Xueyang Zhao 50197145 (zxy4107@gmail.com)
Yue Chen 28166149 (chenyue96@hotmail.com)
Shiyang Li 48753140 (melauria0126@hotmail.com)
Yumeng Chen 35365148 (marsrabbit.meng@gmail.com)

STAT 406 Term Project

1.Abstract

The purpose of our study is to find a model with good prediction power in order to estimate the burned area of forests and illustrate its relationship with various environmental statistics. The methodology includes stepAIC, ridge regression, LASSO regression and random forest to select features and fit models. Using the Mean Squared Prediction Power(MSPE) as our criteria, ridge regression has the smallest MSPE than the other methods. Then, using random forests and out-of-bag error estimates, we ranked the importance of each variable and were able to predict future burned forest area. Even though the MSPE of random forest is larger than that of ridge and LASSO regression, but the difference between those MSPEs is relatively small. In conclusion, the random forest has relatively good prediction power and improves the prediction accuracy comparing to other methods we analyze.

2. Introduction

Every year thousands of acres of forests and residential areas are ruined by the forest fire, leading to a negative consequence that thousands of people are forced to flee their hometown. Predicting the burned area of each forest fire in early time using measured data allows people to prevent the forest fire from affecting a large area. In addition, people can prepare in advance and take measures when forest fire happens. In our study, we aim to find a model with good prediction power to estimate the burned area of forest fire and rank the importance of factors that may affect the area.

In the following report, data description and methodology will be shown in Section 3, including the expected difficulties and challenges we may encounter when processing data. Main Data Analysis will be given in Section 4, and our conclusion will be given in Section 5.

3. Data Description & Methodology

3.1 Expected difficulties and challenges

1. The original data contains 517 observations and it is right-skewed because it has lots of zeros in the response variable. Since our objective is to investigate the burned area of a forest fire, observations with 0 burned area makes no sense to our project. We removed all the data with 0 burned area and log-transform the response variable based on its histogram (Figure 1 and 2).
2. In our dataset, there is a forest fire that causes damage to 1091 ha forest. Based on the scatter plot, it is a rare event and an outlier which can affect our analysis dramatically and may result in a bad prediction. Hence, we removed this row from our dataset as well.
3. There are two categorical variables in our dataset, month and day. Based on our research, day does not affect the forest fire area, so we removed day from our explanatory variables. However, month may have effect on forest fire. For example, in the summertime from June to September, increasing temperature may cause wildfire and forest fire spreading. In order to run ridge and LASSO regression, the categorical variable is transformed into dummy variables.

4. Some of the explanatory variables are correlated, for instance, FFMC and ISI. When deciding on the methodology, ridge regression might perform better than LASSO. In addition, to reduce the negative impact of correlations, we use random forests in order to make bagged trees less correlated.
5. Curse of dimensionality: Our modified data contains 269 observations with 19 explanatory variables. Because of this high dimension, we cannot use methodology like kernel.

3.2 Variables Description

Our modified dataset contains 269 observations and no missing values. The following are response and explanatory variables:

1. **area**: the burned area of the forest (in ha)
2. **X**: x-axis spatial coordinate within the Montesinho park map (1 to 9)
3. **Y**: y-axis spatial coordiante within the Montesinho park map (2 to 9)
4. **9 dummy variables month**: month of the year ('feb' to 'dec')
5. **FFMC**: FFMC index from the FWI system (18.7 to 96.20)
6. **DMC**: DMC index from the FWI system (1.1 to 291.3)
7. **DC**: DC index from the FWI system (7.9 to 860.6)
8. **ISI**: ISI index from the FWI system (0.0 to 56.10)
9. **temp**: temperature in Celsius degrees (2.2 to 33.30)
10. **RH**: relative humidity (in %)
11. **wind**: wind speed (in km/h)
12. **rain**: outside rain (in mm/m^2)

For variable 6 to 9, the data contains some indexes from the FWI system(Candadian Fire Weather Index System, n.d). FWI system stands for Fire Weather Index System which is used to give relative estimates of potential fire behavior using fuel moisture and current weather conditions. In this system, FFMC stands for Fine Fuel Moisture Code which plays a significant role in ignition probability and spread. DMC stands for Duff Moisture Code which contributes to receptivity and over all fire intensity. DC stands for Drought Code which contributes to depth of burn, intensity, and suppression difficulty. ISI stands for Initial Spread Index which combines FFMC and wind speed. The response variable is "area", and the rest are the explanatory variables.

