HKU BUSINESS SCHOOL 港大經管學院

Master of Science in Business Analytics
理科碩士(商業分析)

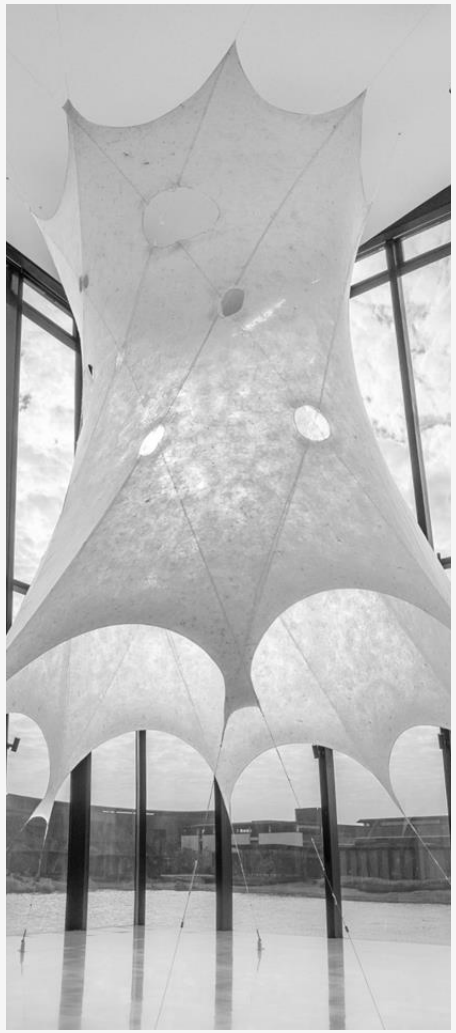
ESQUEL GROUP

Capstone Report

- HBS Sharks

Esquel Group

Leading Textile and Apparel Manufacturer



Project Objective

Apply data analytical skills and develop models to facilitate decision-making in project scheduling, reduce waste and enhance production efficiency.

Predicting yarn-dyeing time

- ✓ Identify the influencing factors of dyeing time and reduce the prediction error

Predicting fabric width

- ✓ Determine the minimum required length of fabric

Predicting fabric shrinkage

- ✓ Determine the range of shrinkage of fabric

Part 1

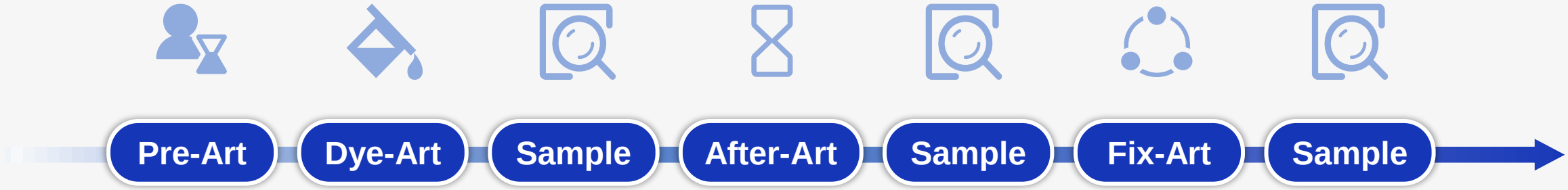
Prediction of yarn-dyeing time

Identify the influencing factors of dyeing time and reduce the prediction error.

Table of Contents

1. Client Problem
2. Objectives and Work Done
3. Exploratory Data Analysis
4. Methodologies
5. Performance Evaluation
6. Commercial Implications
7. Limitations and Future Work

Client Problem



Example 1	12	198	0	99	9	233	233	∑=552min
Example 2	147	252	75	105	21	424	424	∑=937min

Why time varies for similar tasks of similar dye weights?

Objectives and Work Done

Project objective:

Provide estimates of total time needed for dyeing process to facilitate decision-making process and enhance production efficiency

Raw data:

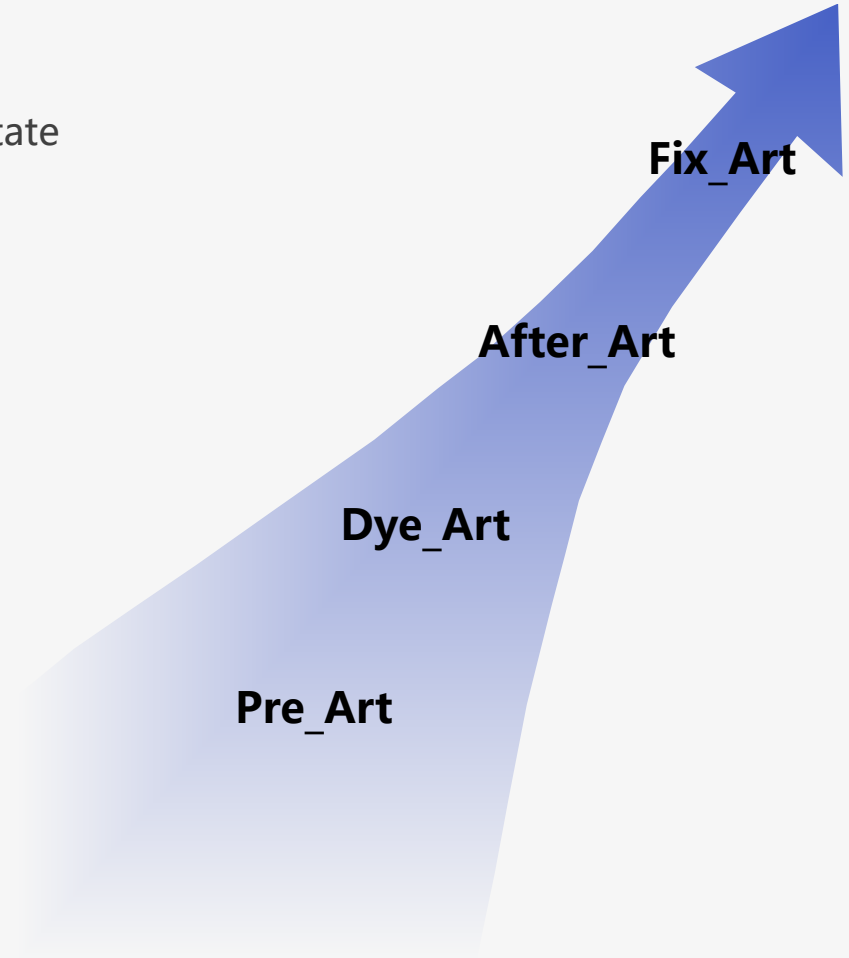
9744 rows, 23 columns

Software employed:

R, Excel and Python

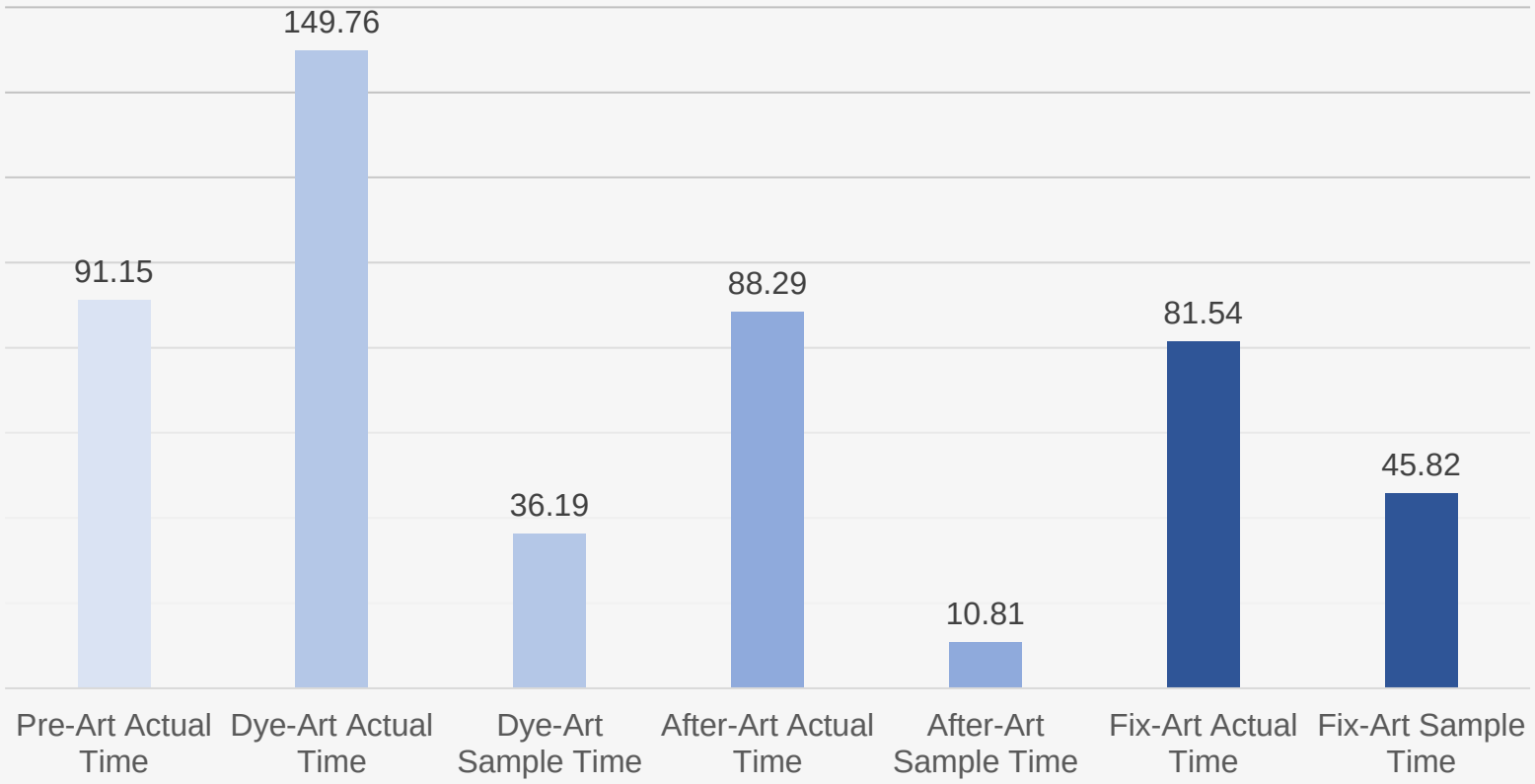
Machine learning algorithms:

Linear Regression, Regression Tree, and Random Forest



Exploratory Data Analysis

Average Value of processing time



Categorical variables

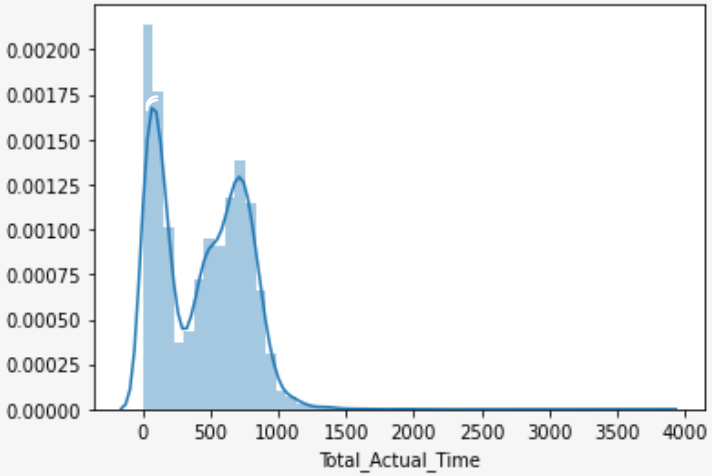
-Machine Name
-Color
-Colorist
-Fabric
-Remark
-Art No

Continuous variables

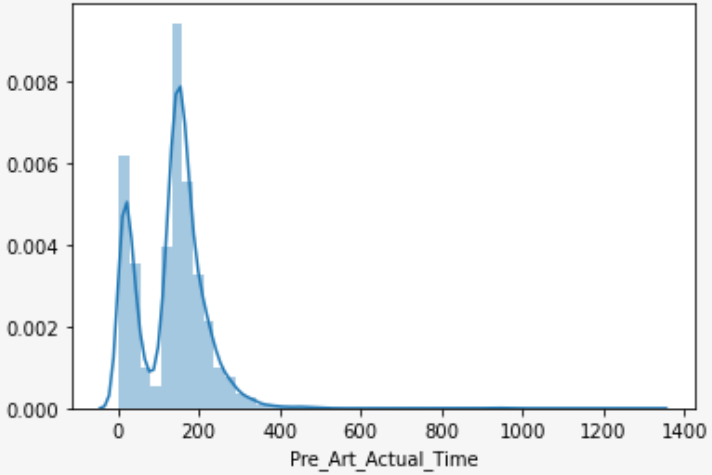
-Dyestuff Usage
-Dyeing_Ratio
-Dye_Weight

Exploratory Data Analysis (Cont' d)

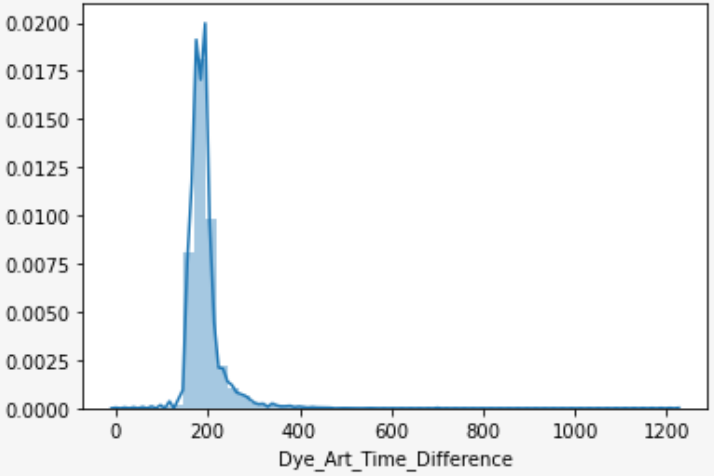
Total Time

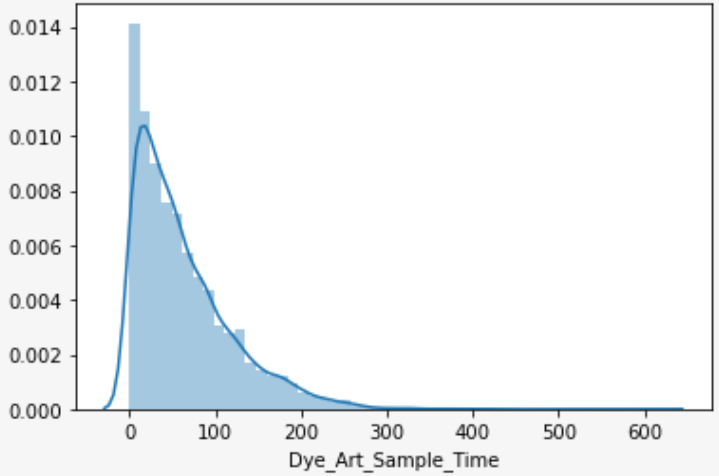


Pre-art Stage



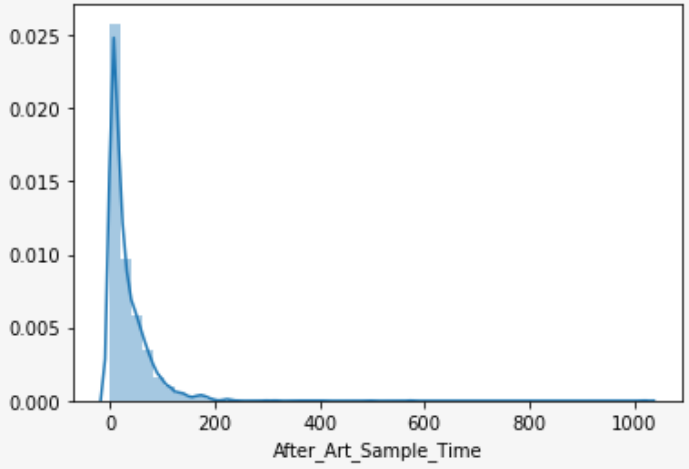
Dye-art Stage

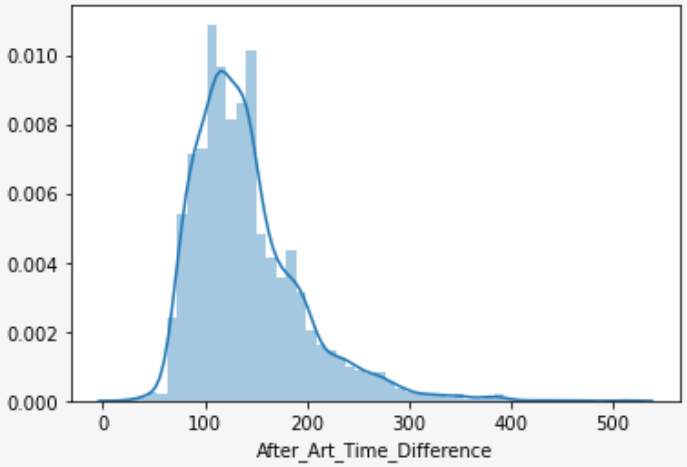




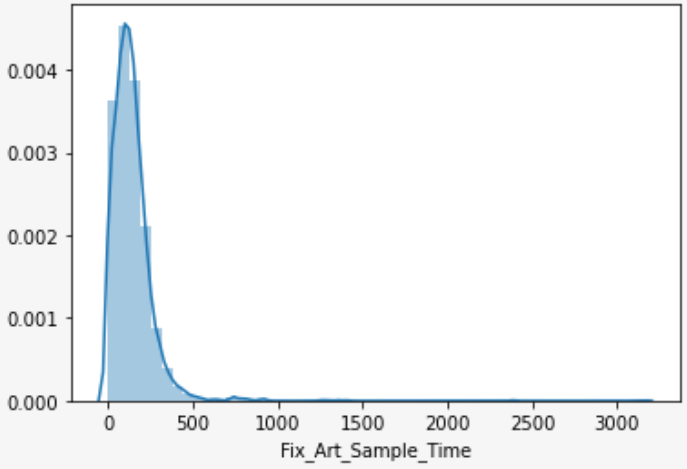
Exploratory Data Analysis (Cont' d)

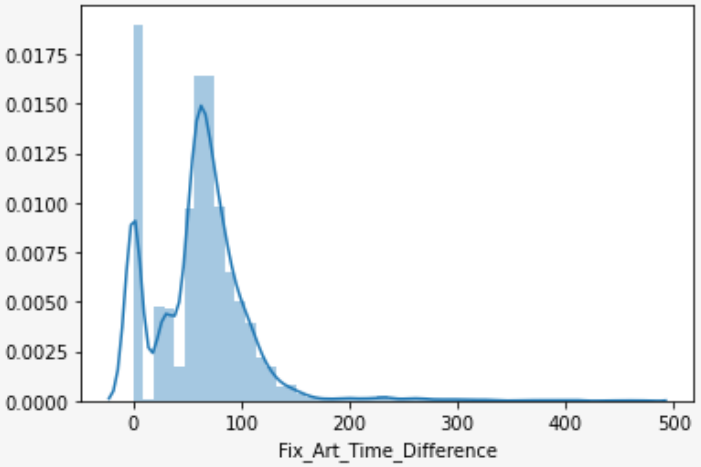
After-art Stage





Fix-art Stage





Methodologies

Step 1: Experimental design and data acquisition

- Raw data: 9744 lots of dyeing assignments
- Dyeing process is divided into 7 stages, namely pre-art, dye-art, dye-art sample, after-art, after-art sample, fix-art and fix-art sample
- Raw data is randomly divided into a training dataset and a test dataset (95% and 5%, respectively)

Step 2: Data pre-possessing

- Label encoding non-numeric or symbolic features
- Imputation missing value using “0”
- Outliers are examined

Step 3: Feature subset selection

- Screen out input features which are unimportant or have lower correlation with the target variables

Methodologies (Cont' d)

Step 4: Model selection

- Performance measurement: root mean square error (RMSE)

No	Linear Regression	Regression Tree	Random Forest	XGBoost
Basic theory	LR allocates weight parameter, theta for each of the training features. The predicted output(h(θ)) will be a linear function of features and θ coefficients.	RT recursively partitions a complex decision into several branched and simplified decisions and may lead to an easier solution.	RF works by creating various decision trees in the training phase and output class labels those have the majority vote.	XGBoost is an implementation of gradient boosted decision trees designed for speed and performance.
Linearity	Only linearity	Supports non linearity	Supports non linearity	Supports non linearity
Pros	-Easy and simple implementation -Space complex solution -Fast training -Value of θ coefficients gives an assumption of feature significance	-Easy -Speed -Transparent process -Handle both numerical and categorical data -Handle both numerical and categorical data	-Powerful and highly accurate -No need to normalizing -Can handle several features at once -Run trees in parallel ways	-Fast execution speed -Working well in small to medium dataset. -Parallel processing
Cons	-Applicable only if the solution is linear -Algorithm assumes the input residuals (error) to be normal distributed -Algorithm assumes input features to be mutually independent	-May overfit -Pruning process large -Optimization unguaranteed -Complex calculations -Deflection high	-Slow -They are biased to certain features sometimes -Can not be used for linear methods -Worse for high dimensional data	-Have to manually create dummy variable/ label encoding for categorical features.

