Analysis of Genetic Traits and Chromosomes' Effects on Potato Tuber Sprouting

Yuzhu Han

B00819905

Supervised by: Dr. Hong Gu

Dalhousie University

April 2024

Abstract

This study explores the genetic determinants influencing potato tuber sprouting. Gradient Boosting Machine (GBM) and Random Forest algorithms are employed to analyze a high-dimensional dataset comprising the Tuber Sprout Index of 100 potato clones with associated integration of 6,248 gene expression transcriptions and 1,602 single nucleotide polymorphism sequences. Through the analysis, this study aims to identify the key genetic markers that impact tuber sprouting and evaluate the predictive performance of these tree-based models. Utilizing 10-fold cross validation, the models are tuned and tested, revealing significant insights into the variable importance within the dataset. The findings show that the gene expression transcriptions TAG5126 and TAG2193 are the most important variables in both models. The majority of TAGs in the top 40 important variables in the two models suggest that gene expression profiles may play a more substantial role in determining tuber sprouting outcomes. The confusion matrices show that the GBM model and Random Forest model have similar classification performance with an accuracy of approximately 70%. Although both models proficiently identify the Tuber Sprout Index of value 2, the limited predictive capability for the Tuber Sprout Index value 1 and 3 highlights the challenges posed by class imbalance.

1.0 Introduction

According to the International Potato Center (2024), potatoes rank as the third most important crop for human diets globally, following rice and wheat. They are a staple food for over a billion individuals across the world, with a yearly production of over 300 million metric tons (International Potato Center, 2024). This significant output highlights the potato's economic importance on a global scale. However, tuber sprouting generates a major challenge during potato storage as it can degrade tuber quality and impact commercial value. Additionally, it may threaten food security due to the toxins produced during the sprouting process (Friedman & Levin, 2016). Conversely, controlled sprouting is essential for utilizing tubers as seed potatoes (Pinhero et al., 2009). Therefore, it is crucial to explore the genetic effects of gene expression and DNA on potato tuber to control dormancy and sprouting.

Gene expression profiles, as well as single nucleotide polymorphism sequences in DNA, often have counts in the thousands or tens of thousands. These types of data usually consist of predictors that are much more than observations, which makes many standard statistical models do not have a good performance in prediction analysis. Gradient Boosting Machine (GBM) and Random Forest are the two tree-based models that can be used to deal with high-dimensional data while requiring careful selection of tuning parameters. This study aims to analyze the genetic determinants of potato tuber sprouting by integrating and examining gene expression and DNA data using GBM and Random Forest algorithms, with the goal of identifying key genetic markers and evaluating the comparative performance of these methods in predictive analysis.

2.0 Data

In this study, the dataset comprises tuber sprouting data from 100 different potato clones cultivated across three separate blocks, combined with the data of 6248 gene expression transcriptions (TAG) and 1602 single nucleotide polymorphism (SNP) sequences. For each clone within each block, five plants were examined. The tubers were harvested from the field and then stored at 7°C. Five tubers were taken from each clone and were bagged separately from each clone and each block. Tuber sprouting in storage was scored at three months after storage for the group of five tubers on a scale of 0 to 5. A score of 0 indicates no sprouting, while scores 1 through 5 represent sprout lengths of 0-0.5 cm, 0.5-1 cm, 1-2 cm, 2-3 cm, and over 3 cm, respectively.

2.1 Serial Analysis of Gene Expression

Counts of 6248 gene transcriptions represent gene expression of each clone of potato in each block, which was measured by Serial Analysis of Gene Expression (SAGE). In order to apply this method, leaf samples of clones were taken at 70 days after planting. SAGE counted the number of transcripts that were expressed from a gene through the sequencing of short tag sequences (Hu & Polyak, 2006). The next generation sequencing platform, Illumina, was used to measure the number of tags. The counts of tags were recorded in the data matric and were used to pre-filtered to retain genes which were expressed at least three times in at least one sample.

3.0 Method

3.1 Gradient Boosting Machine

Gradient boosting machine (GBM), also referred to gradient boosted tree, is a boosting method approach for improving the predictions resulting from a decision tree. The general idea of boosting is creating multiple copies of the training dataset using the bootstrap, fitting a decision tree to one copy, then sequentially growing trees using information from previously grown trees. (James et al., 2013). GBM can be regarded as a slowly-learning method that it can gradually improve the performance of the model with the following three tuning parameters:

- `n.trees`: the number of trees to grow. It may cause overfitting if the number is too large.
- `shrinkage`: the shrinkage parameter is a small positive number that controls the rate at which boosting learns. In order to get good performance, a very small shrinkage parameter may require a very large `n.trees`.
- `interaction.depth`: it controls the number of splits in each tree, influencing the complexity of the boosted ensemble.

3.2 Random Forest

Random Forest is an algorithm that is improved from bagged decision trees and can be used for both regression and classification. In constructing a Random Forest, multiple decision trees are built on bootstrapped subsets of the training data. At each split of a tree, a random subset of the predictors is selected, from which only one predictor is allowed to be used to make the split (James et al., 2013). This method ensures that trees do not rely heavily on any single strong predictor, thus reducing the risk of similar tree structures and promoting the reliability of the result. Therefore, the most important tuning parameter for Random Forest is `mtry`, which represents the number of variables randomly sampled in each split. The default value of `mtry` for classification is the square root of the total number of variables (Liaw & Wiener, 2022).

