Analytics, Regressions and Recommendations for Agriculture-Fishery Industries in Mainland China

SHEN ZHENG, WENZHOU KEAN UNIVERSITY

The fishing industry has been a critical and promising research area. Past studies have been unable to relate the fishery economy geographically to agriculture and freshwater resources. This paper presents analytics, regressions on an agriculture-fishery dataset in mainland China, and we provide suggestions based upon our findings. Firstly, we import the dataset and clean missing and extreme values. Secondly, we conduct various exploratory analyses concerning distribution, correlation, geography to explore the characteristics, connections, and discrepancies for different variables. We also run principal component analysis, independent analysis, and multicollinearity analysis to exploit statistical independence information. Thirdly, we build a carefully-tuned random forest regressor for predicting fishery economy. After that, we verify our model's accuracy using several benchmarking metrics. Finally, we draw significant conclusions and recommendations for the agriculture-fishery industries in mainland China based on our data analytics and machine learning results.

Keywords: Agriculture, Fishery, Correlation, Principal Component Analysis, Random Forest Regression.

1 INTRODUCTION

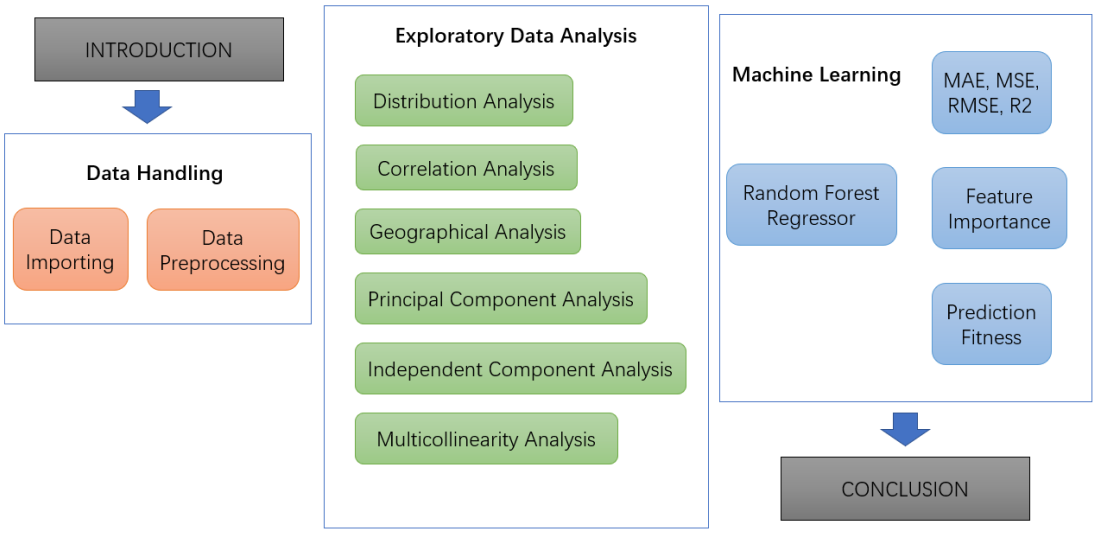
For the recent few decades, the Fishing industry has witnessed explosive development, especially in countries with abundant water resources like China. China has been the world's leading fishery producer. According to Food and Agriculture Organization of the United Nation **[1]**, mainland China in 2015 produced 65.2 million tons of food fish, with 47.6 million tons (73 per cent) from aquaculture and 17.6 million tons (27 per cent) from capture, with an average of 5.4 annual growth rate.

Fishery economy is important for the economy of a country for several reasons. First, the fishing output provides a great quantity of Hyperalimentation food. Take China, for example. China is a country with more than 1.4 billion population. If all Chinese resident relies on land resources, they will be possibly a shortage of land-produced food. Second, fishing industries make significant contributions to other industries. It provides premier fertilizer and forages to agriculture and livestock. It also offers vital raw materials for food, medicine and chemical manufacturing. Third, fisheries are important for export trade, which is a crucial part of gross domestic product (GDP).

Although the fishery economy is a promising study area, little extensive studies have attempted to relate grain agriculture and fishery economy. None is there in-depth studies focusing on freshwater fishery because the majority of fishing activities happens in marine areas. Furthermore, an in-depth examination and comparison of the geographical factors are needed.

In this paper, we analyze an agriculture-fishing dataset in mainland China from 2016 to 2018 using data visualization and machine learning tools. First, we import and preprocessing data. This step includes removing columns and rows that contain missing or abnormal values. Second, we conduct multinomial exploratory data analysis. The analytics includes distribution analysis, correlation analysis, geographical analysis, principal component analysis **[2]**, fast independent component analysis **[3]** and multicollinearity analysis. That analysis not only reveals the underlying data structure but also guide the feature engineering process before model construction. Third, we construct a random forest regressor [4] model, which has carefully tuned hyperparameters, for prediction. After that, we evaluate the accuracy of our model using MAE, MSE, RMSE and Adjusted R squared. Furthermore, we list the feature importance and visualize the predicted value against the actual value to demonstrate our model validity. Finally, we discuss and conclude our findings. The overview of this paper is displayed in the following flowchart.

Figure 1: Overall Workflow Design for This Paper



2 DATA HANDLING

For this paper, I will use Agriculture_Fishing Dataset. This dataset has 19 columns with 97 rows. The dataset mainly contains information about region, date, sown area, grain output and multifarious fishery yields.

First, I check if there is any missing value. I find that columns “seaculture”, “sea_fishing”, “freshwater_fishing”, “sea_culture_yields”, “fishing_yields”, “marine_fishing_yields”,” pelagic_fishing_yields”, “freshwater_fishing_yields” contain significant amounts of missing value. Since we have a small dataset, it is impossible to fill the missing data with an unbiased mean or median in each column. Therefore, I remove those columns that contain missing value. Second, I find three rows with the region “nationwide” are

displaying information as the sum of all province capitals cities rather than an individual one. If I add them in the plot, they will become outliers which harm the visualization and modelling of the data. Therefore, I remove these three rows from the dataset. Last, I drop non-numerical columns including ‘region’ and ‘date’ in preparation for exploratory data analysis in the next section.