Methodologies (Cont’d)

Step 4: Model selection (Cont’d)

	Multi-Stage Model	Single-Stage Model
Definition	Build 7 models for each manufacturing stage separately. The final prediction value will be the sum of the prediction value of each stage	Treat the entire dyeing process as a holistic procedure. Building only one model to forecast the total actual time directly
Advantage	Bring more insights to each manufacturing stage	Simple in execution
Disadvantage	Neglect the dynamics between each stage Complex in execution	The outliers of each stage are ignored

Methodologies (Cont' d)

Step 4: Model selection (Cont' d)

Method of handling △time (Total Actual Time – Sum of time)

1. Delete △time and predict the sum of time of each stages
2. Build a separate model to predict △time by using the same variables as other models
3. Using predicted time of each stage to predict total actual time

Model selection measurement:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

Performance Evaluation

Model Performance Comparison

XGboost | Linear Regression | Regression Tree | Random Forest

RMSE (Predict Total Actual Time)

	XGboost	Linear Regression	Regression Tree	Random Forest
7 steps	119.86	114.87	115.35	113.68
8 steps	108.68	115.11	102.96	99.61
1 steps	105.54	104.01	117.24	99.44

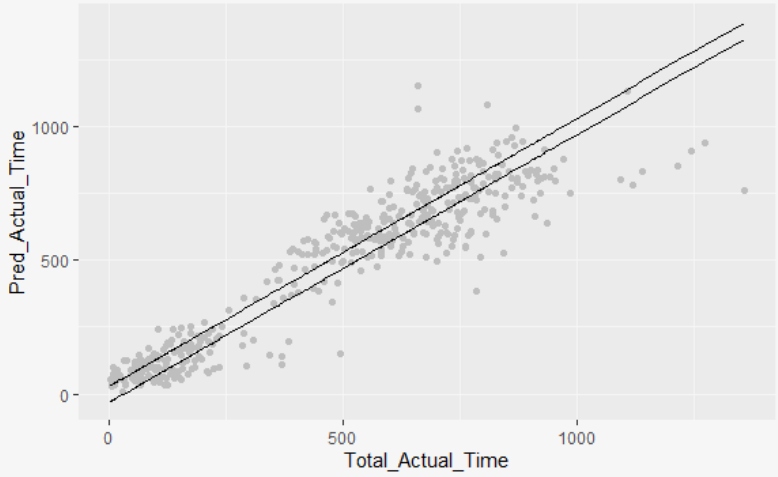
RMSE Delete △time

	XGboost	Linear Regression	Regression Tree	Random Forest
7 steps	99.67	93.7	94.34	91.06
1 steps	93.22	94.99	104.57	91.62

Performance Evaluation (Cont' d)

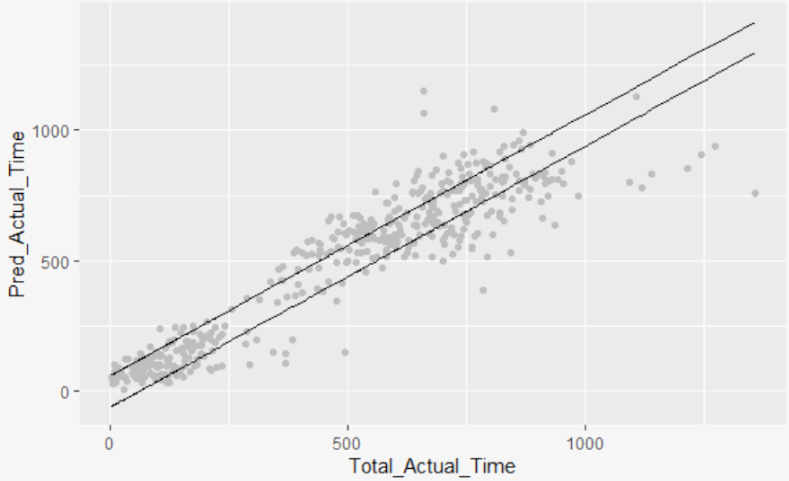
Model Performance Testing

-30min ~ +30min



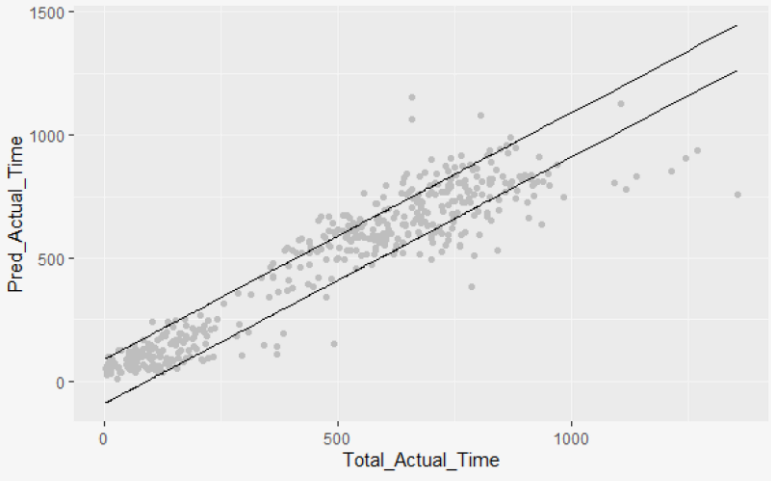
Prob. ≈ 26.43%

-60min ~ +60min



Prob. ≈ 52.05%

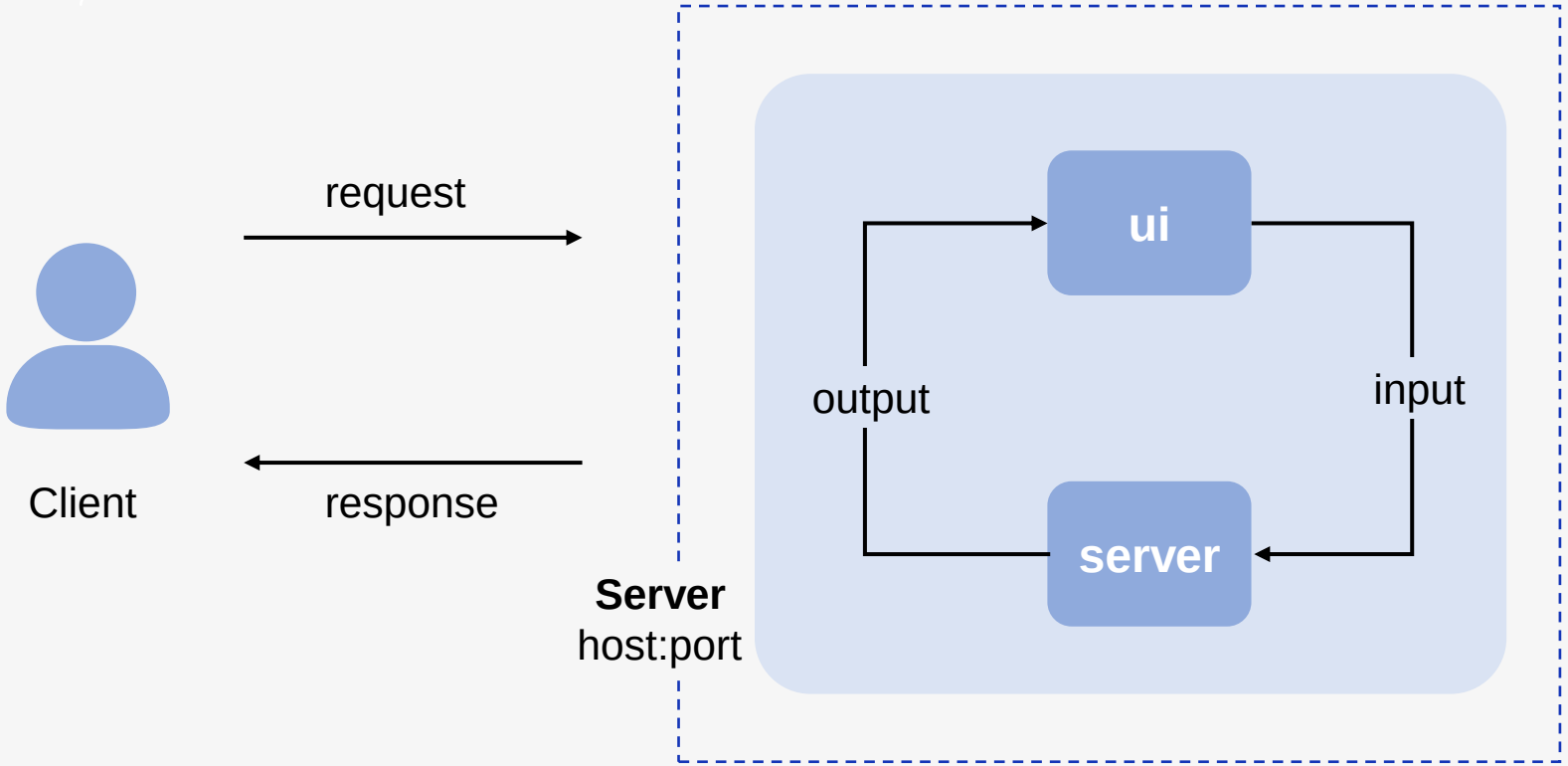
-90min ~ +90min



Prob. ≈ 70.09%

Commercial Implications

User Friendly Web Application



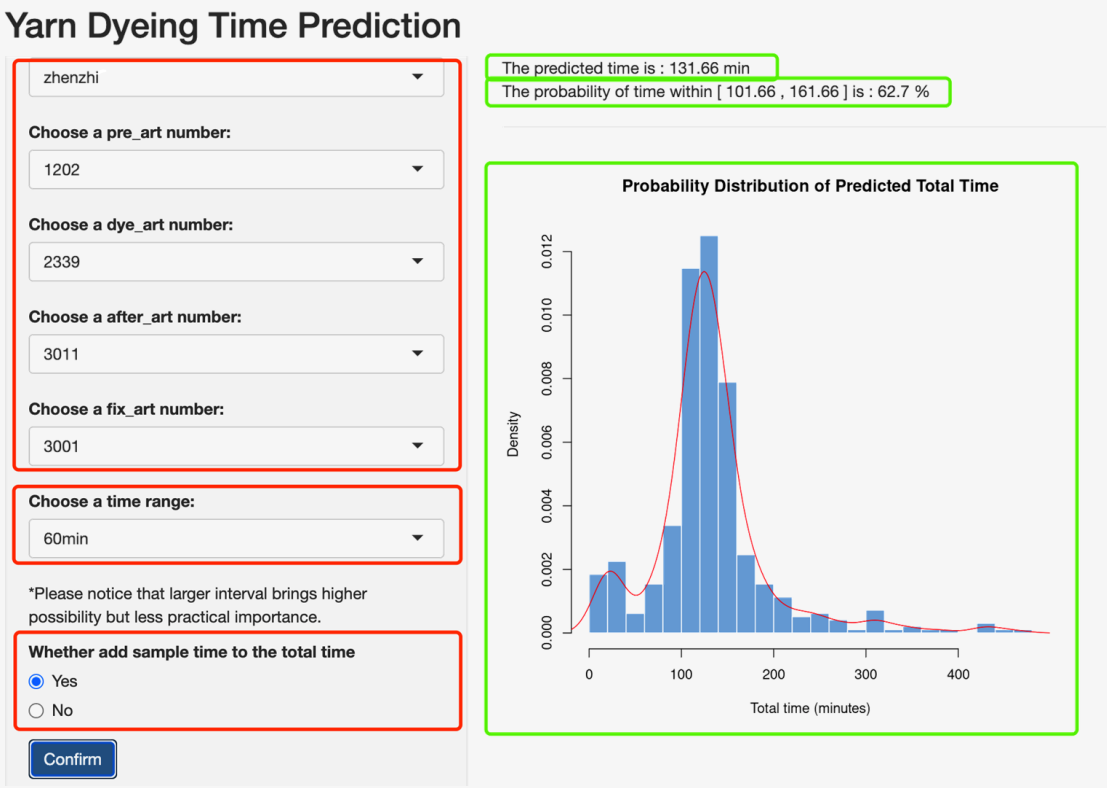
Language
-R

Package
-Shiny

End Product
-Cloud service
-Desktop Application

Commercial Implications (Cont' d)

User Friendly Web Application



URL:
https://irisnan.shinyapps.io/EsquelDyeingTime/

Input

- Various dyeing variables
- Targeted time range
- Time Type
  (Whether add sample time to the total time)

output

- Predicted total time value
- Time range with error percentage
- Probability distribution of total time

Limitations and Future Work

Raw Data

Limitations

- Low volume
- Less variety
- Uncertain accuracy

Future Works

- Historical and real time data
- Set SOP
- Gather more information. (e.g. time)

Methodologies

Limitations

- Few models approach
- Basic selection methods
- Manual parameters tuning process

Future Works

- Feature mining
- Neural network
- Others

Part 2

Prediction of Actual Width

Provide estimates of actual width based on historical data from April 2020 to April 2021 to facilitate decision-making process and enhance production efficiency

Table of Contents

1. Objectives and Data Source
2. Exploratory Data Analysis
3. Methodologies
4. Neural Network
5. Random Forest
6. Comparison of Models
7. Prediction of Production Data
8. Determinants
9. Commercial Implications

Objectives and Data Source

Project objective:

Provide estimates of actual width based on historical data from April 2020 to April 2021 to facilitate decision-making process and enhance production efficiency

Raw data:

308249 rows, 25 columns

Software employed:

Excel and Python

Machine learning algorithms:

Regression, Random Forest, Neural Network

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)

Raw Data Provided By Client

```
table_all.info()
#类型为object的进行离散分析，类型为float的进行连续分析
#进行离散化分析，发现种类数量过多的，可以连续化后再分析
```

```
<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
```

Data Cleansing and Pre-Processing

Feature Subset Selection:

Screen out input features which 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','Actual_width-实际门幅

Data Cleansing:

- Fill in blank entries with 0
- Correcting Typos: 0.8. to 0.8

Data after Cleansing:

- 308249 rows, 19 columns, i.e, 18 independent variables

Variable Type Determination:

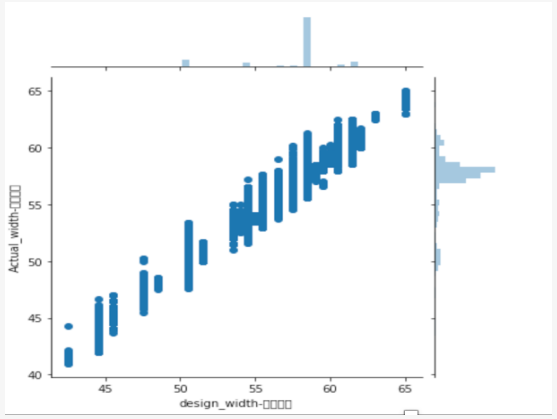
- Numeric: Joint Distribution and Correlation
- Categorical: Box Plot

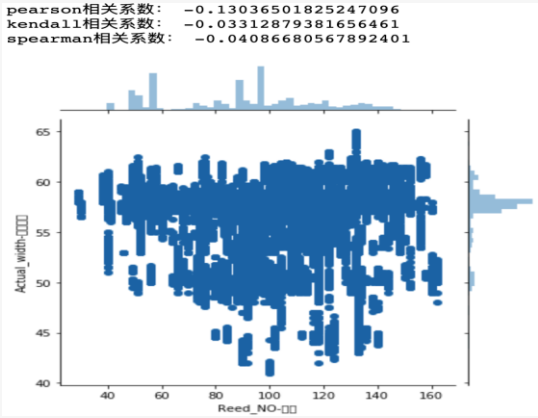
```
design_width-工艺门幅          308249 non-null float64
Actual_width-实际门幅          308249 non-null float64
Fabric_Type-布种代码           308249 non-null object
Warp_Density-经密            308249 non-null float64
Weft_Density-纬密            308249 non-null float64
F_Actual_count-纬向纱支        308249 non-null 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
```

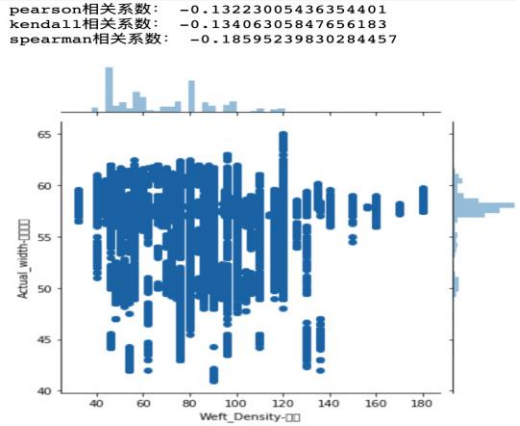
Exploratory Data Analysis

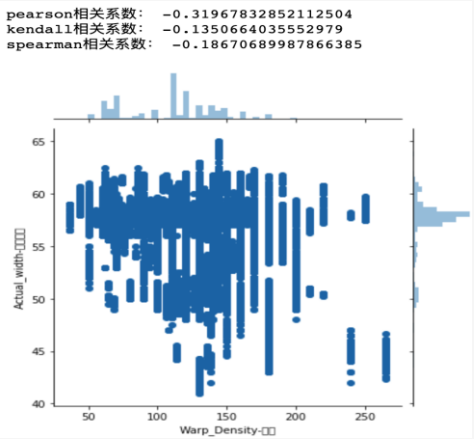
Numerical:

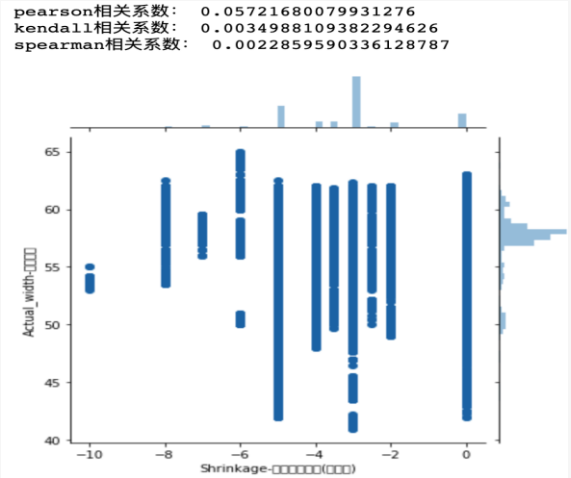
- Only Design Width has strong correlation with Actual Width, Pearson correlation = 0.97
- No pattern was noted for other independent variables

