Iowa Corn Yield Prediction

Group: spv5

December 14, 2015

Data Preprocessing

Model Data

Candidate Predictors

- 9 District as categorical variables;
- Year as continuous;
- 10 year historical yield data;
- 1 year precipitation and temperature of first three PCs;
- 1 year SST of first three PCs.

Response

- Yield of this year.

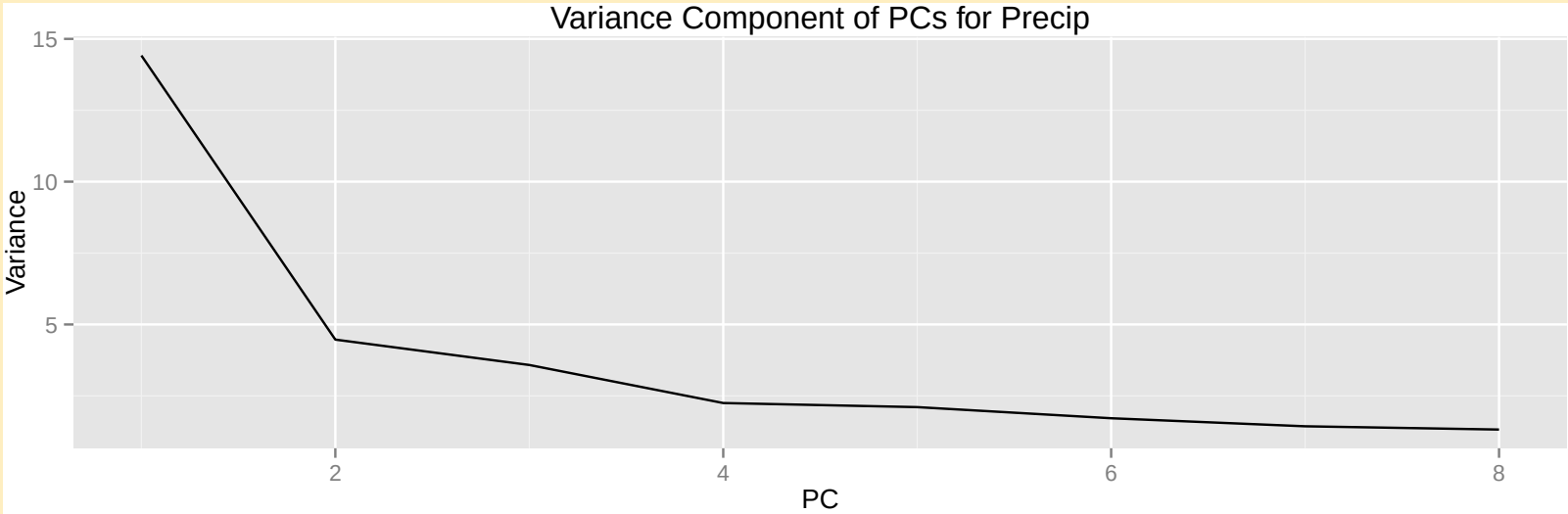
Model Data

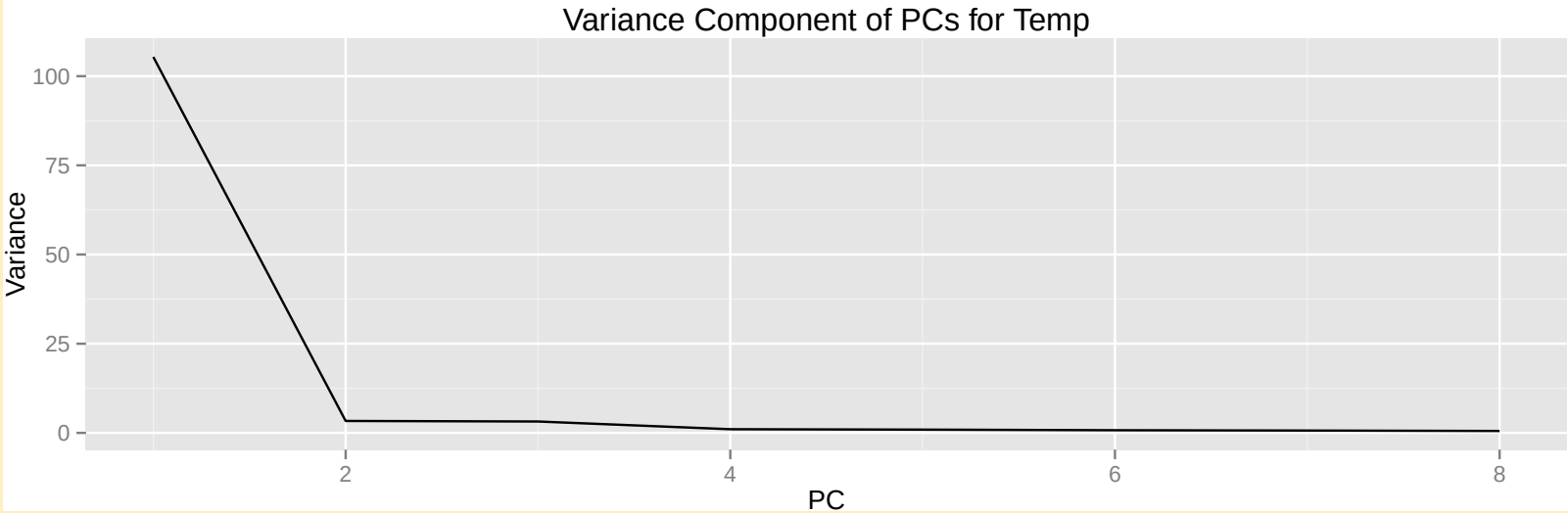
Regression or Classification

- Quantitative prediction model could provide more helpful information with the cost-profit decisions;
- The choice of the cut-off of classification is a controvercial issue.