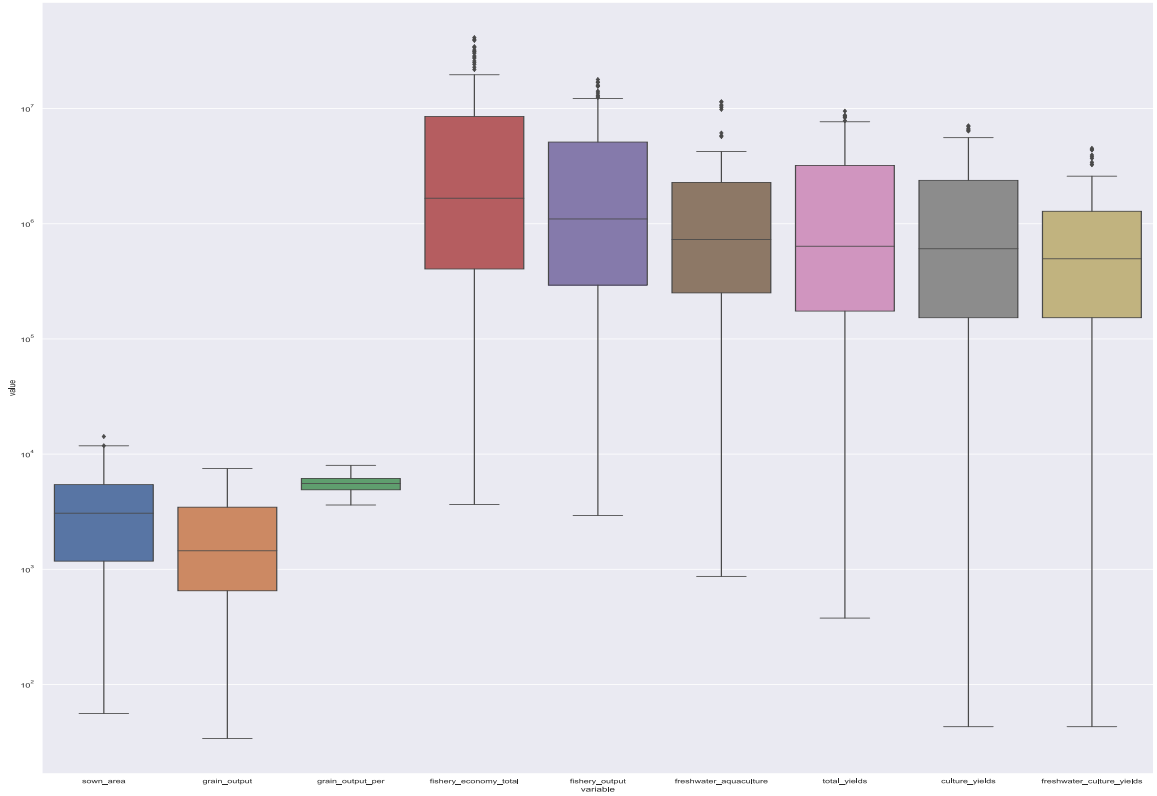
3 EXPLORATORY DATA ANALYSIS

3.1 Distribution Analysis

In this part, I will conduct a correlation analysis at the dataset. First, I plot a boxplot about the dataset, using numerical value at the vertical axis and nine variables at the horizontal axis. We use to boxplot to show the shape of the data distribution, its median value and outliers if detected.

From the plot, we could observe that (1) all variables except "grain_output_per" has dispersed lower parts. (2) All variable except "sown area" has data symmetrically distributed between Q2 and Q3. (3) All variable except "grain_output" and "grain_output_per" has one or more one outliers at Q1.

Figure 2: Boxplot for Nine Numerical Variables



Second, I plot two pair plots about the dataset, using the same nine numerical variables. The first pair plot includes scatterplots, histograms and regression lines. The first pair plots show the cross-variable

distributions with the regression trending line. It also uses a histogram to summarize the distribution of data across different values. The second pair plot includes scattering plots, density plots and contours. The second pair plot also shows the cross-variable distribution. However, it focuses on the underlying structure of the cross-variable distribution. For a single cell of the pair plot, the plot has the value of one numerical variable on the vertical axes and the value of another horizontally.

From the two pair plots, we could observe that (1) Except for "grain_output_per" which is approximately normal distributed, all variables are heavily right-skewed. (2) "Sown area" has a strong positive linear relationship with "grain_output". However, both sown area and grain output have a weak linear relationship with other variables. (3) "fishery_economy_total", "fishery_output", "freshwater_aquaculture", "total_yields", "culture yields" and "freshwater_culture_yields" have strong positive linear relationship with each other. (4) Dense data distributions occur when both the numerical value of two variables are small.