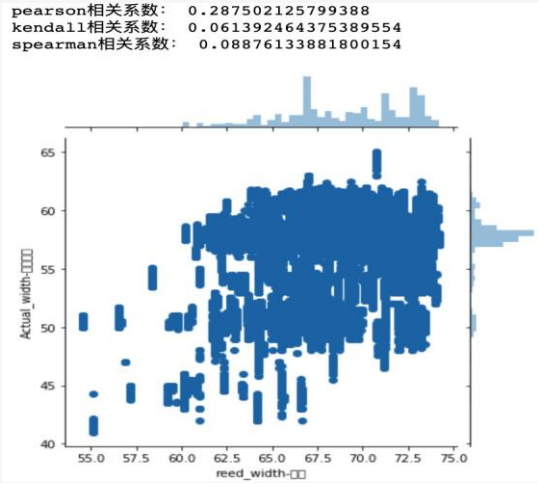


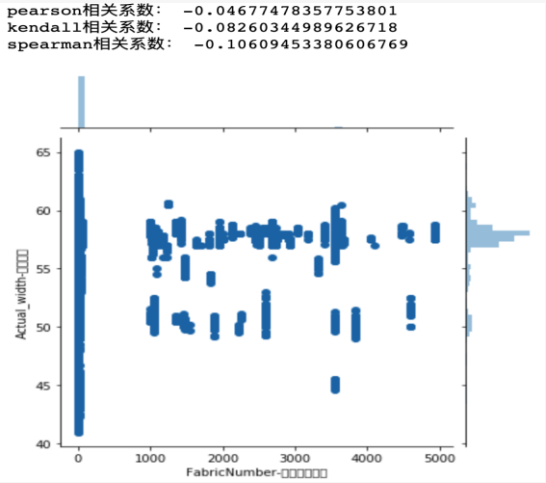








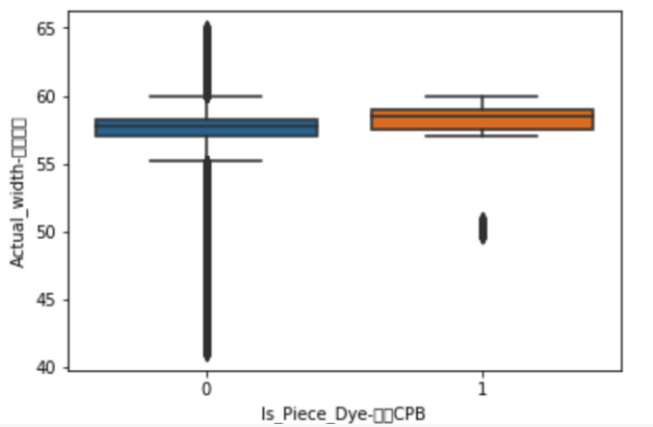


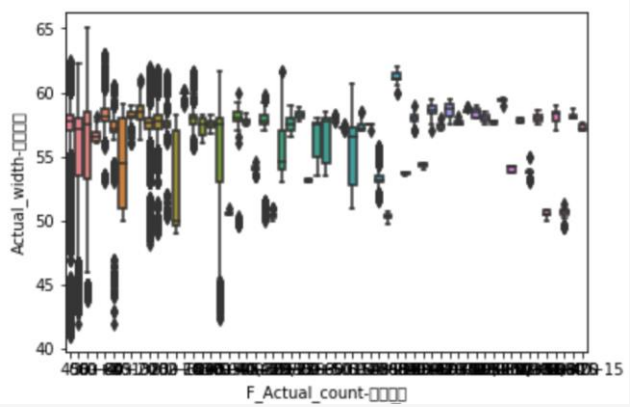


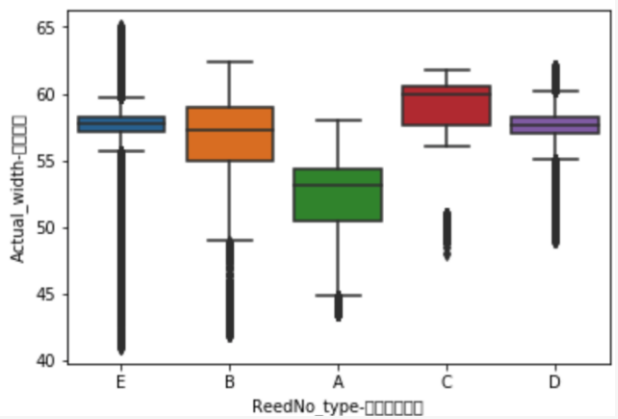
Exploratory Data Analysis

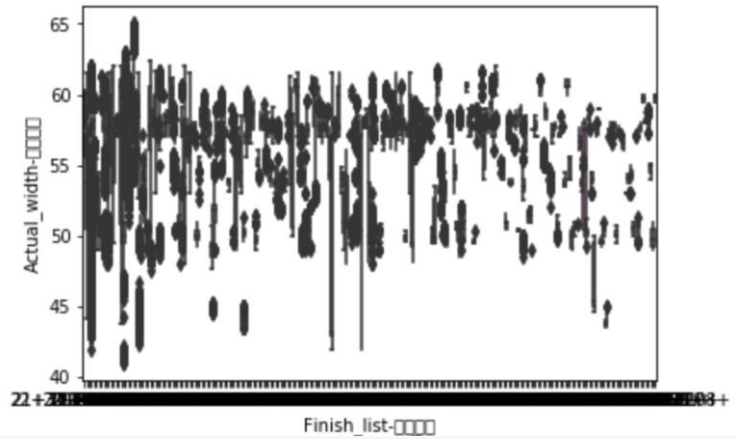
Categorical:

- Some example plots are shown here that indicates there is no pattern was noted for most of independent variables
- For variables like: F_Actual_count-纬向纱支 and W_Actual_count-经向纱支, we will treat them as numerical variables.









Methodologies

Go deeper into categorical variable: one hot_coding with pandas.get_dummies

- Before implementing the function: (308249,18) with 8 numeric and 10 categorical variables

	design_width-工艺门幅	Fabric_Type-布种代码	Warp_Density-经密	Weft_Density-纬密	F_Actual_count-纬向纱支	W_Actual_count-经向纱支	Component-成分	Reed_NO-筘号	SUM_Reeding-入筘数	reed_width-筘幅
0	58.5	D	165.0	95.0	40	40	100.0% CTN	140.0	2#	69.46
1	58.5	D	165.0	95.0	40	40	100.0% CTN	140.0	2#	69.46
2	58.5	D	165.0	95.0	40	40	100.0% CTN	140.0	2#	69.46
3	58.5	D	165.0	95.0	40	40	100.0% CTN	140.0	2#	69.46
4	58.5	D	165.0	95.0	40	40	100.0% CTN	140.0	2#	69.46

- After implementing the function: (308249, 631) without numerical columns and (308249, 639) with numerical columns

	Fabric_Type_C	Fabric_Type_D	Fabric_Type_H	Fabric_Type_J	Fabric_Type_K	Fabric_Type_O	Fabric_Type_P	Fabric_Type_S
0	0	1	0	0	0	0	0	0
1	0	1	0	0	0	0	0	0
2	0	1	0	0	0	0	0	0
3	0	1	0	0	0	0	0	0
4	0	1	0	0	0	0	0	0

5 rows x 631 columns

	design_width-工艺门幅	Warp_Density-经密	Weft_Density-纬密	Reed_NO-筘号	reed_width-筘幅	FabricNumber-织物组织类别	Is_Piece_Dye-是否CPB	Shrinkage-纬向缩水标准(分档前)	Fabric_Type_C	Fabric_Type_D	...	W_Actual_
0	58.5	165.0	95.0	140.0	69.46	15	0	-2.0	0	1	...	
1	58.5	165.0	95.0	140.0	69.46	15	0	-2.0	0	1	...	
2	58.5	165.0	95.0	140.0	69.46	15	0	-2.0	0	1	...	
3	58.5	165.0	95.0	140.0	69.46	15	0	-2.0	0	1	...	
4	58.5	165.0	95.0	140.0	69.46	15	0	-2.0	0	1	...	

5 rows x 639 columns

Methodologies

Data and Model Selection

- Performance measurement: mean square error (MSE)

No	Regressions	Neural Network	Random Forest
Basic theory	Multinomial Logit Regression Polynomial Regression	A neural network is a series of algorithms that endeavors to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates.	RF works by creating various decision tree in the training phase and output class labels those have the majority vote.
Linearity	No no linearity	Supports non linearity	Supports non linearity
Pros	-Easy and simple implementation -Space complex solution -Fast training -Value of θ coefficients gives an assumption of feature significance	-Neural networks are good to model with nonlinear data with large number of inputs; Once trained, the predictions are pretty fast. Neural networks can be trained with any number of inputs and layers. Neural networks work best with more data points.	-Powerful and highly accurate -No need to normalizing -Can handle several features at once -Run trees in parallel ways
Cons	-Applicable only if the solution is linear -Algorithm assumes the input residuals (error) to be normal distributed -Algorithm assumes input features to be mutually independent (no co-linearity)	Neural networks are black boxes. It is computationally very expensive and time consuming to train with traditional CPUs. Neural networks depend a lot on training data. This leads to the problem of over-fitting and generalization. The mode relies more on the training data and may be tuned to the data.	-Slow -They are biased to certain features sometimes -Can not be used for linear methods -Worse for high dimensional data

Neural Network for Prediction of Actual Width

Determining Network Architecture and Loss Function

- Fully connected NN over convolutional: black box; makes no assumptions about the features in the data; no such large volume of data
- MSE over Cross_entropy for classification: Continuous Y
- Network Infrastructure:

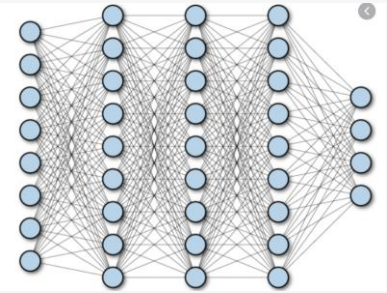
$$(m, n) = (308249,639)$$

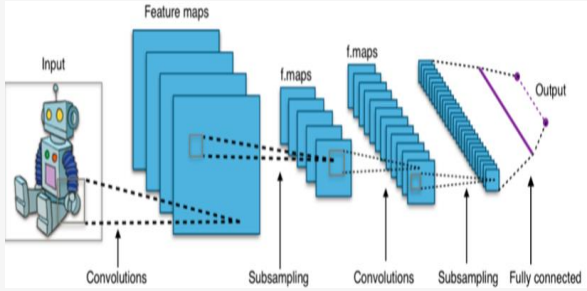
Only 1 hidden layer:

$$h = x \cdot w_{hidden} + b_{hidden}$$
$$y_{pred} = h \cdot w_{output}$$

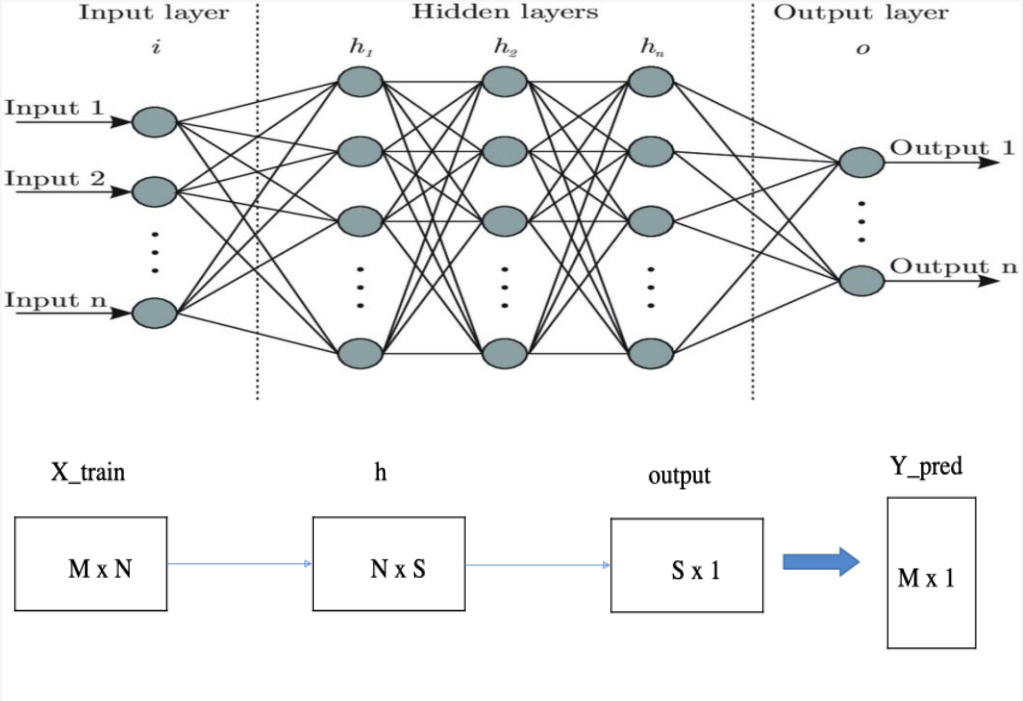
if more hidden layers:

$$h_1 = x \cdot w_{hidden1} + b_{hidden1}$$
$$h_2 = h_1 \cdot w_{hidden2} + b_{hidden2}$$
$$y_{pred} = h_2 \cdot w_{output}$$





CNN vs Fully Connected NN



Network Infrastructure

Neural Network for Prediction of Actual Width

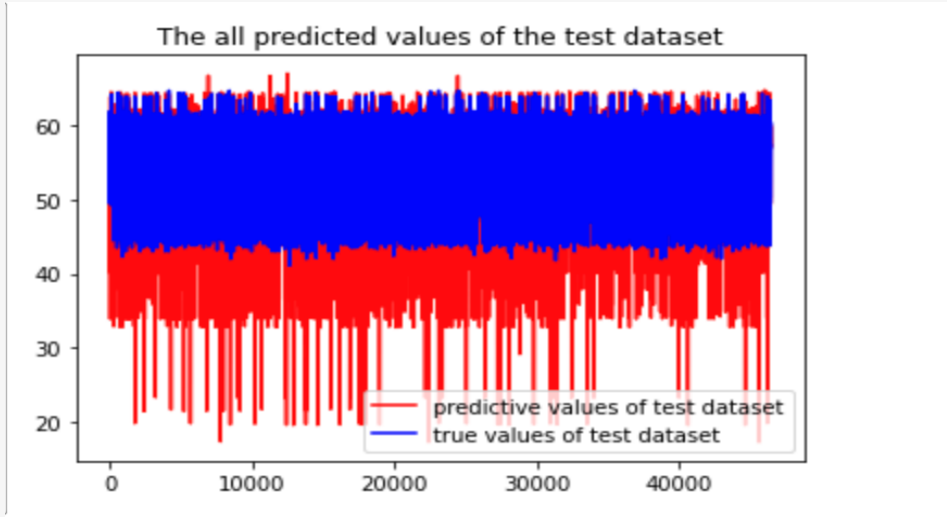
Hyperparameter Tuning

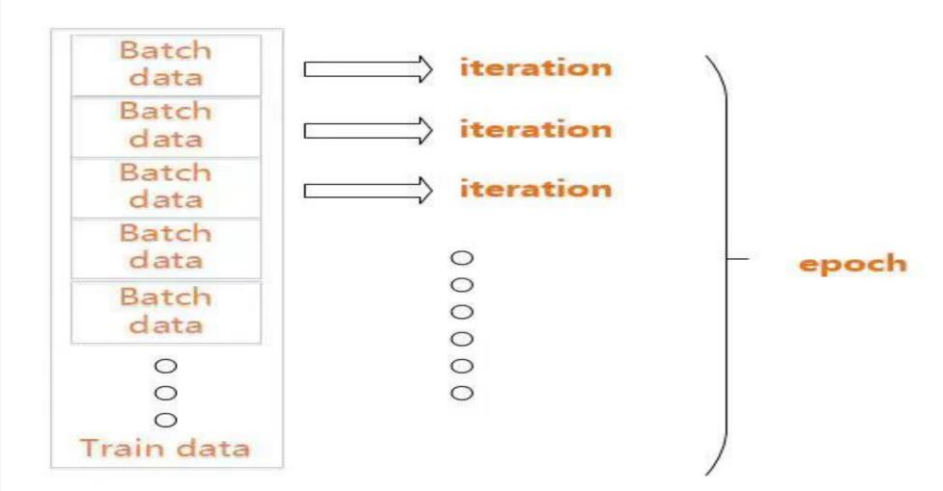
- **Model with 1 layer is not satisfactory**

- **Tuning:**
  - Learning Rate: , [0.002, 0.003, 0.004]
  - Width: no of neurons, [5,10,20], [50,200,2000]
  - Depth: no of layers, range(5,9)
  - Batch_size = 500
  - Steps=100000
  - 1 epoch=308249/500 =616, 162 epochs in total

- **Training vs Test Set: 8:2**

```
Number of row and columns in X train set: (246599, 639)
Number of row and columns in Y train set: (246599,)
Number of row and columns in X test set: (61650, 639)
Number of row and columns in Y test set: (61650,)
```

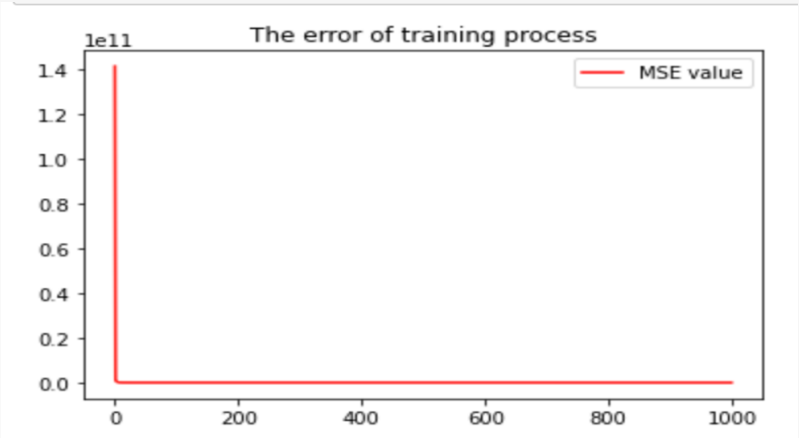


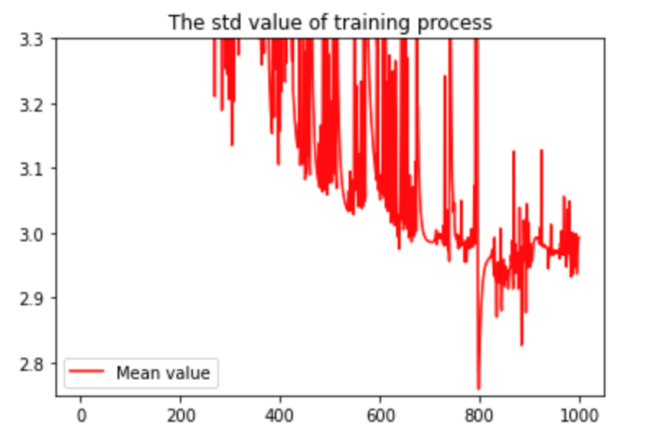


Neural Network for Prediction of Actual Width

Result Overview

- MSE, mean and standard deviation converges to a certain value
- Number of steps is enough and our training is complete