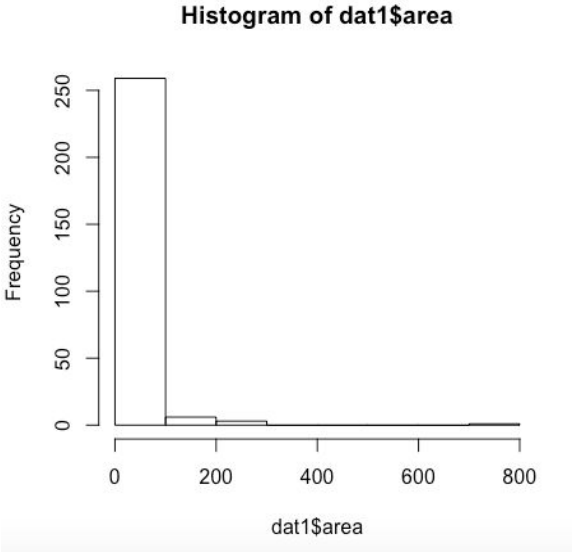

Figure 1

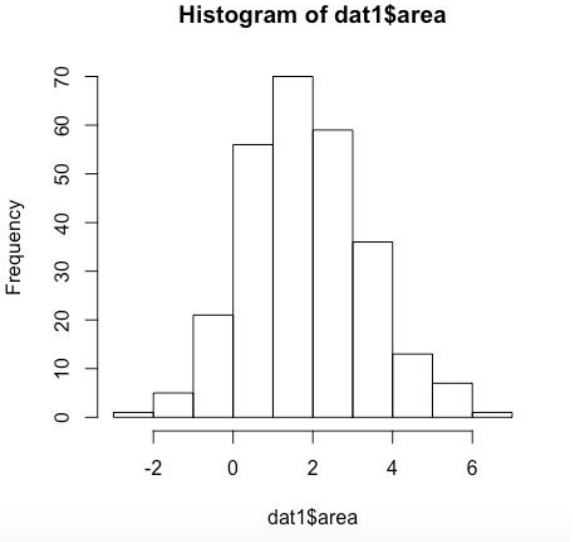

Figure 2

Figure 1: histogram of data after deleting all zeros in area column.
Figure 2: histogram with logarithm transformation after removing zeros.

3.3 Methodology

The metric used in our project is Mean Squared Prediction Error(MSPE) to check whether the model is a good fit. Considering the limitations and challenges, we decided to use ridge and LASSO for linear regression and to compare both of them based on their MSPE, advantages and disadvantages. After linear regression, we used nonparametric method --- random forest to solve curse of dimensionality and reduce the potential correlation when doing regression tree with bagging.

4. Data Analysis

4.1 Full model

Firstly, we fitted the full model to the data using all explanatory variables. The MSPE of full model, which is calculated by 5-fold cross validation, is 2.3261. Obviously, not only is the MSPE of full model larger than other models, but it is also more variable than others in accordance with the box plot(Figure 3). That is, the range of MSPE of full model is apparently much wider than other models. A prediction equation and the significant levels of each variable are shown in the summary of the full model. Even if the full model allows us to predict the burned area, it gives us the worst prediction among these four models. Hence, we will continue our analysis and try to improve our model in the following.

4.2 StepAIC

Another method we used to fit the model is the forward stepwise based on AIC. In the box plot(Figure 3), the stepAIC indeed improve our model; its MSPE(= 2.2970) is smaller than the full model. However, the prediction equation from stepAIC is not good enough because of the highly correlated covariates in this dataset. According to the multicollinearity, if the stepAIC simultaneously includes some correlated covariates in the model, these covariates will mask each other, leading to a consequence that all of them become insignificant. And this will reduce the prediction accuracy of the stepAIC model.

4.3 Ridge Regression

Previous StepAIC can be highly variable. Therefore, in order to achieve more accurate prediction with less variable, ridge regression is a better choice. In addition, ridge regression performs better with potentially correlated explanatory variables.

