Proof Of Concept

Keith Howell (X00072267)

Contents

Project Summary

Background

Data mining or Machine Learning has become increasing popular in recent years. While not a new concept it has been around since the 1950's. When Turing, discussed wheatear or not a machine could think. It is only in more recent years that we have been able to put these theories and other into practice. With the cost of hardware going down enabling amateurs and enthusiasts the change to try out machine learning of there own. It gives greater reach to an exciting new technology. Machine Learning has been applied to almost every industry from spam filtering to shopping trends, fraud detection to Diagnostic assistance. These have all been done and to great success. I am interested in when it is applied to farming how can crops be growing better with less waste and bigger yields. While there are big traditional companies working on machine learning program that will help farmers get the most from there efforts. These model are closed, expensive and often require the use of proprietary hardware or machine to even be able to use the service. This is an excellent option for big industrial farms that are already using the cutting edge machineary. It doesn't help small family farms stay in business. With rising prices in animal feed and stores giving less then what it costs to make produce to farmers any small improvement could help.

Objectives

What this project will aim to do is show how machine learning can be applied to datasets that are available to everyone and show the benefit of using them to the average person. There shouldn't be a need to be a data expert to get the most from data available. To get use from machine learning project data from ever angle shouldn't be necessary. Using the data that the average farmer would already have access to should be enough to show the benefits that are available. The end goal of the project is to show the relationships in a datasets to maximise the yields in a given year.

Prototyping & Testing

The datasets I will be using will be mainly have total values. As such I will first perform regression analysis on the data. The goal with regression analysis is to get the lowest error while seeing the relationships that attributes have on the crops yields for each year. Creating a decision tree will also show how the data is related and how it is effected by each attribute. Using the total values may make the tree very busy and not very useful for this reason I will do some classification analysis on the data. To do this I will do some discretization on

the data to create bins for each of the totals. This should make it easier to see what it effecting yields.

Risk Assessment

As with any data mining of machine learning project, the major risk involved is getting good data. There a many sources for free to use and open data however finding the correct data that could be used within this project will be more difficult. I will widen the range of data rather then focusing on and area or crop. I will use a dataset from different countries and many crops.

The project will mostly be done using weka. Weka is a open source data mining and machine learning tool. It was developed by the University of Waikato in New Zealand for users to get to grips with data mining. While there are many options available to use to create a machine learning model. R and Python are both popular languages which are used. I am not a expert with either of these so I fear it would take me to long to get up to speed with these to create a model that will be useful.

Project Methodology

Introduction

The objective of the project is to see the relationships that effect the yield of a given crop. With the hope that we can change the attributes to improved the yield so it doesn't effect the overall value of the yield. As the first results and a tree are being developed they will be analyzed to see what looks to be causing an effect. If there is something which looks to be causing a effect I will then try to back it up by seeing if there was something that changed around that time period or if there was a change in farming practices then. As the project continues it may become clear that there is a need for more data or data with more attributes which are not in the current dataset. These will be added as necessary and where the data is available.

Project Approach

User Needs

The end users shouldn't be able to use the model without any special equipment. The data used should be data that they would have normally. So while soil data would be very useful and help build a better model. It may not be something that everyone has access to.

Needs and solutions

The end user needs to see a benefit to machine learning for them and there problems. For them to want to continue to use it they will need to convinced that keeping a good account of there data however useless it may seem could be of great benefit for future seasons. The model should show the end user what they can expect from a season assume the weather data is correct. They should be to test it against different crops to see which will give the best yield for a given season. Ideally they could see if adding more or less of a attribute would be of beneficial or if it would be wasted according to other seasons.

Design solution

First results of models for both regression and classification models.

Attributes within data

- year
- country
- area harvested
- yield
- production
- seed
- rainfall
- temp

Regression

In figure one you can see the example of a regression model against the dataset. The tree is has lots of nodes and is very busy its not clear what the relationships are in the dataset.

below is the output of the result you can see that both the mean and root error are quiet high.

Correlation coefficient | 0.8832 |
Mean absolute error | 7606.1471 |
Root mean squared error | 11105.687 |
Relative absolute error | 35.0557 % |
Root relative squared error | 47.2769 % |
Total Number of Instances | 17 |

Classification

Figure two is gives a much simpler example of a tree. As you can see this was done using 3 bins.

