 4 | 55.06% | 35.69% | 52.219 |
| Manufacturing of toys | 1,485 firms | 4 | 76.19% | 42.80% | 64.692 |

The number of firms and the number of years with the response available in each industry varies. This can have a significant impact on the efficacy of training the model using the imposed methodology. It is hypothesized that fewer firms and years present will produce lower accuracies and confidence in the model. Due to the dependence of exports in the response, export market share, it is assumed that high exporting industries will perform well in the methodology. Predictability of the industry may also be impacted by the concentration of the market. With lower variation in the response, highly concentrated industries are anticipated to be more difficult to predict. According to the HHI values, the textile manufacturing, leather shoes manufacturing, and manufacturing of toys industries are the most concentrated industries.

Table 16 tests the hypothesis by training and evaluating a random forest model for each of the group classifications in the selected industries. A relative ranking exhibits the performance of the industry relative to the selected industries.

**Table 16**
Average accuracy and ranking of random forest models for selected industries.

| Industry | Extended Accuracy (Rank) | Standard Accuracy (Rank) | Condensed Accuracy (Rank) | Average Rank |
|---|---|---|---|---|
| Textile manufacturing | 62.22% (2) | 69.59% (3) | 73.56% (3) | 2.33 |
| Electronic parts manufacturing | 59.52% (3) | 67.74% (5) | 71.44% (5) | 4.33 |
| Manufacturing of cotton and chemical fiber knitwear and woven products | 62.36% (1) | 69.40% (4) | 72.54% (4) | 3.00 |
| Leather shoes manufacturing | 56.11% (5) | 71.09% (2) | 77.53% (2) | 3.00 |
| Manufacturing of toys | 59.47% (4) | 76.34% (1) | 81.69% (1) | 2.00 |

The textile manufacturing industry is not the best performing industry. The average relative ranking show the manufacturing of toys has the highest accuracy. The industry is the largest exporter, potentially corresponding to the high accuracy. The HHI value appears to have an impact on the performance of the random forest model. The average relative ranking is worst in the least concentrated industries, electronic parts manufacturing and manufacturing of cotton and chemical fiber knitwear and woven products. This is not the expected relationship between HHI and prediction performance. However, the number of exporting firms and years available are also low for each industry, potentially explaining the underperformance.

The random forest model in the manufacturing of toys industry does not precisely predict the firms with the lowest market share. The relative ranking of the model ranks fourth in the extended classification, but ranks first in the remaining classifications. Prediction in the textile manufacturing industry remained consistent across each classification.

# CHAPTER 5. CONCLUSIONS

## SUMMARY

On average, forty-six features were considered within each of the industries included in the results. However, each of the features are not equally important in predicting superstar firms in their industry. It is hypothesized that most important features in one industry may not be equally important in another. It is important to evaluate which features are important across the selected industries and those that are highly dependent on the industry.

Due to the K-fold training and testing strategy, two metrics are used to evaluate the importance of each feature across the industries.

1. Feature Importance: The percent increase in mean square error (%IncMSE) if a given feature was not included in the model.
2. Percentage of Years Selected: The percentage of K years the feature was selected from the training data.

Each feature is percentile ranked according to the two metrics above, combined using equal weighting, and graded into terciles. A formula is provided below to help conceptualize this grading process.

$$Grade_{n_i} = \begin{cases} 0, \; N_i(t) = 0 \\ \frac{\sum_{t=1}^{Y_i} \frac{Rank(\%IncMSE_{n_i})}{N_i(t)}}{Y_i} + \frac{y_{n_i}}{Y_i}, \; N_i(t) \neq 0 \end{cases} \qquad (4)$$

where:

$Grade_{n_i}$ = *The rating for a selected feature, n, in industry, i, between 1 and 3*

$\%IncMSE_{n_i}$= *Percent increase in Mean Square Error, for a selected feature, n, in industry, i*

$N_i(t)$= *The number of features selected for industry, i, on year y*

$Y_i$= *The number of years present in industry i, between 2002 and 2006*

$y_{n_i}$ = *The number of years the a feature, n, is selected in industry, i*

It is assumed that the formula is evaluated in a single industry. The process would be repeated for the remainder of the industries of interest. The percent increase in mean square error is not reported for features that are not selected in a training year. Formula 4 returns one final grade with values zero, one, two, or three, denoting the least to most important features, respectively. Features with a zero-star feature represents a feature that was never selected in a training year. A three-star feature represents a feature that appeared in multiple training models and was highly important in each.