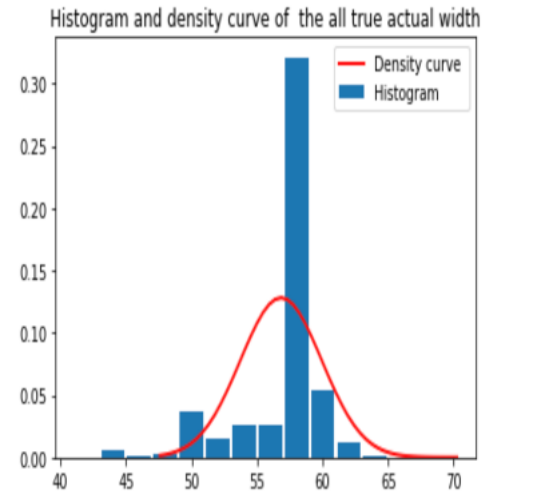


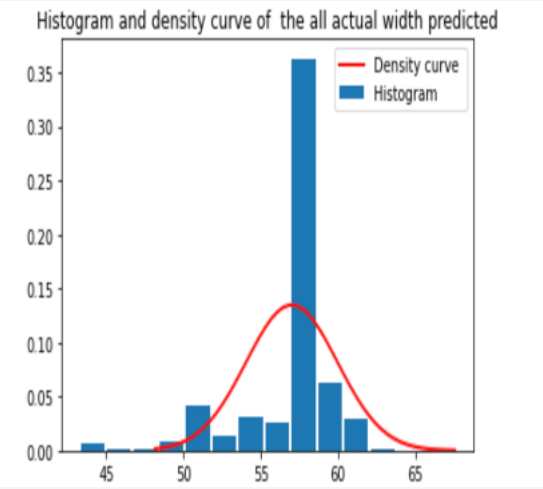


Neural Network for Prediction of Actual Width

Result Overview

- Histogram and Density Curve follows the same pattern

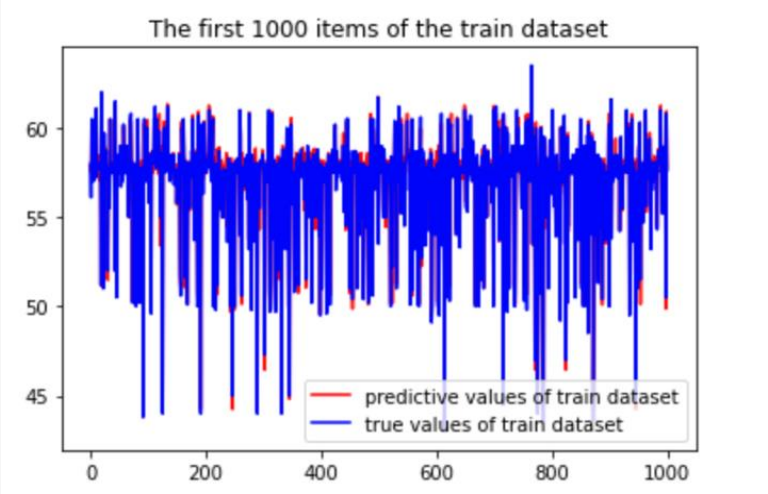


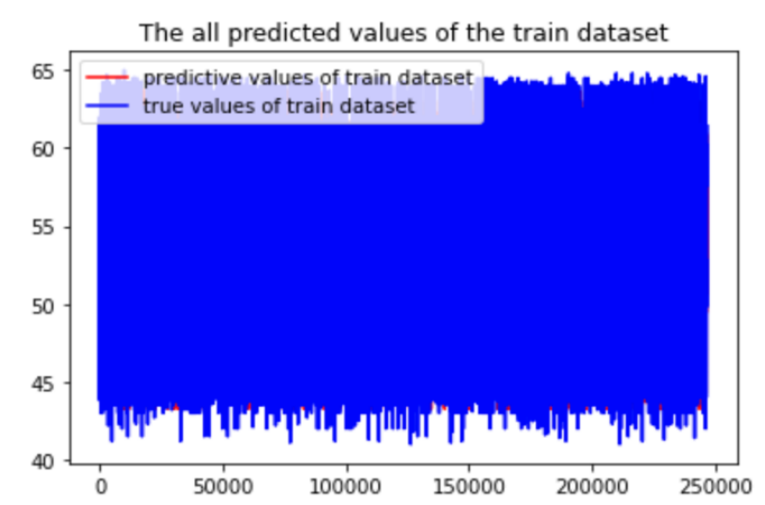


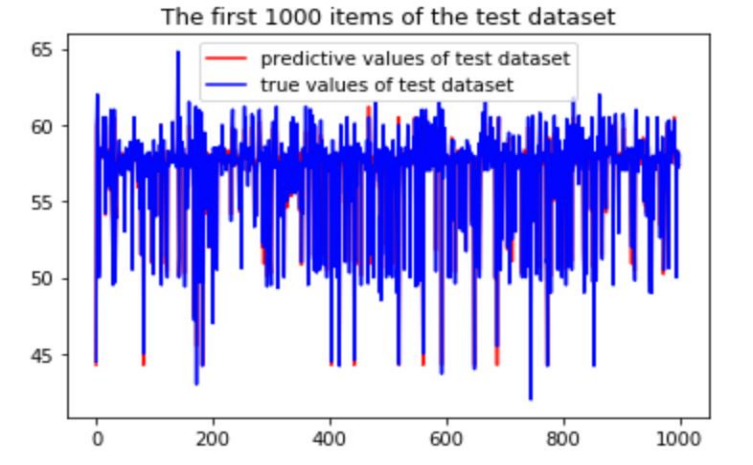
	Mean	Predicted
Ground truth	56.82	3.10
Predicted	57.00	2.95

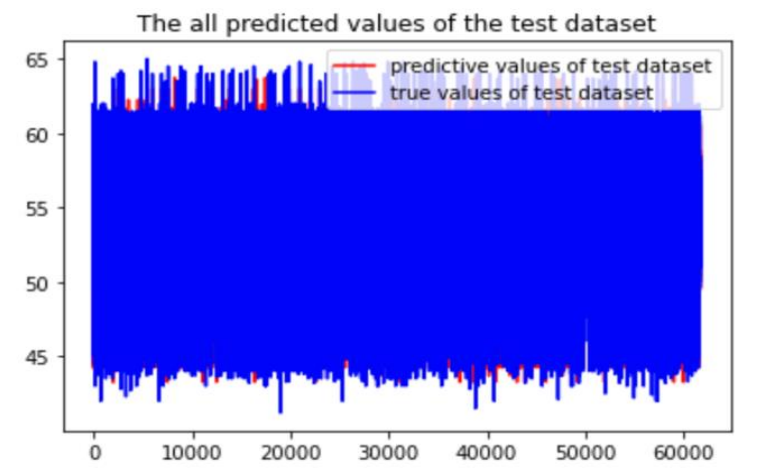
Neural Network for Prediction of Actual Width

Result Overview





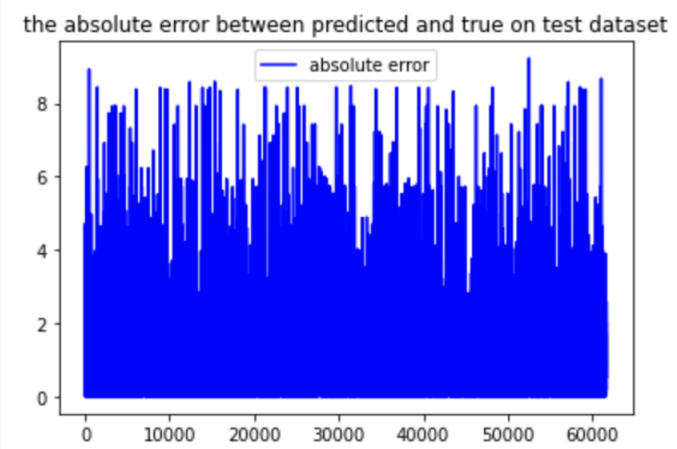


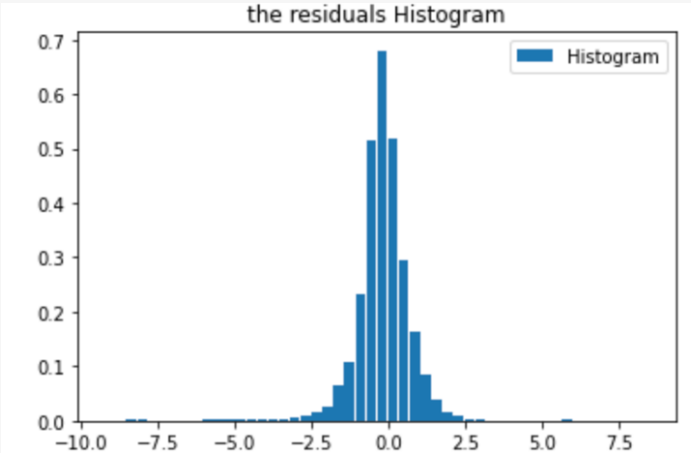


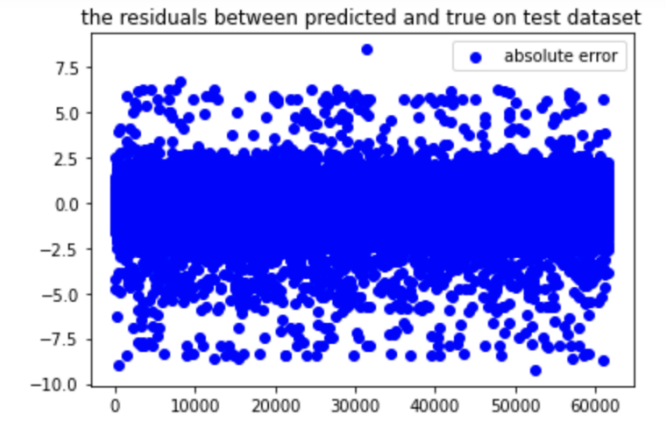
Neural Network for Prediction of Actual Width

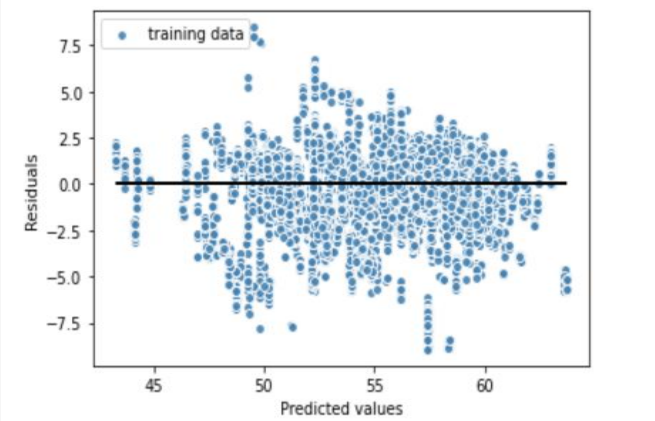
Result Overview - Residuals

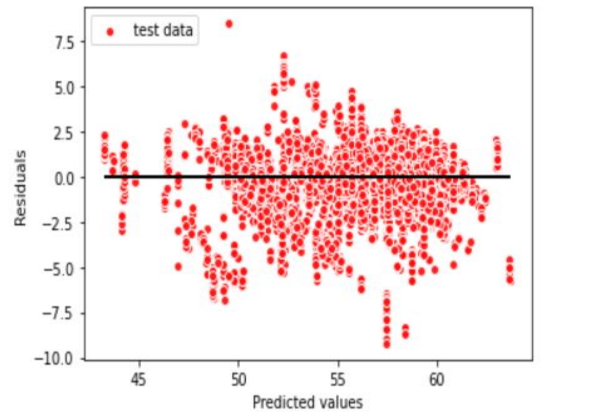
- Residuals fall between [-2.5, 2.5], and follow normal distribution in the histogram











Random Forest and Neural Network

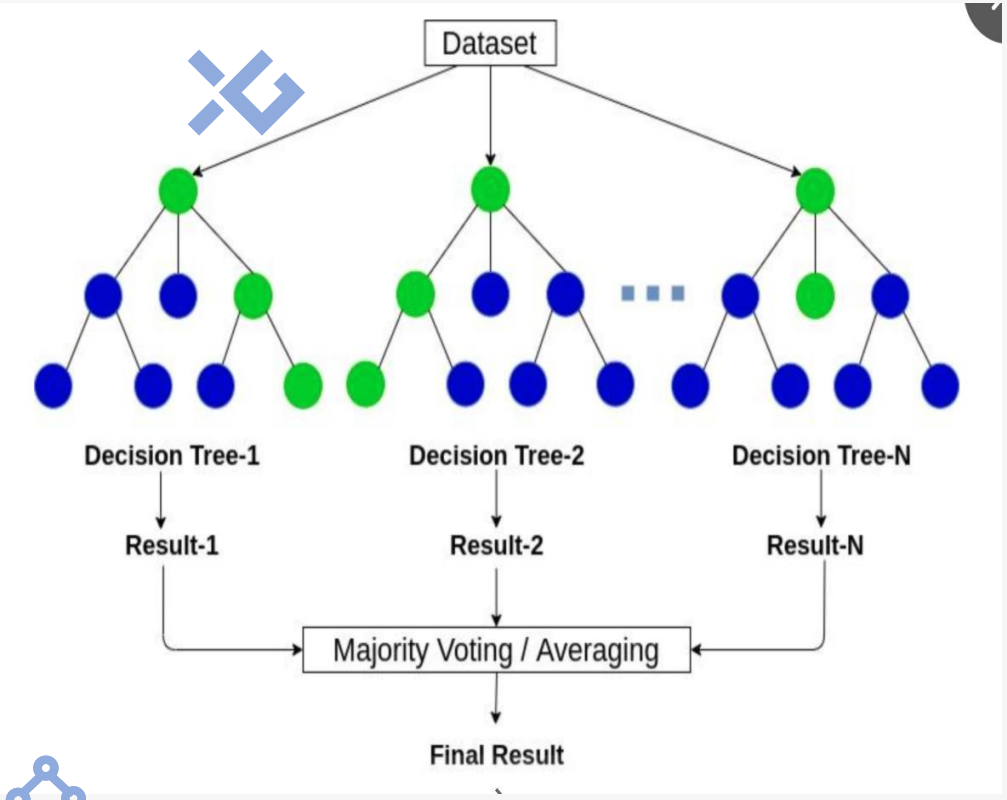
Parameter Tuning

Criterion: MSE, MAE

Parameters:

n_estimators: linspace(start = 200, stop = 2000, num = 10)
max_features: ['auto', 'sqrt']
max_depth: linspace(10, 100, num = 10)
min_samples_split: [2, 5, 10]
min_samples_leaf: [1, 2, 4]
Bootstrap: [True, False]

Final parameter: bootstrap=False,max_features='sqrt', max_depth=20,min_samples_split=10, n_estimators=1200,min_samples_leaf=2

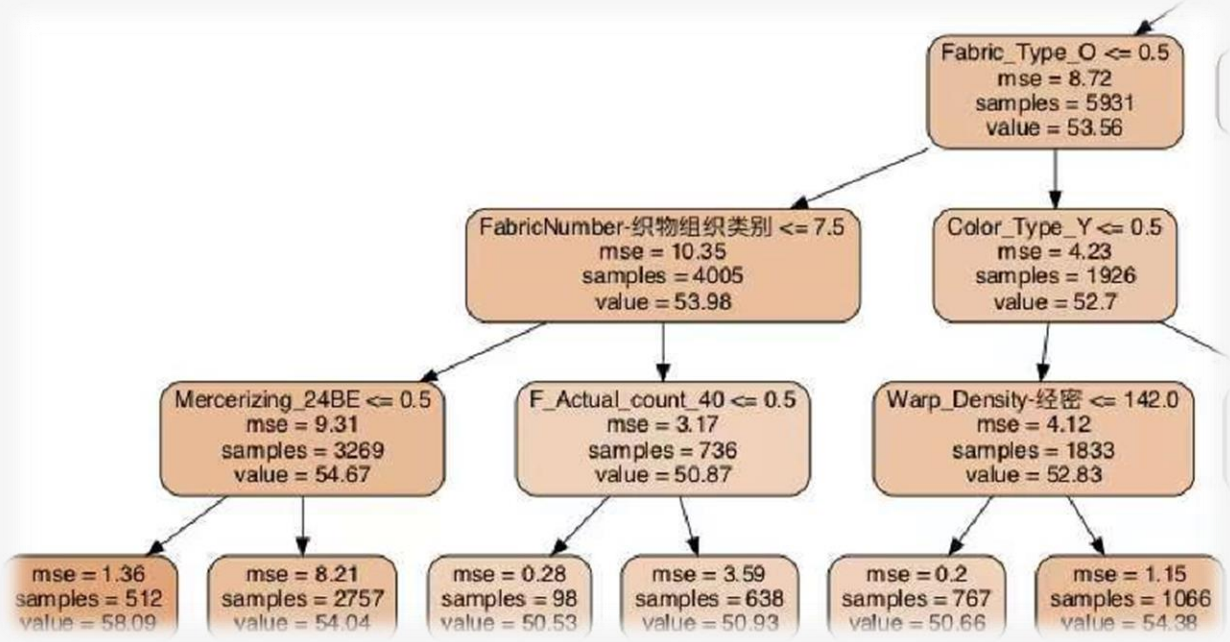


* See the example next slide

Random Forest and Neural Network

Full picture of the Forest

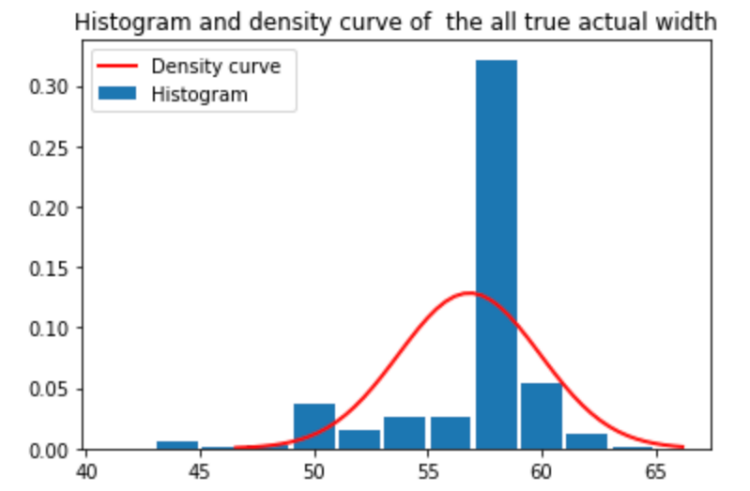
Zoom-in the Forest

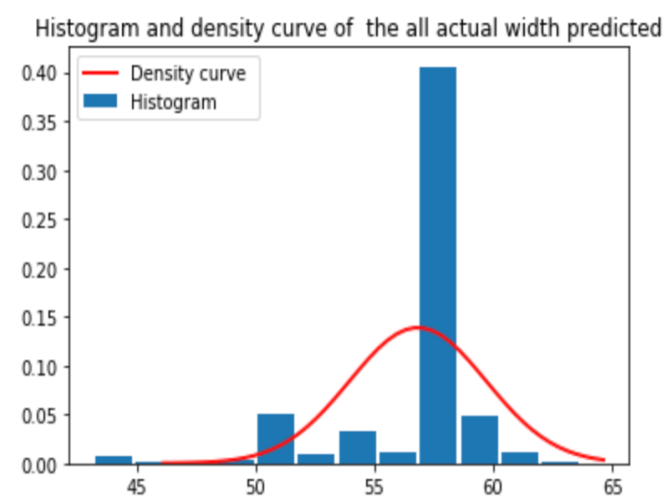


Random Forest for Prediction Actual Width

Result Overview

- Histogram and Density Curve follows the same pattern

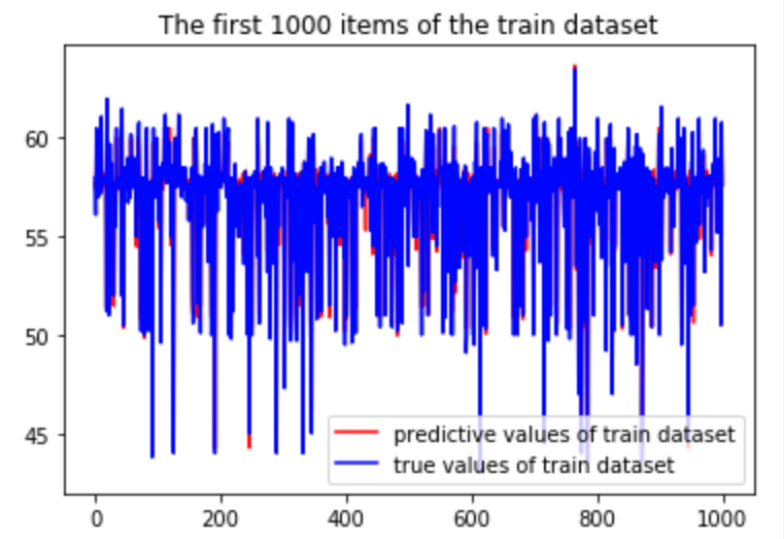


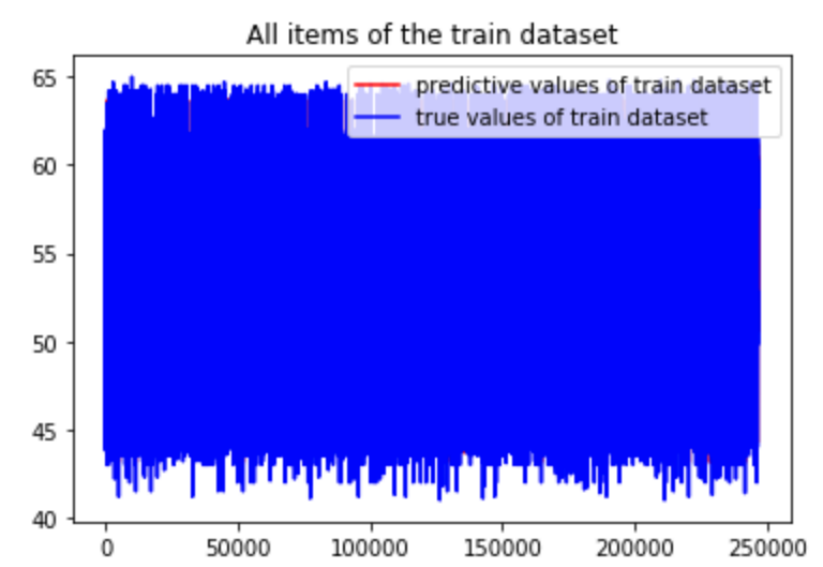


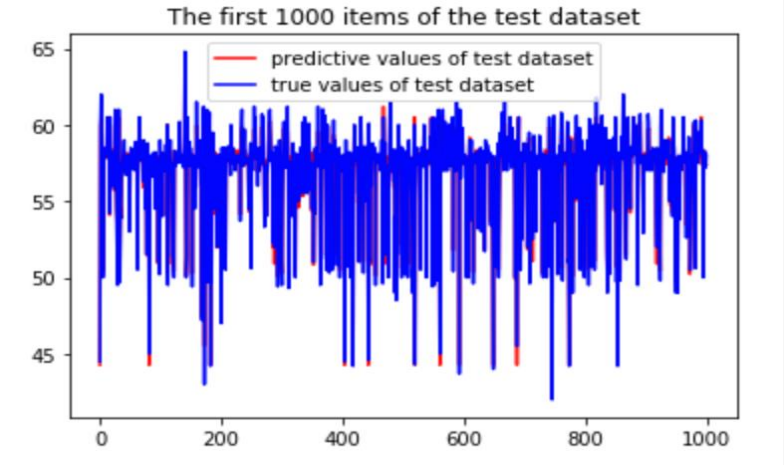
	Mean	Predicted
Ground truth	56.82	3.10
Predicted	56.82	2.87