Figure 3: Regression Pairplot for Nine Numerical Variables

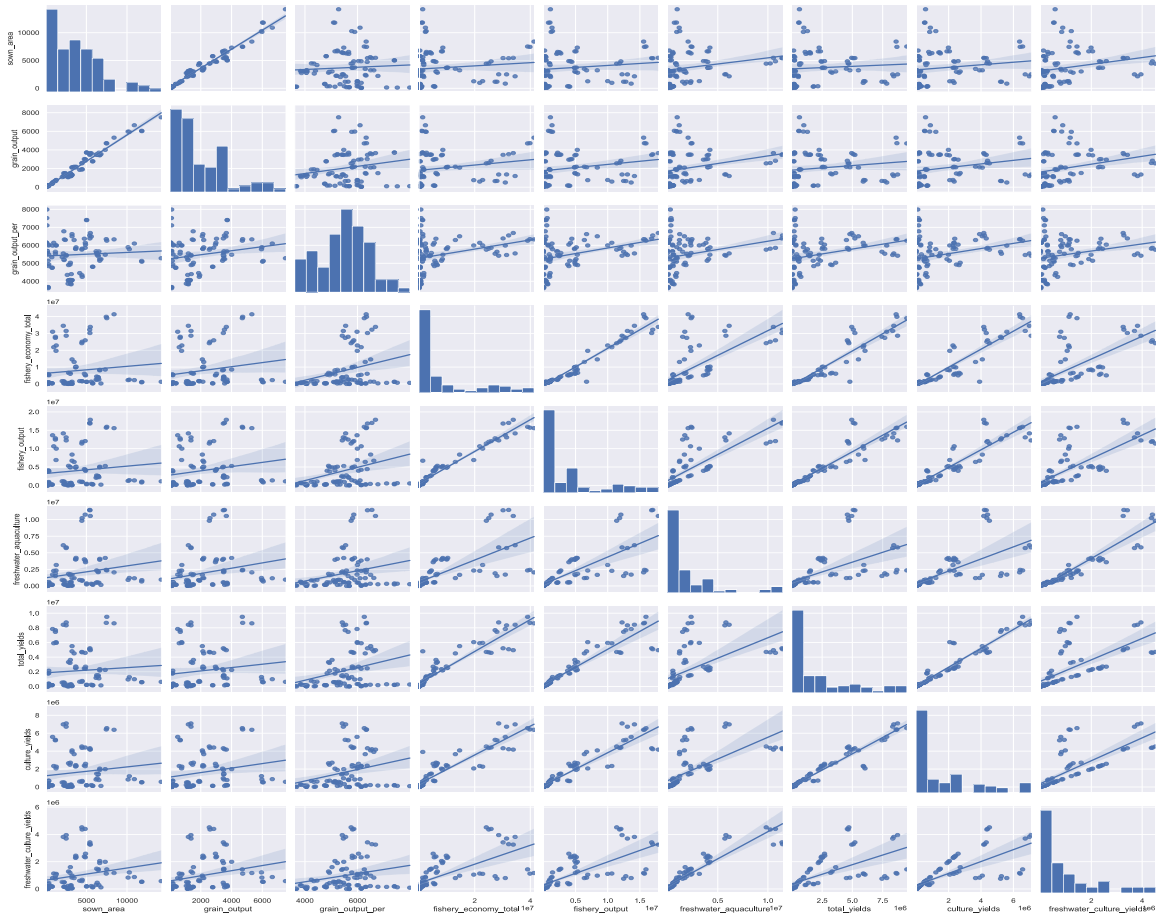
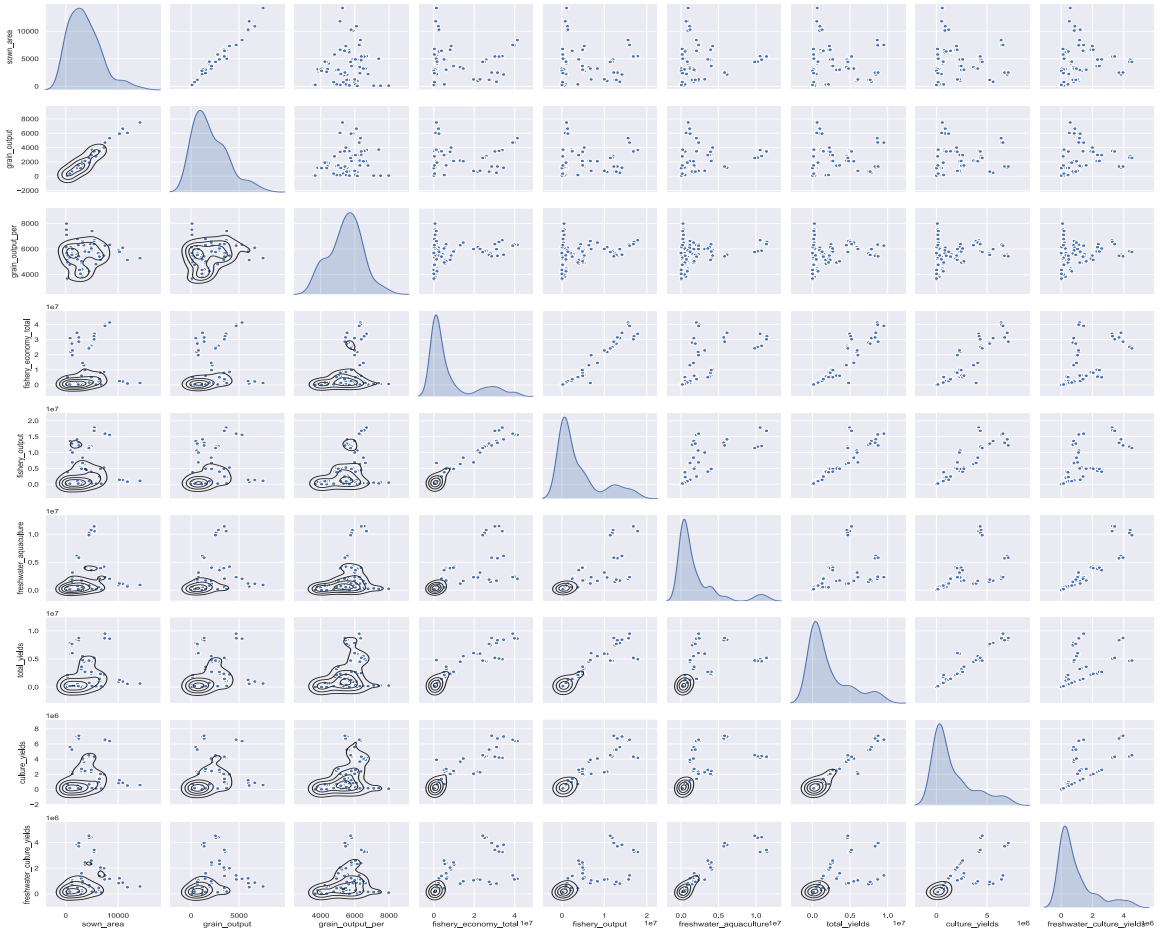


Figure 4: Distribution Pairplot for Nine Numerical Variables



3.2 Correlation Analysis

I plot the correlation plot, using the same nine variables. The correlation plot contains a correlation matrix that visualizes display the correlation coefficient, which is calculated as below:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

