Aerial intel DS Challenge_Zhiyu

High-level summary

We would like to predict wheat yield for several counties in the US. The information available include location (latitude, longitude), time of observation, days in season and all the weather or graphical related information like precipitation, wind speed, etc.

The two datasets available start from 2013 and 2014 separately. After doing some summary statistics as well as exploratory data analysis, I found that the majority of counties in 2013 are also in 2014 dataset and the timeframe covered for both datasets is from 11/30 through 06/03. Therefore, I decided to use 2013 dataset as the data for tuning the model and 2014 dataset as the test dataset to evaluate the model performance.

For this regression problem, I use both Gradient Boosting method and Random Forests method to tune the model based on 2013 data and test the model performance on 2014 data.

Summary of Data

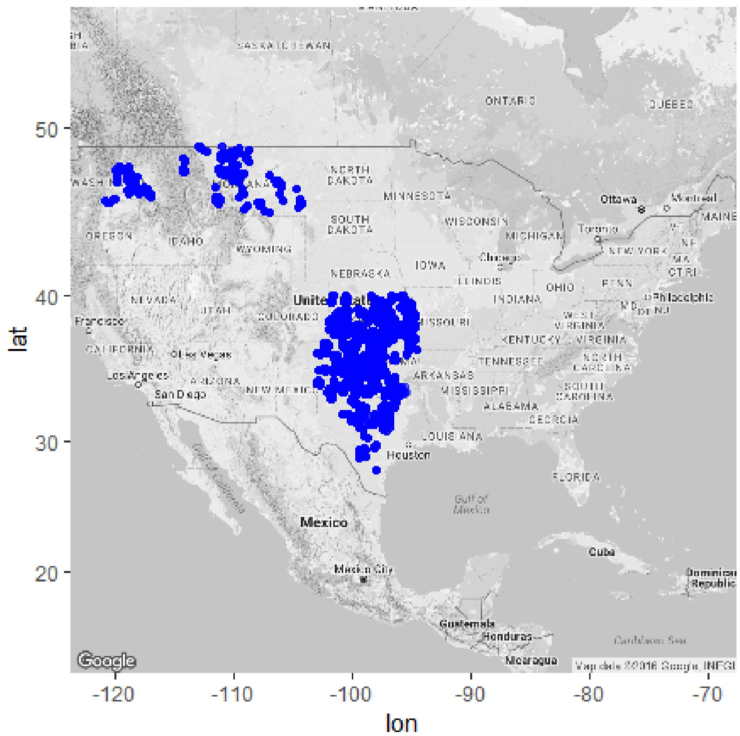
The following summary gives the summary information for all the variables in wheat13 dataset. Except for “precipTypeIsOther” variable, we don’t seem to see other obvious abnormal patterns. I will remove this variable as all the values are 0. We see there are some missing values for “pressure” and “visibility” variables, given that it takes a very small proportion of total observations, I decided to just remove those records with missing values. We could definitely impute the missing values with median or other methods as well, or we could simply leave those values as they are since the tree based methods could take care of missing values.