Dimension Reduction

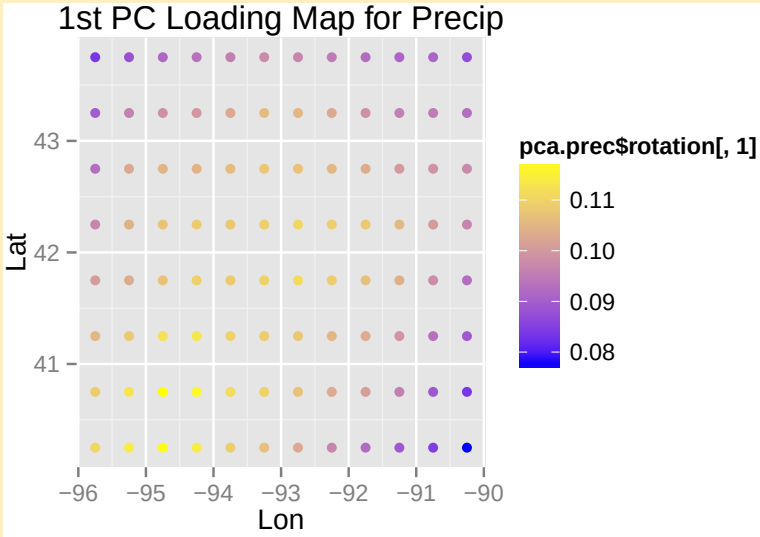
PCA

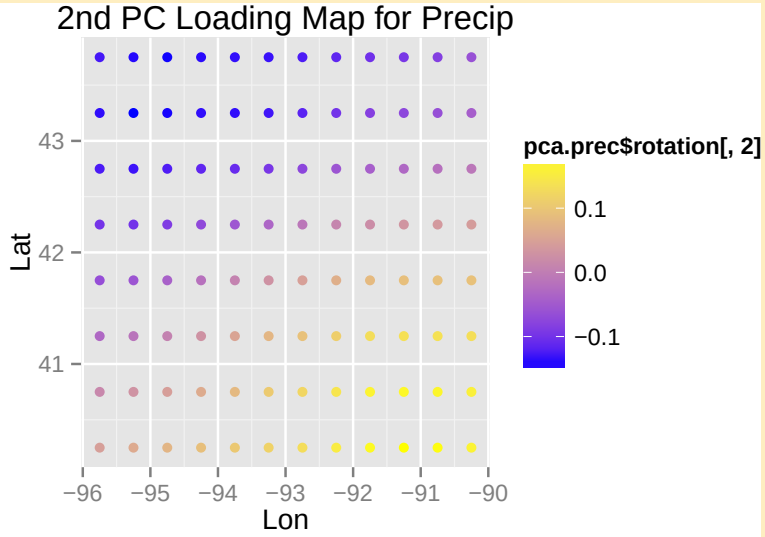


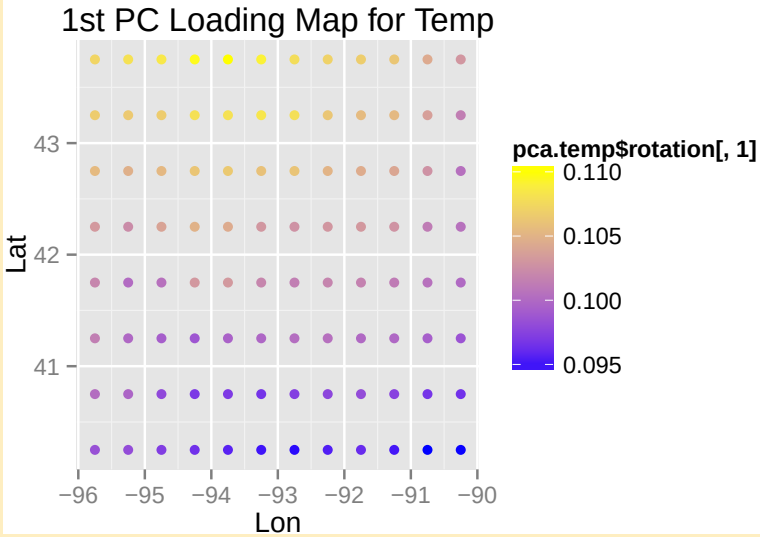


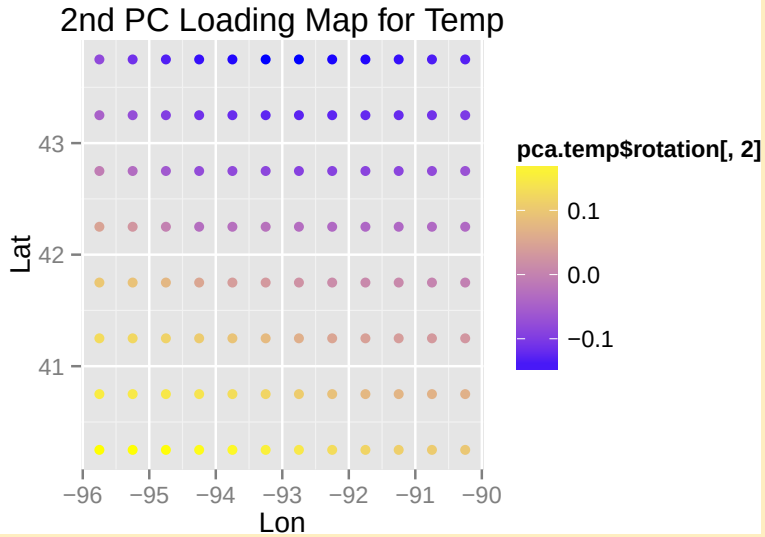
Dimension Reduction

PCA









Model Selection

Feature Selection

We group the features, then try different combinations of the groups to fit models using lasso and ridge regression. From the outputs of this two methods, we compare the MSEs and have following findings:

- SSTs do not help to reduce MSE;
- Temperature itself is not useful, but the interaction of temperature and precipitation is useful;
- When sample size is not large enough, too much noise will harm the predictive power.