Several features were selected in at least one training year in each industry (i) Last Year Export Market Share, (ii) Balance of Current Assets, (iii) Total Assets, (iv) Main Business Revenue, (v) Main Business Cost, and (vi) Intermediate Input. Last Year Export Market Share, Main Business Revenue, Main Business Cost, and Intermediate Input are three-start features in every industry.

In contrast, several features are highly dependent on an industry. The features with the most industry specific dependencies are reported in Table 17.

**Table 17**
Industry dependent important features according to Formula 4.

| Feature | Industries Where Selected | Industry (Feature Grade) |
|---|---|---|
| Net Total Fixed Assets | 1 | Manufacturing of cotton and chemical fiber knitwear and woven fabrics (3) |
| Paid in Capital | 1 | Manufacturing of toys (2) |
| Foreign Capital | 1 | Manufacturing of cotton and chemical Fiber knitwear and woven fabrics (2) |
| Personal Capital | 1 | Leather shoes manufacturing (2) |
| Managerial Tax | 1 | Manufacturing of toys (2) |
| Balance of Fixed Assets | 2 | Manufacturing of cotton and chemical fiber knitwear and woven fabrics (3)<br>Leather shoes manufacturing (2) |
| Financial Expenses | 2 | Electronic parts manufacturing (3)<br>Manufacturing of cotton and chemical fiber knitwear and woven fabrics (2) |
| Other Business Profits | 2 | Textile manufacturing (3)<br>Leather shoes manufacturing (1) |
| Net Receivables | 2 | Textile manufacturing (2)<br>Electronic parts manufacturing (2) |

Various features hindered performance of the models and were never selected in any of the industries: (i) Affiliation, (ii) Year of Establishment, (iii) Business Status, (iv) Corporate Capital, (v) Hong Kong, Marcu, and Taiwan Capital, (vi) Income Tax Payable, and (vii) Main Business Welfare. It is assumed that a superstar firm cannot be effectively identified using the unselected factors.

Each model is able to characterize the top 2000 firms in an industry accurately more than 50 percent of the time. The complex machine learning models outperform the simple naïve and multi-linear regression models. As indicated in Tables 15 and 16, the random forest model tends to perform well in competitive, highly exporting industries. The high performing industries include the manufacturing of toys and textile manufacturing. The lowest performing industry, electronic parts manufacturing, has the fewest exports, is the least competitive, and the response variable is present in only half the testing years.

## LIMITATIONS

Due to data availability, the research question was limited to firms located in China. Although the available data was sufficient to make an assessment of the efficacy of models and features in this research space, it is unknown if the results would applicable in other markets. Likewise, the data source did not provide the exporting sales, the feature used to construct the response variables, in the 2004 dataset. The omitted data potentially limited performance in the K fold prediction of the 2005 export market share.

The implementation in specific Chinese industries, limited other research applications of superstar firm identification to be compared against the machine learning models. The naïve and multi-linear regression benchmark models used in the research were intended to represent trivial, existing applications of the research question. Although the machine learning results largely outperformed the benchmarks, it doesn't represent that a better solution could exists.

Due to supporting evidence on China's value on exporting, the exporting market share was assumed as the response variable. Other features included in the dataset were not considered as alternatives for evaluation of superstar firms. This limited the research, as additional measures of a firm could have been more correlated to the actual performance relative of the firm.

## FUTURE RESEARCH DIRECTIONS

The limitations addressed navigate towards additional research ideas that could be addressed in the future. The following list includes a few of the applications for future research.

- How does the performance of machine learning models compare to researched applications of identifying superstar firms?
- Do any other machine learning models, outside of those studied in this research, demonstrate outperformance of the researched models and benchmarks?

- Are there any universal features across market regimes that are highly important in evaluating future superstar firms?

# REFERENCES

Autor, D., Dorn, D., Katz, L. F., Patterson, C., & van Reenen, J. (2020). The Fall of the Labor Share and the Rise of Superstar Firms*. *The Quarterly Journal of Economics*, *135*(2), 645–709.

