Machine Learning Final Project

Bradley Bush, Sandy Tanwisuth, Jesse St. Amand

May 4, 2017

Introduction

Syngenta is one of the largest seed producers in the world. Currently they have a multi-year process by which they breed, test, and select varieties to sell comercially. They are interested in improving their selection process and have partnered with Idea Connection to host a data mining competition. The purpose of the competition (see:https://www.ideaconnection.com/Syngenta-AI-Challenge/) is to give teams a chance to use AI and data mining techniques (along with their creativity) to come up with ways to improve the selection process and allow Syngenta to 'grow more with less.'

We have decided to use the data from this competition for our project. We are attempting to predict soybean yield from environmental factors such as the amount of sun and precipitation the soybean plant gets during its growing season and from soil properties such as the amount of sand, silt, or organic matter present in the soil at the testing site. A full description of the variables used is given below.

The Data

Here is a description of the variables:

Variable Name:	Variable Description:
`YIELD`	Variety yield in bushels per acre (this is an average because you may have multiple experiments of the same variety at the same location– IE at different parts of the test site)
`AREA`	The number of acres growing soybeans in the segment of about 36 sq miles around the testing location. The information was obtained from http://nassgeodata.gmu.edu/CropScape (NASS USDA, 2014)
`IRRIGATION`	The number of acres with irrigation in the segment of about 36 sq miles around the testing location. The information was obtained from http://nassgeodata.gmu.edu/CropScape (NASS USDA, 2014)
`AWC_100CM`	The available soil water capacity (volumetric fraction) until wilting point. This information was summarized for the first 100 cm.
`TEMP`	Temperature accumulation during the season (in Celsius). Meaning the sum of average daily temperatures of test site from 1 Apr to 31 Oct.
`PREC`	Precipitation accumulation during the season (in millimeters). Meaning the sum of average daily precipitation of test site from 1 Apr to 31 Oct.
`RAD`	Solar radiation accumulation during the season (in watts/meter). Meaning the sum of average daily solar radiation of test site from 1 Apr to 31 Oct.
`SAND_TOP`	Percentages of sand particles according to size (small, less than 0.002 mm; medium, between 0.002 and 0.05 mm; and large, more than 0.05 mm, respectively). These proportions define soil texture. This information was summarized for the first 30 cm.
`SILT_TOP`	Percentages of silt particles according to size (small, less than 0.002 mm; medium, between 0.002 and 0.05 mm; and large, more than 0.05 mm, respectively). These proportions define soil texture. This information was summarized for the first 30 cm.
`CLAY`	Percentages of clay particles according to size (small, less than 0.002 mm; medium, between 0.002 and 0.05 mm; and large, more than 0.05 mm, respectively). These proportions define soil texture. This information was summarized for the first 30 cm.

Variable Name:	Variable Description:
PH	The log of H+ concentration in the soil. Acidic soils have low pH values and high H+ concentration and alkaline soils have high pH values. It impacts soil chemical reactions and the ability of the soil to supply nutrient to plants. Optimum pH range is between 6.5 and 7. This information was summarized for the first 30 cm of the soil profile.
ORGANIC.MATTER	The percentage of the soil consisting of plant and animal residues at various stages of decomposition, soil organisms and their byproducts. It impacts soil chemical reactions and soil structure. It is an important indicator of the ability of the soil to supply nutrient and water to crops. This information was summarized for the first 30 cm of the soil profile.
CEC	The Cation Exchange Capacity (cmol kg-1) quantifies the amount of negative charge in the soil. It impacts soil chemical reactions and the ability of the soil to supply nutrient to plants. It is often associated with clay and organic matter content. This information was summarized for the first 30 cm of the soil profile.

Processing Data

Much of our time was spent in this section of the project. The data is given to us in two different files. The first file had the soil and environmental data and the second file had the information about the soybean varities. We had to first combine the datasets by taking each observation (ie a given soybean variety for a given year) and appending the appropriate weather and soil data so that we could have all of the data in one dataframe. Once we combine the data we have 172,057 observations and 66 variables. Not all of these variables were useful (for example years of weather data for which we did not have soybean yields), so we had to create another dataset with only the variables we wanted to use. When running our models we split the data into three data sets: 50% training data, 25% validation data, and 25% test data.