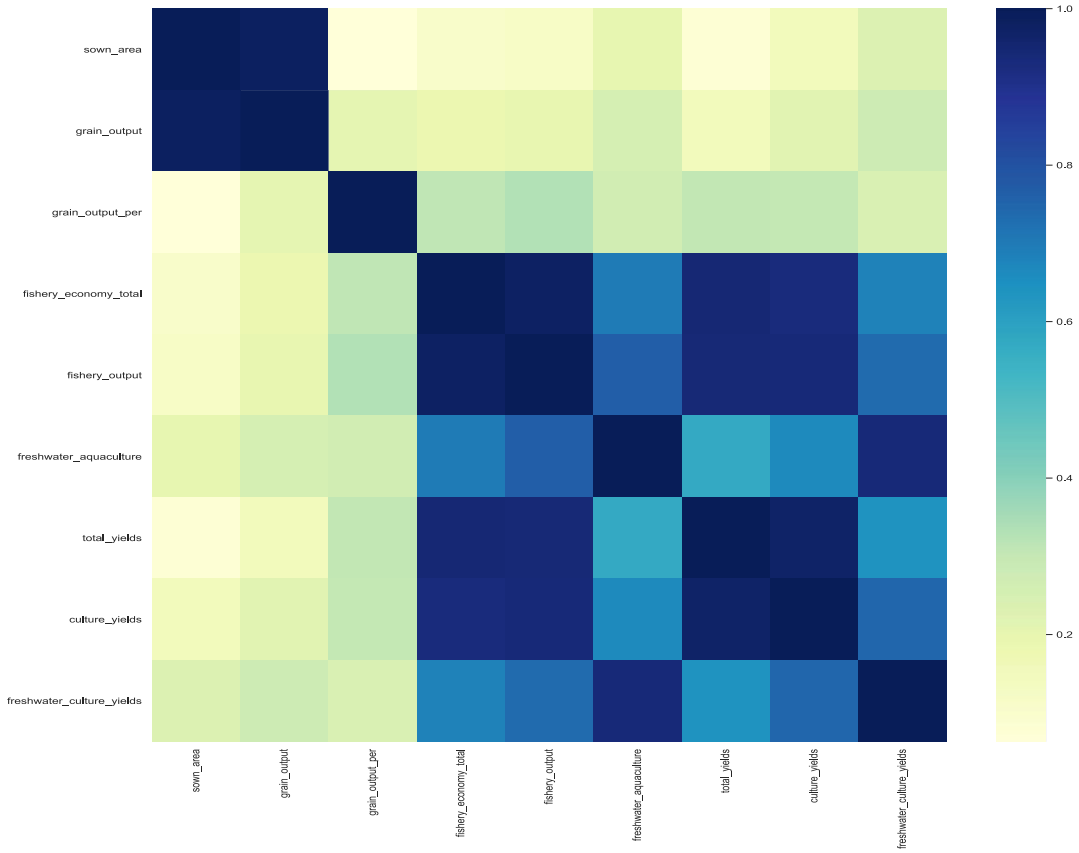
$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$Cor(x,y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

Where x and y are two variables for comparison and n is the numbers of observations for evaluation. The correlation plot shows the correlation coefficient value among different variables. The closer the value to 1.0, the greater the correlation between two variables.

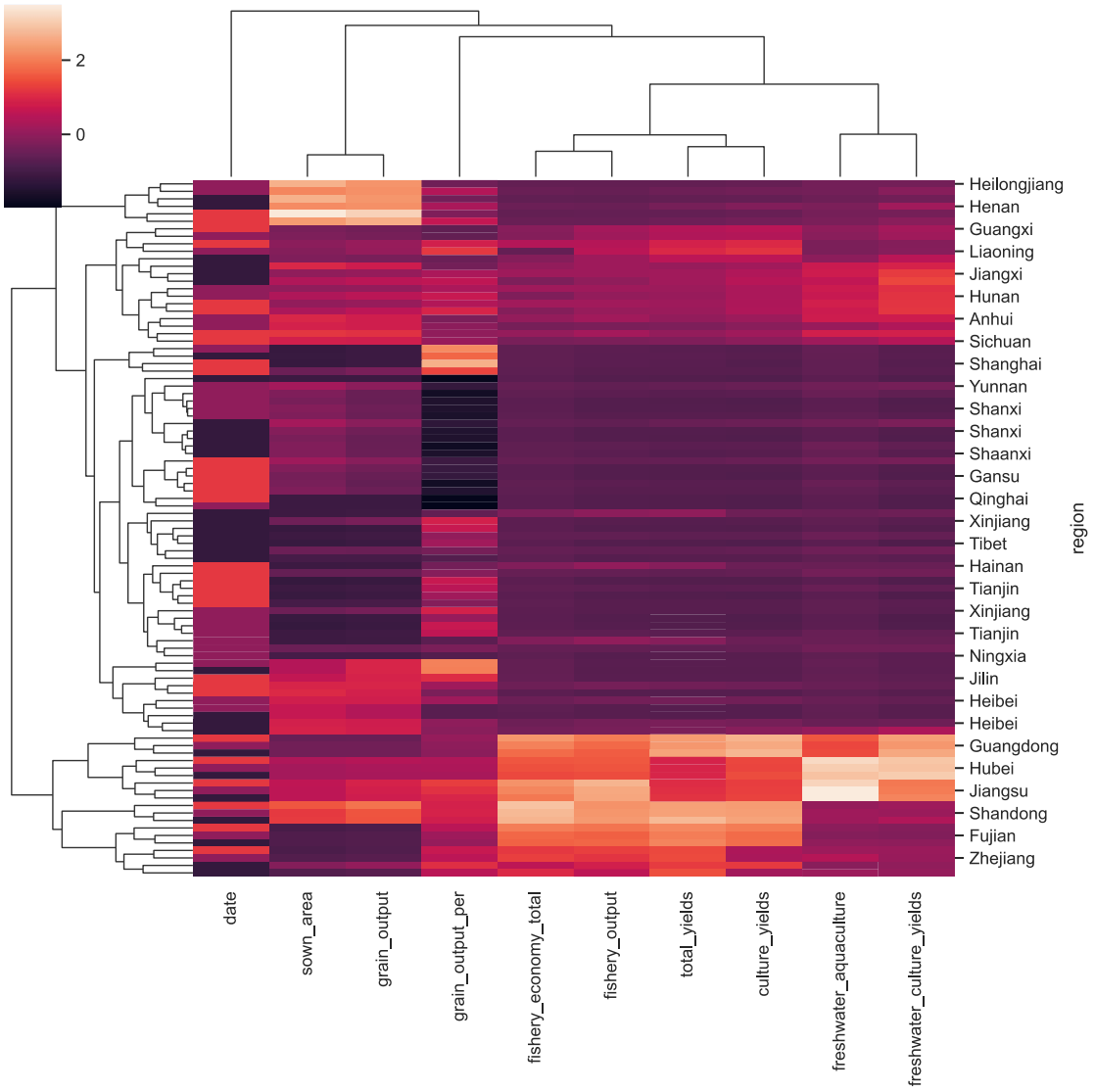
We could observe from the plot that (1) "sown_area" has a strong correlation with "grain_output". (2) "fishery_economy_total" "fishery_output" "culture yields" and "total yields" have a strong positive correlation with each other. (3) "freshwater aquacture" has a strong correlation with "freshwater culture yields".

