# Work until now

As a reminder, the main goal is to predict the variable BAGSOLD, for the current 2014 candidate seeds.

Up to now we have been working on two main streams: **Feature extraction** and **Model Building**

## Feature Extraction

We want to discover what features help predict BAGSOLD. As a starting point, we decided to focus on the data available for **34 varieties** that have a value for BAGSOLD.

Based on this, we created a dataset with the following fields:

|  |  |
| --- | --- |
| Field | Description |
| X1 | Unique ID for the row. |
| VARIETY | Name of variety (only the ones that have a value for BAGSOLD) |
| TEST\_ID | ID for the test that VARIETY was part of. |
| CHECK | Name of Check Variety we want to compare against. Same test ID. |
| RM | Relative Maturity for VARIETY (always constant) |
| VAR\_YIELD | Yield for VARIETY in that TEST\_ID |
| CHECK\_YIELD | Yield for CHECK in TEST\_ID |
| RATIO | VAR\_YIELD / CHECK\_YIELD |
| BAGSOLD | Units sold by VARIETY |

A density plot of this table is attached in the end of the document.

## Model Building

We tried different models to predict BAGSOLD.

The general model we tried is

We have tried different models:

<TABLE WITH RESULTS>

# Discussion and Future Steps

We clearly need to create a model with greater predictive power.

* What other models could work?
* What other features could be used?
* Are there any other approaches we could take?
* How to create more data points? (sampling?)
* What kind of unsupervised learning approaches we could use (e.g. family)
* How can we partition/transform the data to identify patterns in the subgroups

