THE UNIVERSITY OF HONG KONG

MSBA7005: CAPSTONE PROJECT REPORT

ESQUEL GROUP

REPORTED TO:

- **Esquel Group**
- **Prof. Zhang Wei**

PRESENTED BY: **HBS SHARKS**

- Tang Yilan 3035141712
- Cheng Chuhang 3035104099
- Liu Ningbo 3035814670
- Zheng Zhikang 3035813901
- Zhao Yuedian 3035331604
- Chen Mengsha 3035675476
- You Yi 3035029237
- Nan Yun 3035811563

TABLE OF CONTENTS

1. **EXECUTIVE SUMMARY**
2. **INTRODUCTION**
   - COMPANY BACKGROUND
   - CURRENT BUSINESS OPERATIONS AND FOCAL BUSINESS UNIT
3. **PROBLEM THE COMPANY IS FACING**
   - THE PROBLEMS
   - DESIRED OUTCOME
4. **ANALYTICAL CHALLENGE**
   - IF RELEVANT, WHY CAN USE THE SIMPLE/TRADITIONAL METHOD
   - HOW THE PROBLEM IS RELATED TO BUSINESS ANALYTICS
5. **THE ANALYTICAL SOLUTIONS**
   - THE DATA USED (OR AVAILABLE FOR ANALYSIS)
   - THE ANALYTICAL METHODS
6. **THE RESULTS**
   - POTENTIAL IMPROVEMENTS
   - BUSINESS IMPLICATIONS
   - COMMENTS FROM CLIENT
7. **APPENDIX**
   - TECHNICAL DETAILS OF THE MODEL
   - ADDITIONAL TABLES OR GRAPHS
   - SAMPLE (OR TOY) DATA WITH DATA DICTIONARY
   - WORK ALLOCATION

1. EXECUTIVE SUMMARY

Esquel group has developed and extended its business rapidly on a global scale by progressively constructing its vertically integrated supply chains or operating its 'Integral' project and in this condition, the productivity and the efficiency have already improved significantly because of the innovative development for the past few years. The Company is a market-driven company that progressively and dynamically innovates itself, aspiring to make a revolutionary difference in the traditional textile and apparel industry so that it can keep the leading position in the industry as well. Therefore, one of the ways for the Company to further increase its productivity and efficiency is to have better control in the production process of products, thus in this project, we mainly focus on the predictions of dyeing time, actual width, and shrinkage rate based on the historical data given by the Company.

Having a statistical analysis on the process of production, the Company can have a systematical understanding of the production process so that the Company would know what are the key factors that influence the production process, and thus, have better quantitative control on them. By doing so, the Company can maximize its utility while reducing the cost of production so that it can have a much better financial performance for future perspectives. With the given dataset of dyeing time, actual width, and shrinkage rate, we have applied a variety of model types in this project such as XGBoost model, random forest model, neural network model and etc. on the data that are separated into the training dataset and the test dataset.

In the dyeing time prediction problem, our team used the root mean square error as the measurement of model selection and, as a result, the random forest in the one-step model has generated the best results among all. For the actual width prediction, we used the difference between actual width and design width as the response variable; then, we compared multiple models and it turned out that the random forest model with the difference as the response is the optimal one as well. Lastly, for the shrinkage prediction part, we ruled out two predictors in the data that have a weak effect on the response variables and the random forest turned out to be the optimal model that could provide us a higher accuracy.

2. INTRODUCTION

a. Company Background

Esquel group (Esquel or the "Company"), founded in 1978 by family patriarch Yang YuanLoong, and in 1995, Marjorie YANG took charge of the Company after that, is one of the world's largest cotton-based textile and apparel manufacturing company and also the leading producer of premium cotton shirts that with more than 5,700 employees and multiples manufacturing facilities across four continents. The Company supplies its high-quality innovative products and services for plenty of top brands and department stores all around the world, including Hugo Boss, Nike, Tommy Hilfiger, Lacoste, etc. Esquel is a market-driven company that progressively and dynamically innovates itself, aspiring to make an innovative difference in the traditional textile and apparel industry while advocating the positive development of innovation, environmental sustainability, and responsibility to society.

i) Market Position & Comparative Advantages

Esquel aims to become a dynamic and progressive global-scale company that can provide its high-quality innovative products and services to the market and customers. To attain the goal, the five-e business principles, which stands for ethics, environment, exploration, excellence and education, has been developed within the Company organization as a part of the company culture that becomes one of the driving factors for the continuous improvement of a company in order to become the leading company in the industry.

The Company with its long-term goal and strategy, over the years, has gradually expanded its scope of operations, and eventually constructed a successful vertical integration set-up for textile and apparel business, starting from the scratch of design, high-quality cotton farming, ginning, yarn spinning, weaving and to the production of the fabrics, garments, and the delivery of the final products to customers after producing it, with its internal integrated supply chain that makes Esquel a unique company that is distinguishable from other garment manufacturing company in the global industry.

With its own vertically integrated supply chain, the Company has a much larger comparative advantage than the others in the industry since the cost of producing or assembling one high-quality shirt is much lower compared to others' products. Furthermore, the vertical integration helps the Company to better deal with the fluctuations in the price of a variety of inputs such as cotton so that it can relatively be easier to control the margin compared to other companies, which makes the Company more competitive as well.

The other gain from having an internal vertically integrated supply chain is that it can provide a shorter cycle time, which not only increases efficiency but also can make the Company have a quicker response to the market or to the customers for providing better products or services.

b. Current Business Operations and Focal Business Unit

The vertically integrated supply chain is a revolutionary development that is crucial to the Company, and it is what makes the Company unique in the traditional industry.
However, due to the special characteristics of the textile and apparel industry, the supply chain is much vulnerable to the change of natural environments such as climate change that would immediately affect the input of cotton and so the production of fabric or garments. Therefore, in order to reduce the unpredictable impact of climate change and increase the sustainable growth for the future, the research & development department developed a brand-new cotton seed variety of Sea Island Cotton that is reinforced in resistance to pests, disease, and climate change with its strong vitality so that it could provide a steady supply to satisfy the future production.

Besides the supply chain, currently, the Company also focuses on the production side since a more productive and efficient production process can meet the future demand and increase the profitability of the Company.

Therefore, by applying the statistical model in the process of production such as the dyeing of dye, production of fabric, etc, the Company has the ability to analyze the production process and have a quantitative control on it for cost reduction and better production performance.

3. PROBLEM THE COMPANY IS FACING

a. The Problem

i) Internal pain point

Yarn Dyeing Time Prediction

The total yarn dyeing time varies for similar dyeing assignments and can be different from the sum of the single identifiable stages, which indicates some unexplained delay that can be detected and avoided. The prediction result is influenced by the colorist, specific dyeing task, art selection, dyestuff usage, and etc. Each carries different significance weights and influences the yarn dyeing time in a collective way for the Company to identify and prioritize.

Fabric Actual Width & Shrinkage Prediction

To produce any cotton shirts, the material dosage should be minimized. However, due to various factors, the width of fabric rolls is uncertain. Hence the minimum required length of fabric is difficult to determine and it involves the interaction of a large number of variables, especially in a categorical way that needs to be optimized in the modeling process. At the same time, fabric shrinkage rates are also hard to predict in the same sense involving large numbers of factors. For example, theoretically, the fabric shrinkage ratio is affected by the following variables but to an unknown extent: 1. The type, count and density of the warp yarns. 2. The type, count and density of the weft yarns. 3. Weave structure (Kadi and Karnoub,2015: The Effect of Warp and Weft Variables on Fabric's Shrinkage Ratio) The Company needs to discover a trained model to reliably predict the fabric dimensional change with bulk raw data as input.

ii) External pain point

Yarn Dyeing Time Prediction

Yarn dyeing has multiple stages, and the random duration caused by various factors complicates the scheduling of downstream processes. This may further influence labor cost and labor hours, order shipment time, etc. In nowadays international trade environment, shipment date is defined as an important point of sales to calculate day sales outstanding, Accounts receivable collections and related interest cost.