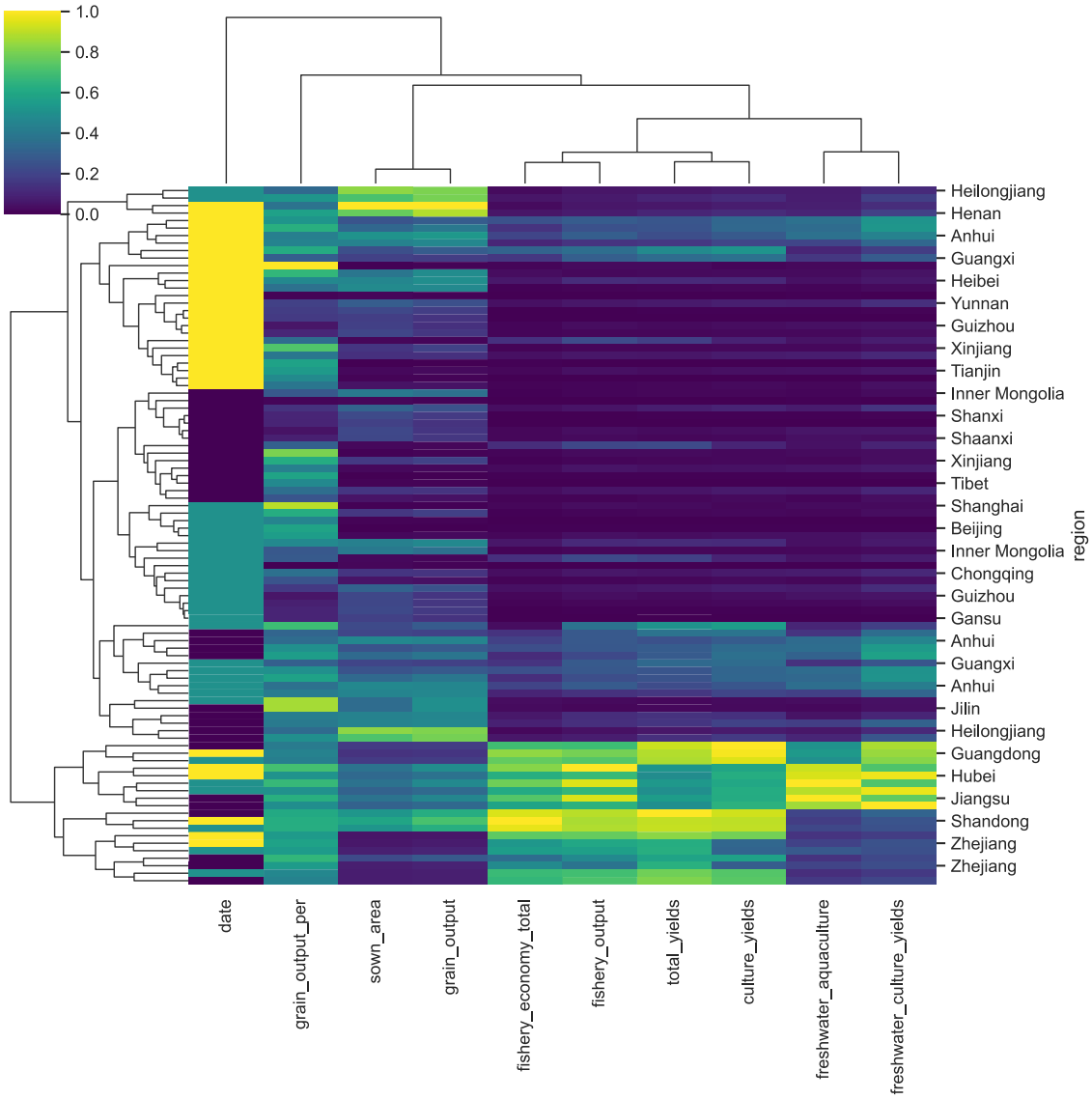
Figure 5: Correlation Matrix Plot for Nine Numerical Variables



I also plot the dendrograms with heatmaps, using the aforementioned numerical variables, the date and the region. In preparation of the plot, I regather the date and region column. The first figure uses heatmap with normalization, which means the variables are scaled to values between 0 and 1. The second figure uses heatmap with standardization, which means the variables are transformed to have a mean of zero and a standard deviation of one. The purpose of these plots is to cluster and visualize the variables according to their similarities with each other.