Modeling the Data:

We tried 3 different methods from our Machine Learning class. We focused on Random Forests, Deep Neural Nets, and KNN. We mention Linear Regression (as a reference baseline) and Boosting (as a method to explore in the future). We tuned Random Forests, Neural Nets, and KNN, and discuss each method with its corresponding results below.

Baseline: Generalized Linear Models

To give us a baseline for OOB performance, we chose a linear regression model using the package `glm`. We use stepwise eliminination to find our model using the `step` function. This gave us our baseline RMSE of 13.0193 bushels per acre. Our baseline model is given below.

$$Yield = 32.61 + 0.00056 * area + 0.00074 * irrigation + 2.330 * organic.matter + 0.325 * silt.top - 0.449 * cec + \varepsilon$$

Random Forests

With Random Forests we tried to optimize using a spread of values for `mtry` and `ntree` as well as using an optimization function called `tuneRF` from the `randomForest` package, which optimizes over `mtry` in a more thorough fashion. Our results didn't vary much over `mtry` values and were fairly consistent over `ntree` values.

We have included a table of values along with a graph below. Our final Random Forests model choice had an OOB RMSE of 7.9042 bushels per acre with `mtry` set to 6 and `ntree` set to 1500.

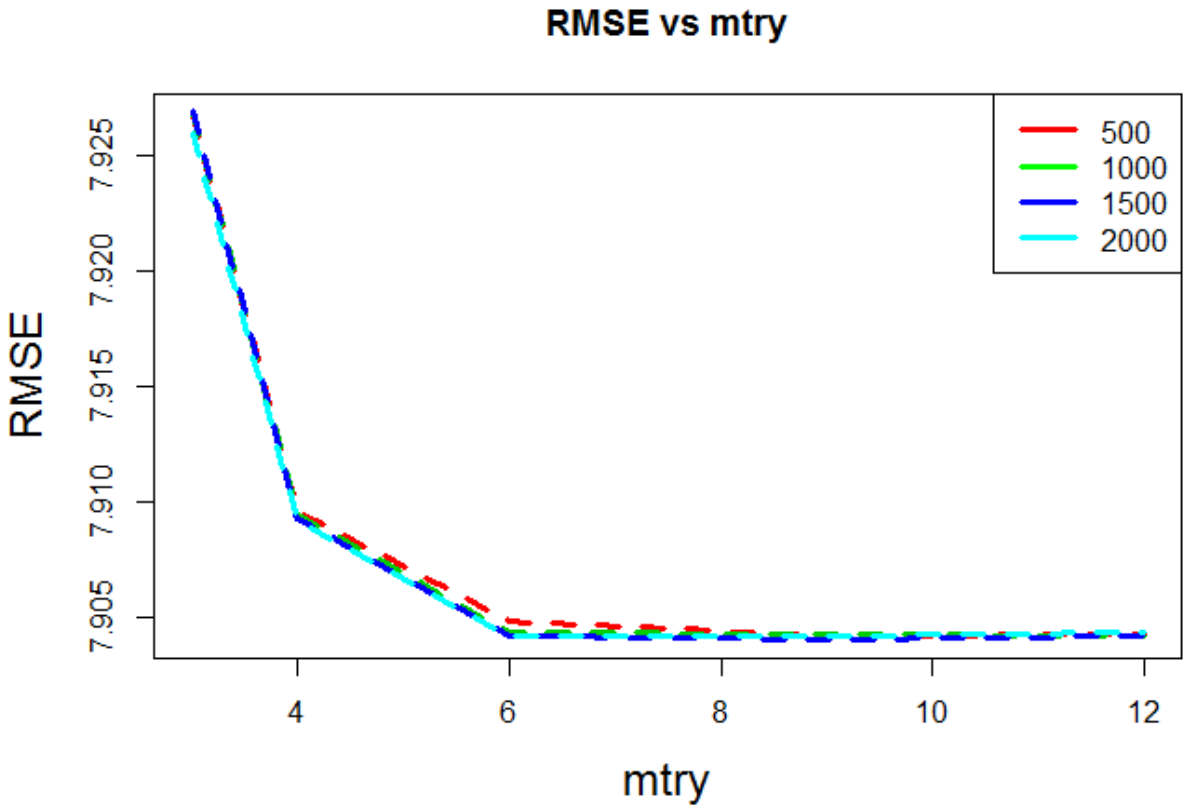

Figure 1: ntree values of 500, 1000, 1500, and 2000 are used with mtry values of 3, 4, 6, 9 and 12. The results do not vary much for any of the settings.