Model Selection

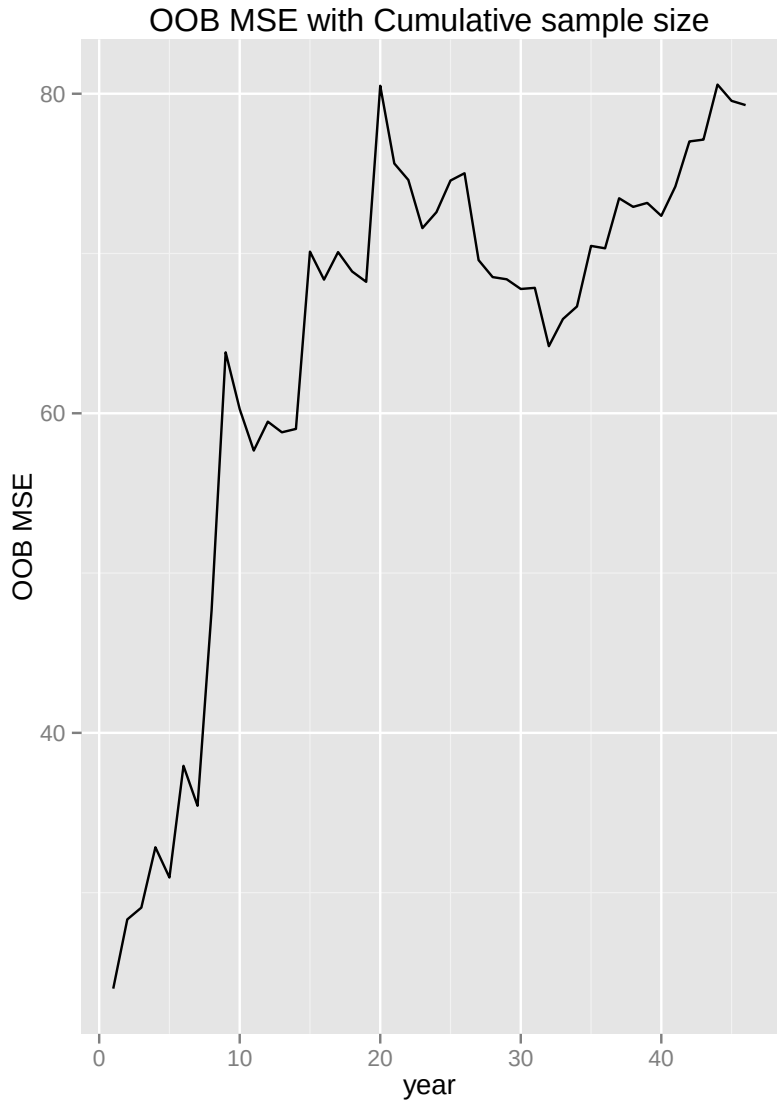
Candidate Model

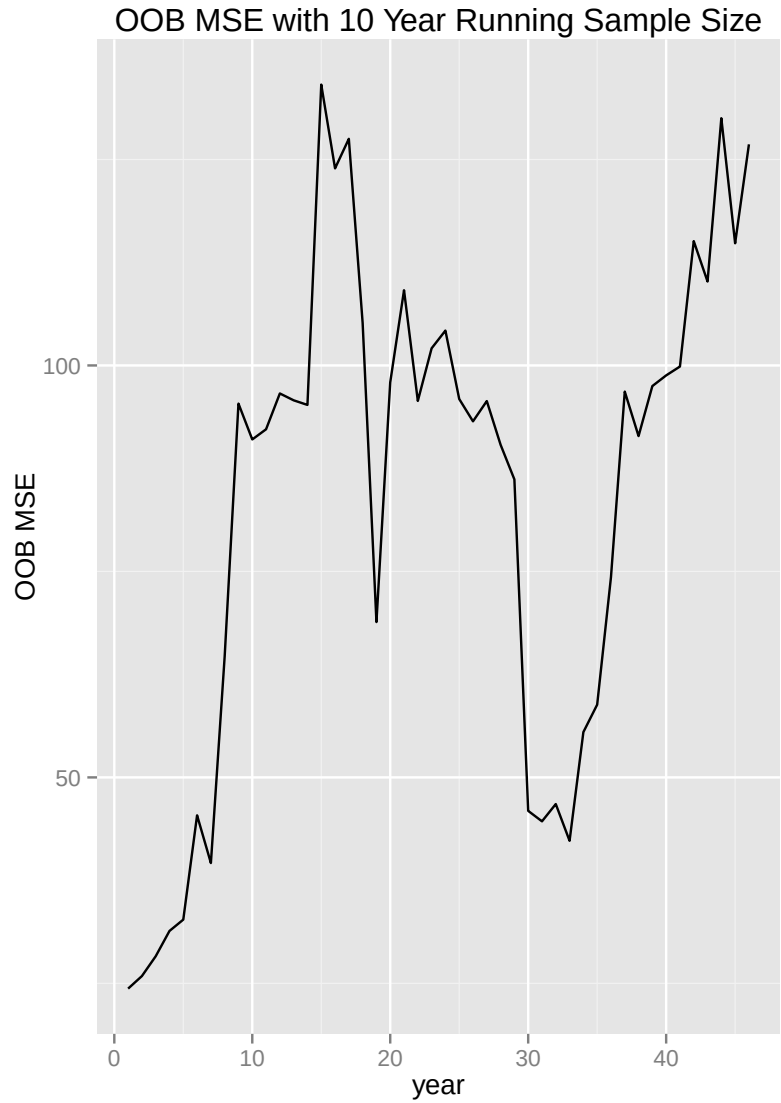
- RandomForest
- Deeplearning
- Boosting
- Lasso/Ridge

Training Sample Size

- Cumulative sample size (all historical data)
- 10 year running sample size (sample size is always 10)