Fabric Actual Width & Shrinkage Prediction

Fabric shrinkage is the major concerning point for the ready-made garment's suppliers, as fabric shrinkage after washing may affect the size of the garments. In the competitive markets where quality is accessible at low cost, companies are requesting the low shrinkage to create value from reliable suppliers to satisfied end buyers. In other words, manufacturers need to know that the fabrics they are producing are of excellent quality. Customers need to have an idea of the right treatment of the fabric given shrinkage expectation. And brand reputation is built in the lowest wasted materials and energy saved in remaking things.

b. Desired Outcome

Yarn Dyeing Time Prediction

By predicting the total time needed for the dyeing process in a holistic way can facilitate the decision-making process and enhance production efficiency. Moreover, by analyzing the significance of the dyeing parameters, the Company can put more control on the key parameters to make prediction outcomes stable and reliable to transform into time, energy and cost savings.

Fabric Actual Width & Shrinkage Prediction

The prediction of shrinkage of cotton knitted fabrics is substantially beneficial to garment manufacturers. Prediction of any property of materials has attracted the attention of many scientists all over the world to produce better products. The deviation from the fabric width will either lead to more consumption of raw materials or affect the Company's profit in both the short and long run. And estimation of dimensional change before manufacturing could allow us to maintain control over dimensional change.

4. ANALYTICAL CHALLENGE

a. If Relevant, Why the problem cannot be solved by a simple model

i) Computational Challenges

Difficulties in processing large volumes of data, especially categorical data

The existing data processing tool Esquel uses, namely Excel, is not powerful and efficient enough to handle the large and increasing volumes of production data. Excel maxes out at a million rows and its limitations in computation capability become more profound in processing categorical data. We are provided with around 0.8 million rows of fabric width and shrinkage data with a huge amount of categorical information. Considering the aggressive expansion and prosperous future of Esquel, Excel is reaching its breaking point to work on the scale of big data.

In order to obtain more insights from these business-critical data, we deployed Python and R to construct models and perform statistical analysis. Reproducibility of Python and R source codes is much more advanced and less time-consuming than Excel formulas or VBA source codes.

Wrangling of categorical data

Data wrangling is a critical foundation of data analytics. It presents particular challenges to wrangle categorical data in our project as many variables have high cardinalities and are often coded with texts and numerical values. It is easy to break the relationship between category numbers and category labels without realizing it, thus overlooking the information encoded in a variable. For example, in the raw dataset for yarn-dyeing assignments, information regarding dyeing color, the number of yarn count and ply are combined in one single column named “ColorCodes”. Raw data extracted directly from the front-line operation is not in an ideal format and might hide potentially important information for the rest of the data analysis workflow.

To facilitate the data exploration and modeling process, we deployed regular expressions in Python and Excel formulas to transform raw data into features suitable for machine learning pre-possessing. By wrangling categorical data, we could dig up embedded information without affecting Esquel’s existing data management system.

Dimensionality deduction of the high-cardinality categorical variables

The dimensionality of the feature matrix is substantially increased by high-cardinality categorical variables, which increases the computational cost of data analytics. In the yarn-dyeing topic, the cardinality of certain categorical variables, such as "Colorists", "Dye_Art_No" and "After_Art_No", ranges from 56 to more than 97 levels, which exceeds the computational capability of the regression tree algorithm in R. Moreover, the raw dataset has at least one level that appears only once, thus, some levels will likely be present only in the testing set when the data is split into a train and test set. All these pose a great challenge to in-depth data analysis and application in production scheduling.

To tackle this problem, we managed to reduce data dimensionality by distributed random forest algorithms and n-numeric methods. Unlike the one-hot coding method, which designates a feature vector on a reduced vector space, these two methods do not have the risk of losing information. In the real-world application, our programs could better predict most, if not all, production combinations.

ii) Endogeneity

Endogeneity is one of the key problems in a statistical model that usually refers to a situation when there is an explanatory variable correlated with the error term in the model, in this case, the explanatory variable is called endogenous variable. Optimally, the independent variable in a regression model should be exogenous variables that should not be affected by other observed or unobserved variables so that the response variables in that model could be explained by the explanatory variables and the rest of part that cannot be explained by predictors should be included in the error term. However, if there is an unobserved variable, which is included in the error term, that can affect the response variable but also has a correlation with the explanatory variables, the problem of endogeneity occurs in this case.

There are several reasons and conditions that can cause the occurrence of endogeneity problems, and commonly the problem happens when:

1) there is a simultaneity, the dependent variables have a causality relationship with independent variables that means dependent variables can affect predictor so that in a way affects itself, and eventually, it would lead to a loop of causality relationship problem;

2) there is an omitted variables in the regression model, which results in the independent variables being correlated with the error term, and the error term can explain part of the response variable;

3) there is a selection bias, in the case that the data is sampling with bias (not random) so that we cannot correctly measure the effect;

4) there is a measurement error, and in this case, it is included in the error term of the regression model resulting in having endogeneity issues.

There are three main target estimations in our Esquel project:

1) prediction of dyeing time; 2) prediction of actual width; 3) prediction of shrinkage rate. The datasets that we used in our statistical model consisting of three target variables (the response variables in our model) dyeing time, actual width and shrinkage rate are historical data that is from April 2020 to April 2021 that are all given by the Company, which all of the observation are randomly chosen from their database so that the problem of selection bias doesn't seem to exist in our model.

Secondly, the omitted variables problem doesn't seem to be in our model either, the reason is that all of the independent variables are given by the Company, which are accurately controlled by the clients so that this is not the source that would cause endogeneity. Lastly, the problem of simultaneity that means the response variables (dying time; actual width; shrinkage rate) can influence explanatory variables while the explanatory variables can also affect the response variables and the measurement error do not exist either because the result in our model seems to be accurate; therefore, for the current model we have, we believe that problem of endogeneity seems to not be relevant.

b. How is the problem related to business analytics?

The predictions of actual widths, shrinkage rate and yarn dyeing time in Esquel's production process are based on relevant business data, historical business performance and past events, while these assumptions are consistent with the essences of business analytics. Business analytics is forward-looking and based on the analysis of relevant business data and task drivers that have a strong tracking linkage to financial results and business performance. Task drivers can be either financial or operational, external or internal. There are plenty of such drivers: some reflect changes over time, such as volume of property sales, new product, loan defaults, and foreclosures; some reflect changes over a specific time period, such as the number of new births, new car sales and new jobs; while others reflect changes at a specific point in time, such as the rate of change in interest rates, fuel prices, tax rates and sales commission.

What is the instructive significance of business analytics? The first role: making decisions with data. Use data to quantify the current situation and eliminate ambiguity with clarity. This includes direct-recording, multiple-processing of data, and establishment of a system of business analytics indicators. The second role: judging with data. The second biggest role of business analytics is to make decisions with data instead of feelings. All business judgments cannot be made without standards. Use data to quantify, test, and track judgment standards. The third role: using data to figure out reasons. Find causes of problems from a data perspective. The fourth role: evaluating with data and using data to predict business trends. The fifth role: using data to forecast and make comprehensive evaluations of operational results.

The business analytics process includes predictive analysis, business evaluation, and driver-based projections, etc. The process integrates these analyses and predictions to make them effectively apply, transform and disseminate within organizations, and further promotes the plan and implementation.

Machine learning can improve efficiency and boost differentiation. As options for platforms, solutions, and vendors proliferate and improve, becoming more astute consumers of big data technologies will likely become increasingly important for companies to integrate and scale those technologies.

For example, for the prediction of shrinkage, which contains more than 113k rows and 20 columns of raw data, we tried several models with Random Forests (RF) and Neural Network (NN) worth mentioning. Both support nonlinearity prediction. RF is the ensemble of decision trees. Each decision tree, in the ensemble, processes the sample and predicts the output label (in the case of classification). Decision trees in the ensemble are independent. Each can predict the final response. In our model, we prefer distributed RF due to: Improved ability to train on categorical variables. By default, DRF builds half as many trees for binomial problems. For multiclass problems, a tree is used to estimate the probability of each class separately. RF is comparatively powerful and highly accurate for not very high dimensional data but can be slow and biased towards certain features.