mtry:	ntree=500	ntree=1000	ntree=1500	ntree=2000
3	7.926718	7.926793	7.926908	7.925924
4	7.909572	7.909412	7.909228	7.909150
6	7.904852	7.904379	7.904189	7.904188
9	7.904154	7.904245	7.904059	7.904156
12	7.904267	7.904182	7.904173	7.904311

Table 1: List of values for `ntree` and `mtry`

KNN

We used the r package `kknn` for our knn models. Tuning these models was extremely time consuming (read: computationally expensive), but they ended up performing surprisingly well. Below is a table of the RMSE values for various k values.

k	RMSE
5	8.592175
8	8.341477
11	8.217457
14	8.172575
17	8.126609
20	8.102771
23	8.089667
26	8.076027
29	8.074170
32	8.067071
35	8.062278
38	8.055297
41	8.054643
44	8.055504
47	8.057311
50	8.056076
53	8.062764
56	8.067178
59	8.066341
62	8.069128
65	8.075273
68	8.076966

Figure 3. Table of k values and the RMSE of each model as depicted in figure 2. The best RMSE is ~8.055 found at a k of 41.

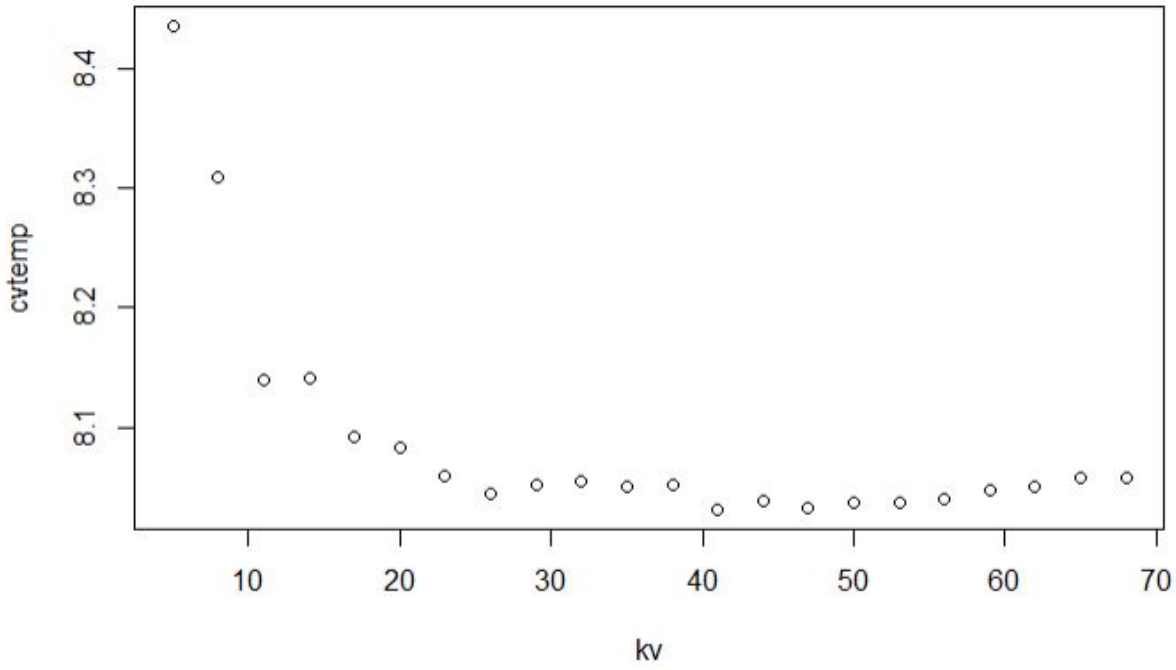

Figure 2: Preliminary figure of k verses cvmean, which compares the paramter k at values between 5 and 68 with the RMSE of a single model. This was used to test the approximate range of appropriate k values before moving onto the more computationally expensive model below.