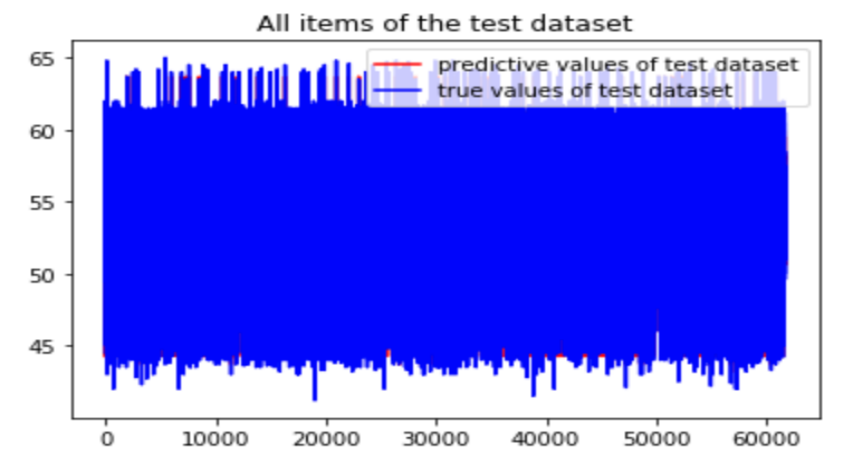
Random Forest for Prediction Actual Width

Result Overview





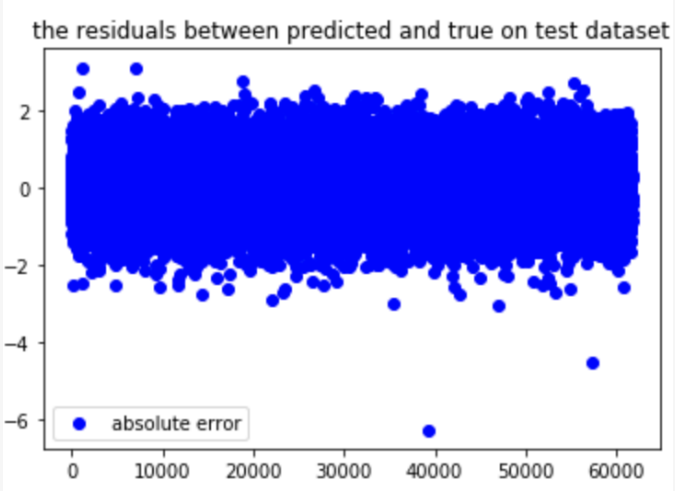


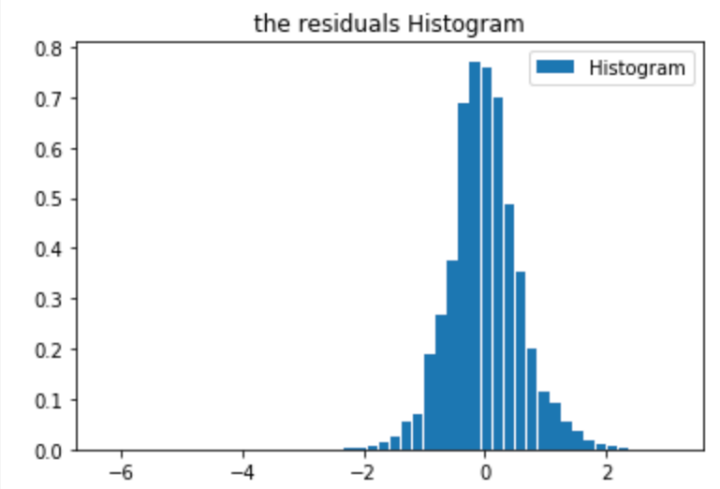


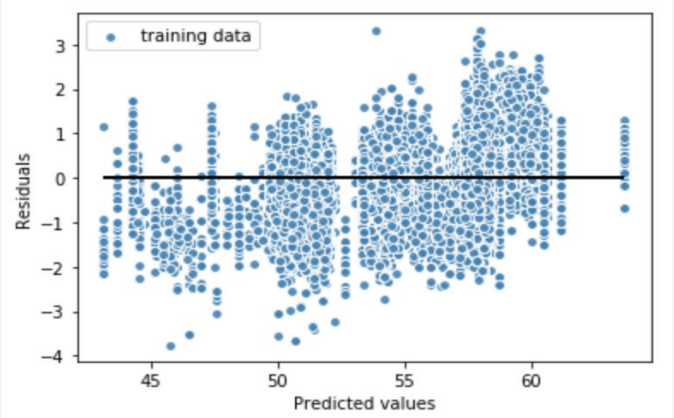
Random Forest for Prediction Actual Width

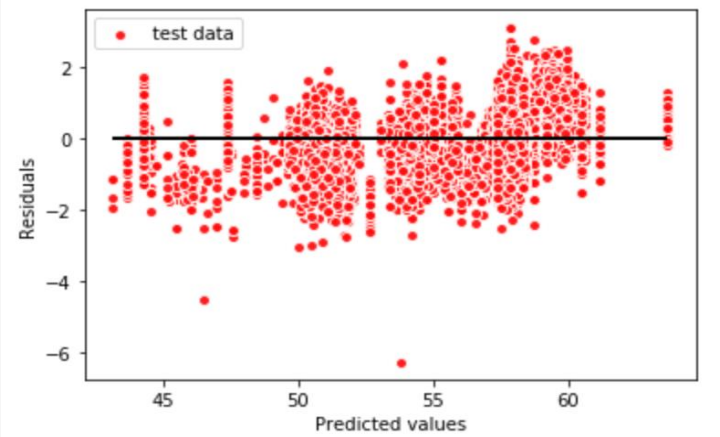
Result Overview - Residuals

- Residuals fall between [-2, 2], and follow normal distribution in the histogram



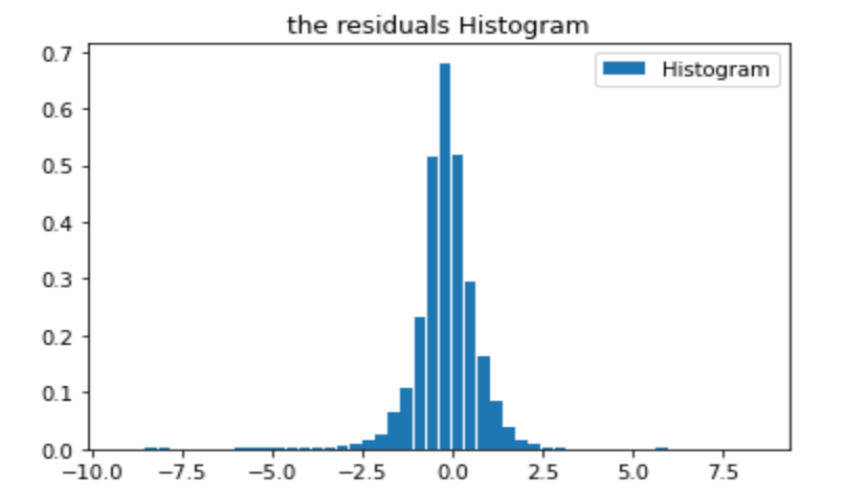


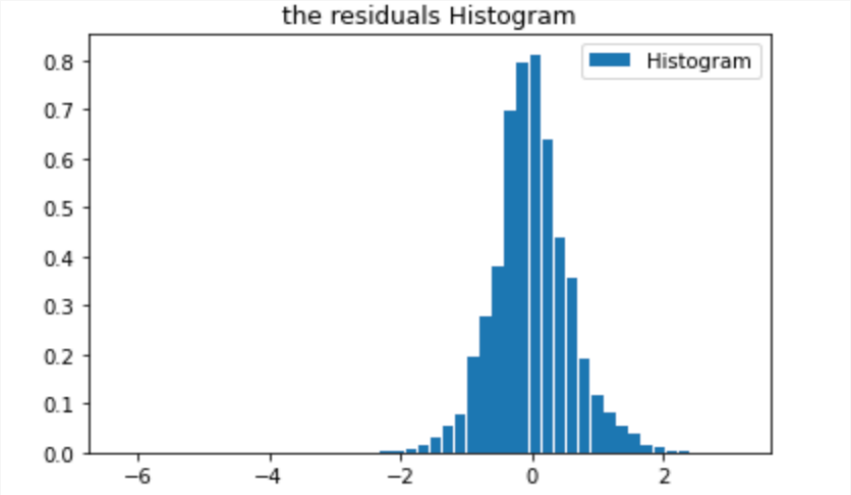




Comparison of Models for Prediction Actual Width

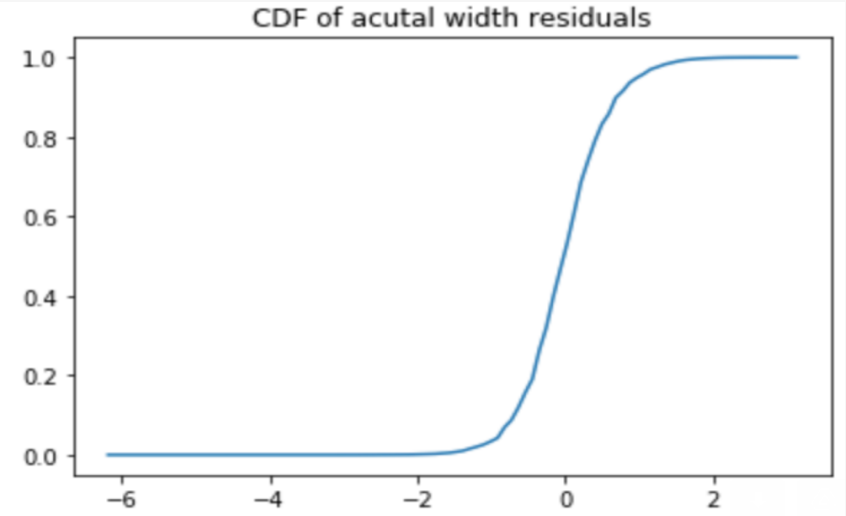
1. Use of the same random seed, so we compare the MSE of the test set, it is 0.42 for NN and 0.33 for RF
2. We compared Residual Plot



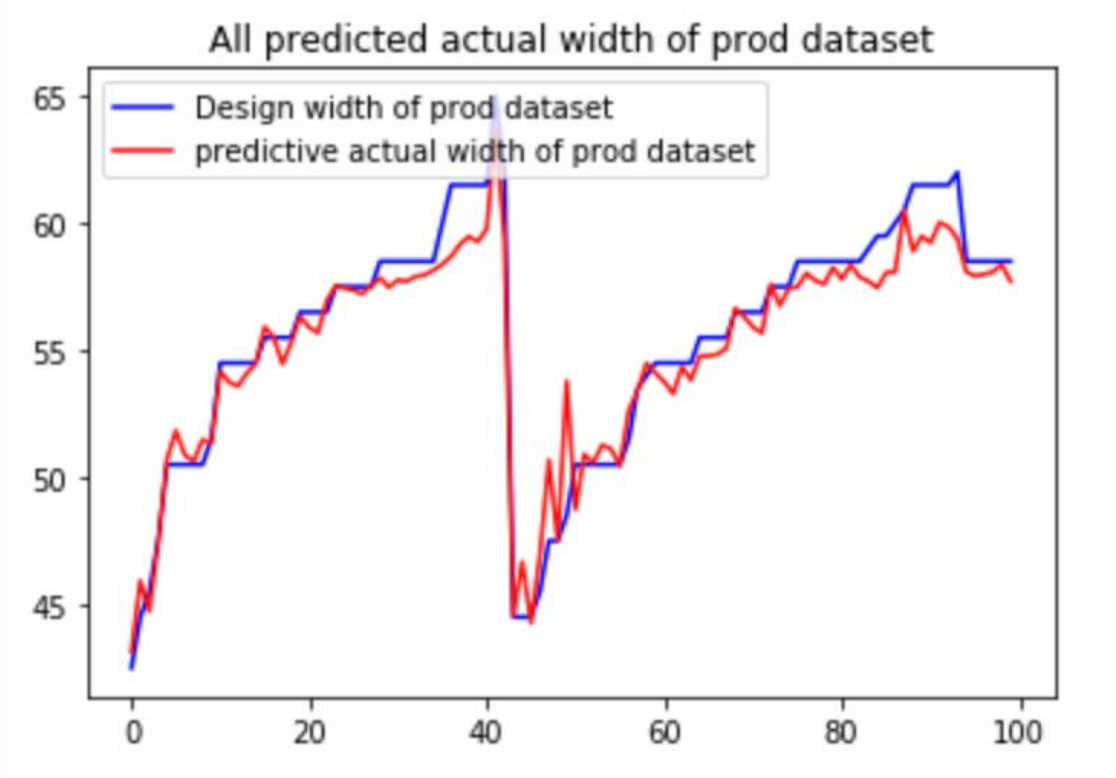


3. Performance of Random Forest

Interval	Probability
[-0.25,0.25]	41%
[0.5,0.5]	69%
[0.75,0.75]	83%
[-1,1]	92.5%
[-1.25,1.25]	97%
[1.5,1.5]	98%



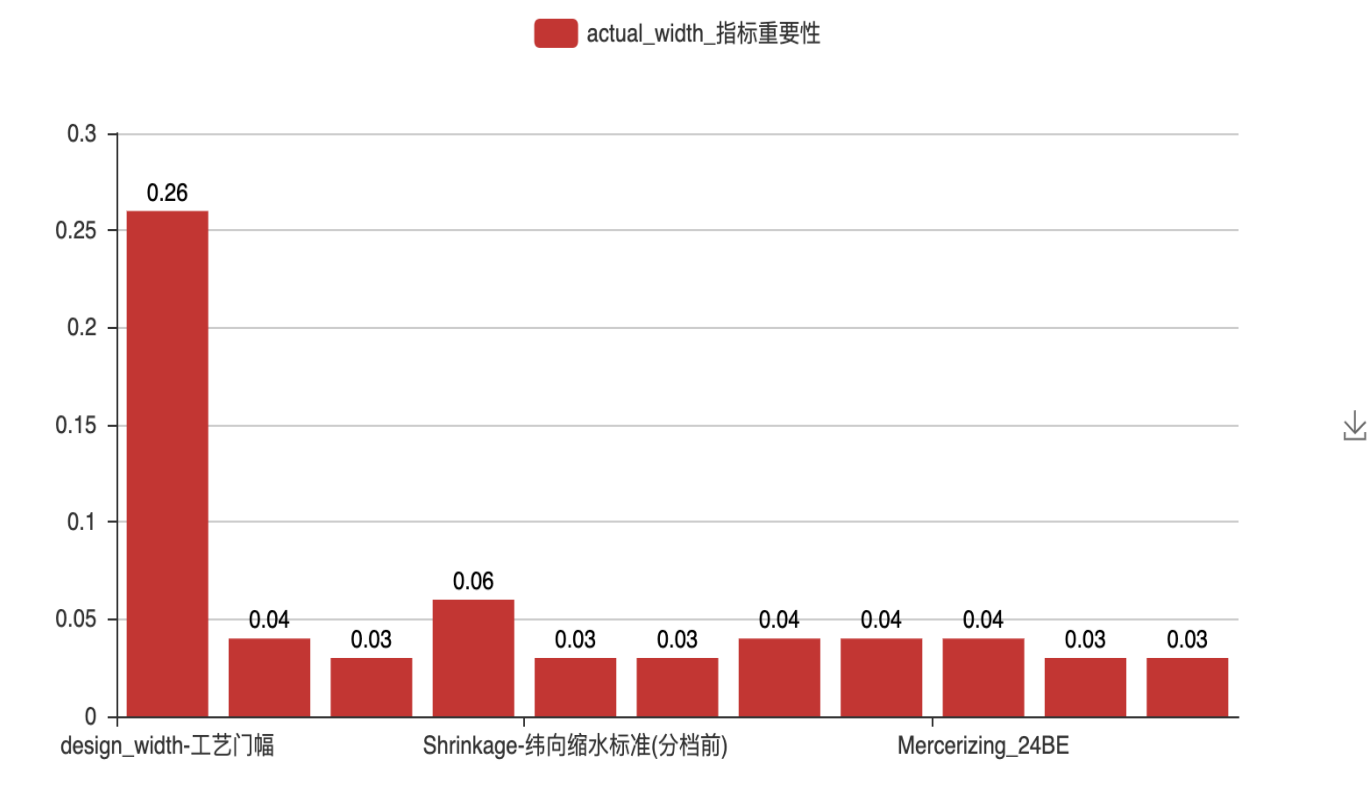
Prediction of Production Data for Prediction Actual Width



Determinants for Prediction Actual Width

HKU BUSINESS SCHOOL 港大經管學院 | Master of Science in Business Analytics 理科碩士(商業分析)

- Design Width 0.26
- Reed Width 0.06
- Component 0.04
- ReedNo_Type 0.04
- Mercerizing: 0.04
- Warp_density 0.04



ESQUEL GROUP

Commercial Implications

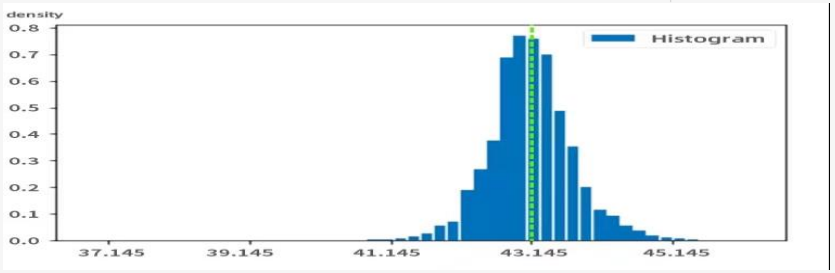
How to apply ?