The Distributed Random Forest (DRF) is one of the most useful regression models for the purpose of data classification. Unlike the regular random forest, when modeling the data, a distributed random forest model usually generates a bunch of trees instead of a single tree, and each tree is a weaker learner that is only based on a subset of data (rows and columns) but not the whole dataset. By constructing the model in this way, the model can reduce the variance so that it probably could provide a better performance in the final prediction since the final result is based on the average prediction of all trees for both classification and regression. Furthermore, the distributed random forest model is good for tackling the problem of high cardinality because of its unique structure.

The Neural Network is a network of connected neurons. The neurons cannot operate without other neurons - they are connected. Usually, they are grouped in layers and process data in each layer and pass forward to the next layers. The last layer of neurons is making decisions. The number of hidden layer neurons varies from problem to problem and depends on the number and quality of training patterns. If too few neurons in the hidden layers are selected, the network can result in large errors. Whereas if there are too many hidden layers, the network can result in overfitting. NN provides a fast and black-box solution that works out well for high-dimensional data but depends a lot on training data.

5. ANALYTICAL SOLUTION

Yarn Dying Time Prediction

a. The data used (or available for analysis)

The data we have used for yarn dyeing process time prediction is provided by the Esquel group. It is stated that the data is real manufacturing data with some transformation applied to remove sensitive information.

i) Context and content

In total, in our dataset, there are 9744 rows of yarn dyeing records with 23 columns describing different kinds of information about the yarn dyeing process. As has mentioned above, the yarn dyeing process consists of seven stages(pre art, dye art, dye art sample, after art, after art sample, fix art, fix art sample), thus, 7 columns(Pre_Art_Actual_Time, Dye_Art_Actual_Time, Dye_Art_Sample_Time, After_Art_Actual_Time, After_Art_Sample_Time, Fix_Art_Actual_Time, Fix_Art_Sample_Time) in the dataset are used to display duration of each manufacturing stage and 1 column is used to display total duration of the yarn dyeing process.

Among the 15 rest columns, except the first column(DyelotRefNo) is only sequence number without any meaningful information, the other 14 columns(Machine_Name, Batch_Type, Batch_NO, ColorCode, Colorist, Remark, DyestuffUsage, Dyeing_Ratio, Dye_Weight, Is_Repair, Pre_Art_No, Dye_Art_No, After_Art_No, Fix_Art_No) all contain some information correlated with the duration of the manufacturing process.
Due to different production requirements, not every process would go through all stages. Therefore, there are many blank values in our dataset. For instance, if the process doesn't use any dyestuff, ColorCode and Colorist will all be blank.

ii) Definition of variables

In the prediction models, there are two types of variables. The first type of variable is the original value of the raw data, and the other type of variables are generated or encoded from the raw data. The table below shows the variables we used and the elaboration of the definition of the variables.

Name	Type	Definition
Machine_Name	Categorical variable	The name of the machine used in the yarn dyeing process.
ColorOnly	Categorical variable	The color of the dyestuff, contained in the raw data column "ColorCode"
Fabric	Categorical variable	The type of fabric, contained in the raw data column "ColorCode"
Colorist	Categorical variable	The name of the colorist who is responsible for the yarn dyeing process
Remark_label	Categorical variable	The encoded label of the column "Remark" indicating different types of the production process
DyestuffUsage	Continuous variable	Dyestuff dosage, the unit is %. It is Weight of dye (kg) /weight of yarn(kg) *100%. When there is no dyestuff used, the value is 0.

Dyeing_Ratio	Continuous variable	Weight of yarn (kg) / volume of liquid (including dyestuff and water) (L). When there is no dyestuff used, the value is 0.
Dye_Weight	Continuous variable	Weight of yarn(kg)
Is_Repair	Binary variable	"1" indicates the process is to repair yarn, "0" has a converse meaning
Pre_Art_No	Categorical variable	Codenames of pre art crafts
Dye_Art_No	Categorical variable	Codenames of dye art crafts
After_Art_No	Categorical variable	Codenames of after art crafts
Fix_Art_No	Categorical variable	Codenames of fix art crafts

iii) Summary statistics (tables, graphs)

Name	Summary
Machine_Name	53 levels
ColorOnly	18 levels
Fabric	3 levels
Colorist	29 levels

Remark_label	10 levels
DyestuffUsage	Min.: 0 Mean: 1.6431 Max.:17.9120
Dyeing_Ratio	Min.: 0 Mean: 7.251 Max.:62.750
Dye_Weight	Min.: 1.2 Mean:224.3 Max.:999.0
Is_Repair	2 levels
Pre_Art_No	97 levels
Dye_Art_No	35 levels
After_Art_No	63 levels
Fix_Art_No	12 levels

b. The Analytical Methods

i) Description of the math/conceptual model and assumptions

Data Pre-processing

We pre-processed the raw data for the model training purpose. Firstly, we impute all missing values with value 0, since we assume that the missing data depends on a hypothetical value (e.g. missing data of Colorist means that there is no colorist involved in the process).

Secondly, as we have mentioned above, we have several categorical variables with very high cardinality, which works greatly to the disadvantage of machine learning model training with a small data size. Thus, we combined some infrequent levels together to another single level.

Thirdly, we marked the outliers in our dataset with the external information provided by the Company. In the following steps, we trained all models twice with outliers and without outliers.

Finally, we explored some hidden information of our ultimate predicted variable, total actual time. According to the earliest information we have received, the total actual time is equal to the sum of time of 7 individual stages. However, based on our observation, there is a gap between the total actual time and the sum of time for 94% of the records. We named this gap in time as "delta_time" and analyzed it with different methods.

Model Training and selection

To predict yarn dyeing process time, we randomly divided 95% and 5% of raw data into the training set and test set. As mentioned in part 3b, we trained four different machine learning algorithms (Linear regression, Regression tree, Regression tree, XGBoost) with the same training dataset. And during the training process, we used the grid search method to find the optimal hyperparameters for our learning algorithm.

To compare the prediction accuracy of the trained models in the test set, we used root-mean-square error as our measurement.

Fabric Actual Width Prediction

a. The data used (or available for analysis)

The data we used for actual width prediction is internal data provided by Esquel group. The company would like to provide estimates of actual width based on the rest features to facilitate the decision-making process and enhance production efficiency. Four Excel files containing historical data from April 2020 to April 2021 were provided by the company as our raw data.

2020.04-2020.07(fabric width for HKU).xlsx	18 May 2021 at 10:24 PM	11 MB	Microso...k (.xlsx)
2020.07-2020.10(fabric width for HKU).xlsx	11 May 2021 at 7:14 PM	8.2 MB	Microso...k (.xlsx)
2020.10-2021.01(fabric width for HKU).xlsx	11 May 2021 at 7:15 PM	7.3 MB	Microso...k (.xlsx)
2021.01-2021.04(fabric width for HKU).xlsx	6 Jun 2021 at 6:07 PM	5.6 MB	Microso...k (.xlsx)

i) Context and content

After concatenation, there are in total 308249 rows and 25 columns of actual width related records. Among them, actual width is our dependent variable, while the rest of the 24 columns of data would be the dependent variables describing the potential determinants of the actual width statistics. As instructed by the client, the actual width figure will fluctuate against design width, and design width will be considered as the major contributor to the actual width statistics. The rest of the variables, for instance, reed_no will influence how much the actual width statistics deviate from the design width statistics