Figure 6: Dendrograms with Heat maps for Numerical Variables, Date and Region. Upper: Scaled. Lower: Standardized





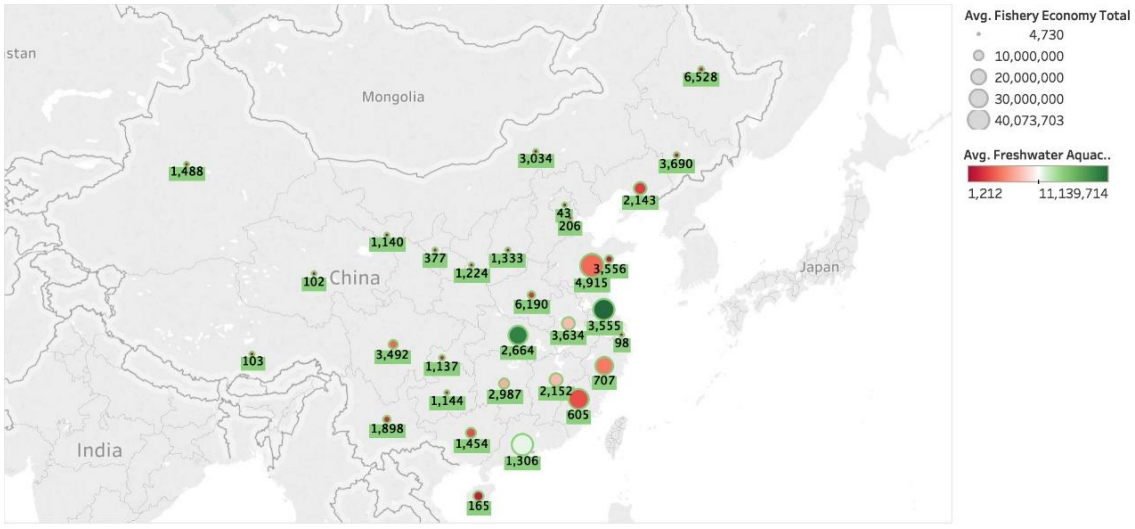
3.3 Geographical Analysis

In this part, I will conduct geographical analysis at the dataset. First, I obtain the geographical location (longitude and latitude) of every province capital city in mainland China. Next, I take the average data from the year 2016 and 2018. The colour, size, marks show the amount of Freshwater Aquaculture, the average of Fishery Economy Total and the grain output, respectively

We could observe the plot tat from the year 2016 to 2018 (1) Provinces in southeast china that is beside the Yangtze river has advantages in freshwater aquaculture. Two major cities and provinces are Wuhan and

Shanghai. (2) Provinces on the east coast of China has advantages in Fishery Economy. Two provinces and cities are Shandong, Shanghai and Wuhan. (3) Provinces in north and northeast China has advantages in grain output. However, Yunan province in southwest China also has advantages in grain output.

Figure 7: Geographical Maps for Fishery Economy, Freshwater Aquaculture and Grain output



Map based on Longitude and Latitude. Color shows average of Freshwater Aquaculture. Size shows average of Fishery Economy Total. The marks are labeled by average of Grain Output.

3.4 Principal Component Analysis

In this part, I will leverage a dimensionality reduction method called principal component analysis (PCA) to analyze the dataset. Before conducting the analysis, I scale the data using Standard Scaler (Z-score normalization). It scales the data such that they have a mean of zero with unit variance. Scaling the data is essential for PCA because variables with large variance might dominant other features, thereby affecting the accuracy of the principal component axis. Furthermore, scaling the data increase the convergence speed of the gradient descent algorithms. The purpose of PCA is to reduce dimension and improve interpretability for data visualization results. To achieve this goal, PCA finds orthogonal linear transformation that maximizes the variance of the variable along the principal axes, which represent the original variables with minimum accuracy loss.

To found out if there are dominant features in the dataset, I also plot the explained the variance concerning the numbers of components. The first plot shows that there are no principal axes that capture the most variance of the data. The second plot shows that there is no dominant feature that could explain the variance of the data. Therefore, if we use the principal axes to replace the original axes, we will loss significantly amount of information about the feature variations. Therefore, there is no dominant variable that could represent the features, and we should not use PCA axes in the modelling part.

Figure 8: Principal Component Analysis for Numerical Variables

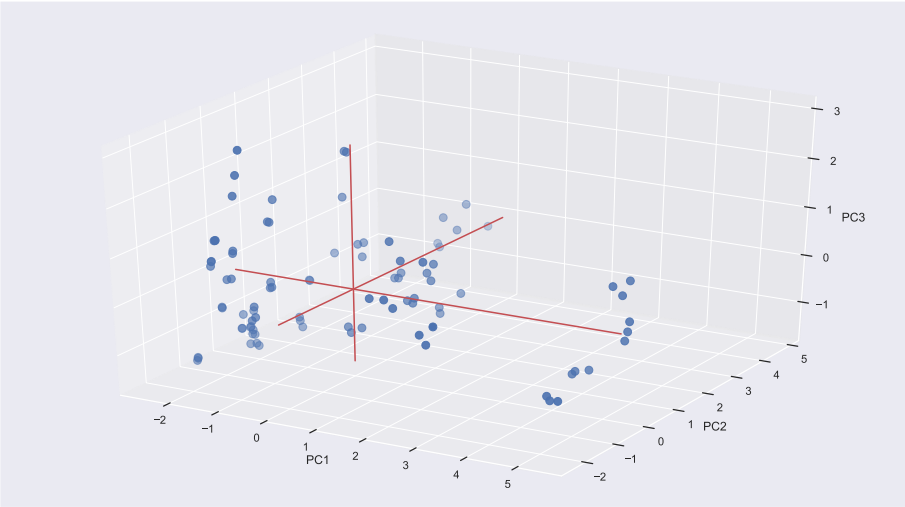
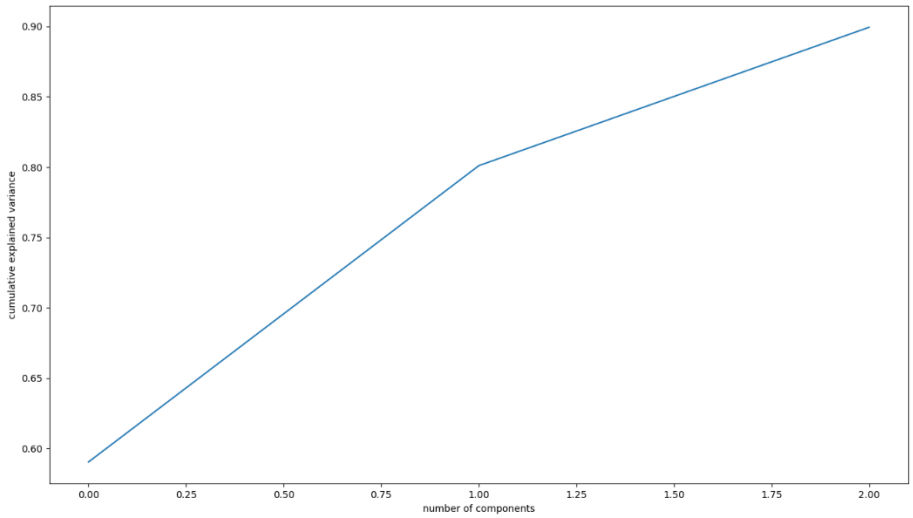