For Reuse

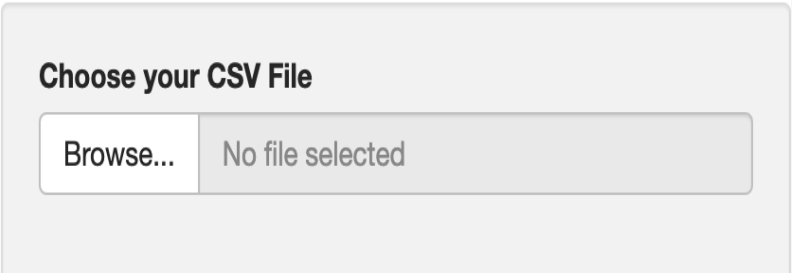
- SAAS
- Application
- Database
- Server

actual_width_random_forest.joblib

For Singular Output



For Multiple Outputs



Actual_width-实际门幅
43.14515991
45.9232869
44.73385269
47.39191595
50.7139825
51.83112461
50.87931597
50.62951708
51.46518341
51.35948776
54.16917493
53.73861251
53.58659824
54.09409496
54.4864018
55.91363805
55.52921351
54.46031665
55.31662213
56.32276447
55.90250818
55.67645542
56.95461562
57.53444371
57.43730963
57.35541664
57.217976
57.53708932
57.84290462
57.47659187
57.76596475
57.70944648
57.8986802

Part 2

Prediction of shrinkage

Build a model to accurately predict the fabric shrinkage rates

Table of Contents

1. Objectives and Work Done
2. Exploratory Data Analysis
3. Methodologies
4. Neural Network
5. Random Forest
6. Model Evaluation
7. Determinants
8. Production Test Result
9. Commercial Implications

Objectives and Work Done

Project objective:

Try to predict the fabric shrinkage rates accurately based on historical data from the past Shrinkage data to facilitate decision-making process and enhance production efficiency.

Raw data:

113481rows, 20 columns

Software employed:

Excel and Python

Machine learning algorithms:

Regression, Random Forest, Neural Network

Raw Data Provided By Client

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 113481 entries, 0 to 113480
Data columns (total 20 columns):
gf_id
品种ID                    113481 non-null int64
gf_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
GEW STD warp
GEW经向缩率审单标准         113481 non-null float64
GEW STD weft
GEW纬向缩率审单标准         113481 non-null float64
Mercerizing
丝光碱浓                    113481 non-null object
```

Exploratory Data Analysis

Feature Subset Selection:

Unlike the actual width part, the input features of Shrinkage all seem to be important. In this case, there is no need to screen out input features.

Data Cleaning:

- Typo in Weft-Shrinkage data
- Correcting Typos: -0.8. to -0.8

Data after Cleaning:

- 113481 rows, 20 columns

Variable Type Determination:

- Numeric: Joint Distribution
- Categorical: Box Plot

Data Typo

```
for a in table_2['weft_shrinkage\n纬向缩率']:
    if not isinstance(a,float):
        if not isinstance(a,int):
            print(a)
    else:
        ;
```

```
-0.8.
```

Data after Cleaning

```
table_3=pd.read_excel("Shrinkage Data for HKU updated 5_4.xlsx")
table_3.info()
#FG1909250049处的纬向缩率数据被污染(-0.8->-0.8.), 此处已手动修改
#触发点是纬向缩率
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 113481 entries, 0 to 113480
Data columns (total 20 columns):
gf_id
品种ID                    113481 non-null int64
gf_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 float64
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
GEW STD warp
GEW经向缩率审单标准           113481 non-null float64
GEW STD weft
GEW纬向缩率审单标准           113481 non-null float64
Mercerizing
丝光碱浓                    113481 non-null object
```

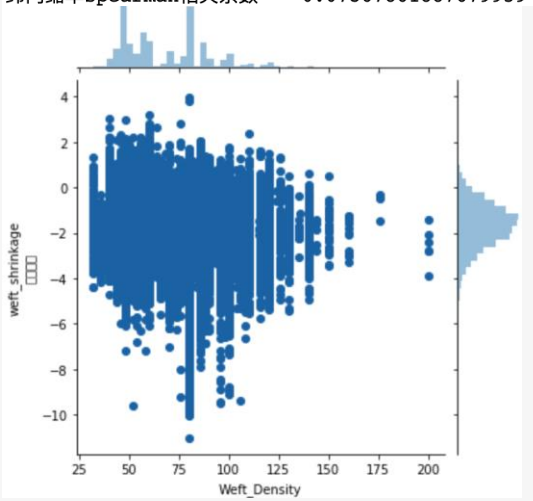
Exploratory Data Analysis

Numerical:

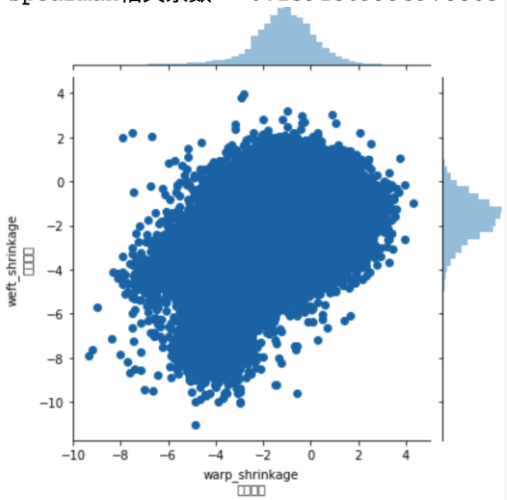
- No clear correlation (or weak correlation) between the response variables (Weft Shrinkage & Warp Shrinkage) and the rest of the independent variables.

- No pattern was noted for other independent variables.

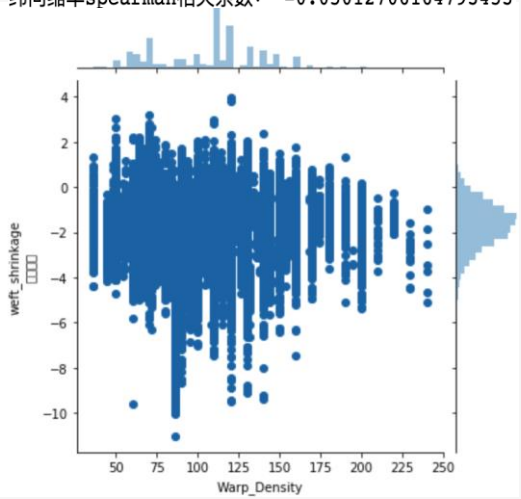
纬向缩率pearson相关系数: -0.11985893614296182
纬向缩率kendall相关系数: -0.04964466789341022
纬向缩率spearman相关系数: -0.07367861887679959



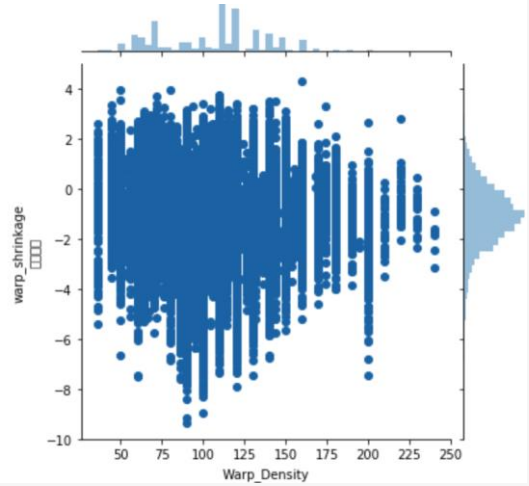
pearson相关系数: 0.3616227764001237
kendall相关系数: 0.17759696883304182
spearman相关系数: 0.2591309993978863



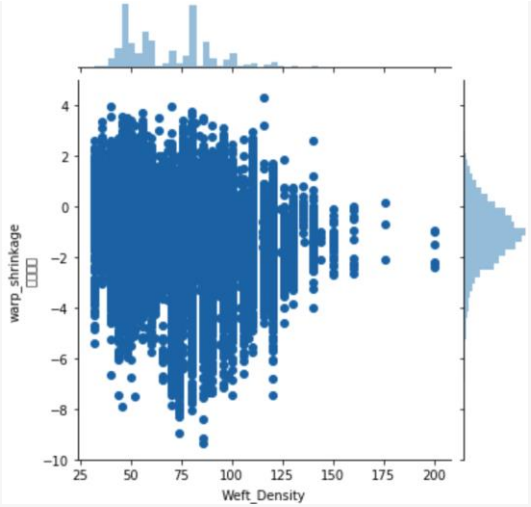
纬向缩率pearson相关系数: -0.030709888208449145
纬向缩率kendall相关系数: -0.03310514188988799
纬向缩率spearman相关系数: -0.05012706164793453



经向缩率pearson相关系数: -0.09594290730462028
经向缩率kendall相关系数: -0.08810780911123295
经向缩率spearman相关系数: -0.12697600942146395



经向缩率pearson相关系数: -0.21541698463860623
经向缩率kendall相关系数: -0.13813553871620235
经向缩率spearman相关系数: -0.19890635527222683

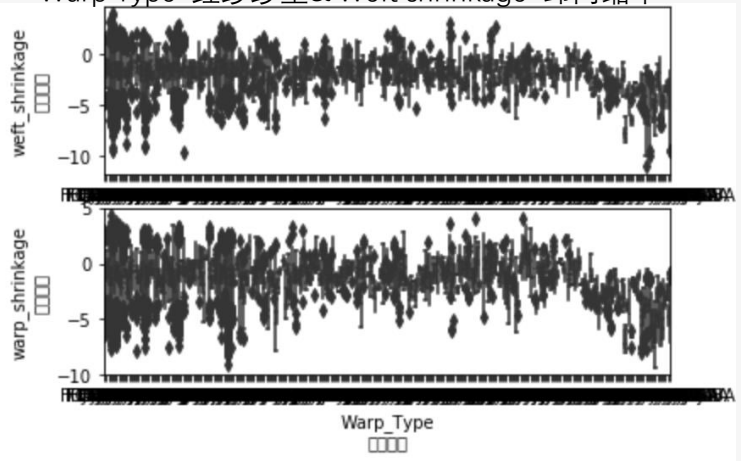


Exploratory Data Analysis

Categorical:

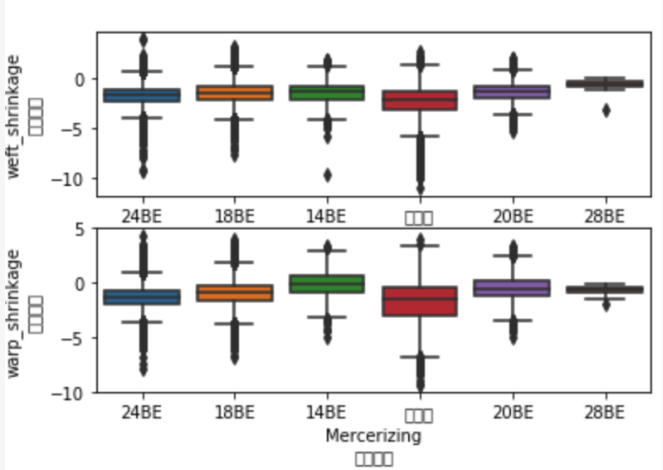
- Some examples of plot that show there is no pattern was noted for other independent variables

Warp Type-经纱纱型& Weft shrinkage-纬向缩率



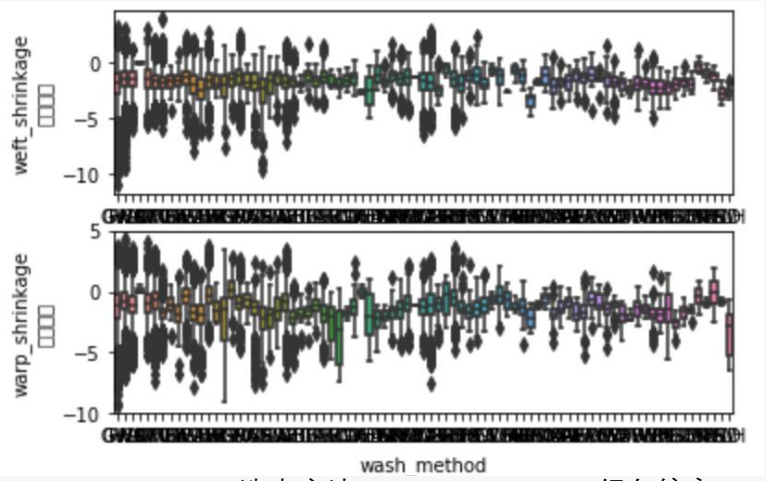
Warp Type-经纱纱型& Warp shrinkage-经向缩率

Mercerizing-丝光碱浓& Weft shrinkage-纬向缩率



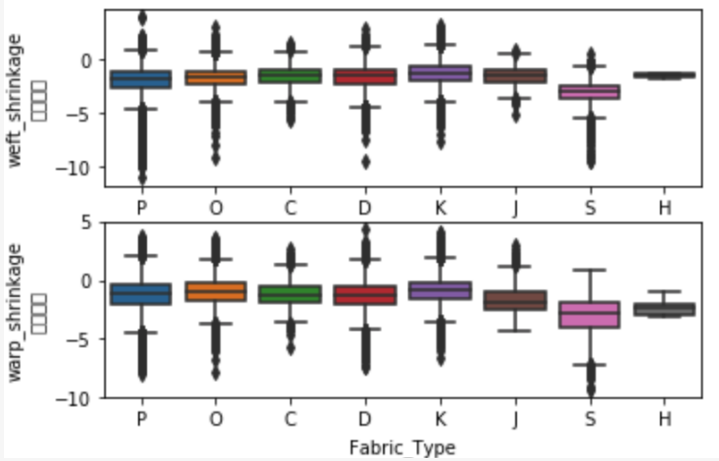
Mercerizing-丝光碱浓& Warp shrinkage-经向缩率

Wash method-洗水方法& Weft shrinkage-纬向缩率



Wash method-洗水方法& Warp shrinkage-经向缩率

Fabric Type-织法& Weft shrinkage-纬向缩率



Fabric Type-织法& Warp shrinkage-经向缩率

Methodologies

Go deeper into categorical variable: one hot coding with pandas.get_dummies

- Before implementing the function: (113481,12) with 4 numeric and 8 categorical variables

	wash_method\n 洗水方法	Fabric_Type\n 织法	Warp_Type\n 经纱纱型	Weft_Type\n 纬纱纱型	Warp_Count\n 经纱支	Weft_Count\n 纬纱支	Warp_Density\n 经纱密度	Weft_Density\n 纬纱密度	Finishing_Method\n 整理方法	warp\nGEW 经向缩率审单标准
0	GW	P	CF+JC	CF+JC	50+100/2	50	120	100	1+	2.0
1	GW	P	CF+JC	CF+JC	50+100/2	50	120	100	1+	2.0
2	GW	O	JC+CF	JC	40+80/2	20	110	54	1+	2.0
3	GW	C	JC+CF	JC+CF	50+100/2	50	150	126	1+	2.0
4	GW	C	JC	JC	50	50	150	126	1+	2.0

- After implementing the function: (113481, 1175) with numerical columns

	Warp_Density\n 经纱密度	Weft_Density\n 纬纱密度	GEW STD warp\nGEW 经向缩率审单标准	GEW STD weft\nGEW 纬向缩率审单标准	wash_method_ABGA	wash_method_AGWA	wash_method_BGW	wash_method_BIP
0	120	100	2.0	2.0	0	0	0	0
1	120	100	2.0	2.0	0	0	0	0
2	110	54	2.0	2.0	0	0	0	0
3	150	126	2.0	2.0	0	0	0	0
4	150	126	2.0	2.0	0	0	0	0

5 rows × 1175 columns

Neural Network for Prediction Shrinkage

Determining Network Architecture and Loss Function

- Same for the Shrinkage part: Fully connected NN is used over convolutional: black box (makes no assumptions about the features in the data; no such large volume of data)

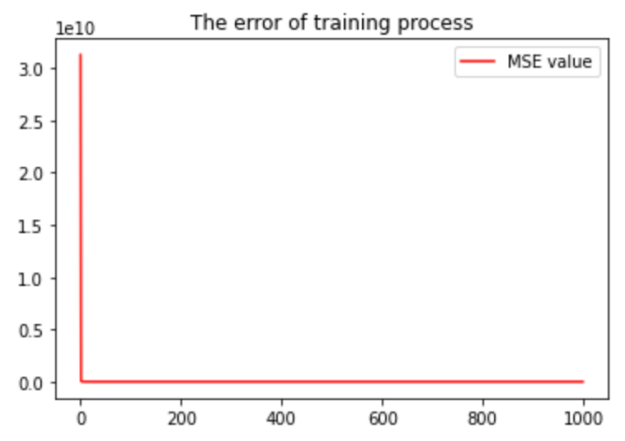
Hyperparameter Tuning

- Similar to the Actual width part but in this case, but need to do for two estimation: 1) Weft Shrinkage 2) Warp Shrinkage
- Same tuning structure for both Weft Shrinkage part and Warp Shrinkage part.
- Then, we train the model corresponding to the dataset.

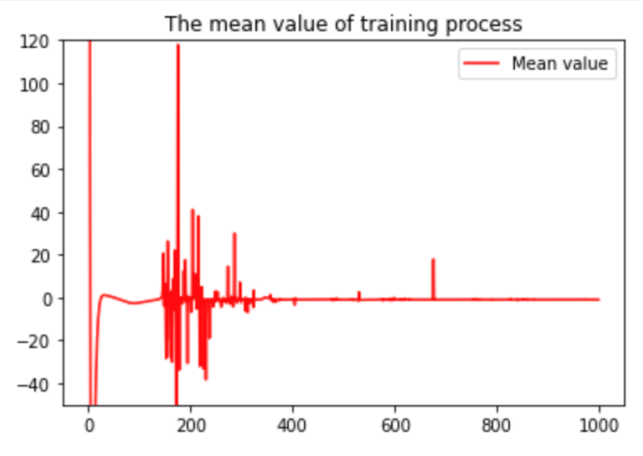
Neural Network for Prediction of Shrinkage

Result Overview- Training Part

✓ The Training Result for Warp Shrinkage

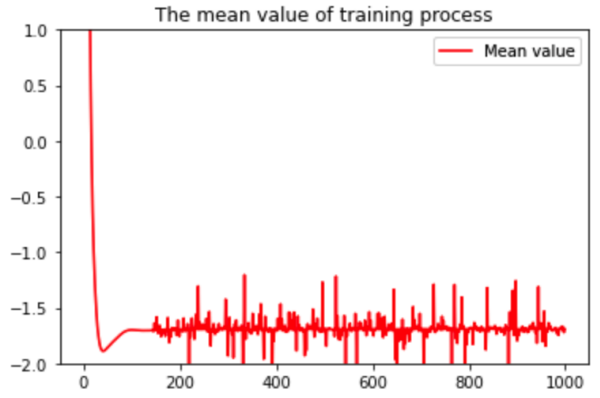


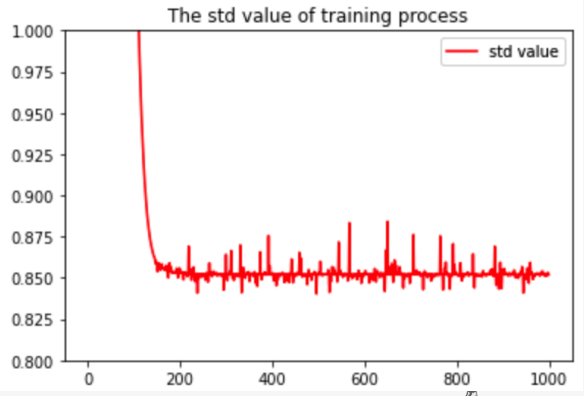

```
predicted mean value of train datasets is  -1.0634663
predicted std value of train datasets is  1.5679386
predicted mean value of test datasets is  -1.0620855
predicted std value of test datasets is  1.5781986
groundtruth mean value of train datasets is  -1.1219961515134809
groundtruth std value of train datasets is  1.2845598193671337
groundtruth mean value of test datasets is  -1.1386154728299223
groundtruth std value of test datasets is  1.2728678541335832
```



✓ The Training Result for Weft Shrinkage