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 308249 entries, 0 to 67380
Data columns (total 25 columns):
ETD-出货日期                    308249 non-null object
GF_NO-品名                     308249 non-null object
job_no-排单号                   308249 non-null object
fabric_no-布号                 308249 non-null object
fnsend_no-后整送布单             308249 non-null object
Quantity-卷长                  308249 non-null float64
design_width-工艺门幅            308249 non-null float64
Actual_width-实际门幅            308249 non-null float64
Fabric_Type-布种代码             308249 non-null object
Warp_Density-经密              308249 non-null float64
Weft_Density-纬密              308249 non-null float64
F_Actual_count-纬向纱支          308249 non-null object
W_Actual_count-经向纱支          308249 non-null object
Component-成分                 308249 non-null object
Reed_NO-筘号                   308249 non-null float64
SUM_Reeding-入筘数               308249 non-null object
reed_width-筘幅                308249 non-null float64
ReedNo_type-筘号整理类别           308249 non-null object
Finish_list-整理方式             308249 non-null object
Mercerizing-丝光                308249 non-null object
Wash_Method-水洗方式             308249 non-null object
FabricNumber-织物组织类别          308249 non-null int64
Color_Type-是否本白              308249 non-null object
Is_Piece_Dye-是否CPB            308249 non-null int64
Shrinkage-纬向缩水标准(分档前)      308249 non-null float64
dtypes: float64(8), int64(2), object(15)
memory usage: 61.1+ MB
```

ii) Definition of variables

During the exploratory data analytics, we analyzed data sets to summarize their main characteristics, with statistical graphics data visualization methods. The first thing we noted is that there are two types of variables, i.e, numeric variables and categorical variables. Numerical data are expressed in numbers and are continuous. Categorical data is, however, discrete, and is a type of data that is stored into groups or categories with the aid of names or labels. As for categorical data, there are two subtypes of data, i.e, nominal data where data sets are names or labels. Ordinal data includes elements that are ranked, ordered. Our categorical data all fall into the type of nominal data as they are not ranked and ordered. We distinguish data with Python command table.info(), where float number stands for numeric values and object stands for categorical variables. We then exclude those variables that don't have a business meaning, and are just a way to distinguish different rows of data, which hence are unimportant or have lower correlation with the target variables including 'ETD-出货日期', 'GF_NO-品名','job_no-排单号', 'fabric_no-布号', 'fnsend_no-后整送布单', 'Quantity-卷长'.

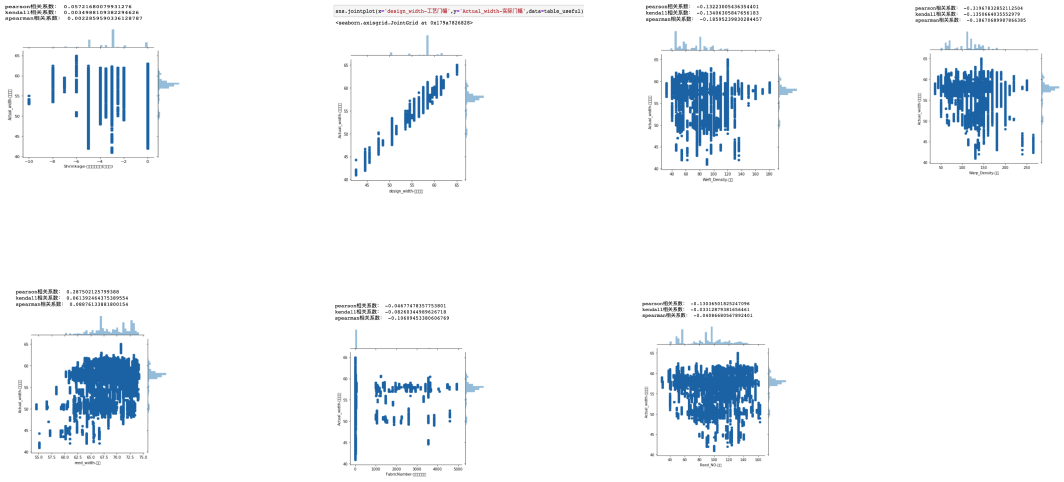
The statistics are summarized as below:

```
design_width-工艺门幅          308249 non-null float64
Actual_width-实际门幅          308249 non-null float64
Fabric_Type-布种代码           308249 non-null object
Warp_Density-经密             308249 non-null float64
Weft_Density-纬密             308249 non-null float64
F_Actual_count-纬向纱支        308249 non-null object
W_Actual_count-经向纱支        308249 non-null object
Component-成分                308249 non-null object
Reed_NO-筘号                  308249 non-null float64
SUM_Reeding-入筘数             308249 non-null object
reed_width-筘幅               308249 non-null float64
ReedNo_type-筘号整理类别         308249 non-null object
Finish_list-整理方式           308249 non-null object
Mercerizing-丝光              308249 non-null object
Wash_Method-水洗方式           308249 non-null object
FabricNumber-织物组织类别        308249 non-null int64
Color_Type-是否本白            308249 non-null object
Is_Piece_Dye-是否CPB          308249 non-null int64
Shrinkage-纬向缩水标准(分档前)     308249 non-null float64
```

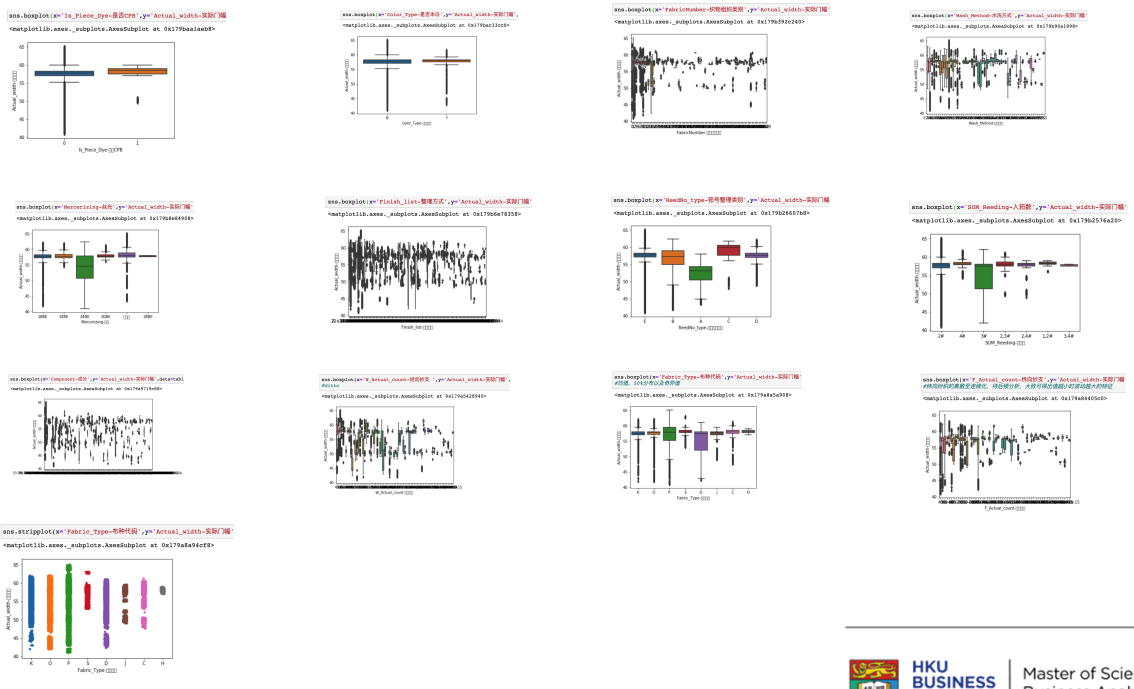
iii) Summary statistics (tables, graphs)

We utilized different methods in evaluating numeric and categorical variables. For numeric variables, we utilized the correlation method and scatterplot. For categorical variables, we utilized box-plot to display the distribution of data based on quantile, minimal, maximal, and outliers.

Characteristics of numeric variables are shown below. For the dependent variable actual width, we noted that it is a weighted average of several normal distributions. For independent variables, we noted that design width represented a strong positive correlation with actual width, with Pearson correlation = 0.97. Other variables are randomly distributed, and no pattern was noted for the rest of the numeric variables.



For categorical variables, no pattern was noted when they are plotted against actual width individually. Further correlations and relationships can only be dug when they are explained collaboratively.



b. The analytical methods

i) Description of the math, conceptual model and assumptions

a) Data Pre-processing

We pre-processed the raw data for the model training purpose. Firstly, we impute all missing values with value 0, since as discussed with the client, the missing data represent null values.

Secondly, in order to make categorical variables understandable for computers, we apply one hot encoding method with Python command pandas.get_dummies that derives the categories based on the unique values in each feature. For instance, to represent Fabric Type in a binary sense, we derived 8 columns for each fabric type and the value will be assigned to 1 if the row belongs to a certain fabric type and 0 if it does not belong to a certain fabric type. Before implementing the one-hot encoding function, the dataset matrix is(308249,18) with 8 numeric and 10 categorical variables. After implementing the function, the data set matrix is (308249, 631) without numerical columns and (308249, 639) with numerical columns.

b) Model Training and selection

To predict actual width, we randomly divided 80% and 20% of raw data into the training set and test set with seed 1234. As mentioned in part 3b, we trained four different machine learning algorithms (Linear regression, Neural Network, Random Forest) with the same training dataset. Also, since actual width and design width have an extremely high correlation, we noted that the one with the dependent variable being actual width generates unsatisfactory results. Hence, we trained two sets of models, one with target variables being actual width, another with target variable being the discrepancy between actual width and design width, i.e, actual width - design width. We will further compare the variance between the two and determine which model can make better predictions. To compare the prediction accuracy of the trained models in the test set, we used mean-square error as our measurement.

Model details of our trained models are shown as below:

Number	Details
1	Random Forest, Dependent Variable Being Actual Width
2	Random Forest, Dependent Variable Being Discrepancy (actual width and design width)
3	Neural Network, Dependent Variable Being Actual Width
4	Neural Network, Dependent Variable Being Discrepancy (actual width and design width)
5	Linear Regression, Dependent Variable Being Discrepancy (actual width and design width)

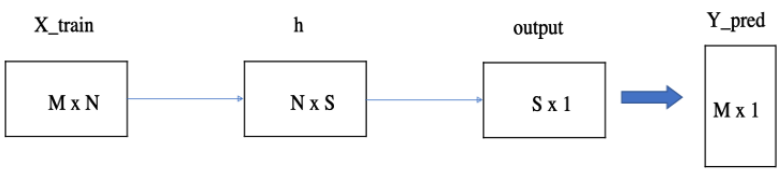
c) Neural Network architecture:

For Neural Network, we chose the following parameters:

i. for network architecture, we chose fully connected NN over Convolutional as we are facing a black box situation and make no assumptions about the features in the data.

ii. Loss function: we chose MSE over Cross_entropy for classification as we have a continuous dependent variable

iii. For the network input matrix, we determine that (m,n)=(308249,639). When there are only 1 hidden layer, h=x·w_$h$idden+b_$h$idden and y_pred=h·w_output. When there are more hidden layers: h_1=x·w_$h$idden1+b_$h$idden1, h_2=h_1·w_$h$idden2+b_$h$idden2, and y_pred=h_2·w_output.



vi: Parameter Tuning: we used the following parameters in the tuning process:

Learning Rate: , [0.002, 0.003, 0.004]

Width: no of neurons, [5,10,20], we originally chose [50,200,2000], but found out the result to be unsatisfactory.

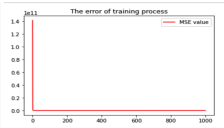
Depth: no of layers, range(5,9)

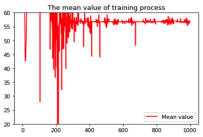
Batch_size = 500

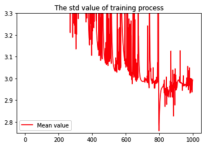
Steps=100000

1 epoch=308249/500 =616, 162 epochs in total

After the tuning process, we noted that MSE, mean and standard deviation converges to a certain value, which means the number of steps is enough and our training is complete.







d) Random forest architecture:

We utilized the following parameters in the tuning process.

Criterion: MSE, MAE

n_estimators, which stands for number of trees: linspace(start = 200, stop = 2000, num = 10)

max_features which stands for the max feature for each tree: ['auto', 'sqrt']

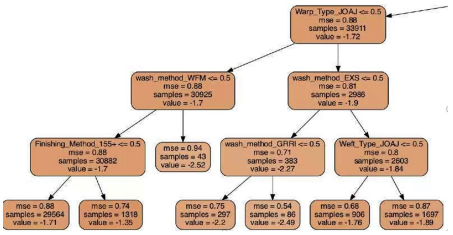
max_depth which stands for the max depth for each tree: linspace(10, 100, num = 10)

min_samples_split which stands for the minimum number of samples required to split an internal node: [2, 5, 10]

min_samples_leaf The minimum number of samples required to be at a leaf node: [1, 2, 4]

Bootstrap: [True, False]

We noted that the final parameter: bootstrap=False,max_features='sqrt', max_depth=20,min_samples_split=10, n_estimators=1200,min_samples_leaf=2. If we take one node to look deeper, we noted that the result of the leftmost node would be -1.71.



Shrinkage Prediction

a. The data used (or available for analysis)

The data we used for shrinkage prediction is internal data provided by Esquel group. The company would like to provide estimates of shrinkage in warp and weft direction based on the rest features to facilitate decision-making process and enhance production efficiency. Historical data are provided for our training purpose.

i) Context and Content

There are in total 113481 rows and 20 columns of shrinkage related records provided by the client. Among them, warp shrinkage and weft shrinkage are our dependent variables, while the rest of the 18 columns of data would be dependent variables describing the potential determinants of the shrinkage statistics. Two models are going to be constructed in order to predict warp shrinkage and weft shrinkage accordingly.

ii) Definition of variables

Similar to actual width prediction, we analyzed data sets to summarize their main characteristics, with statistical graphics data visualization methods. The first thing we noted is that there are two types of variables, i.e, numeric variables and categorical variables.

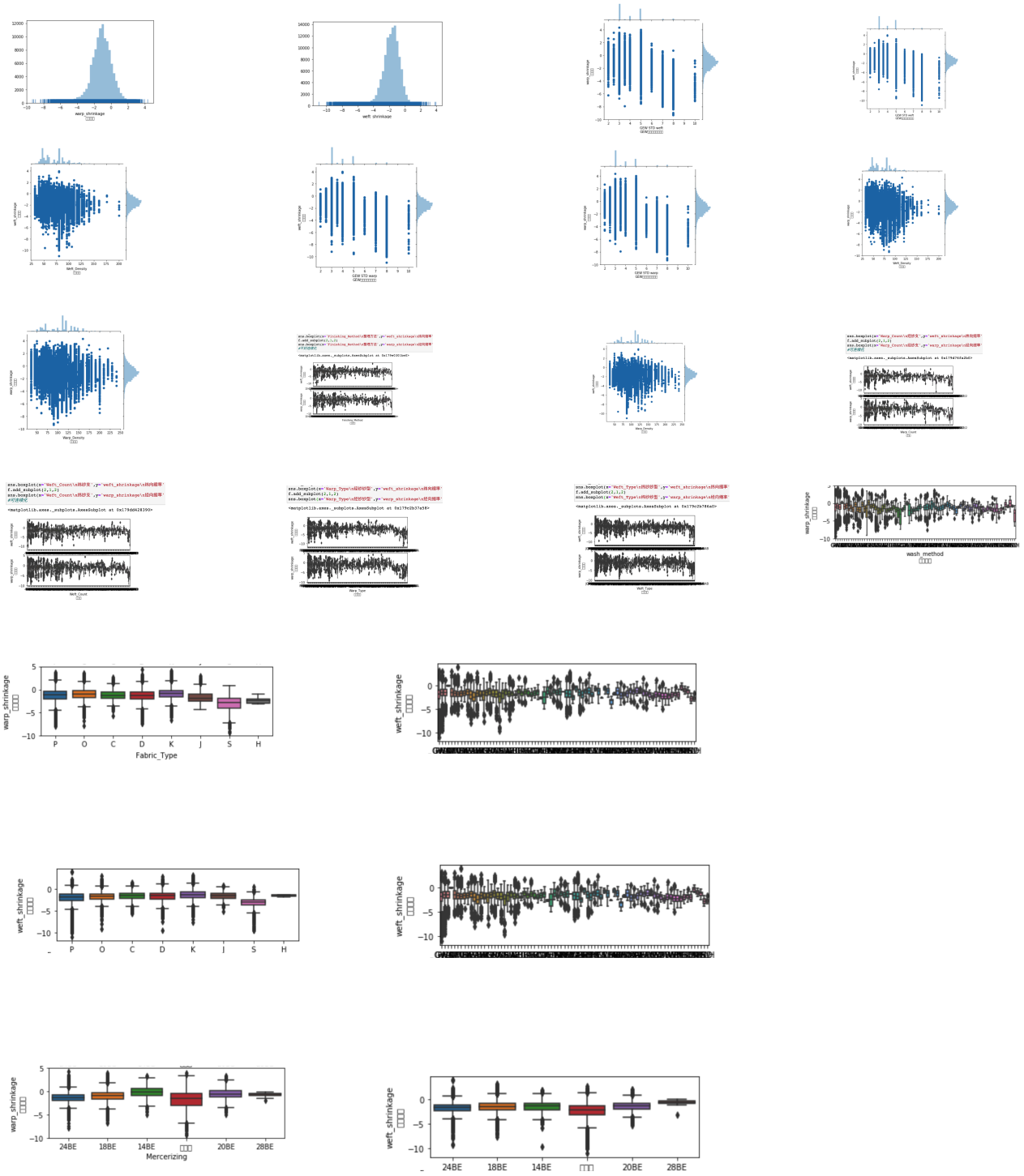
The statistics are summarized as below:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 113481 entries, 0 to 113480
Data columns (total 20 columns):
qf_id
品种ID                    113481 non-null int64
qf_no
品种号                     113481 non-null object
PPO_no
ppo号                     113481 non-null object
fabric_no
布号                      113481 non-null object
send_no
送布单号                    113481 non-null object
warp_shrinkage
经向缩率                  113357 non-null float64
weft_shrinkage
纬向缩率                  113357 non-null object
wash_method
洗水方法                  113481 non-null object
test_date
测试时间                    113481 non-null datetime64[ns]
Fabric_Type
织法                      113481 non-null object
Warp_Type
经纱纱型                    113481 non-null object
Weft_Type
纬纱纱型                    113481 non-null object
Warp_Count
经纱支                     113481 non-null object
Weft_Count
纬纱支                     113481 non-null object
Warp_Density
经纱密度                  113481 non-null int64
Weft_Density
纬纱密度                  113481 non-null int64
Finishing_Method
整理方法                  113481 non-null object
GSM STD warp
GSM经向根容量单标准         113481 non-null float64
GSM STD weft
GSM纬向根容量单标准         113481 non-null float64
Mercerising
丝光碱液                  113481 non-null object
```

