Flowering Time Prediction for the Maize Dataset

Sanjay Shanbhag
Dept of Computer Science
University of Central Florida
Oviedo, United States of America
Sanjayshan@knights.ucf.edu

Asst Prof Rui Xie
Dept of Statistic, Data Science
University of Central Florida
Oviedo, United States of America
Rui.Xie@ucf.edu

***Abstract*— This project is mainly about the flowering time prediction for the male type in the maize plant. It will majorly involve the complete utilization of the given independent variables and intelligent variable selection for an accurate prediction.**

Keywords—Statistical Analysis, Regression, Variable Selection

I. INTRODUCTION

The project was worked on as a part of NSF funded project called 'Biology of Alleles of maize and it's wild relatives". The purpose of this funded project was to derive the connection between the phenotypes and genotypes. As the name suggests, phenotype is just the physical appearance whereas genotype is the genetical characteristic. This maize diversity project was the result of a combined efforts from 6 different labs. But the scope of this paper is restricted to flowering time prediction, hence we will not be covering any biological analysis of the given maize data. Going further, we will discuss various methodologies to select variables and its effects on the prediction accuracy.

II. DATA ANALYSIS

A. Dataset and Descriptive Analysis

Before we even enter the data analysis part, it is important to understand the size and the characteristics of the dataset. Overall, the dataset consists of 7393 independent variables out of which there are 7389 variables representing the SNP marker, whereas the remaining represent the class/variety of maize. It is also known to us that there are a total of 25 crosses, each with about 200 recombinant inbred lines, popularly known as RIL's, which accounts for a total of 5000 unique variety of maize's in this dataset.

Now that we know the characteristics of the dataset. The Next step was to determine the quality of the dataset, which meant it was necessary to check if all the required variables were in the right datatype format, which would make it ready for model training. On looking for the unique datatypes in the dataset, it was found out that there were 3 different datatypes namely string, int64 and float64. Upon further investigation it was found out that the SNP marker variables ranging from m1 - m7389 included data of mixed types and invalid strings (making it impossible to convert to float directly). Hence a thorough data type conversion was done considering every value in the dataset. In the process of cleaning the dataset, the formatted dataset was saved for every step to ensure that we have a backup of the latest dataset.

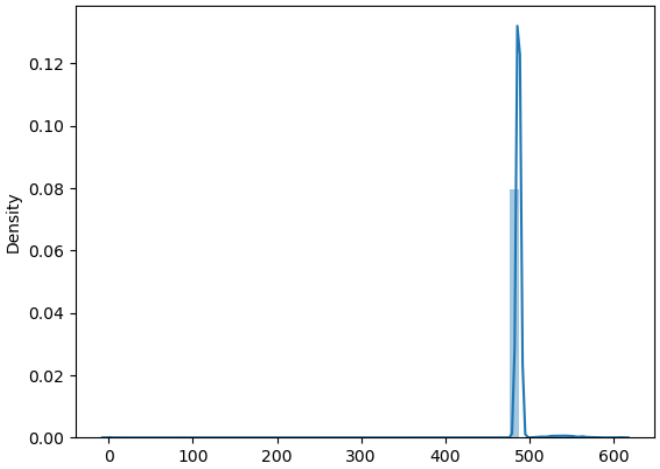

Fig 1. Distribution of NaN's across the columns in the dataset

The next step in the data analysis process was to check for cleanliness of the dataset, which meant the investigation of Nan's in dataset. On Visualizing the plot in Fig 2. It was found out that there were on an average of 488 NAN values in the dataset, which had to be removed. On cleaning the dataset, a total of 2146 entries were left out in the dataset, which accounted for less than 50% of the dataset.

B. Visualization and Analysis

In the previous section, a brief overview of the dataset was

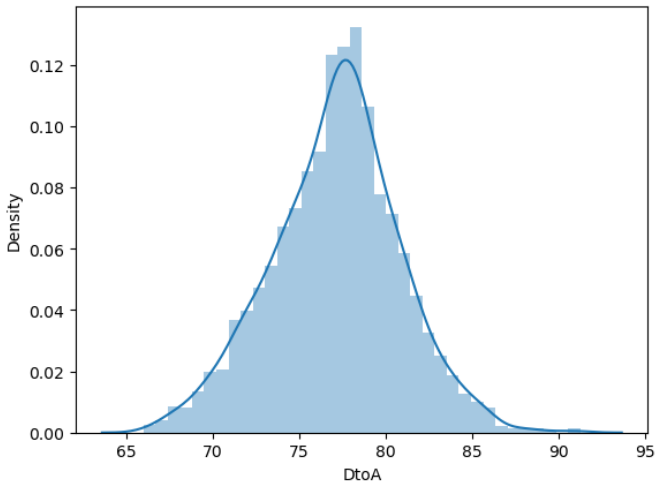

Fig 2. Distribution of DtoA values across the dataset

Identify applicable funding agency here. If none, delete this text b

given. Going ahead in the direction of data modelling , it was very important to thoroughly visualize the data and the relation with the respone variable. Due to the presence of 7393 variables in the dataset. It was not possible to visualize any kind of relation between the independent variables and the reponse variable. But here a we have tried to show the distribution of average flowering time of the maize , and looking at the plot in Fig 2. ,it was calculated on an average the maize plant took of 77 days to flower.

