**Target attributes analysis of white wines**

**Methods used:**

1. Naïve Bayes
2. Generalised Linear model
3. Decision Tree
4. Random Forest
5. Gradient Boosted Trees
6. Target variable – level of preservatives

In this particular exercise we are doing analysis to figure out how changing the average of one of measurement affects another.

This attributes has been created for this purposes – Level of preservatives in white wine. It is categorical attribute.

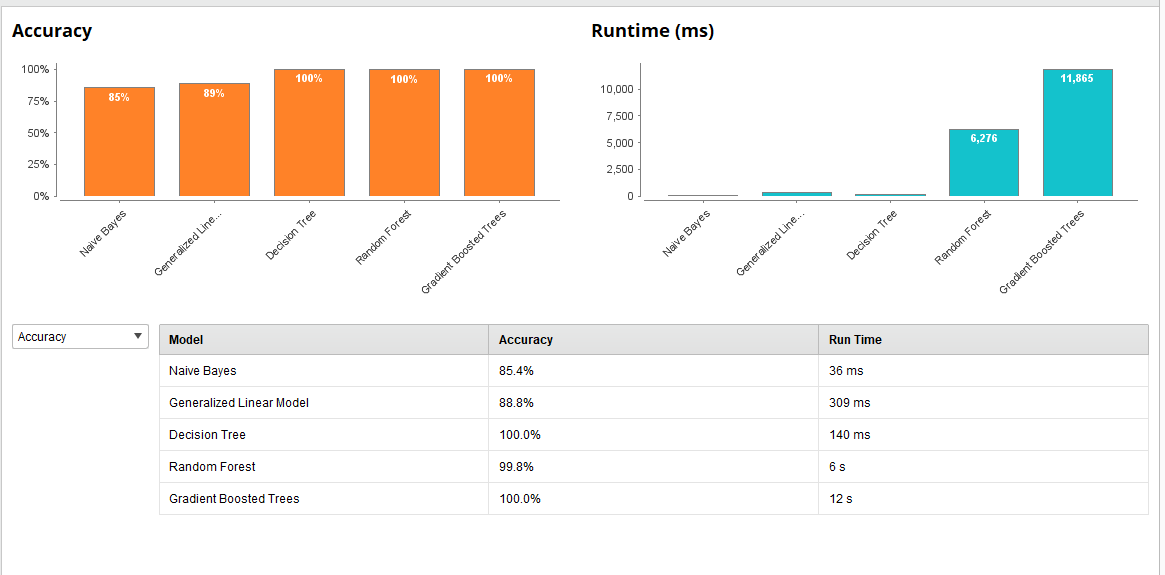
Rule : =IF(AND(J44<=0.45,E44<=0.045),"Low",IF(AND(J44<=0.6,E44<=0.06),"Medium","High"))

There is 3 kind of output Low, Medium and High. Ordinal Polynomial attributes.

The aim of this analysis is to check what will happen with one variable if we will change or delete the other.

Training set was extracted at 70% of dataset and 30% was a test set.

Training:

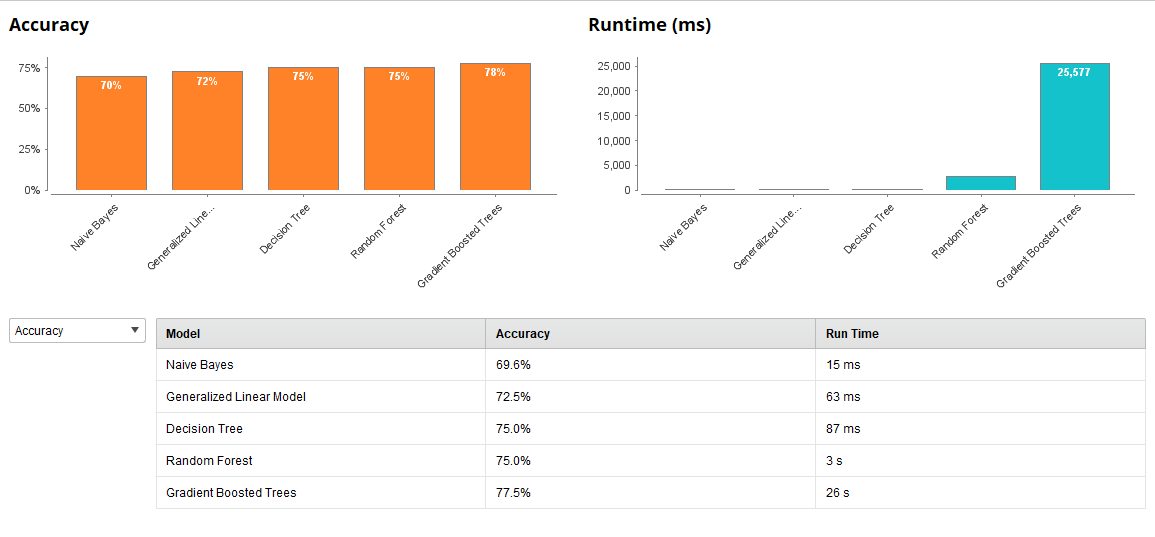


Summary of model outputs with given attribute.

In test set we have delete chloride attribute and then try to predict the outcome of target variables based only on other attributes.

Cross – validation:

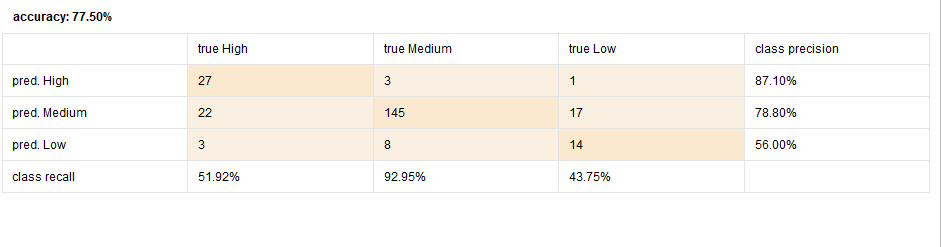
Test:



Summary of model outputs with deleted attribute - chloride.

As we can see the Gradent Boosted Trees method had the highest accuracy at 77.5%. But it was the slow as it takes 26 seconds compared to others.

That’s mean that even without chloride attribute the other attributes of white wines can give us information about level of preservatives at 77.5% chances that we are correct.



Let’s see the outcome of Gradent Boosted Trees method.

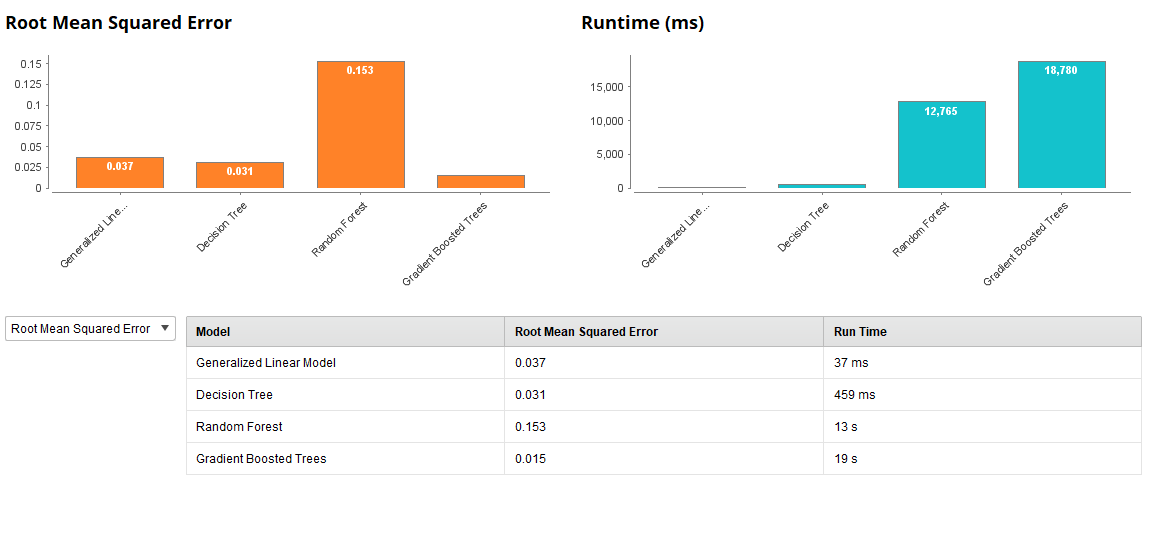
We can see that model had no problem with medium level of preservatives but had problem with the High and Low level.