Basuchoudhary, A., Bang, J., & Sen, T. (2017). Machine-learning Techniques in Economics New Tools for Predicting Economic Growth / by Atin Basuchoudhary, James T. Bang, Tinni Sen. (1st ed. 2017.. ed., SpringerBriefs in Economics).

Françoise Lemoine. (2003). China and its Regions. Economic Growth and Reform in Chinese Provinces. Perspectives Chinoises, (76), 70-72.

Freund, C., & Pierola, M. (2015). EXPORT SUPERSTARS. The Review of Economics and Statistics, 97(5), 1023-1032.

Freund, C., & Pierola, M. (2020). The Origins and Dynamics of Export Superstars. The World Bank Economic Review, 34(1), 28-47.

Gogas, P., & Papadimitriou, T. (2021). Machine Learning in Economics and Finance. Computational Economics, 57(1), 1-4.

Hsieh, C., & Klenow, P. (2009). Misallocation and Manufacturing TFP in China and India. The Quarterly Journal of Economics, 124(4), 1403-1448.

Huang, B., Zhao, J., Geng, Y., Tian, Y., & Jiang, P. (2016). Energy-related GHG emissions of the textile industry in China. Resources, Conservation and Recycling, 119, 69-77.

Jarreau, J., & Poncet, S. (2009). Export sophistication and economic growth: Evidence from China. Journal of Development Economics, 97(2), 281-292.

Koopman, R., Wang, Z., & Shang-Jin, W. (2008). How Much of Chinese Exports is Really Made In China? Assessing Domestic Value-Added When Processing Trade is Pervasive. Cambridge, Mass: National Bureau of Economic Research.

Manasse, P., & Turrini, A. (2001). Trade, wages, and ‘superstars’. Journal of International Economics, 54(1), 97-117.

Rashid, S., Sadaqat, M., Jebran, K., & Memon, Z. (2018). Size premium, value premium and market timing: Evidence from an emerging economy. *Journal of Economics, Finance and Administrative Science*, *23*(46), 266-288.

Sarkar, D., & Natarajan, V. (2019). Ensemble Machine Learning Cookbook Sarkar, Dipayan. (1st ed.).

Song, Z., Storesletten, K., & Zilibotti, F. (2011). Growing Like China. The American Economic Review, 101(1), 196-233.

## APPENDIX

**Table A1**
The industries considered in the analysis, their key characteristics, and performance in the condensed group ranking random forest model.

| Industry Name | Industry Size | Years Present | Percent of Exporting Firms | Ratio Between Exports and Total Sales | Average HHI | Prediction Accuracy |
|---|---|---|---|---|---|---|
| Textile manufacturing | 10,782 firms | 4 | 58.38% | 32.53% | 10.779 | 73.56% (3) |
| Electronic parts manufacturing | 2,514 firms | 2 | 50.14% | 31.23% | 502.667 | 71.44% (5) |
| Manufacturing of cotton and chemical fiber knitwear and woven products | 3,783 firms | 2 | 50.41% | 30.42% | 345.35 | 72.54% (4) |
| Leather shoes manufacturing | 2,717 firms | 4 | 55.06% | 35.69% | 52.219 | 77.53% (2) |
| Manufacturing of toys | 1,485 firms | 4 | 76.19% | 42.80% | 64.692 | 81.69% (1) |