```
predicted mean value of train datasets is  -1.3158735
predicted std value of train datasets is  1.5451525
predicted mean value of test datasets is  -1.3213041
predicted std value of test datasets is  1.5594366
groundtruth mean value of train datasets is  -1.7500117218944697
groundtruth std value of train datasets is  1.2019767175426859
groundtruth mean value of test datasets is  -1.7590099241354975
groundtruth std value of test datasets is  1.2066587686431909
```

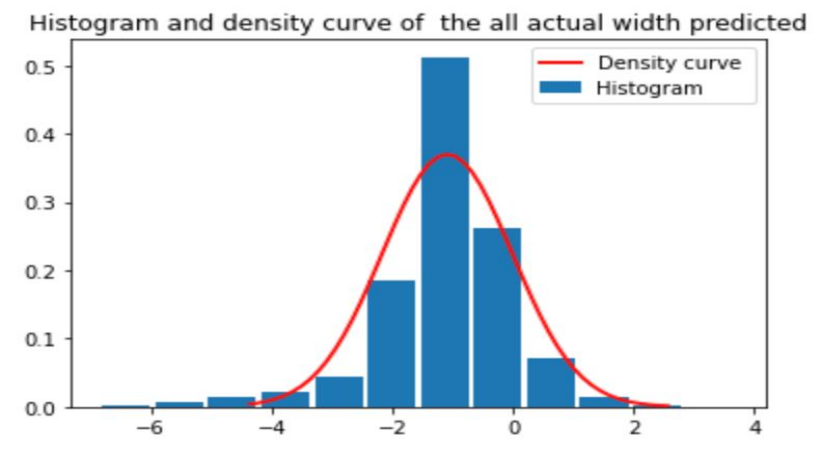


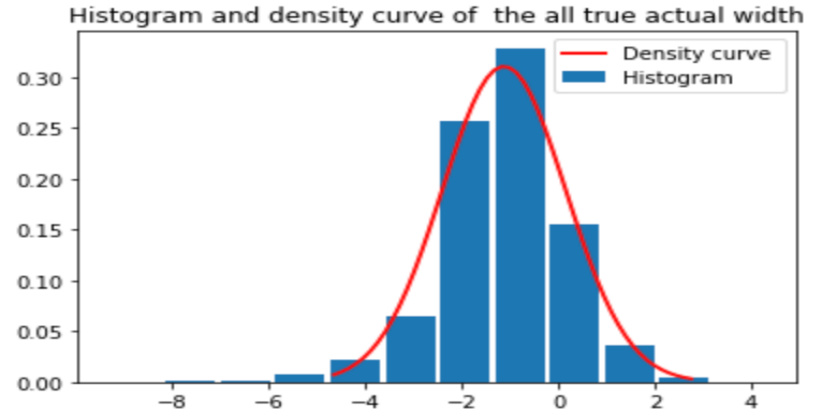


Neural Network for Prediction Shrinkage

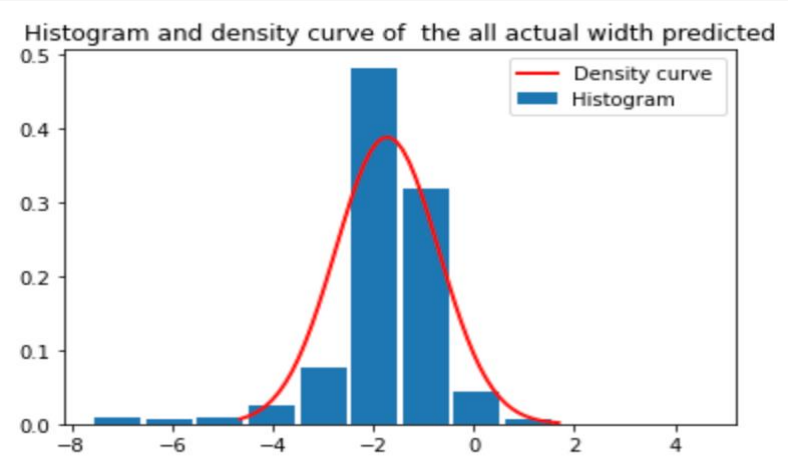
Result Overview- Histogram Plots

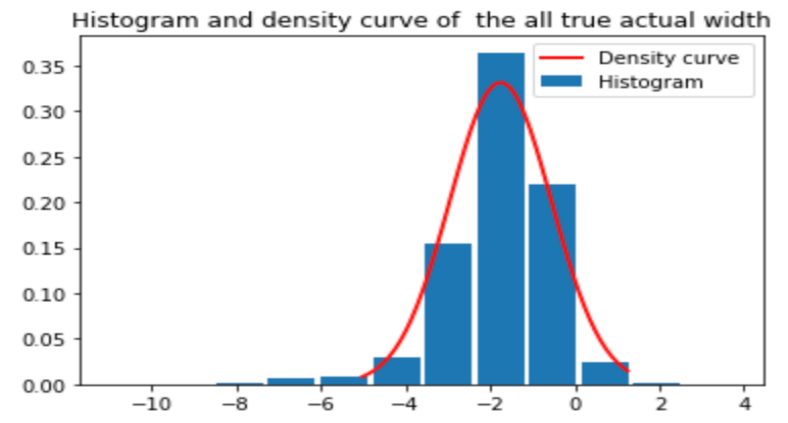
✓ **The Result for Warp Shrinkage**





✓ **The Result for Weft Shrinkage**



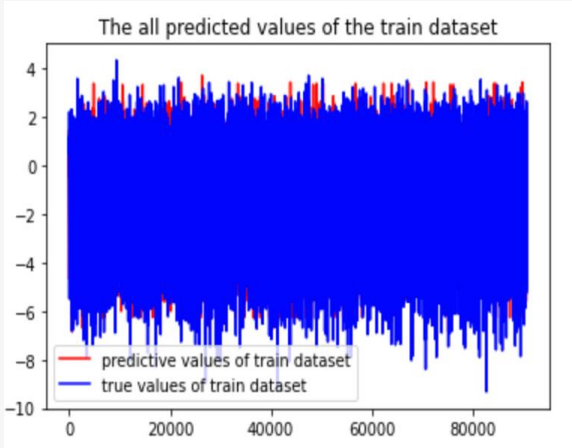


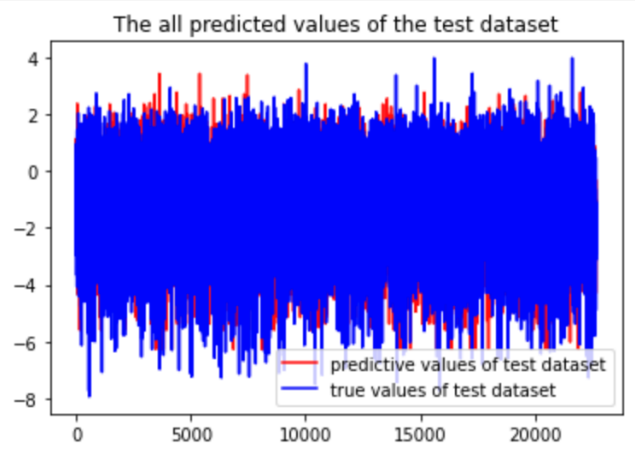
Neural Network for Prediction of Shrinkage

Result Overview- Prediction Part

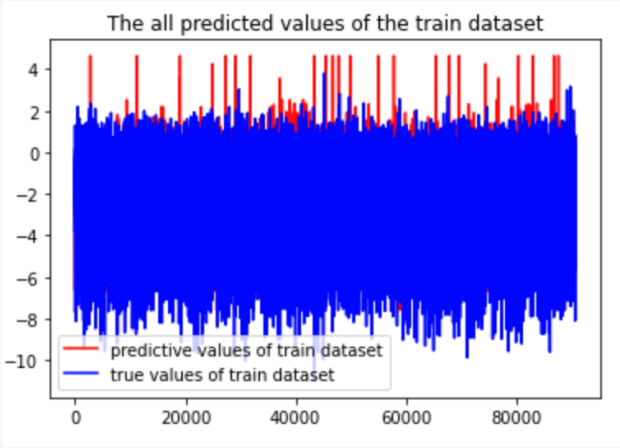
- Although we have the best tuning parameter, the plots both from the Warp & Weft Shrinkage show that the prediction is accurate but not as good as the result from actual width part
- Here are the result plots from these two parts:

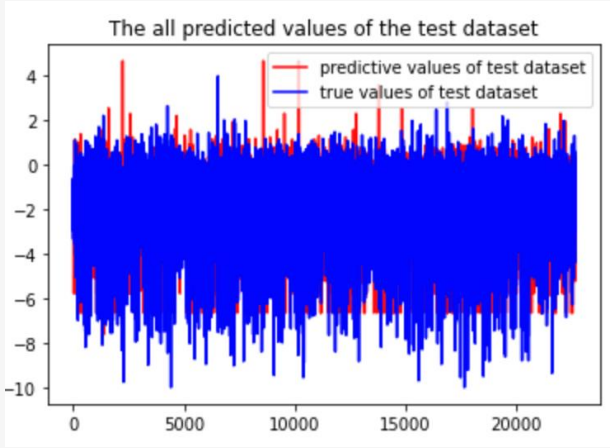
✓ **The Prediction Result for Warp Shrinkage**





✓ **The Prediction Result for Weft Shrinkage**



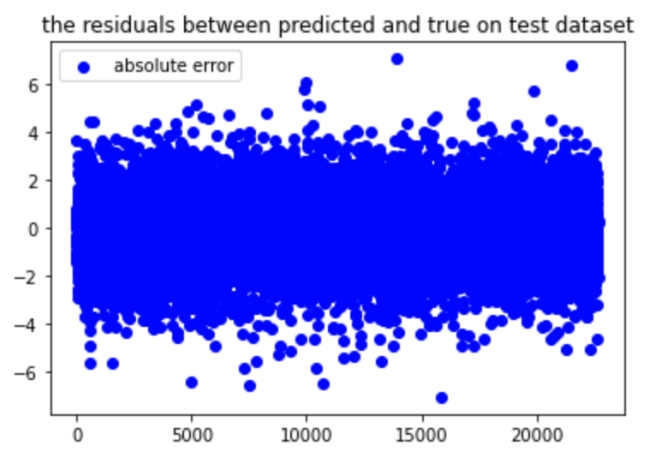


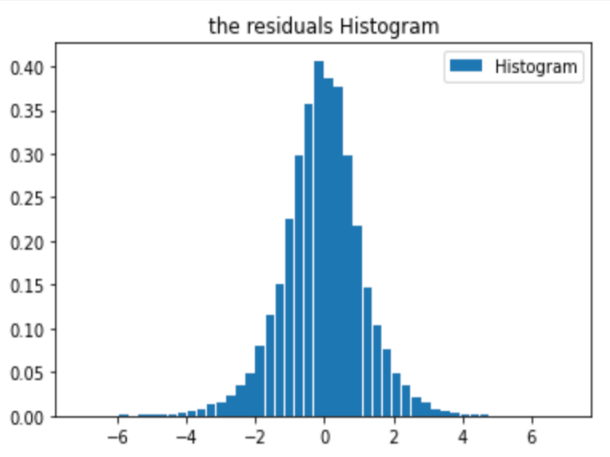
Neural Network for Prediction of Shrinkage

Result Overview- Residual Plot

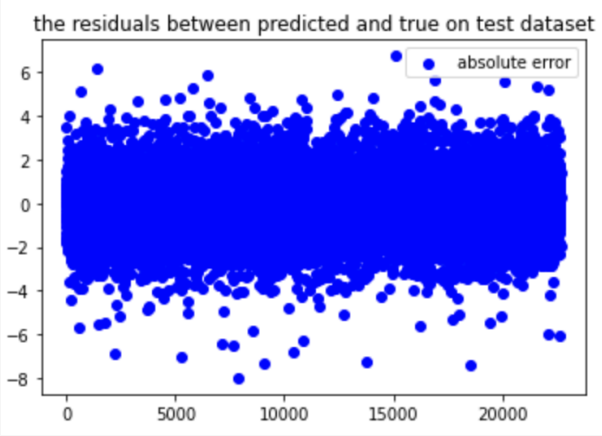
- Although we have the best tuning parameter, the plots both from the Warp & Weft Shrinkage show that the prediction is accurate but not as good as the result from actual width part
- Here are the result plots from these two parts:

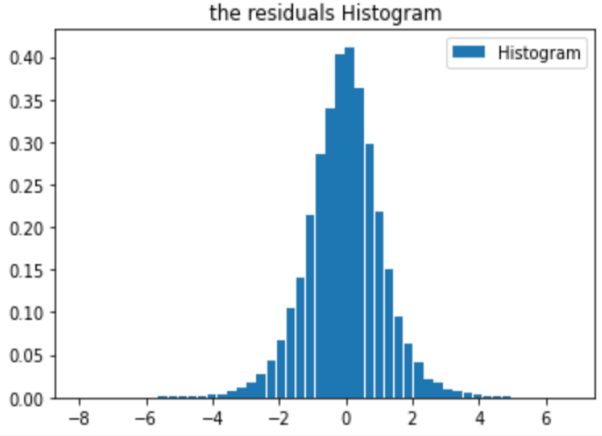
✓ **The Prediction Result for Warp Shrinkage**





✓ **The Prediction Result for Weft Shrinkage**

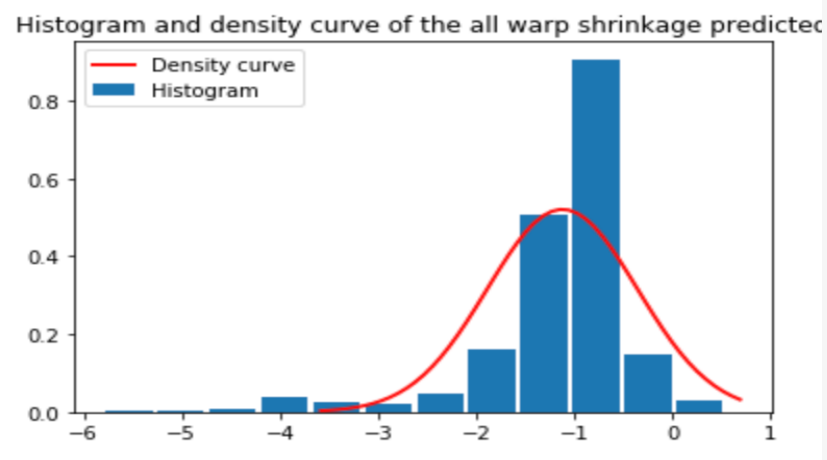


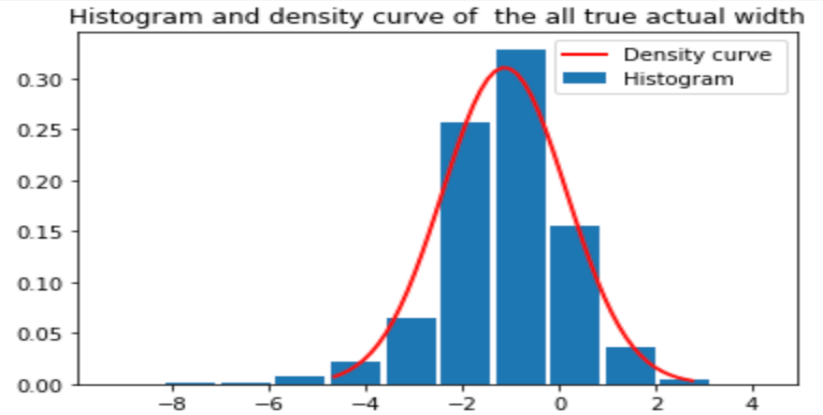


Random Forest for Prediction Shrinkage

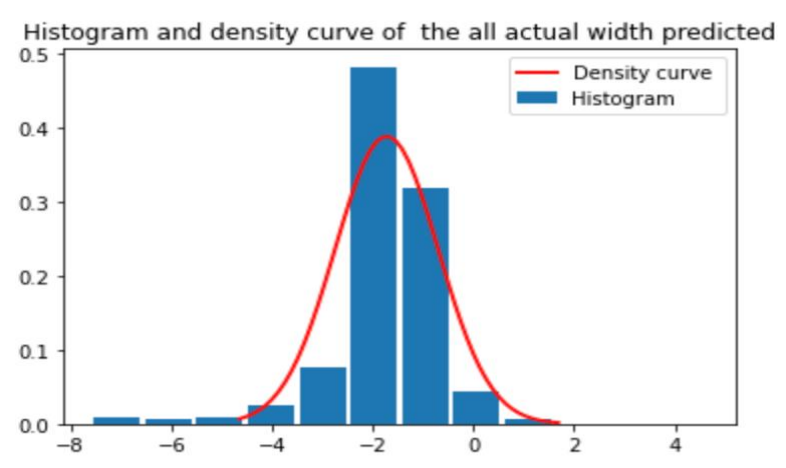
Result Overview- Histogram Plots

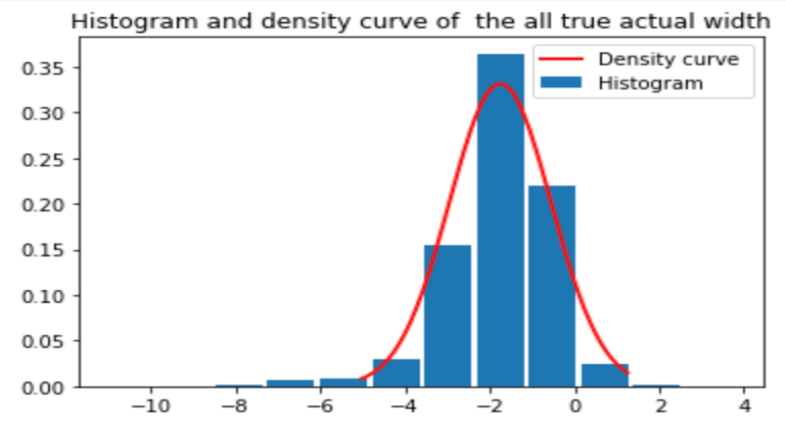
✓ **The Result for Warp Shrinkage**





✓ **The Result for Weft Shrinkage**



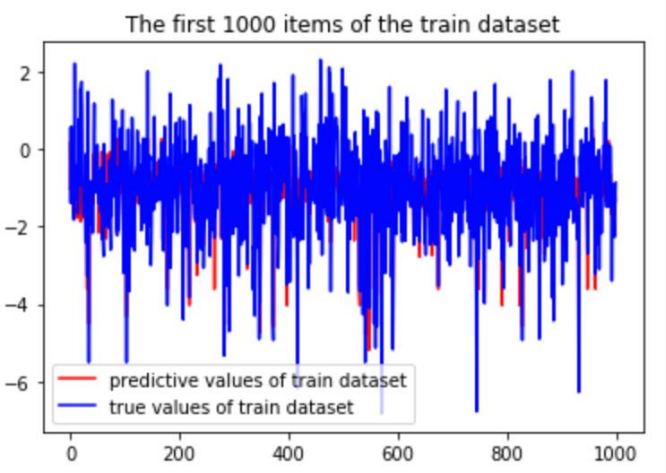


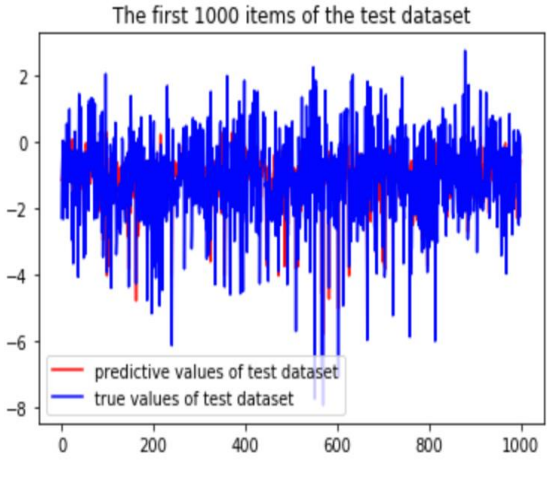
Random Forest for Prediction of Shrinkage

Result Overview- Prediction Part

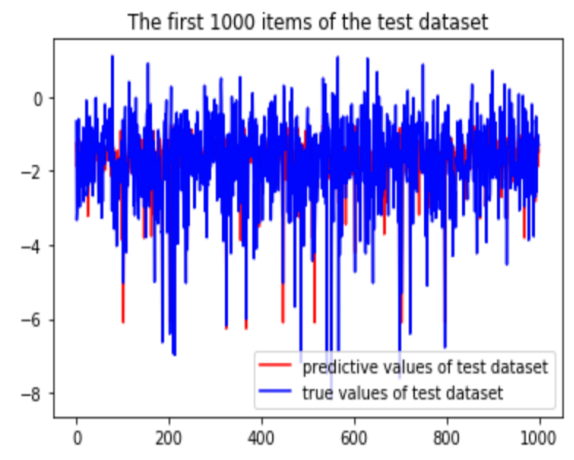
- Although we have the best tuning parameter, the plots both from the Warp & Weft Shrinkage show that the prediction is accurate but not as good as the result from actual width part