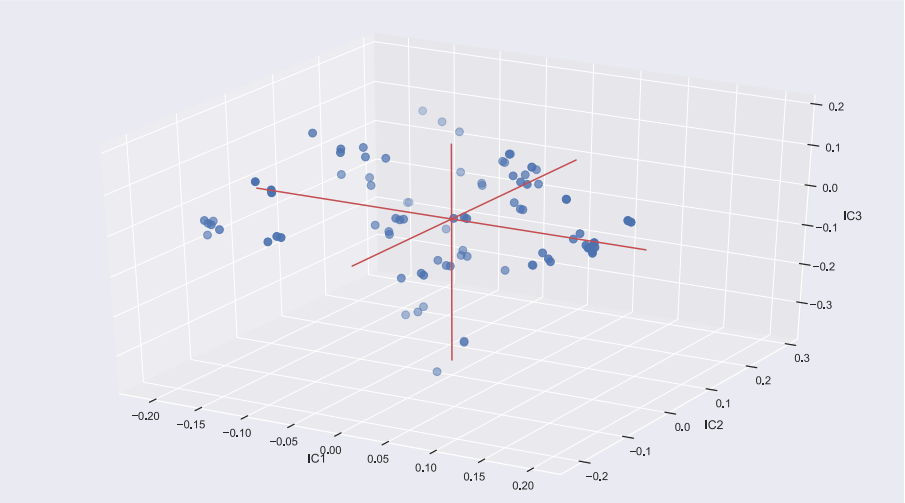
Figure 9: Principal Component Analysis Cumulative Explained Variance



3.5 Independent Component Analysis

In this part, I will leverage a dimensionality reduction method called Independent Component Analysis (ICA) to analyze the dataset. We use the same standardization process of PCA for ICA because that technique could minimize the singularity in the covariance matrix and makes the algorithm converge faster and better. The purpose of ICA is similar to PCA—to remove redundant dimension and to boost data interpretability. While ICA also seeks orthogonal linear transformations, it is different from PCA in a way that it aims to maximize the statistical independence of the variables among the independent component axes which separate the original variables with minimum accuracy loss. From the ICA plot, we could observe that there are no axes that can explain the independence of the dataset. Therefore, there is no significant statistical independence between different variables, and it is not a good idea to use independent component analysis.

Figure 10: Fast Independent Component Analysis for Numerical Variables



3.6 Multicollinearity Analysis

In this part, I will utilize the variance inflation factor (VIF) score to analyze the XXX variables. VIF is used to detect multicollinearity in regression studies. The purpose of VIF is to find the dependency between independent variables we select, thereby guiding the feature selection for building a machine learning model. The formula for VIF is as below:

$$VIF = \frac{1}{1 - R_i^2}$$

where R_i^2 is the coefficient of determination for different variables. Typically, a VIF score under 5 means the independent variables are not significantly correlated, where VIF score above 5 suggest strong correlation. The lower the VIF value, the better the variable is for model construction. The table for VIF score is shown as below:

Table 1: VIF score for numerical variables

Variables	**VIF**
sown_area	79.453343
grain_output	83.823201
grain_output_per	2.633046
fishery_output	91.071524
freshwater_aquaculture	65.740459
total_yields	58.552583
culture_yields	33.058529
freshwater_culture_yields	43.602514

From the table, we could observe that there is strong multicollinearity for all selected variables except "grain_output_per". Since multicollinearity does not affect the prediction accuracy, which is our goal, we do not plan to solve that problem in the data analysis part. Instead, we will choose a sophisticated machine learning model which can accommodate different multicollinear variables.

4 EXPLORATORY DATA ANALYSIS

In this part, we will construct a regression model called random forest regressor with the same standardized variables for PCA and ICA as an independent variable, except that "fishery_economy_total" as the dependent variable. Random Forest regressor is an ensembled supervised machine learning algorithm for regressions. It is a meta estimator that leverages numbers of decision trees and takes the average of those predictors to improve forecasting accuracy and reduce prediction fluctuations. We choose a random forest regressor for three reasons. First, the random forest is simple and interpretable. Compare with black-box algorithms like neural networks, random forest methods could directly feedback the feature importance for model evaluation and hyperparameter tuning. Second, the random forest is less sensitive to outliers and independent variables with high multicollinearity because the mean prediction from all decision trees mitigates the influence of those small fractions of undesired data. The data will split into training data and testing data with a split ratio of 2:1. After that, we will fit the training and testing data into the random forest regressor.

We also carefully tune the hyperparameters of random forest regressor using Randomized SearchCV. The hyperparameters we select in a random grid for tuning is n_estimators, max_depth, min_samples_split and min_samples_leaf. We restrict our selection to these four hyperparameters to obtain maximum improvements from limited computational resources. Specifically, we first create a random grid to include the selected hyperparameters. Secondly, we run a threefold cross-validation for 100 iterations. In this way, we

can find the group of hyperparameters minimizing the loss. Last, we report the best params and to use it to construct the optimized random forest regressor as the table below.

Table 2: Best Hyperparameters for random forest regressor

Metrics	Value
n_estimators	288
min_samples_split	3
min_samples_leaf	1
max_depth	90

Our efficient hyperparameter tuning gives us great model accuracy, which we will explore in the next subsection.

5 MODEL EVALUATION

In this part, we evaluate our machine learning model on several benchmarking metrics including mean absolute error (MAE), mean squared error (MSE), and root means square error (RMSE) and coefficient of determination (R2). The first three errors measure the difference between the predicted and actual value. The last one measure how well the predicted value explains the variation of the actual value. Suppose n represent the numbers of data points. The formula for MAE, MSE, RMSE and R squared value is as below.