iii) Summary statistics (tables, graphs)

We utilized different methods in evaluating numeric and categorical variables. For numeric variables, we utilized correlation method and scatterplot. For categorical variables, we utilized box-plot to display the distribution of data based on quantile, minimal, maximal and outliers.

Characteristics of variables are shown below: We noted that both warp density and weft density are normally distributed. However, independent variables are randomly distributed against dependent variables, and have little correlation with dependent variables. No pattern was noted for the independent variables.



Hence, from our exploratory data analytics, we can note that we cannot simply apply regression methods to the datasets. Instead, machine learning methods like random forest and neural networks should be utilized in our case.

b. The analytical methods

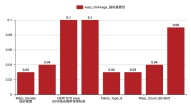
i) Description of the math/conceptual model and assumptions

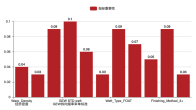
- Data Pre-processing

We pre-processed the raw data for the model training purpose. Firstly, we impute all missing values with value 0, since as discussed with the client, the missing data represent null values. Secondly, we corrected typos from the original dataset. For example, we changed the value 0.8. to 0.8. Thirdly, we applied one hot encoding method with Python command pandas.get_dummies as is in the actual with part. Before implementing the function: we had a matrix of (113481,12) with 4 numeric and 8 categorical variables. After implementing the function, the data set matrix is (113481, 1175) with numerical columns

- Model Training and selection

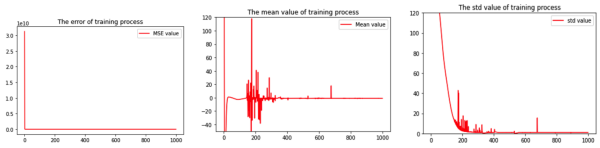
To predict shrinkage, we randomly divided 80% and 20% of raw data to the training set and test set with seed 1234. As it mentioned in part 3b, we trained four different machine learning algorithms (Linear regression, Neural Network, Random Forest) with the same training dataset. Also, we noted that the dataset has extremely high cardinality, which will obscure the order of feature importance and make the accuracy lower, we also used alternatives like distributed random forest. Other methods can include a single feature with n numeric values and log n Boolean features representing n values. After that, to further deduct cardinality, we found out from the first model and also after discussion with client as for the business sense importance that warp yarn count and weft yarn count represents lower significance, and we also tried sets of models without the two variables.





Number	Details
1	Random Forest, With Independent Variables Being All the Variable
2	Random Forest, With Independent Variables Excluding warp yarn count and weft yarn count
3	Neural Network, With Independent Variables Being All the Variable
4	Distributed Random Forest With Independent Variables Being All the Variable
5	Linear Regression, With Independent Variables Being All the Variable

We used the same sets of parameters as in actual width prediction thought the below parameter tuning process and we will not repeat it here.After the tuning process, we noted that MSE, mean and standard deviation converges to a certain value, which means number of steps is enough and our training is complete.For Warp shrinkage, it is shown as below:



For Weft shrinkage, it is shown as below:

For Random Forest: we utilized the same parameters in the tuning process as in actual width prediction. We noted that the final parameter: bootstrap=False,max_features='sqrt', max_depth=20,min_samples_split=10, n_estimators=1200,min_samples_leaf=2.

6. THE RESULTS

a. Potential Improvement

Yarn Dyeing Time Prediction

As we use RMSE as model selection measurement, comparing multi-stage model and single-stage model with XGboost, linear regression, regression tree and random forest algorithms, we finally selected the best model with the smallest RMSE of total time, which is random forest in one step modeling.

Among all multi-stage models and single-stage models, random forest keeps showing the best performance. However, in the one-step modeling process, the random forest gives a slightly smaller RMSE value than the multiple steps process, which is 99.44 and 99.61 respectively. Apart from that, we should see that almost all one-step models tend to have better performance in training data sets with four algorithms. It is due to the dynamics between each stage which are neglected in multi-stage models. Moreover, from the time efficiency perspective, single-stage models are less time-consuming and simpler in execution. For example, the processing time to train a regression tree is five times more than that to train a random forest.

It is also worth mentioning that we applied models to predict the total time delete delta time, which means the sum of time in all stages rather than the total actual time. We achieved the same conclusion to select a random forest as the final model. Although the RMSE values are all smaller than before, in the real business scenario the delta time can not be omitted, we still used the below models for application.

The details are shown in the table below.

RMSE(Predict Total_Actua_Time)				
	XGB	lm	rt	rf
7 steps	119.86	114.87	115.36	113.68
8 steps	108.68	115.11	102.96	99.69
1 step	105.54	104.02	117.25	99.44
RMSE(Predict sum of 7 stages time)				
7 steps	99.67	93.70	94.34	91.07
1 step	93.22	94.99	104.57	91.62

After we determined the effective model, we applied it to the test data set to evaluate the time prediction performance. 30 minutes, 60 minutes and 90 minutes are used as our benchmarks to check the possibility of predicted values lying between actual values plus and minus benchmark numbers. On average, there are 52.05 percent predicted time values that lie within the range of true value minus 60 minutes and plus 60 minutes. When it comes to 30 mins and 90 mins, the percentages are 26.43% and 70.29%.

To have better visualization, we plot and use two lines to respectively represent the relationship between the predicted value and actual value plus and minus a certain number. Those dots lie between two lines means that we correctly predict them with a certain error range. The three graphs below shared the same result that when the total time is shorter, the possibility to predict well is larger. In addition, larger intervals bring higher possibility but less practical importance.