The penalty parameter will typically yield estimators with varying predictive accuracies. Therefore, finding the optimal lambda is important. We used the cv.glmnet package in R with 5 folds to find the optimal penalty parameter based on MSPE. We found out the optimal lambda equals to 18.304, and also the model produced by ridge regression with this lambda. In addition, the MSPE of ridge regression is 2.2632 and the effective degree of freedom is 12.1402. Compared with stepAIC, we can see that ridge regression has a smaller MSPE which means more accurate prediction power. However, ridge regression uses all the explanatory variables, so it does not produce a sparse model. We can improve this problem by doing LASSO regression.

4.4 LASSO Regression

In LASSO regression, we also used the glmnet package in R with alpha=1 and used function cv.glmnet with 5 folds to find the optimal lambda. The procedure is similar to ridge regression. LASSO Regression can give us a sparse model, which means it can do variable selection. Therefore, it can eliminate the correlation between the variables. However, LASSO might not choose the most important one among correlated variables since the method will always randomly choose one variable from all correlated variables. Also, if two or more correlated variables need to be considered in the model, LASSO regression might have a worse performance than ridge regression. The optimal lambda

for LASSO Regression is 0.141721 and the MSPE of this model is 2.2986 which is larger than that of ridge regression. At this time, LASSO would select two most important variables for us.

4.5 Random Forest

Random forest was then used to further improve the stability of the predictor. Bagging can also be used to improve the stability, but trees may be correlated with each other when using bagging. Random forest is usually used to de-correlate bagged trees, so we omitted the bagging step and used random forest directly.

1000 was the first choice of the forest size. After plotting the error versus number of trees graph, we observed that error became stable when the number of trees was greater than 300. Hence, 500 was then used as the forest size, and 6 explanatory variables were used at each split. The resulting mean of squared residuals is 2.328157 which is smaller than that of full and stepwise models. Therefore, 500 is a reasonable forest size.

According to the variable importance ranking plot (Figure 6), DMC is considered to be the most important variable in this model, followed by other meteorological conditions and fire indexes except rain. All months are considered to be unimportant which is a reasonable result. Rain is ranked as the least important variable which is reasonable because there are a lot of zeros in the rain column.