- Here are the result plots from these two parts:

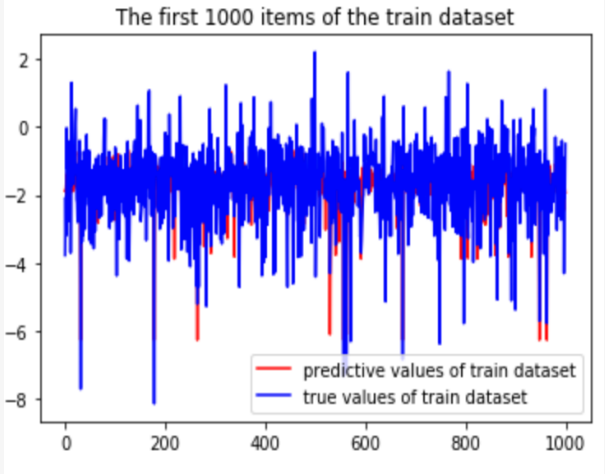
✓ **The Prediction Result for Warp Shrinkage**





✓ **The Prediction Result for Weft Shrinkage**



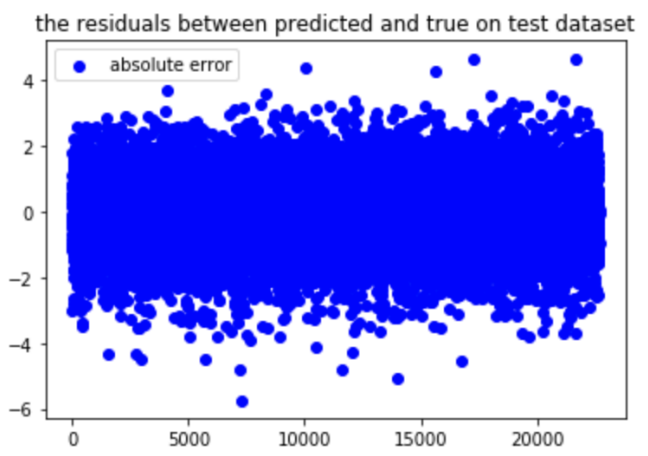


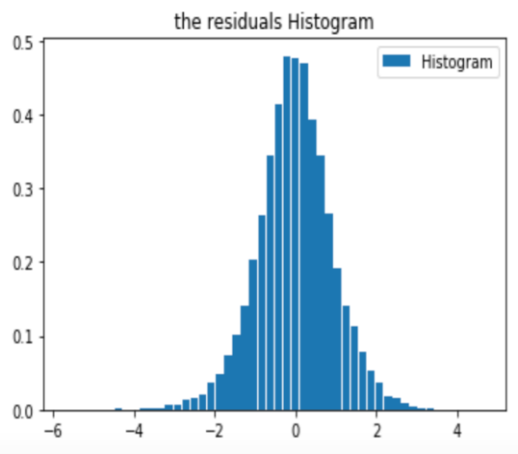
Random Forest for Prediction of Shrinkage

Result Overview- Residual Plot

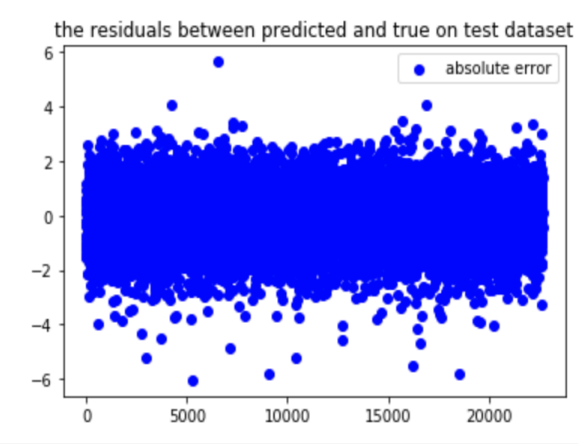
- Although we have the best tuning parameter, the plots both from the Warp & Weft Shrinkage show that the prediction is accurate but not as good as the result from actual width part
- Here are the result plots from these two parts:

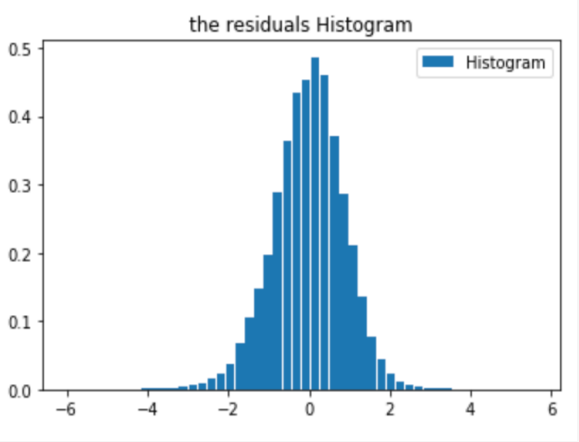
✓ **The Prediction Result for Warp Shrinkage**





✓ **The Prediction Result for Weft Shrinkage**





Model Evaluation

Model Deficiencies

- Higher Cardinality: 1175, compared to 1200 decision trees
- Obscure the order of feature importance
- Alternatives: distributed random forest, a single feature with n numeric values, log n Boolean features representing n values

	Fabric_Type(Text)	Fabric_Type(Numerical)
0	P	0
1	O	1
2	C	2
3	D	3
4	K	4
5	J	5
6	S	6
7	H	7

	Fabric_Type_C	Fabric_Type_D	Fabric_Type_H	Fabric_Type_J	Fabric_Type_K	Fabric_Type_O	Fabric_Type_P	Fabric_Type_S
Fabric_Type_C	1	0	0	0	0	0	0	0
Fabric_Type_D	0	1	0	0	0	0	0	0
Fabric_Type_H	0	0	1	0	0	0	0	0
Fabric_Type_J	0	0	0	1	0	0	0	0
Fabric_Type_K	0	0	0	0	1	0	0	0
Fabric_Type_O	0	0	0	0	0	1	0	0
Fabric_Type_P	0	0	0	0	0	0	1	0
Fabric_Type_S	0	0	0	0	0	0	0	1

Warp Shrinkage

- Neural Network • 0.69
- One Hot Encoding Random Forest • 0.86
- N Numeric Random Forest • 0.51
- ★ Distributed Random Forest • 0.41

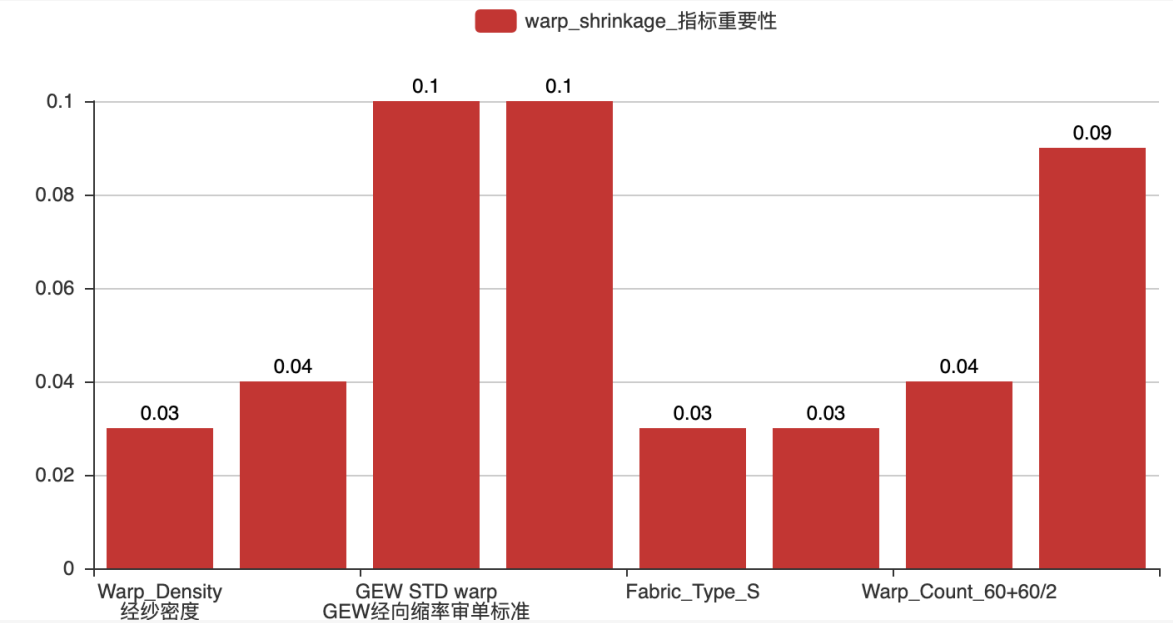
Weft Shrinkage

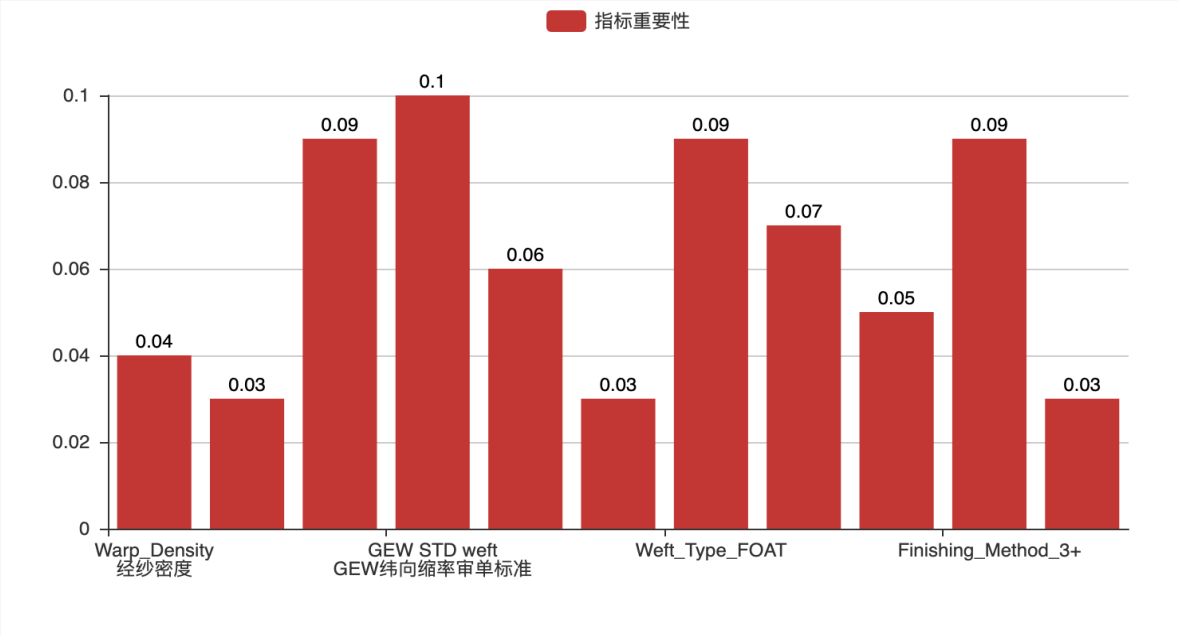
- Neural Network • 0.58
- One Hot Encoding Random Forest • 0.77
- N Numeric Random Forest • 0.48
- ★ Distributed Random Forest • 0.31

Determinants for Shrinkage

HKU BUSINESS SCHOOL 港大經管學院 | Master of Science in Business Analytics 理科碩士(商業分析)

- **GEW STD**
- **Finishing Method**
- **Weft and Warp Density**





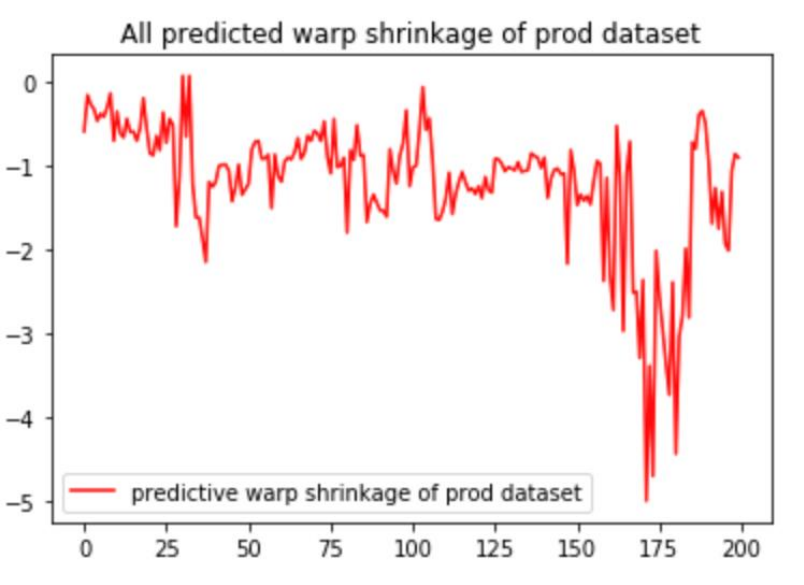
ESQUEL GROUP

Production Test Result for Prediction Shrinkage

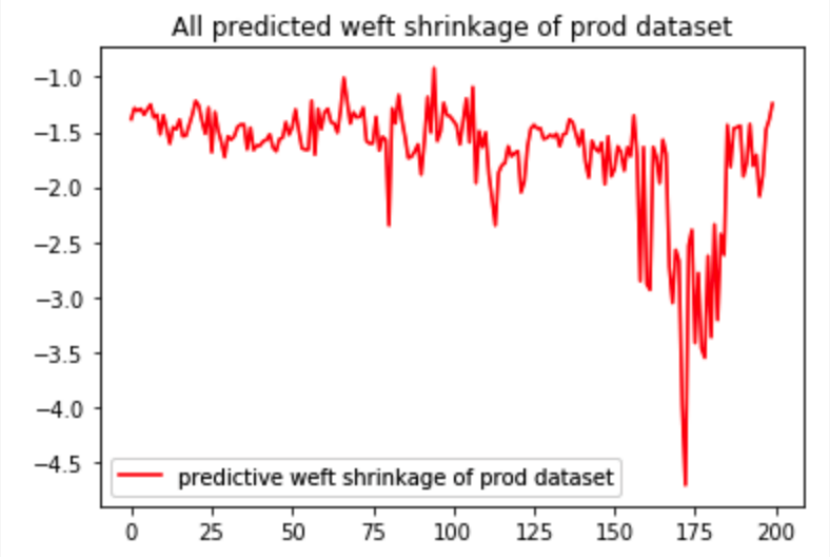
Result Overview- Prediction Results

- We predicted 200 results from the production environment.

✓ **The Production Result for Warp Shrinkage**



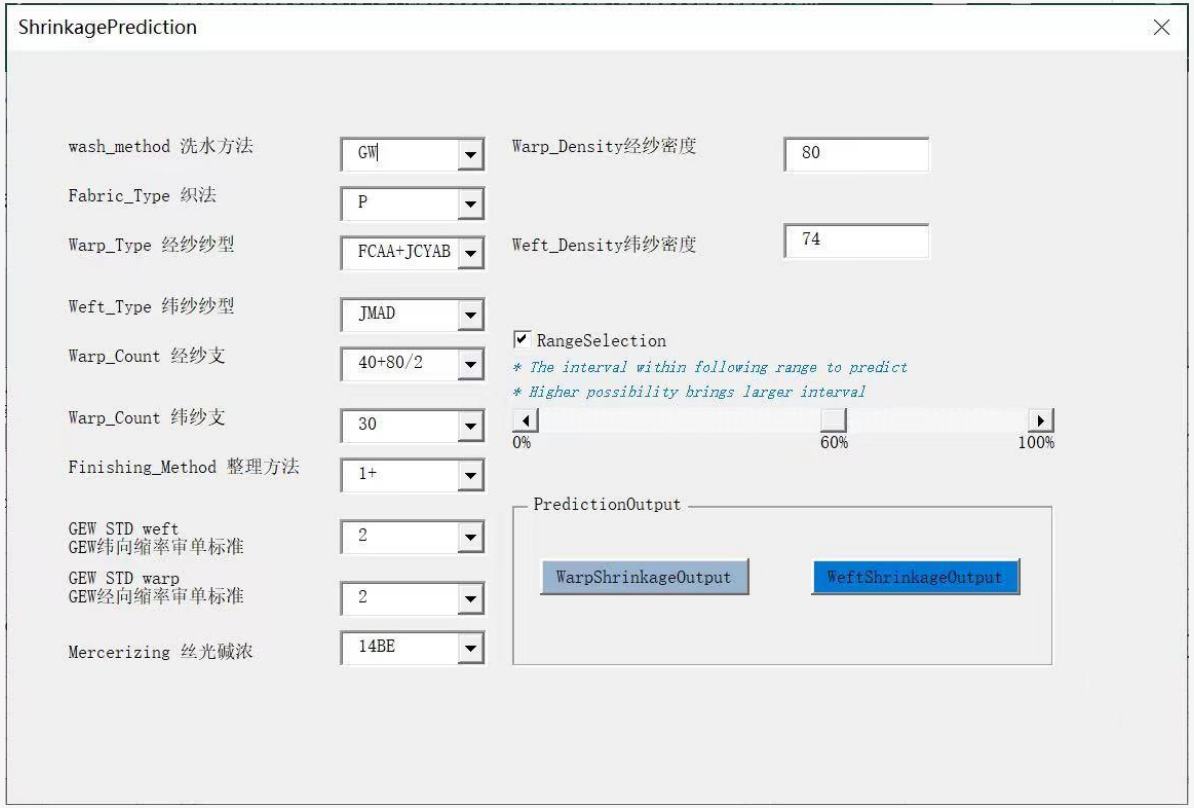
✓ **The Production Result for Weft Shrinkage**



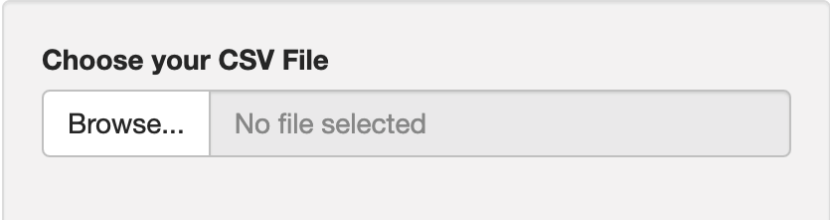
Commercial Implications

How to apply ?

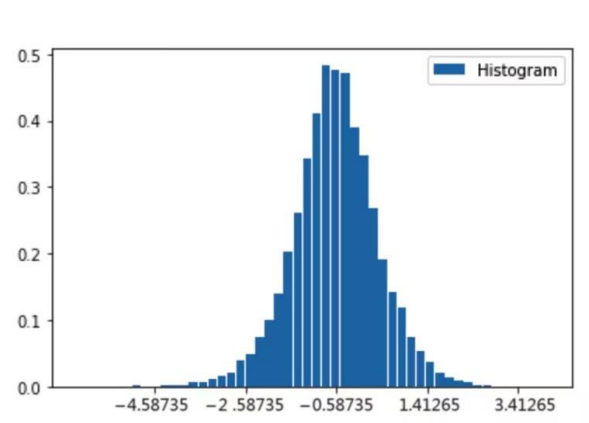
For Singular Output



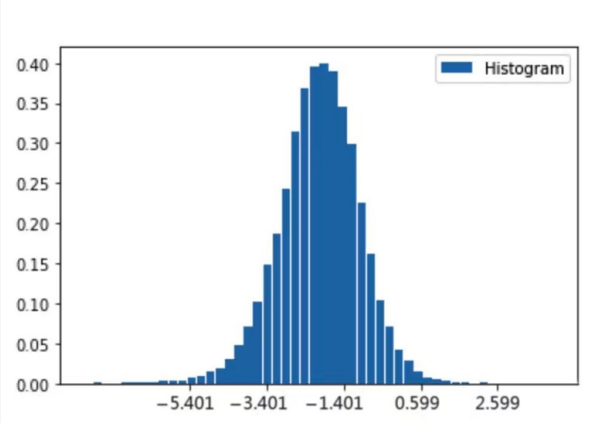
For Multiple Outputs



Warp Shrinkage



Weft Shrinkage



Thank You

HBS Sharks