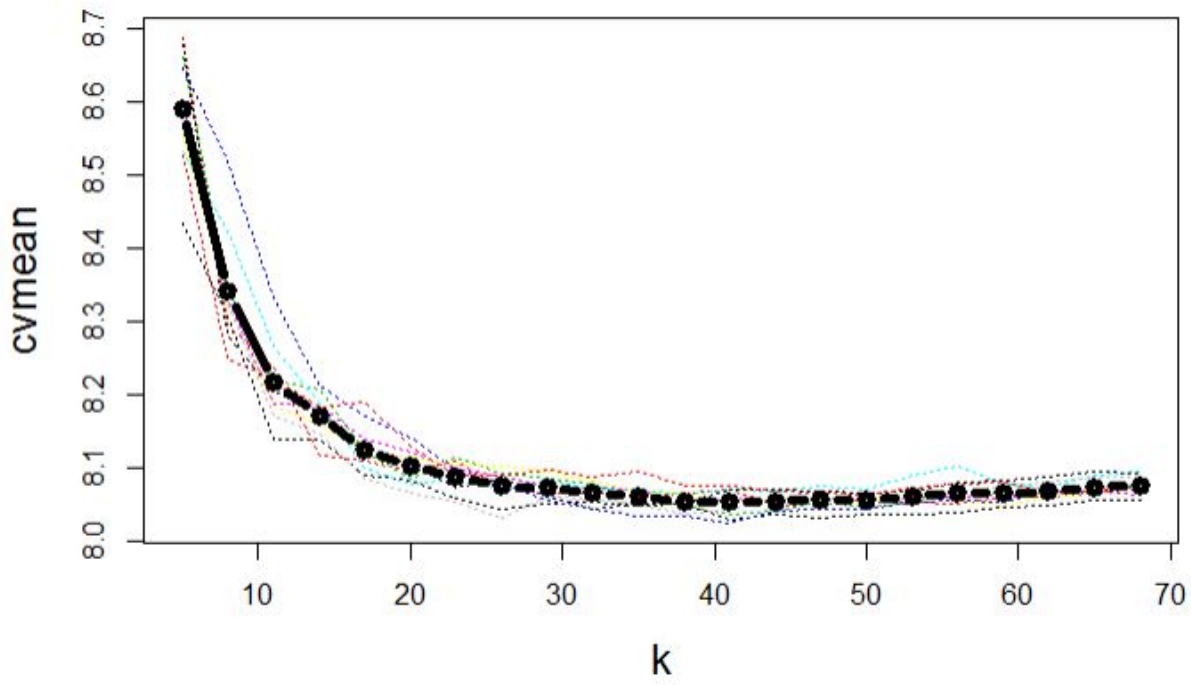

Figure 3: Plot of k verses cvmean, which compares the paramter k at values between 5 and 68 with the RMSE of each model (calculated by comparing against the testing data). Each colored line represents a different subset of the training data that, when averaged together to produce the black line, reduces bias and varience in the model.

