ORIE 4741: PROJECT MIDTERM REPORT

DAVID LEE (DYL44), CLARA ONG LISHAN (LO88), ZILONG WANG (ZW243)

NOTE: Our project topic is based on the INFORMS competition. Much of the information was only just released by the organizers. Therefore, most of our time was spent on trying to understand the data. There are some parts of the data which we still do not understand, and we are waiting for the organizer's clarification. In addition, we are not allowed to publicly display the .xlsx and .csv files so we apologise beforehand.

Abstract.

In order for a crop variety to become commercialised, it often has to undergo several rounds of strict testing and experimentation to ensure that it offers good yield. Throughout these multiple phases of testing, crop varieties are often benchmarked against their "peers" and successfully commercialised varieties that came before by their yield. Every year, varieties that fail to make the cut are discontinued, while those that survive compete again in the following year with newly introduced varieties.

The primary goal of the INFORMS competition is that given past data, we should develop a prediction model that can be used to predict the volume of potential sales for each soybean variety in the class of 2014, and based on those predictions, determine if we should commercialise them or not. This report describes our preliminary data analyses conducted, interpretation of results, and our future plan of attack.

Motivation.

Despite numerous tests that suggest that a commercialised variety should have performed well, the reality is that some underperforming varieties actually slipped past the checks (Type I Errors). It thus behooves us to find a new learning model that can predict sales with a greater degree of accuracy and thus minimise our number of misclassified (mistakenly commercialised) varieties.

Understanding the Variables.

Before using any sophisticated procedures, we first performed exploratory data analysis to get acquainted with the dataset. Our data consists of 11 predictors, and we provide a description of each predictor variable below.

(1) **Year** (Integer): When the experiment was conducted.
(2) **Experiment** (Categorical): Consists of experimental varieties of relative similar maturity that are tested together.
(3) **Location** (Categorical): Where the experiment was conducted.
(4) **Variety** (Categorical): Groups of soybeans that are genetically identical.
(5) **Family** (Categorical): Sharing the same family means the varieties have the same parents.
(6) **Check** (Boolean): Whether the commercial soybean varieties are used as performance benchmarks in yield trials.
(7) **RM** (Float): Soybean relative maturity. Every 0.1 stands for 1 day.
(8) **REPNO** (Categorical): Replication number. A variety under a specific experiment and location is tested more than once.
(9) **Yield** (Float): This refers to the amount of grain per unit of land that a soybean variety produces.

Date: October 27, 2016.

(10) **Class** (Integer): The batch the soybean variety belongs to. Takes on the value -1 if it is not part of a class.
(11) **Grad** (Categorical): Whether the soybean variety graduated from the last round of yield test and proceeded to be commercialised. Takes on the value -1 if it is not part of a class, 0 if it is part of a class but did not graduate, 1 if it is part of a class and graduated.

Our response variable is **Bags Sold** (Float). This is the number of bags of seed sold in the second year after commercialisation. High relative sales volume in the second year of sales is associated with the superiority of a variety relative to other choices in the marketplace. Bags Sold takes on the value -1 if the variety is not part of a class.

Panel Data.

The type of data given to us is called panel data, which refers to multi-dimensional data involving multiple measurements over multiple time periods.

(1) In an experiment, multiple varieties (the same varieties) are tested over multiple locations. Within each location, some varieties are benchmarks.
(2) Within the same experiment and location, the test for a variety is replicated. Replication is necessary because the yield can vary even within the same field due to different soil conditions for instance.
(3) One row of the data represents a particular replication of a specific experiment, location and variety.
(4) For each variety, the $Family$, RM, $Grad$ and $BagsSold$ are fixed. $Yield$ varies.

Splitting the Data.

We split our dataset into a training set, test set and prediction set. The prediction set contains all the unlabelled observations from the Class of 2014.

The training set and test are split using 5-fold cross validation for time series, based on the year of the experiments. This is slightly different from the usual cross validation for non-time series data.

Fold	Training Set	Test Set
Fold 1	2009	2010
Fold 2	2009, 2010	2011
Fold 3	2009, 2010, 2011	2012
Fold 4	2009, 2010, 2011, 2012	2013
Fold 5	2009, 2010, 2011, 2012, 2013	2014

Understanding the Data.