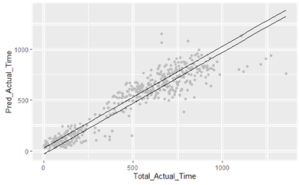


Table1: predicted value = actual value ± 30 (minutes)

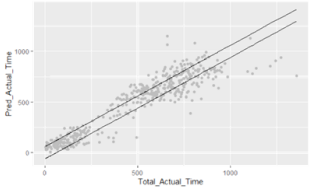


Table2: predicted value = actual value ± 60 (minutes)

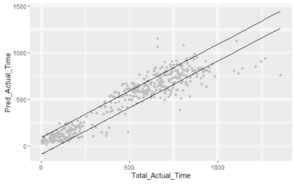


Table3: predicted value = actual value ± 90 (minutes)

Fabric Actual Width Prediction

To summarize, our optimal model predicted accurate results and is notably better than the client's current model based on blind test results.

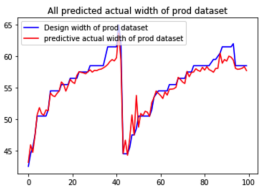
As we use MSE as model selection measurement, comparing models with discrepancy being the dependent variable and actual width being the dependant variable, and using linear regression, regression tree and neural network, we finally determined that the best model with the smallest MSE would be random forest with discrepancy being the dependent variable.

MSE statistics are shown as below:

MSE	Model
0.33	Random Forest, Dependent Variable Being Actual Width
0.21	Random Forest, Dependent Variable Being Discrepancy (actual width - design width)

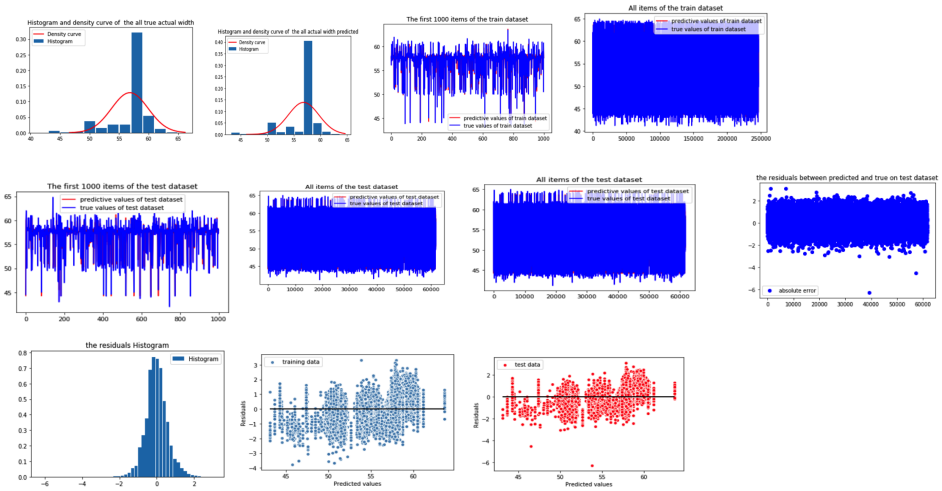
0.42	Neural Network, Dependent Variable Being Actual Width
0.36	Neural Network, Dependent Variable Being Discrepancy (actual width and design width)
0.49	Linear Regression, Dependent Variable Being Discrepancy (actual width - design width)

After we determined the effective model, we finished blind test based on the client's requirements, i.e, we made predictions of actual width based on 300 production data without any knowledge of the ground-truth value, and the client would compare the residuals with their current model to determine whether our model outperform their current model. With this blind test, we have no chance of manipulating data and should be an effective way of evaluating model efficiency. We used residual interval probability distribution to determine model accuracy and noted that our optimal model's residuals are far more concentrated in the low residual area, i.e, (-0.25, 0.25) than the company's model. For further validation results, please refer to the attachment.



Range	Company's version	Optimal Model's version
[-0.25, +0.25]	41%	64%
[-0.50, +0.50]	72%	83%
[-0.75, +0.75]	83%	92%

Also, to have better visualization, we plot the comparison between the predicted value and actual value in terms of histogram, tables and residuals. We noted that the predicted value has a similar mean and standard deviation versus the ground-truth value from the histogram, and the predicted value shares the same trend with the ground-truth value. Also, as for residuals, we noted that residuals fall between [-2, 2], concentrated around 0, follow a normal distribution in the histogram and are randomly distributed in the scatterplot. We can conclude that our model generates considerably accurate results.



Also, feature importance can be generated from the random forest model. We noted that design width, reed_no, component are the most important factors in determining actual width.

Shrinkage Prediction

To summarize, our optimal model predicted accurate results and is notably better than the client's current model based on blind test results.

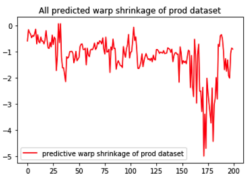
As we use MSE as model selection measurement, comparing models with discrepancy being the dependent variable and actual width being the dependant variable, and using linear regression, regression tree, distributed random forest and neural network, we finally determined that the best model with the smallest MSE would be random forest with discrepancy being the dependent variable.

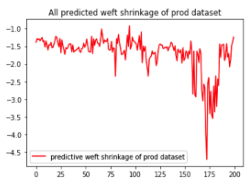
MSE statistics are shown as below:

Warp	Weft	Details

0.86	0.77	Random Forest, With Independent Variables Being All the Variable
0.41	0.32	Random Forest, With Independent Variables Excluding warp yarn count and weft yarn count
0.69	0.58	Neural Network, With Independent Variables Being All the Variable
0.58	0.52	Distributed Random Forest With Independent Variables Being All the Variable
0.98	0.92	Linear Regression, With Independent Variables Being All the Variable

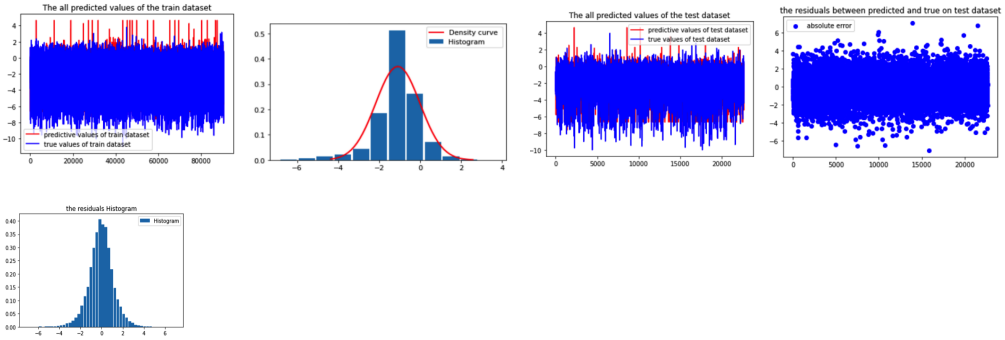
After we determined the effective model, we finished blind test based on the client's requirements, i.e, we made predictions of shrinkage based on 200 production data without any knowledge of the ground truth value, and the client would compare the residuals with their current model to determine whether our model outperform their current model. With this blind test, we have no chance of manipulating data and should be an effective way of evaluating model efficiency. We used residual interval probability distribution to determine model accuracy and noted that our optimal model's residuals are far more concentrated in the low residual area, i.e, (-0.5, 0.5) than the company's model. For warp shrinkage, the probability of residual falling into the range of (-0.5, 0.5) is 68% for our optimal model and 67% for the company's current model. For weft shrinkage, the probability of residual falling into the range of (-0.5, 0.5) is 72% for our optimal model and 63% for the company's current model. For further validation results, please refer to the attachment.