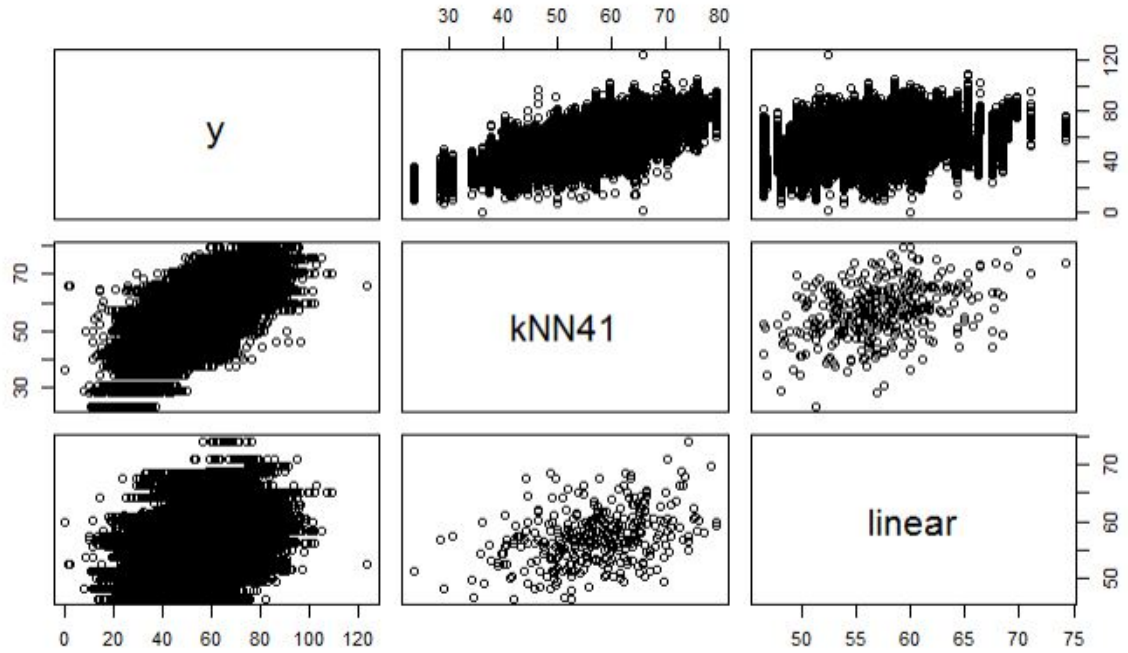

Figure 4: Pairwise comparison of kNN at a k of 41 against a linear model. The greater linearity in the y versus kNN41 plots over the y vs linear plots give a visual representation that the kNN model is a better predictor for the data than the linear model.

Neural Net

We used the `h2o` package for running our neural nets. It was very different from the other packages we have worked with (as mentioned in class) and took some time to get used to.

The following table was created by running iterations of the neural network model with different parameters. In attempts 1-10, the RMSE (number of bushels per acre) was calculated by comparing the training data with the validation set. The final attempt found the RMSE of the best conditions (9), applied to the second training set (the first training data plus the validation data) and the test data. The greatest differenses in RMSE occured with changes to the hidden layers, with the best values coming out of the 400x400x400 set. Although attempt number 10 used more layers, it appears to overfit the data and therefore produce a higher RMSE.

Attempt Number	Hidden Layers	Epochs	Activation	L1	RMSE
1	10	1000	RectifierWithDropout	1.00E-02	9.609
2	10x10	500	RectifierWithDropout	1.00E-03	10.185
3	150x150	300	RectifierWithDropout	1.00E-04	8.74
4	200x200	500	RectifierWithDropout	1.00E-04	8.567
5	100x100x100	1000	RectifierWithDropout	1.00E-04	8.545
6	250x250	1000	RectifierWithDropout	1.00E-05	8.437
7	400x400	1000	RectifierWithDropout	1.00E-05	8.41
8	50x50x50x50	1000	RectifierWithDropout	1.00E-04	10.06
9	400x400x400	1000	RectifierWithDropout	1.00E-05	8.07
10	500x500x500x500	1000	RectifierWithDropout	1.00E-05	8.14

Attempt Number	Hidden Layers	Epochs	Activation	L1	RMSE
final	400x400x400	1000	RectifierWithDropout	1.00E-05	8.115

Boosting: Future Work

We were curious how our 3 methods would compare to boosting with trees, so we used the `gbm` package with settings `distribution` set to gaussian, `interaction.depth` set to 4, `n.tree` set to 5000, and `shrinkage` set to 0.01. We were surprised to see that our Boosting model (with almost no tuning) was very competitive with an RMSE of 7.9159 bushels per acre. This is a model we would like to tune in future work as it shows promise.

Conclusion

Neural Nets, KNN, and Random Forests were all fairly close in performance, and all better than Linear Regression. Our best OOB performance was from our Random Forests model with an OOB RMSE of 7.9042 bushels per acre. Boosting with trees showed promise and should be considered in future work.