In the direction of undersatnding the data, it was important to understand the distribution of the critical temperature. From the above Fig 3. It is clear the distribution is more biaed towards the 0K cases as compred to the varied range of critical temperatures. If it were a classification, it would have meant that the model would be more inclined towards the 0K class and hence would be inefficient in predicting other classes.

Apart the distribution plot, various other visualizations were also made to understand the data well. Few of them include Principle component analysis of the attributes in the dataset. The theory states that the components corresponding to the Eigen vector with larger eigen value are nothing but the principle components. The eigen values obtained were normalized to make all the components visible in the graph in the sorted order.

III. Data Modelling Methodologies

In this section we will mainly talk about the modelling methodologies required to predict the accurate flowering time for the maize plant. The plan is to utilize various variable selection methods to consider only the best related features for training a regressor. In the first approach, we will be going with the statistical approach namely Pearson's correlation and Spearman's Correlation. Finally, we plan to consider the more supervised methodologies namely Ridge regression, Lasso regression and Elastic net which is a combination of both Lasso and Ridge Regression.

Before we enter the phase of data modelling methodologies, lets examine the training pipeline for the above-mentioned algorithms.

1. Train test split is mainly used to validate the model on the test set.
2. R^2 Statistic and MSE are being calculated to understand the effectiveness of the algorithm.
3. In case of Statistical methods, top K features are considered for training using Random Forest Regression and Linear Regressors.
4. Since there are different values of K being used for Statistical methods, a plot has been provided to visualize the improvement.

A. Supervided variable Selection

As mentioned earlier, 3 algorithms are used in this category. The main reason for utilizing this approach is that they are known to severely penalize the irrelevant variables, which in turn helps generalize a model. It has to be also understood that these models are mainly useful to avoid overfitting, which happens when there is less amount of quality data. On experimenting it was found out that, all three models failed to perform well on the given dataset. The below table Table I. summarizes the results obtained from 3 of the models.

Table I. Outcome of supervised Variable selection

Algorithm	Alpha value	MSE	R^2 Statistic
Lasso	1.0	13.28	0.089
Ridge	1.0	34.24	-1.32
Elastic Net	1.0	13.28	0.089
Lasso	0.5	13.28	0.089
Ridge	0.5	35.10	-1.40
Elastic Net	0.5	12.77	0.1245

From the table it is clear that these methods are not giving their best towards the dataset. But it can be equally noticed that Lasso and Elastic Net are almost equal or comparatively better than that of the ridge regression. The reason could be that there may be certain features which might not be relevant to the data and it is very much required to be removed from the data. So from my understanding, ridge cannot purely get rid of the data and hence the poor performance, whereas the lasso and Elastic Net have the capability to get rid of those variable hence performing comparatively better.

B. Statistical Approaches for variable Selection

The Variable selection approaches mainly includes the basic correlation approaches to understand the relation between the response and the independent variable. So, as mentioned in the training pipeline, all these statistically filtered features are finally trained on the Random Forest Regressor.

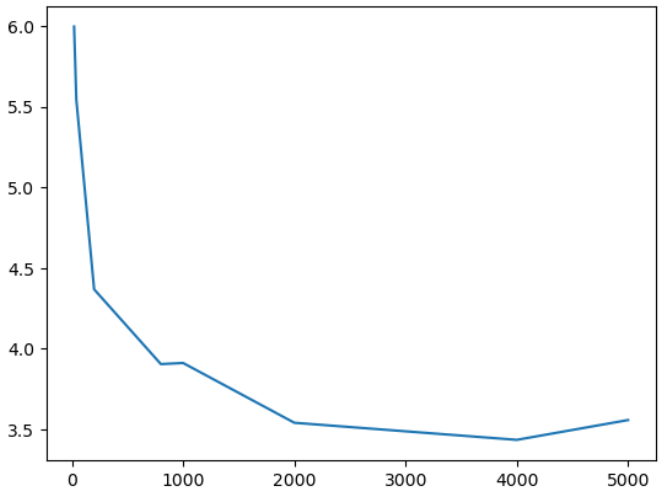

Fig 3. MSE (y-axis) vs K-values (x-axis) for Pearson's Correlation

The Basic idea behind considering different K values is to see on how far we can go in the direction of either increased

R^2 Statistic or decreased MSE values. By now it must be clear that the statistical method is outperforming the supervised methodology both in terms of the R2 and MSE. Side by it must