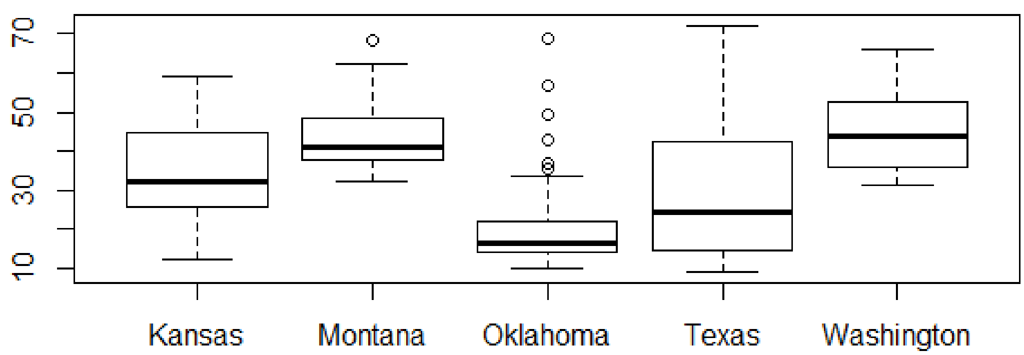
```
  CountyName           State              Latitude        Longitude
 Length:177493      Length:177493      Min.   :27.80   Min.   :-120.91
 Class :character   Class :character   1st Qu.:34.14   1st Qu.:-101.29
 Mode  :character   Mode  :character   Median :36.81   Median : -99.13
                                       Mean   :37.53   Mean   :-100.88
                                       3rd Qu.:38.95   3rd Qu.: -97.35
                                       Max.   :48.98   Max.   : -94.61

     Date           apparentTemperatureMax apparentTemperatureMin   cloudCover
 Length:177493      Min.   :-39.97         Min.   :-58.42         Min.   :0.00000
 Class :character   1st Qu.: 37.83         1st Qu.: 14.31         1st Qu.:0.00000
 Mode  :character   Median : 58.88         Median : 26.56         Median :0.01000
                    Mean   : 54.84         Mean   : 27.92         Mean   :0.07148
                    3rd Qu.: 73.10         3rd Qu.: 42.20         3rd Qu.:0.09000
                    Max.   :177.32         Max.   : 77.18         Max.   :1.00000
```

```
    dewPoint          humidity      precipIntensity     precipIntensityMax precipProbability
 Min.   :-36.09   Min.   :0.080   Min.   :0.000000   Min.   :0.00000    Min.   :0.0000   
 1st Qu.: 19.60   1st Qu.:0.470   1st Qu.:0.000000   1st Qu.:0.00000    1st Qu.:0.0000   
 Median : 27.85   Median :0.600   Median :0.000000   Median :0.00000    Median :0.0000   
 Mean   : 29.71   Mean   :0.594   Mean   :0.001158   Mean   :0.01063    Mean   :0.1335   
 3rd Qu.: 38.89   3rd Qu.:0.720   3rd Qu.:0.000200   3rd Qu.:0.00280    3rd Qu.:0.0900   
 Max.   : 75.18   Max.   :1.000   Max.   :0.152900   Max.   :2.05490    Max.   :0.9600   
                                  NA's   :1          NA's   :1          NA's   :1        
 precipAccumulation precipTypeIsRain precipTypeIsSnow  precipTypeIsOther    pressure     
 Min.   : 0.00000   Min.   :0.0000   Min.   :0.00000   Min.   :0         Min.   : 942.5  
 1st Qu.: 0.00000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0         1st Qu.:1011.2  
 Median : 0.00000   Median :0.0000   Median :0.00000   Median :0         Median :1016.7  
 Mean   : 0.05747   Mean   :0.2107   Mean   :0.09037   Mean   :0         Mean   :1017.1  
 3rd Qu.: 0.00000   3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0         3rd Qu.:1022.9  
 Max.   :19.48700   Max.   :1.0000   Max.   :1.00000   Max.   :0         Max.   :1048.1  
                                                                         NA's   :254     
 temperatureMax   temperatureMin     visibility      windBearing      windSpeed     
 Min.   :-22.00   Min.   :-39.79   Min.   : 0.600   Min.   :  0.0   Min.   : 0.040  
 1st Qu.: 43.35   1st Qu.: 23.42   1st Qu.: 9.180   1st Qu.:127.0   1st Qu.: 4.760  
 Median : 58.88   Median : 33.25   Median : 9.890   Median :192.0   Median : 7.670  
 Mean   : 57.55   Mean   : 34.39   Mean   : 9.286   Mean   :191.2   Mean   : 8.437  
 3rd Qu.: 73.10   3rd Qu.: 46.07   3rd Qu.:10.000   3rd Qu.:275.0   3rd Qu.:11.530  
 Max.   :105.20   Max.   : 77.18   Max.   :10.000   Max.   :359.0   Max.   :31.730  
                                   NA's   :30                                       
      NDVI         DayInSeason         Yield      
 Min.   :117.0   Min.   :  0.00   Min.   : 9.00  
 1st Qu.:137.9   1st Qu.: 46.00   1st Qu.:17.30  
 Median :147.2   Median : 93.00   Median :31.10  
 Mean   :146.3   Mean   : 92.63   Mean   :31.44  
 3rd Qu.:152.9   3rd Qu.:139.00   3rd Qu.:43.10  
 Max.   :206.0   Max.   :185.00   Max.   :72.20  
```

