Aerial Intelligence - Data Challenge

Thomas Hossler

1. Objectives

We want to predict the wheat yield in several counties in the US. The counties are located in the states of Washington, Oklahoma, Kansas, Texas and Montana. Two data set are available, one for 2013 and one for 2014. Each data set has over 170,000 observations and 26 variables. The variables are geographic (latitude, longitude) and climatic (temperature, pressure etc).

2. Exploratory Data Analysis

2.1 Distributions

In the first time, an individual analysis of both dataset has been performed. The variables *CountyName*, *State* and *Date* have been dropped (risk of overfitting) as well as the variable *precipTypeIsOther* (empty). The missing values have not been treated and simply removed due to their small amount. The distribution for each predictor and each year have been plotted (figure/distributions.pdf - red for 2013, blue for 2014). Moreover, due to the nature of the predictor, a particular attention has been given to multicollinearity (figures/corrplot.pdf). Indeed, we could expect some variables to be correlated (for example latitude and temperature).

The study of the distributions allows us to note the (dis)similarities between the two years. The following observations can be made:

- Most of the variables have similar distributions (shape, number of modes etc).
- For a few variables, the distributions are shifted from one year to another (difference in means)
- For the forecasting variable *Yield*, the distributions are quite different. The spread is smaller for 2014. Moreover, 2014 displays some high isolated values (above 70).

Some of the variables are heavily skewed positively (*cloudCover* for example) and negatively (*visibility*). These variables will not be transformed in a first model but will probably require additional transformations in a second time.

Some of the variables are binary (*precipTypeIsRain* for example), already encoded with 0’s and 1’s.

2.2 Correlation

Due to the nature of the predictors, we could expect some correlation between the predictors.

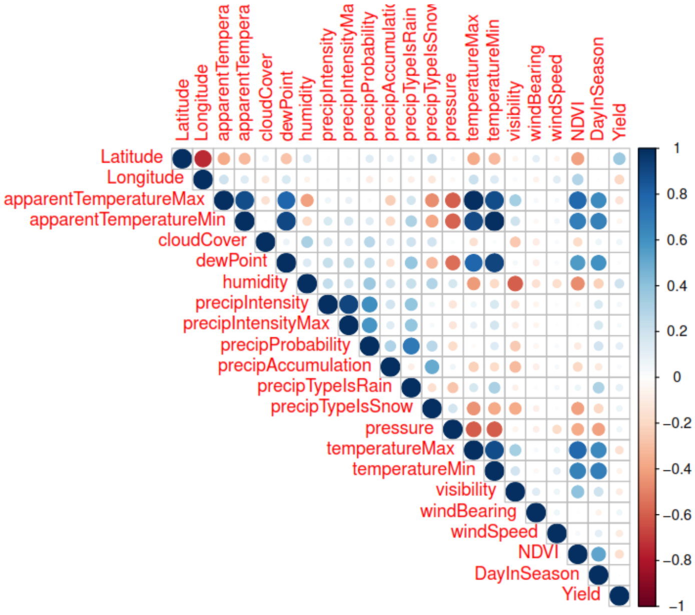

Figure 1: Correlation plot of the 22 predictors. The value of the correlation coefficient is indicated both by the color and the size of the circle.

Figure 1 shows a correlation plot of the predictors. As expected some predictors show a strong correlation (*dewPoint* and *apparentTemperatureMin* for example). This needs to be kept in mind for the forecasting part.

3. Forecast

After the EDA phase, we are now aware of the followings:

- The forecasting variable is continuous.
- Most of the predictors are also continuous. Some predictors are categorical (for example, *precipTypeIsRain*) but they are binary.
- Some of the predictors are correlated
- 2014 dataset might have some outliers

In the spirit of producing simple models that are easier to interpret and implement, two approaches have been chosen: multiple linear regression with regularization and principal component regression. The two approaches both have a way of reducing the number of predictors to facilitate the interpretation of the model. Cross-validation has been used to study the stability of the models. The 2013 data set will be used for training and validation whereas the 2014 dataset will be used for testing. The main code is located in the **code** folder.

4. Results

Cross-Validation has been performed using the k-fold method (10 folds) at two levels: to select the training and the validation sets and to pick the regularization parameter (number of principal components or lasso coefficient). The CV estimates on the 2013 data set is less for lasso than for PCR. The multiple linear regression model with lasso has therefore been picked. A quantile-quantile plot is presented Figure 2. The mean squared error for the forecast is 86.75.

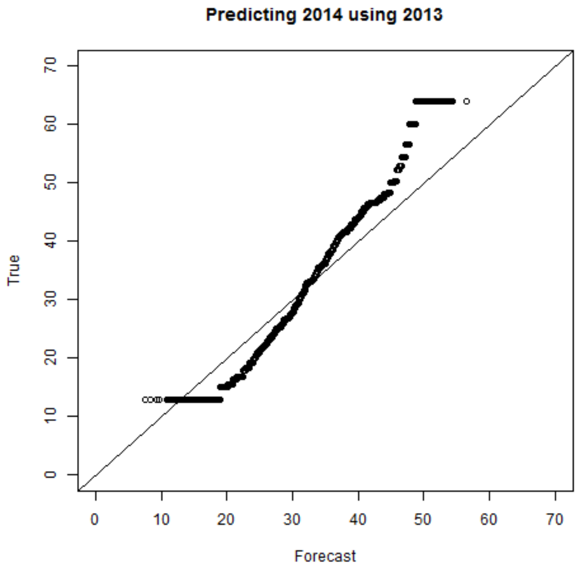

Figure 2: Quantile-Quantile plot of the true yield in 2014 versus the forecasted yield using a multiple linear regression model with lasso.

Looking at the lasso coefficients, it appears that no predictors coefficients has been shrunk to zero. The lowest coefficients are for *precipProbability* and *windBearing*. Outliers have been identified using a Tukey test and the high values of Yield for 2014 have been removed, which reduced the MSE.

5. Next steps

The following tasks will be performed to improve the model predictive power:

- Transformation of the skewed predictors
- Basis expansions