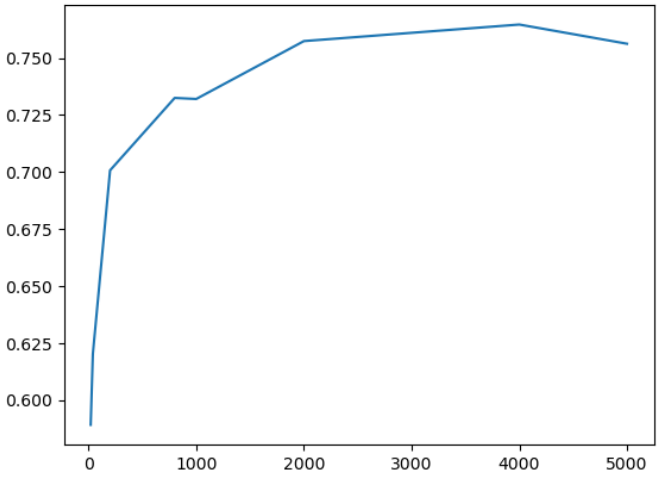

Fig 4. R^2 Statistic (y-axis) vs K-values (x-axis) for Pearson's Correlation

must also be noticed that in the Fig 4. The R^2 Statistic is consistently increasing until K=4000, after which we see a dip in the graph. A similar pattern is observed in the plot Fig 3. For MSE. Now that we have understood the performance of Pearson's correlation on the maize dataset. Let's try to see the performance of Spearman's correlation. On reading about the same, it was found out that Spearman's algorithm considers the monotonic pattern in the data and finds the relation between the given 2 variables. To understand mathematically, it simply calculates the Pearson's correlation on the ranked data, where the order of the data is given preference.

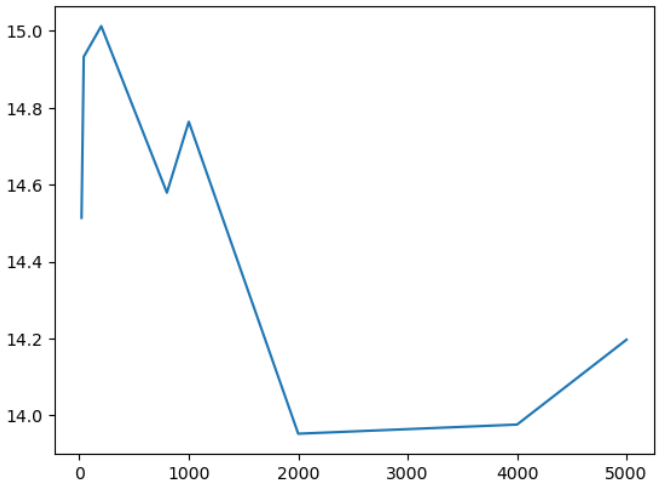

Fig 5. MSE (y-axis) vs K-values (x-axis) for Spearman's Correlation

Now that we can see the plot in Fig 5. and Fig 6. It is clear that the above method is underperforming as compared to the Pearson's method. One of the possible reasons for this poor performance as I can think about being is that due to the linear and gaussian nature of the dataset which are the main assumptions of the Pearson's Correlation, where as Spearman's method is mainly used towards Non-linear datasets.

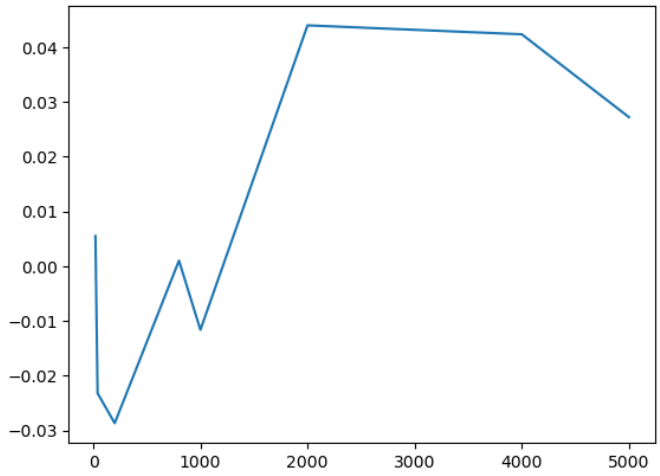

Fig 6. R^2 Statistic (y-axis) vs K-values (x-axis) for Spearman's Correlation

C. Conclusion

A clearcut data analysis couldn't be done due to the huge size of the dataset, which clearly created lot of barriers on the way. But the models that were trained in the process where to an extent were capable enough to understand the pattern in the data. This paper to an extent shows that statistical methods are equally effective as compared to supervised methods when it comes to variable selection.

D. Future Scope

A further deep analysis is required to understand the reason behind the failures of the other methods like lasso, ridge and elastic Net. Adding on to this, a few line fitting plots would have given a better understanding of the fitness of the data.

REFERENCES.

[1] Waldmann, Patrik, Gábor Mészáros, Birgit Gredler, Christian Fuerst, and Johann Sölkner. "Evaluation of the lasso and the elastic net in genome-wide association studies." *Frontiers in genetics* 4 (2013): 270.].

APPENDIX

1. The code that was utilized perform the various analysis and the modelling is attached in the form of link below.

   Maize_Data_Analysis.ipynb

2. Other online resource that was used as reference for this paper.

   a. https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/

   b. https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php