**Table A2**
Features that have insufficient coverage across each year of interest, 2002 – 2007, excluding 2004.

| Removed Features | | |
|---|---|---|
| Accounting System | Product Sales Profit | Cash From Financing |
| ID in Source | Labor Insurance | Employment Female |
| Region | Total Loss | Research & Development Cost |
| Phone | Total Tax | Operating Intermediate Input |
| Zip | Payable Profit | Operating Cash In |
| Construction Level | Tax Processing | Operating Cash Out |
| Industrial Units | Real Estate Industry | Investment Cash |
| Agriculture | Other Industry | Investment Cash Out |
| Industry | Operating Income | Financing Cash |
| Construction Industry | Short Term Investments | Financing Cash Out |
| Transportation Industry | Accounts Payable | Current Assets |
| Wholesale and Retail Trade | Other Income | Organization Type |
| Catering | Office Fee | Operating Tax |
| Other | Staff Education Fee | Asset Impairment Loss |
| Business Scale | Investment Income | Changes in Fair Value |
| Light and Heavy Industry | Non-Operating Income | Panel ID |
| Gross Output Constant | Non-Operating Expenses | Industrial Sales Output |
| Gross Output Current | Advertising Fee | Month of Establishment |
| New Product Output Value | Pension Insurance | Value Added Tax Payable |
| Industrial Added Value | Housing Accumulation | Main Business Tax |
| Total Current Assets | Direct Material | Management Fee Input |
| Intangible Deferred Assets | Manufacturing Cost Investment | Operating Expenses |
| Product Sales Fee | Cash From Operations | Cash From Financing |
| Cash From Investment | | |

**Table A3**

List of features after data pre-processing, including the response variable Expert Market Share.

| Included Features | | |
|---|---|---|
| Last Year Export Market Share | Balance of Fixed Assets | Property Insurance |
| Managerial Tax | Intangible Assets | Financial Expenses |
| Depreciation This Year | Total Assets | Interest Expense |
| Registration Type | Current Liabilities | Operating Profit |
| State Owned Holdings | Long Term Liabilities | Subsidy Income |
| Affiliation | Total Liabilities | Total Profit |
| Year of Establishment | Owner Equity | Income Tax Payable |
| Business Status | Paid in Capital | Payroll Payable |
| Employment | National Capital | Main Business Wages |
| Net Receivables | Collective Capital | Welfare Payable |
| Inventory Value | Corporate Capital | Main Business Welfare |
| Finished Product | Personal Capital | Input Tax |
| Balance of Current Assets | Hong Kong, Marcu, and Taiwan Capital | Output Tax |
| Long Term Investment | Foreign Capital | Intermediate Input |
| Net Total Fixed Assets | Main Business Revenue | Management Cost |
| Total Fixed Assets | Main Business Cost | Accumulated Depreciation |
| For Production | Other Business Profit | Export Market Share |
| Market Share Rank | Group Rank | |

**Table A4**
Performance comparison of selected models for each group classification in the textile manufacturing industry.

| Model | Group Classification | Performance Rank | Accuracy Difference to Random Forest |
|---|---|---|---|
| Random forest | Extended | 1 | -- |
| | Standard | 1 | -- |
| | Condensed | 1 | -- |
| Naive model | Extended | 4 | -13.80% |
| | Standard | 4 | -15.31% |
| | Condensed | 4 | -15.55% |
| Multi-linear regression | Extended | 3 | -11.63% |
| | Standard | 3 | -12.42% |
| | Condensed | 3 | -12.79% |
| XG boost | Extended | 2 | -2.24% |
| | Standard | 2 | -2.63% |
| | Condensed | 2 | -2.94% |

**Table A5**
The initial randomly generated importance results for an example industry.

| Feature | Training Year 1 | | Training Year 2 | | Training Year 3 | |
|---|---|---|---|---|---|---|
| | %IncMSE | Selected | %IncMSE | Selected | %IncMSE | Selected |
| A | 3 | True | 3 | True | 3 | True |
| B | | False | | False | | False |
| C | | False | 1 | True | 1 | True |
| D | | False | 4 | True | 2 | True |
| E | 1 | True | 1 | True | 1 | True |
| F | 2 | True | | False | | False |

**Table A6**
The percentile ranked importance values on the randomly generated example results.

| Feature | Training Year 1 | Training Year 2 | Training Year 3 | Average Percentile Rank |
|---|---|---|---|---|
| | %IncMSE Percentile Rank | %IncMSE Percentile Rank | %IncMSE Percentile Rank | |
| A | 1.000 | 0.750 | 0.750 | 0.833 |
| B | | | | |
| C | | 0.250 | 0.250 | 0.167 |
| D | | 1.000 | 0.500 | 0.500 |
| E | 0.333 | 0.250 | 0.250 | 0.278 |
| F | 0.667 | | | 0.222 |