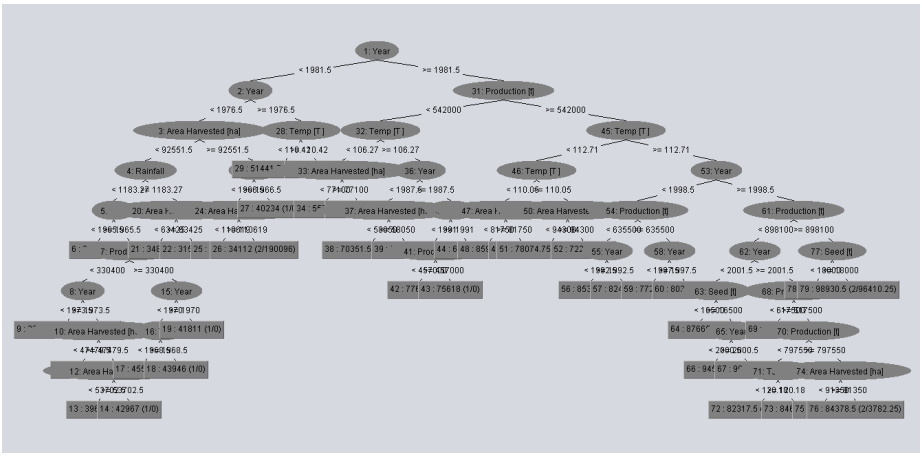

Figure 1: Random Tree

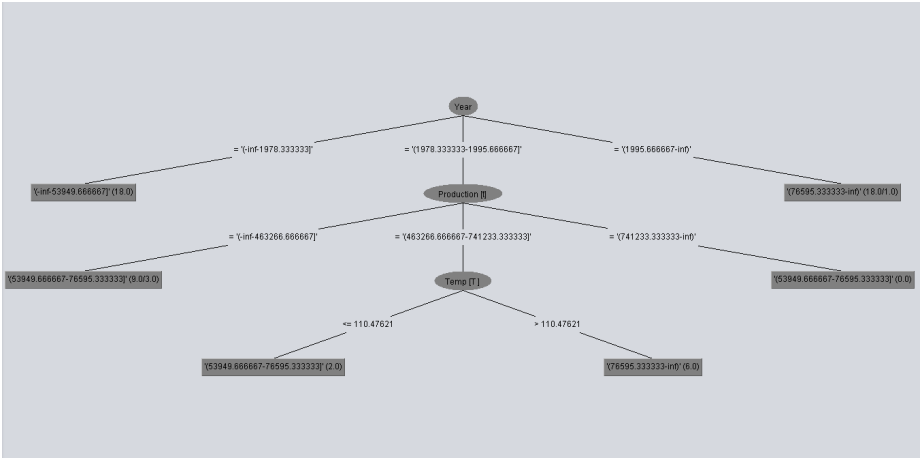

Figure 2: J48

Below the output table from running the model. 88% Correct is OK but I would like to see this above 90%. Also as its only classification with 3 bins its not good to be as useful. It will really only give you and idea if its a good, normal or bad year.

Correctly Classified Instances | 15 | 88.2353 % |
Incorrectly Classified Instances | 2 | 11.7647 % |
Kappa statistic | 0.8132 |
Mean absolute error | 0.0915 |
Root mean squared error | 0.265 |
Relative absolute error | 21.7703 % |
Root relative squared error | 57.6042 % |
Total Number of Instances | 17

Modify

In the above section I show what machine learning can do with the limited data gathered so far. The Classification model using J48 is much clearer and allows the user to see what is having more of an effect on the yields easier then with the regression model. Added more bins to the Classification model will allow the model to be more useful to the end user showing what makes a good yield into a great yield or a bad yield into a good yield. These bins need to be defined so its clear what a 'good' yield is. What was a good season one year may not be the next year. The data itself will also have to changed as the model is being built to ensure that one attributes is not outweighing the others. As the model is being built more attributes may be added as it is reviewed as from viewing it may become clear there something happened in the time period to effect the yield and it needs to be added.

Report

It the models above its clear that the year is the biggest factor in what was a successful year of not. In the regression model it shows 1982 as the year that had a big effect on the yields. It may be worth investigating what the what happened around this time to see if there is something that is missing. From the Classification model you can see that it uses the year to split the data. Its interesting that after 1995 the yields are always high and there wouldn't have be a 'bad' year. This could be down to GMO seeds but this would need to validated first. I would be interested in comparing this to another country or with different crop grown in Ireland.

Project Plan

Project Phases

The project will go through several different phases which will be repeated and re-worked as the model grows.

Data Preparation

During this phase the data will be loaded and prepare for a machine learning model. Making sure that there are no missing values and if there are deciding how to deal with the missing values. Any data that isn't useful or could cause issues will be removed such as unique id's. These server no purpose in a machine learning model and can cause problems with the model. Switching from a regression to a classification model the data will also need to be prepared so it can be used.

Choosing Model

I will have to spend more time researching the different models currently available in weka and create models in the them to build. This will involve a lot of trail and error. At the moment J48 has given the best results however this may change as the project continues

Training

In the training section there are two main ways to test. Either cross validation or using a training set. Within weka there is a ability to decide which to use at the time of testing. It won't be necessary to spent a lot of time deciding which to use as the model can be tested using both methods.

Evaluation

This will be a are area in which I can easily make a mistake. As I'm not a domain expert looking at something that appears to be interesting within the model may not be relevant to the users. I can compare how the model does against the yields that the model was tested against to ensure that the relationship it does find are valid.

Parameter Tuning

Once a model has been chosen and is giving result that are acceptable I will tuning it to better handle the data. This might involve changing the number of bins used in the classation model. This will probably be the most time consuming part as it will be mainly trail and error and rerunning the test again.

Predication

This will be where we get to see if the model will can answer the question of what increases the crops yield for a given season.

Deliverables and Milestones

I hope to deliver a model that can accurately show what effects yields in a given season. I hope to show this over several crop types and in different regions. First I will build a model on for Ireland and review the results.

Conclusion

Being able to see what is causes a yields to go up or down during the year would have great value to a farmer or other user. As knowing what happened in a previous year would help avoid the same mistakes in the next year. Using the data gathered it should be able to show what is having the biggest effect or point to the what was a large factor in the data points to. A farmer will often talk about how 'its a bad year' for a certain crop it would be good to be able to define this project so we can say what was the cause for it and what could we do next time to try and offset its effect. It may not be possible to find out the exact deals with the data. Some data we would need to work that out would require senors which the average user wouldn't have access to.