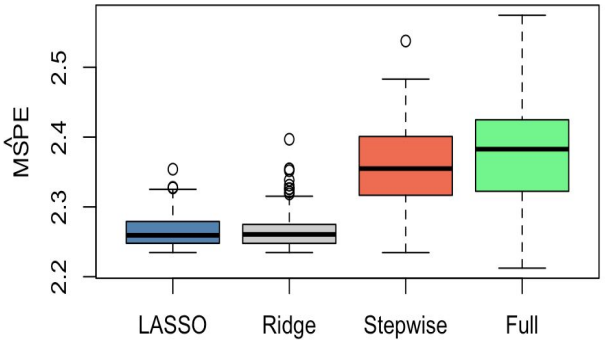

Figure 3

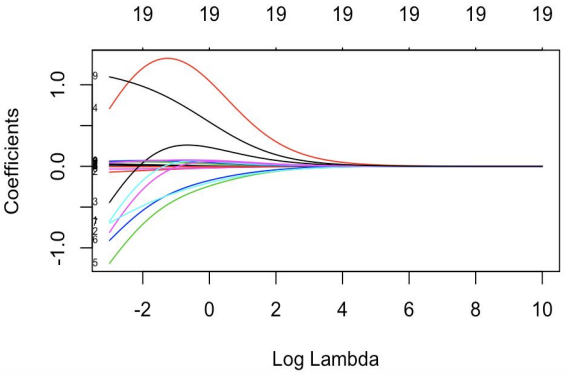

Figure 4

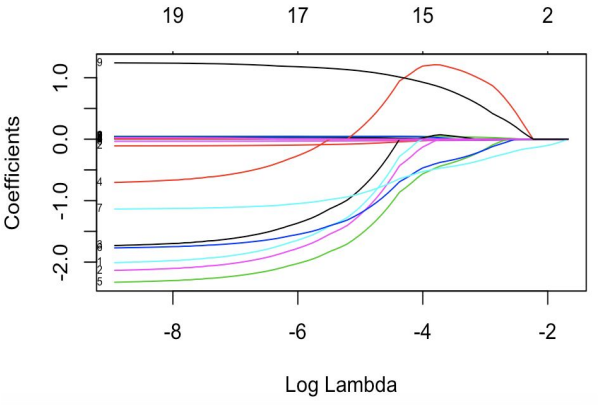

Figure 5

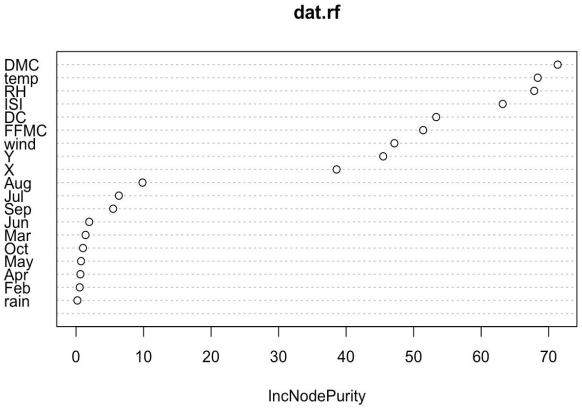

Figure 6

Figure 3: boxplot of MSPE of the full, stepwise AIC, ridge, and LASSO models

Figure 4: plot of the estimated regression coefficient for each possible value of the regularization parameter in ridge regression

Figure 5: plot of the estimated regression coefficient for each possible value of the regularization parameter in LASSO regression

Figure 6: variable ranking plot of random forest method

5. Conclusion And Further Discussion

5.1 Conculsion

In conclusion, LASSO and Ridge Regression has the smallest MSPE. The equations for these two models are:

LASSO:

log(area)= 1.8654066208 + -0.0006397549(ISI) + -0.0919697785*(Aug)*

Ridge Regression:

log(area)= 1.82+ 6.43e-4(X) - 1.83e-3*(Y) - 1.24e-3*(FFMC) + 7.82e-5*(DMC) - 1.52e-5*(DC) - 2.78e-3*(ISI) - 6.73e-4*(temp) - 3.03e-4*(RH) + 3.38e-3*(wind) + 4.27e-3*(rain) + 9.69e-3*(Feb) + 1.27e-2*(Mar) + 3.45e-2*(Apr) + 0.14*(May) - 2.74e-2*(Jun) - 1.91e-2*(Jul) - 2.74e-2*(Aug) + 1.33e-2*(Sep) + 6.41e-2*(Oct)*

However, Random Forest may have a better prediction power for future data by applying the out-of-bag error estimation to predict each observation using those trees in which this observation was not used. For this dataset, the variable ranking from Random Forest makes more sense than LASSO and Ridge Regression — all months and rain are ranked as unimportant variables while ridge regression and LASSO consider several months to be even more important than meteorological variables and fire indexes. In addition, even though Random Forest has a larger MSPE than ridge regression and LASSO, the differences are relatively small (less than 1). Hence, it is reasonable to use Random Forest even its MSPE is larger than ridge regression and LASSO.

5.2 Further Discussion

In this study, we discussed about the influencing factors for the burned area of forests, so we only cared about those observations with burned area of forests larger than 0 , and deleted all observations with area equals to 0. For future studies, we can use this dataset to predict whether a place would have a wildfire or not given influential variables. In order to achieve this, the area column in dataset can be changed into a categorical variable with values “0” and “1”. “0” means there is no wildfire, so the burning area is 0; “1” means there exists wildfire, which indicates burned area is greater than 0. Then, classification tree, boosting, random forest, and other methods can be applied to predict the model. Furthermore, the area column can be classificated into more clusters by considering the size of burned areas. For example, all observations with no wildfires detected are classificated as “0”; all observations with burned area between 0 and 50 can be classificated as “1”; all observations with burned area between 50 and 100 can be classificated as “2”, and so on. This allows us to have a deeper study about forest fires burning area. Such knowledge is useful for forest fire management and resources plan.

6. Reference

1. Candadian Fire Weather Index System. (date unknown) Retrieved October 15, 2017, from https://www.frames.gov/files/6014/1576/1411/FWI-history.pdf
2. [Cortez and Morais, 2007] P. Cortez and A. Morais. A Data Mining Approach to Predict Forest Fires using Meteorological Data. In J. Neves, M. F. Santos and J. Machado Eds., New Trends in Artificial Intelligence, Proceedings of the 13th EPIA 2007 - Portuguese Conference on Artificial Intelligence, December, Guimarães, Portugal, pp. 512-523, 2007. APPIA, ISBN-13 978-989-95618-0-9. Access online at: http://archive.ics.uci.edu/ml/datasets/Forest+Fires