I've included the following density plot to show where the observations are located in the United States. There are five states covered in 2013 dataset.

For the five states included in the 2013 dataset, I have also included the following boxplot. We see that Washington has the highest overall yield while Oklahoma has the lowest overall yield.





Variables included in the model

For all the weather or graphical related variables, I decided to put all of them in the model as we don't have too many to start with and since the model form is tree based, we don't need to do much additional treatment for the purpose of prediction as the method could take care of the possible complicated interaction among different variables.

For the location information, I created a factor variable to indicate the combination of latitude and longitude information for all the records. Given that we may have too many levels, I rounded the latitude and longitude to integers. For example, if the latitude is 46.81169 and longitude is -118.6952, I rounded the numbers with lat = 47 and long = -119 and the level for this specific record would be 47:-119. There are 115 different levels for 2013 dataset.

One additional variable to handle is "date", I used the month of the date as a variable to incorporate the time information. Therefore, if the date is 12/01, the month will be December.

Modeling methodology

Given this regression problem, I decided to use both gradient boosting method and random forest method and compare their model performance.
Given the limited computing power of my laptop, and the 2013 dataset has around 180K records, I randomly selected 10% of the records to fit the model. The cross validation method has been used to find the optimal parameters.

Then two models are compared on the test dataset, which is 2014 dataset. The mean squared error is used to evaluate the model performance. Please note that the prediction is generated based on the location level, while the actual yield is the same for different locations in a same county. I have also calculated the prediction for different counties using the average of predictions of various locations in the same county and tested the model performance, the result is similar to the following result. We see that Gradient boosting model produces better prediction than the Random forest model. While the difference between the minimum actual yield and maximum actual yield is around 70, 9.4 is pretty good prediction error range.

Method	Gradient boosting model	Random Forest model
Sqrt of MSE	9.41	10.8

The variable importance result belows shows the top 20 most important variables by Gradient boosting model. We see that location information takes up the majority of most important variables, in addition, NDVI and humidity are also important predictors.

```
  only 20 most important variables shown (out of 147)

                    Overall
location.rd133:-96   100.00
location.rd148:-114   79.35
location.rd146:-118   60.64
location.rd131:-97    48.71
location.rd139:-95    47.81
NDVI                  46.14
```

```
location.rd138:-95    43.51
location.rd137:-98    42.19
location.rd135:-99    42.00
location.rd135:-100   36.62
location.rd134:-100   36.52
location.rd134:-96    34.53
location.rd132:-97    34.51
location.rd132:-99    34.11
location.rd146:-117   31.52
humidity              27.80
location.rd136:-98    27.12
location.rd134:-97    26.76
location.rd136:-99    26.41
location.rd137:-95    25.73
```

Further discussion

Based on the analysis above, we see that Gradient boosting model could provide a pretty good fit to the wheat yield prediction. The location variables and some graphical or weather related variables are important in predicting the wheat yield.

Given the limited computing power of my laptop, I selected 10% of data when fitting the model. Although I also tested 30% of data for fitting GBM model and it didn’t show much difference in prediction error, I would still like to include more data and more folds in cross validation methods in tuning the parameters if more resource and time is given.

Given more time, I would also try k nearest neighbor methods and some possible spatial methods. In addition, another possible way to approach this problem is as follows: since the wheat yield is given based on the county level, we could probably build a model on the county level. The corresponding features could include certain summary statistics like minimum, maximum, quantile, etc. of different variables on the county level.