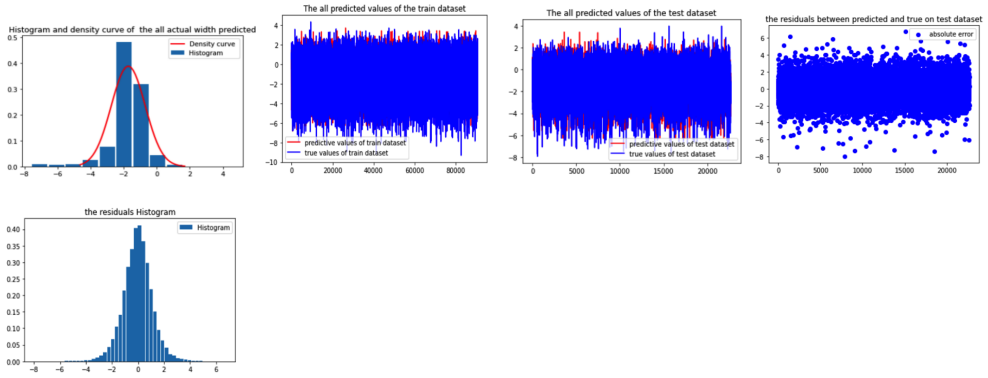


Also, to have better visualization, we plot the comparison between predicted value and actual value in terms of histogram, tables and residuals. We noted that the predicted value has similar mean and standard deviation versus the ground truth value from the histogram, and the predicted value shares the same trend as ground truth value. Also, as for residuals, we noted that residuals fall between [-2, 2], concentrated around 0, follow a normal distribution in the histogram, and are randomly distributed in the scatterplot. We can conclude that our model generates considerably accurate results.

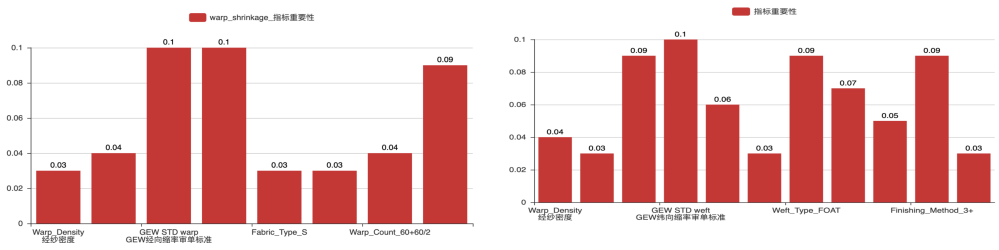
Warp shrinkage graphs are shown as below:



Weft shrinkage graphs are shown as below:



Also, feature importance can be generated from the random forest model. We noted that GEW STD, Finishing Method, Weft and Warp Density are the most important factors in determining shrinkage.



b. Business Implications

Yarn Dyeing time prediction

a) Time prediction capability

We use our final models to provide time prediction capability for the dyeing process. Using the capability, our clients received benefits in the following ways. Firstly, for employees, It could assist inexperienced employees to predict the dyeing time and provide reference to increase their working efficiency. Secondly, for managers, it could improve production scheduling and staff management, leading to reduced resource waste and labor cost of the Company in the end. Finally, for the industry, with higher stability and predictability, it could bring better efficient performance for both upstream and downstream.

b) Web application

Based on the capability, we also provide an interactive web application by the shiny package with r language. By inputting various variables related to the dyeing process, such as fabric type, colorist name, machine name, etc. and targeted time range on the left side of the panel, it could automatically update and show the predicted total time and the possibility of respectively time range with the plot of probability distribution on the right side. With this application, the operators could predict dyeing time for certain tasks anytime and anywhere by opening the URL link (https://irisnan.shinyapps.io/Esquel_DyeingTime/) with any browser

Fabric Actual Width and Shrinkage Prediction

Since actual width and shrinkage predictions collaboratively solve the problem of what fabric length should be chosen, we will illustrate the business implications together.

a) Highly Accurate Predictions

We provided highly accurate estimates of actual width as well as shrinkage, which could be used in the reduction of uncertainty of material. In this way, the fabric length is more certain and material usage can be minimized. It can lead to a reduction in material waste and boosts higher efficiency.

b) Web application and batch data processing tools

Based on the capability, we can provide solutions for creating an interactive web application. By inputting various variables related to actual width, such as fabric type,reed_no and etc. and predicted actual width and shrinkage can be generated.

We also allow the processing of batch data. When a batch of data is imported as input, the batch result of predicted actual width and shrinkage can be generated.

c) Data Analytics

As is mentioned in the potential improvement part, we made the client aware of the potential factors in determining actual width and shrinkage. The client can further optimize their manufacturing process based on the feature importance given.

c. Comments from the Client

Yarn Dyeing time prediction

a) Higher accuracy requirement

As we discovered in the EDA process, the average total dyeing time is over 440 mins, so we choose 30min,60min, 90min as the error range to increase the accuracy of prediction. The overall possibility of a plus or minus 30 minutes error range is roughly 26% and the possibility is higher (up to 60%) when the total time is shorter. Based on that, our clients present requests for higher level accuracy from that to nearly 80%.

The current limitations are obvious. On one hand, our data have less than ten thousand records with less variety and uncertain accuracy. On the other hand, due to the time limitation, we only apply some machine learning modeling approaches and basic model selection methods. There are still lots of works that could be done on feature mining and modeling such as with the neural network approach. Future work could be contributed on large-scale data sets and various information such as date, temperature, reports to the abnormal, as well as more modeling methods to be applied.

b) Capacity of batch prediction

For batch use scenarios, our clients expect to predict multiple records and do not need to manually choose variables one by one.

The present web application is designed for operators to have a user-friendly experience without understanding the program and coding to achieve the results. For mass usage, we provide two following solutions. Firstly, algorithm engineers could upload a data file with multiple records with our open-source modeling codes and run the programs to get all results. Secondly, the IT department could modify the application codes to get inputs from the interface of the internal system and do the calculation automatically. All related codes and detailed instructions are kept.

Fabric Actual Width and Shrinkage Prediction

Given that actual width and shrinkage predictions collaboratively solve the problem of what fabric length should be chosen, we will illustrate the business implications together. As is mentioned in the business implication, blind tests are done for both actual width and shrinkage and the prediction results are far more accurate than the company's current model. It is very impressive and can be recognized as an extremely successful implementation of the knowledge we gained from the MscBA curriculum, given the large volume of data, and high cardinality categorical variables. In order to generate results with high efficiency, we finished 15 models in total using various advanced machine learning methods. The client highly appreciates our efforts and professionalism. Also, business application design instructions are provided in order that the client can truly apply our model into their production environment. Last but not least, original codes are provided, so that the client can easily apply our logic, and make amendments whenever necessary.

7. APPENDIX

a. Technical Details Of the Model, especially particular aspects/adjustments made specifically for the project

Linear regression: default settings

Regression tree:

Method: "anova" (indicates the outcome is continuous)

Minsplit: 1 (minimum number of objects for a split)

Minbucket: 1, (minimum number of objects in terminal nodes)

Random forest:

Ntree: 1000

XGBoost:

Input type: sparse matrix (We have transformed all categorical variables to dummy variables)

Booster: gbtree

Tree_method: auto: Use heuristic to choose the fastest method.

b. Additional Tables and Graphs

Table of yarn dyeing prediction outputs(test dataset)

prediction_outputs.xlsx in folder yarn_dyeing_models

c. Sample (or toy) Data With Data Dictionary

EDA results of dyeing time prediction data

Codes file

Web application for dyeing time prediction

Codes file and instruction file

Work Allocation

GROUP MEMBERS	WORKS DONE
Tang Yilan	Finished Exploratory Data Analytics; Fine-tuned Neural Network Model, built LM, RF, DRF model of Shrinkage and Actual Width; Blind test and validation; Prepared slides contents; Drafted Report Part 5&6
Cheng Chuhang	Pre-processed yarn-dyeing assignments data; Performed literature review; Built and trained yarn dyeing time models; Prepared presentation slides; Drafted report.
Liu Ningbo	Model Brainstorming, Team Coordination, Slides Fine Tuning, Report Drafting, Report Refining, File Consolidation
Zheng Zhikang	Model discussion and data pre-processing of actual width & shrinkage prediction; Prepared for slides; Formatting and finalized the report; Report writing.
Zhao Yuedian	Exploratory data analysis; Built and trained dyeing time prediction model; Prepared presentation slides; Drafted report.
Chen Mengsha	Model discussion and brainstorming; Built Neural Network for Actual Width and Shrinkage Predictions; Team Coordination, Slides Fine Tuning, Report Refining, File Consolidation
You Yi	Model discussion and brainstorming for prediction of Shrinkage and Actual Width; Slides preparation; Report writing;
Nan Yun	EDA; Web application development; Slides preparation and presentations; Drafted Report Part 3