1. Target attribute Sugar/PH ratio.

This ratio was picked not without discussion and further research. It is proven that this ratio should be the lowest to make sure that wine will be tasty.

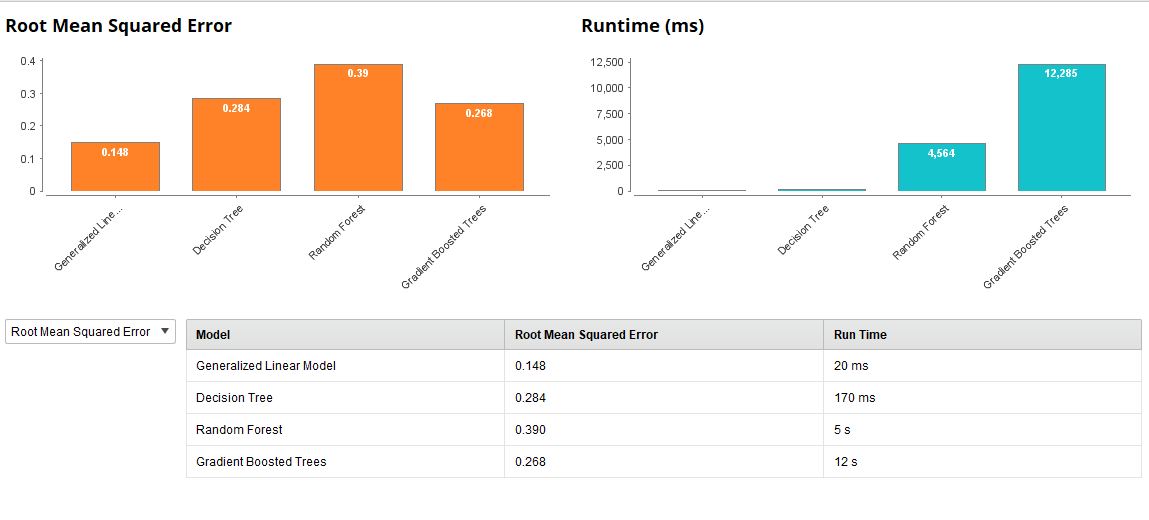
Are we able to prove it with our sample of wines?

Methodology used as in previous example: training 70%, test 30% and for test we have deleted PH attribute.

Training:

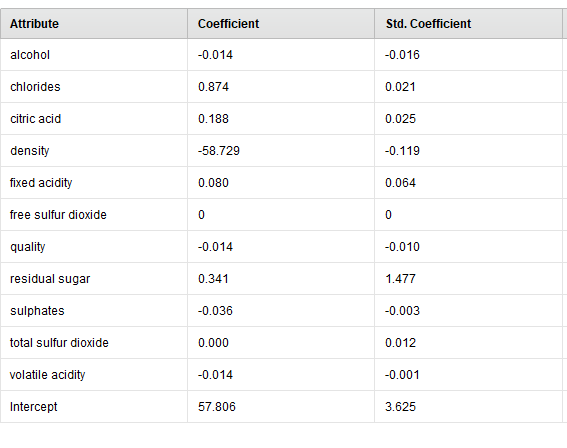
Summary of model outputs with given attribute. – root mean squared error

Cross -validation:

Test:

Summary of model outputs with deleted attribute – Ph.

In this case the best was simple Generalised Linear Model with the smallest error. Output below:



That’s is very interesting finding. We can see here that even without given PH of wine we are able to predict sugar/Ph ratio in almost 90% accuracy.

The biggest impact on that ratio will have density -58.72. But few others are also very important like chlorides or residual sugar.