**Table A7**
The final grade for each of the randomly generated example features.

| Feature | % Years Present | Average Percentile Rank (%IncMSE) | Combined Rank | Tercile (Grade) |
|---|---|---|---|---|
| A | 1.000 | 0.833 | 0.9165 | 3 |
| B | | | | 0 |
| C | 0.667 | 0.167 | 0.417 | 2 |
| D | 0.667 | 0.500 | 0.584 | 2 |
| E | 1.000 | 0.278 | 0.639 | 3 |
| F | 0.333 | 0.222 | 0.278 | 1 |

**Table A8**
The performance of various model types in each group classification for the textile manufacturing Industry.

| Model | Group Composition | Train: 2002<br>Test: 2003 | Train: 2003<br>Test: 2005 | Train: 2005<br>Test: 2006 | Train: 2006<br>Test: 2007 | Average |
|---|---|---|---|---|---|---|
| Random forest | Extended | 63.47% | 56.32% | 64.62% | 64.49% | 62.22% |
| | Standard | 70.87% | 65.44% | 71.22% | 70.84% | 69.59% |
| | Condensed | 74.51% | 70.71% | 75.09% | 73.93% | 73.56% |
| Naive model | Extended | 51.54% | 34.98% | 54.15% | 53.03% | 48.43% |
| | Standard | 57.44% | 41.67% | 59.27% | 58.76% | 54.29% |
| | Condensed | 60.64% | 46.53% | 62.57% | 62.30% | 58.01% |
| Multi-linear regression | Extended | 51.29% | 39.00% | 57.28% | 54.83% | 50.60% |
| | Standard | 57.56% | 47.28% | 62.74% | 61.12% | 57.18% |
| | Condensed | 61.08% | 51.46% | 66.38% | 64.16% | 60.77% |
| XG boost | Extended | 62.52% | 50.96% | 62.57% | 63.88% | 59.98% |
| | Standard | 68.86% | 59.92% | 69.34% | 69.72% | 66.96% |
| | Condensed | 72.38% | 64.77% | 72.75% | 72.58% | 70.62% |

**Table A9**

Average sample statistics for the selected features in the textile manufacturing model.

| Feature | Mean | Median | Standard Deviation | Minimum | Maximum |
|---|---|---|---|---|---|
| Other Business Profits | 297.4 | 0.0 | 2137.2 | -15293.2 | 45917.0 |
| Net Receivables | 10242.9 | 3522.0 | 29988.8 | -5136.8 | 709913.0 |
| Current Liabilities | 35547.4 | 13441.9 | 117797.4 | -1353.2 | 2749902.8 |
| Total Profit | 5227.9 | 1410.6 | 26214.8 | -45527.6 | 745420.0 |
| Total Liabilities | 39391.9 | 14358.2 | 136982.5 | -334.8 | 3135071.4 |
| Management Costs | 4347.0 | 2173.9 | 10912.4 | 0.0 | 301147.0 |
| Input Tax | 9541.0 | 5015.0 | 29783.1 | -7776.0 | 810340.4 |
| Owner Equity | 32094.1 | 10795.7 | 131653.0 | -99166.2 | 3198079.0 |
| Operating Input | 5325.9 | 1491.7 | 26170.8 | -41684.4 | 778538.0 |
| Payroll Payable | 9344.4 | 5526.8 | 20405.0 | 48.6 | 596864.4 |
| Main Business Wages | 8850.6 | 5286.4 | 17374.4 | 0.0 | 406753.2 |
| Balance of Current Assets | 40453.9 | 16148.7 | 138839.9 | 104.0 | 3216215.8 |
| Total Assets | 71486.0 | 27373.3 | 259710.2 | 394.2 | 6046024.4 |
| Last Year Export Market Share | 0.03 | 0.02 | 0.06 | 0.00 | 1.16 |
| Main Business Revenue | 106213.3 | 54676.1 | 278372.5 | 2901.0 | 7110147.4 |
| Intermediate Input | 82885.3 | 42285.4 | 221332.6 | 127.4 | 5571624.6 |
| Output Tax | 4228.4 | 1693.0 | 16552.1 | -617.4 | 910426.4 |
| Net Total Fixed Assets | 8153.0 | 2513.0 | 41008.0 | 0.0 | 2415598.8 |
| For Production | 8792.7 | 2757.0 | 44538.3 | 0.0 | 2126719.0 |