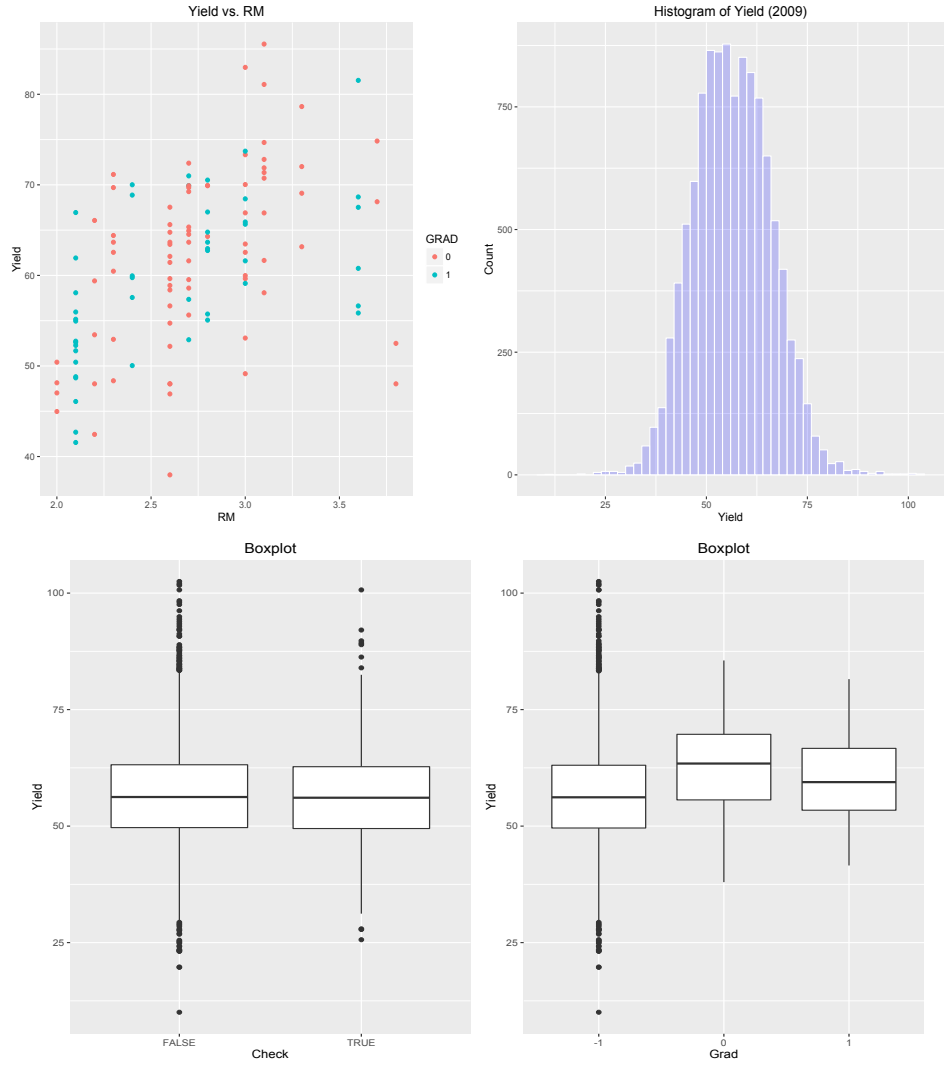
From our own Excel analysis and the clarification provided by the organizers, we understood the following:

(1) There are more locations per experiment as the stage of evaluation increases. For instance, there can be as few as 3 in the first year and as many as 26 in the last year.
(2) Every experiment has at least one benchmark.
(3) If $Grad = 0$, $\iff$ $BagsSold = 0$.
(4) If $Grad = 1$, $\iff$ $BagsSold > 0$.
(5) For varieties which are not part of a class, $Class = -1$, $Grad = -1$ and $BagsSold = -1$.
(6) Using set intersection, we found out that the varieties do not overlap across the classes $Class = -1$, 2011, 2012, 2013, 2014.
(7) It is possible that some varieties belonging to a certain class skip some year(s) of testing.
(8) Benchmarks need not be part of a class, and benchmarks may or may not graduate.

(9) Some benchmarks do not have bags sold, probably because the benchmarks are from other seed companies and sales data is not available.

Data Visualization.

All visualizations were done for experiments in 2009 only. We plotted $Yield$ against RM and saw that there was a positive trend in the data. From this scatter plot, we saw no observable difference between varieties that graduated and those that did not. We also plotted a histogram of $Yield$, and it appears that there are no outliers. From the boxplots, the yield for benchmarks and non-benchmarks are similar, and graduates have a lower yield than non-graduates.



Choice of Language and Possible Methodology.

We will be conducting our analysis mainly in R, though we may use some other scripting languages, such as Python and Julia, to automate the I/O and data cleaning beforehand.

Even though we are to predict the sales volume, we can still do classification beforehand to do a sanity check. This means that we will most likely be using SVMs for classification, as they are a very robust classification methods.

As for now, we are still unsure about what feature transformations or kernels we will be choosing for fitting the data to a linear model, but further investigations later on should give us a better understanding on this issue.