Model Selection





Model Evaluation

MSE by h2o package

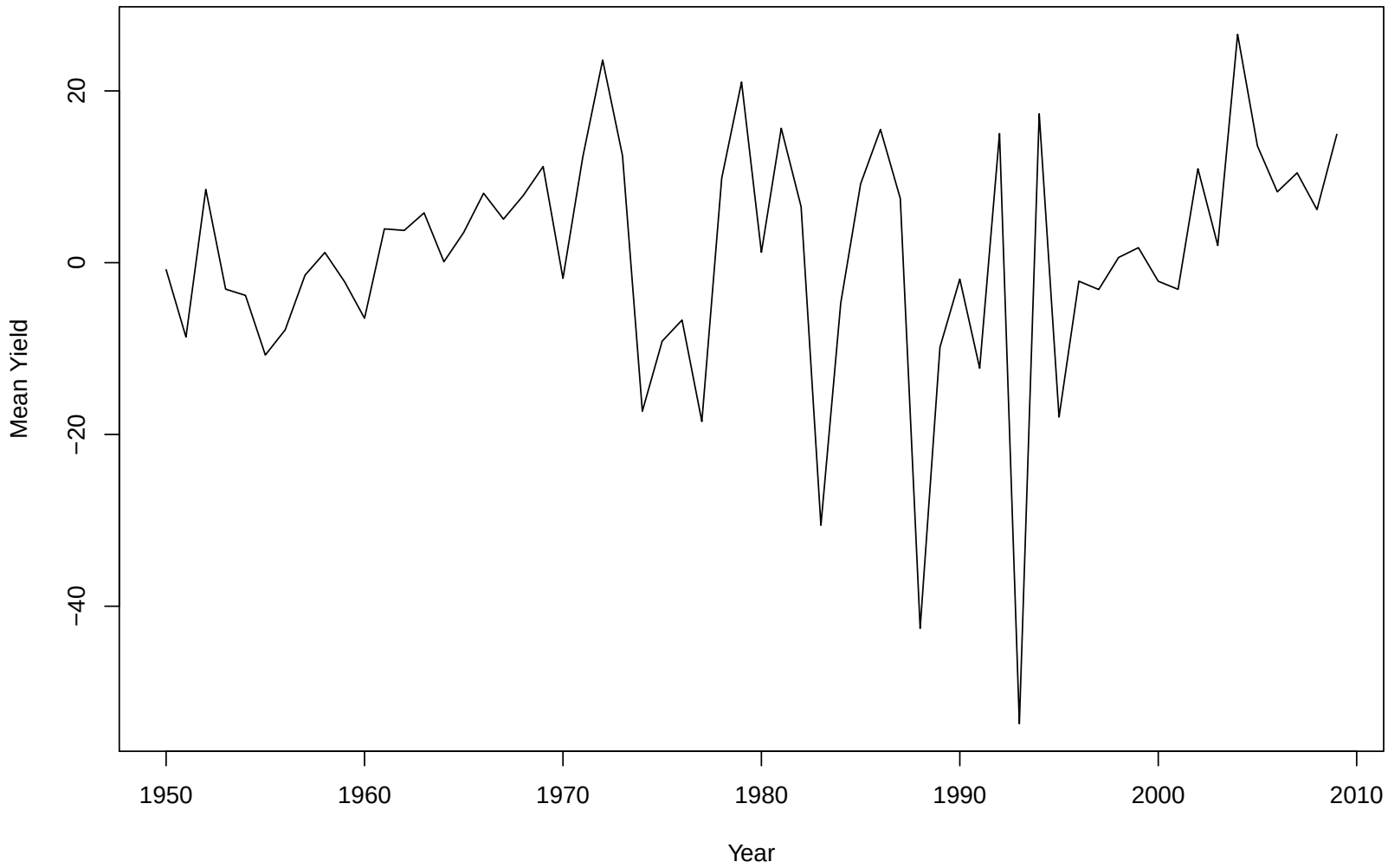
Baseline

- Average Estimate: MSE=364.0061
- Last Year Estimate: MSE=701.7978

Model

- RandomForest: MSE=393.2477(June), MSE=393.7953(March)
- Ridge: MSE=434.0248(June), MSE=413.9061(March)

Model



Conclusion

Conclusion

1. For small sample of observations, the noise may overwhelm the effect of useful variable, even for random forest, the variable selection procedure cannot extract the key information of data;
2. The result indicates that the interaction between the temperature and precipitation does have some effects on the corn yield. The extreme values of yield which have very poor prediction usually occur when the temperature and precipatation also become the local extreme value. But the relationship may not be predominant, so it cannot lead a low error rate;
3. To improve the predictive power of this model, we need more information which is beyond the scope of data we obtained.