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n}$$

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

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

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} \left(y_i - \frac{1}{n}\sum_{i=1}^{n} y_i\right)^2}$$

Where y_i represent the actual function value and $\hat{y}_i$ represent the predicted value. We also summarize the model performance using various metrics in the table below.

Table 3: Random Forest Regressor Model Performance

Metrics	Value
Mean Absolute Error	0.04076977
Mean Squared Error	0.004522260
Root Mean Squared Error	0.06724775
R2	0.9915118

We could observe from the result that both the MAE, MSE and RMSE are significantly lower than the standard deviation of the scaled data. Therefore, the difference between predicted and actual one is caused by random noise rather than the model misfit. Furthermore, the R squared value is extremely close to 1. The

result indicates that our model well captures and explain the variance of the actual value. From the analysis above, we can tell that our random forest regressor model did an excellent job in predicting the variable.

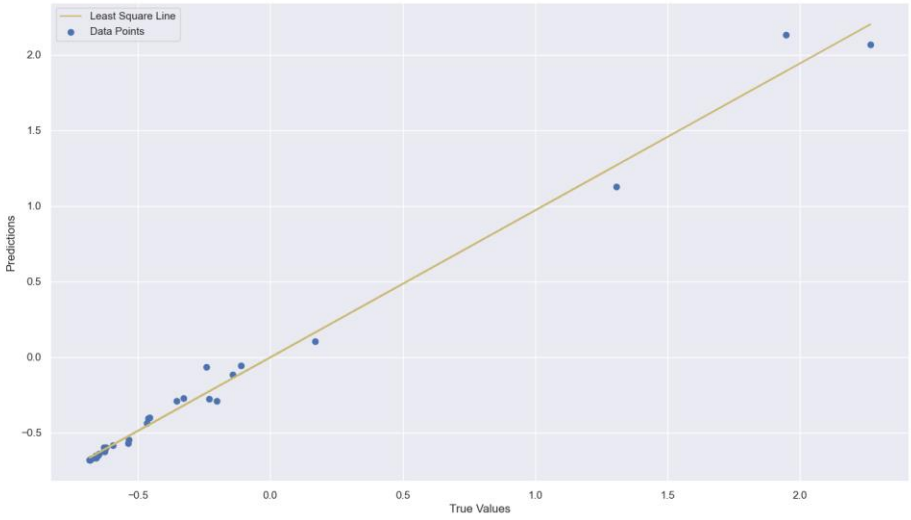
To assess the robustness of our model, we display the feature importance table and the Least Square Line between the true and predicted values. The feature importance score shows the four features, which is "fishery_output", has the most governing impact on the random forest regressor. Besides, the variable "total_yields" also have some significant influence on the prediction output.

Table 4: Random Forest Regressor Feature Importance

Features	Importance Scores
1	0.00767
2	0.00693
3	0.00462
4	0.75914
5	0.00784
6	0.14272
7	0.06398
8	0.00710

We also plot scatter points with a least-square line, using the true values of "fishery_economy_total" at horizontal axes and the predicted value at the vertical axes. Our least square line show there is a strong positive linear relationship between the true and predicted value. Therefore, our prediction well fit the actual values.

Figure 10: Prediction Fitness against Actual Values



7 CONCLUSION

In this paper, we have visually and quantitatively analyzed an agriculture-fishing dataset from 2016 to 2018 in mainland China. First, we clean missing and unrepresentative observations. Second, we exploit various data analysis tools to exploit the distribution and structure of different variables. Third, we construct an optimized random forest regressor for prediction the fishery economy total. Last, we demonstrate our model's accuracy using several benchmarking metrics.

Based on our implementations, we conclude the following major findings for mainland China from 2016 to 2018. Firstly, agriculture development is geographically more balanced than fishery development. Although north and northeast China has advantages in the agriculture economy, southwest and south China also have a reasonable amount of agriculture output. In contrast, the east coast of China has dominant advantages in fishery economy. Besides, southeast China at the Yangtze river region has overwhelming advantages in freshwater aquaculture. Secondly, provinces in southeast China are strongly correlated for marine fishery and agricultural industry. Besides, Provinces in the southwest and northwest China are strongly correlated for fishery and agriculture. However, their relationship in unstable concerning time. Thirdly, agriculture features are strongly correlated with agriculture ones but no other features. So are fishery. Although there is no significant correlation between agriculture features and fishery features, there is some relationship between them when their numerical value is extremely small. Fourthly, the fishery economy is more importantly concerned with fishery output and yields. However, since there are neither statistically dominant nor independent features, and since there is strong multicollinearity for most variables, more statistical experiments are required to say that other features are trivial.

We also offer several major suggestions for mainland China Agriculture-Fishing industry. Firstly, provinces in the north and south part in China should strength trade cooperation in agriculture and fishery to bridge regional imbalance. For example, provinces in the east coast of China in the Yangtze River region could utilize freshwater aquaculture fertilize for provinces in northeast China's grain planting. And those provinces excelled at agriculture could provide grain as fishery feeds to eastern coastal provinces. Secondly, the western provinces of China should establish a more stable agricultural and fishery cooperation. For instance, province leaders could put forward Long term policy, schematization and memorandum for cross-province collaboration. Thirdly, Provinces with underdeveloped agriculture and fishery industries should exploit the cooperation between agriculture and fishery to promote win-win development. For example, they may process agricultural, and fishery produces together and to export integrated goods to other provinces or countries. Fourthly, provinces with backward or stagnant fishery growth should prioritize increasing fishery output and yields. For instance, they should award the individuals and companies that produce large amounts of fishery output.

REFERENCES

[1] Jason Jerald. 2015. The VR Book: Human-Centered Design for Virtual Reality. Association for Computing Machinery and Morgan &

[2] Pearson, Karl. "LIII. On lines and planes of closest fit to systems of points in space." *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11 (1901): 559-572.

[3] Hyvarinen, Aapo. "Fast and robust fixed-point algorithms for independent component analysis." *IEEE transactions on Neural Networks* 10.3 (1999): 626-634.

[4] Ho, Tin Kam. "Random decision forests." Proceedings of 3rd international conference on document analysis and recognition. Vol. 1. IEEE, 1995.