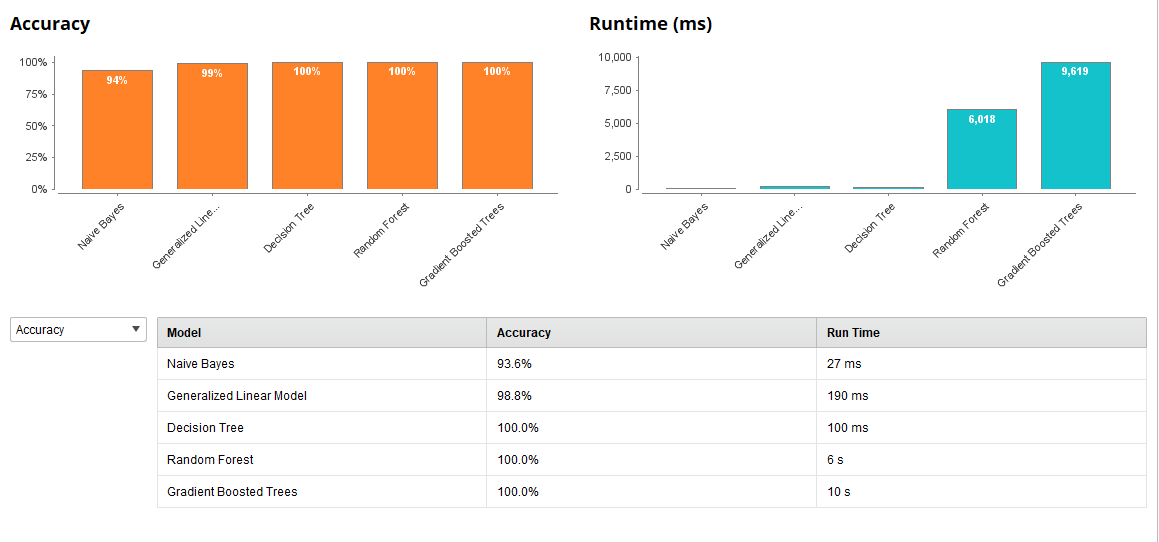
1. Target attributes - level of alcohol.

That is a very interesting one. Are we able to predict level of alcohol without given alcohol content?

Level of alcohol was set up under rule:

=IF(K2<=9,"Low",IF(K2<=11,"Medium","High"))

There is 3 kind of output Low, Medium and High. It is Ordinal Polynomial attributes.

Methodology used as in previous example: training 70%, test 30% and for test we have deleted PH attribute.

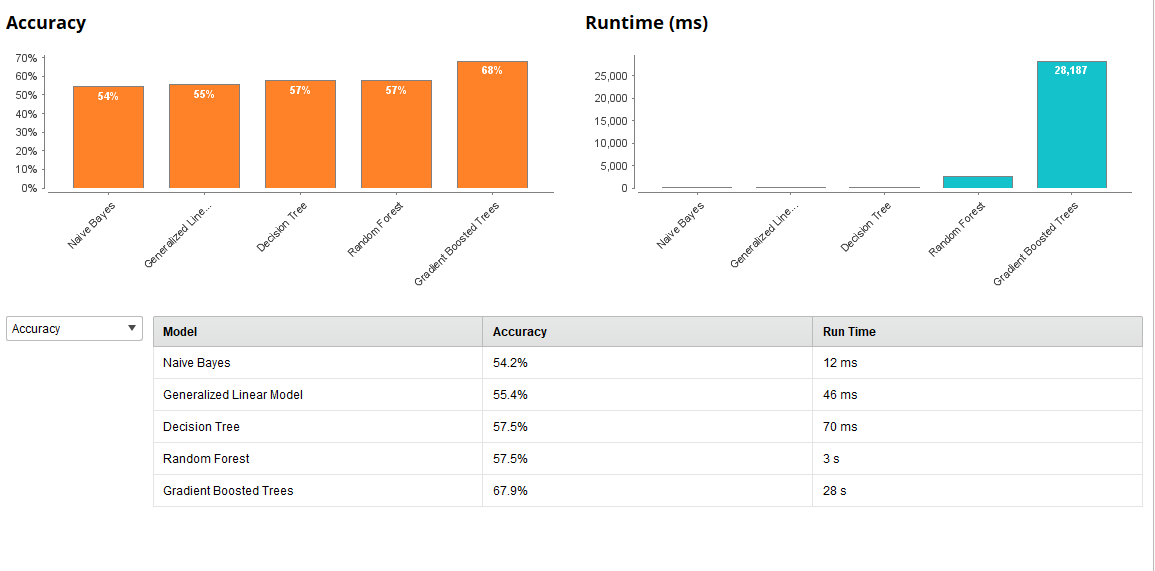
Training:

Summary of model outputs with given attribute.

Test:

Cross-validation:

Summary of model outputs with deleted attribute – alcohol content.



Only Gradient Boosted Trees could give us 68% of accuracy.