3.3 Cross Validation

Cross-validation is a resampling technique used to assess the performance and generalization ability of a predictive model. This method splits the dataset into multiple subsets, known as "folds." In this technique, the model is trained on several subsets (the training set) and evaluated on the remaining subset (the test set). This process is iterated multiple times, with each subset serving as both training and test data at different points.

In this study, 10-fold cross validation is used, where the dataset is divided into 10 equal-sized folds. The model is trained 10 times, each time using nine folds for training and the remaining fold for validation. This process helps in tuning model hyperparameters and selecting the best performing model, thereby reducing the risk of overfitting and providing more reliable performance estimates.

4. Confusion Matrix

A confusion matrix is a tool used to evaluate the performance of classification models. It assesses the accuracy of predictions by comparing them against actual observed outcomes. Besides, the matrix offers insights into the model's effectiveness, facilitating a straightforward interpretation of the prediction performance for each class.

4.0 Analysis

4.1 Data cleaning and processing

To investigate the impact of gene expression and SNP data on tuber sprouting, observations with missing values are removed instead of imputed in order to keep the reliability and completeness of the data. After data cleaning, 253 observations are retained for further analysis. After data cleaning, the Tuber Sprout Index exhibits high concentration around the value of 2, with a few occurrences at index value 0, 4, and 5 (Figure 1.) For the purposes of model fitting, the Tuber Sprout Index is reclassified to create a more focused dataset with fewer extreme values. Index values originally recorded as 0 and 1 are combined and labelled as index value 1. Similarly, original index values of 3, 4, and 5 are grouped into index value 3. As a result of regrouping, the final Tuber Sprout Index consists of 37 observations categorized as index value 1, 179 observations as index value 2, and 37 observations as index value 3.

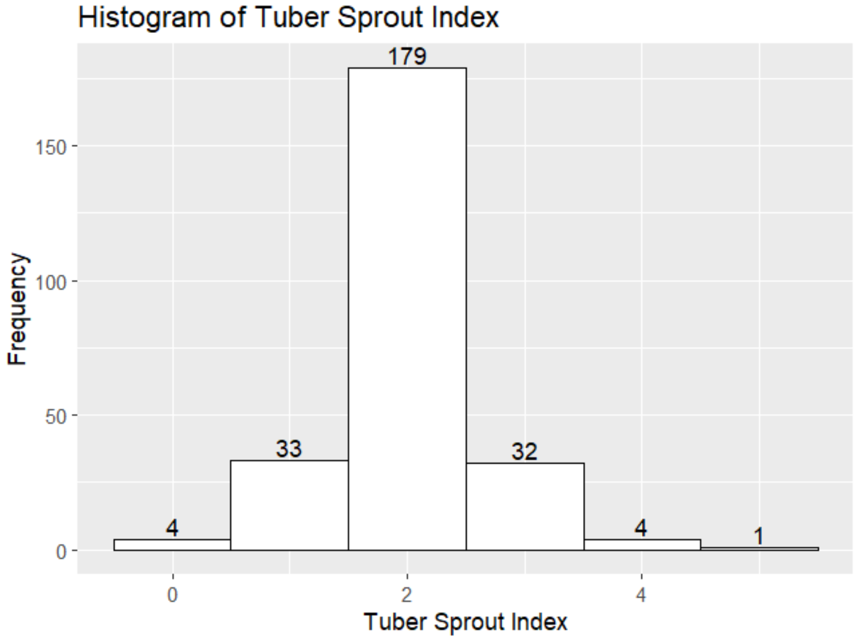

Figure 1. Histogram of Tuber Sprout Index after data cleaning.

4.2 Model analysis

Both GBM and Random Forest are applied to assess the most important TAGs and SNPs that influence tuber sprouting. For both algorithms, the first step is shuffling the data and randomly assigning 10 equal-sized folds. Then ten GBMs and ten Random Forests are used to fit the data using each fold as a validation set once.

Before fitting the model each time, a grid search with10-fold cross validation is used to select the optimal tuning parameters for the training set. For the GBM model, `n.trees` is selected from values of 200, 300, 400, 500, and 600; `shrinkage` is chosen from 0.1, 0.01, and 0.001; and `interaction.depth` is tuned from 1, 2, and 3. For the Random Forest model, the default value of `mtry` is set at 88.6 (the square root of 7850), so the tuning range for `mtry` is set from 60 to 120.

After fitting the models, the importance of variables is ranked to identify the top 50 most influential variables for each model. In the GBM model, variable importance is assessed using the relative influence (`rel.inf`), where a higher `rel.inf` value indicates a greater impact on the prediction. In the Random Forest model, variable importance is interpreted from the `MeanDecreaseGini`, which is derived from the Gini impurity index used during the decision tree construction. A variable with a higher `MeanDecreaseGini` value is considered more important for making accurate predictions.

The top 50 important variables from each of the ten models are consolidated to determine the most important variables across all ten folds. Graphs of the sum importance scores (Figure 2) reveal a high skewness, with the majority of these variables appearing only once across the folds.

Consequently, only the top 40 variables with the highest sum importance from both the GBM and Random Forest models are selected for analysis.

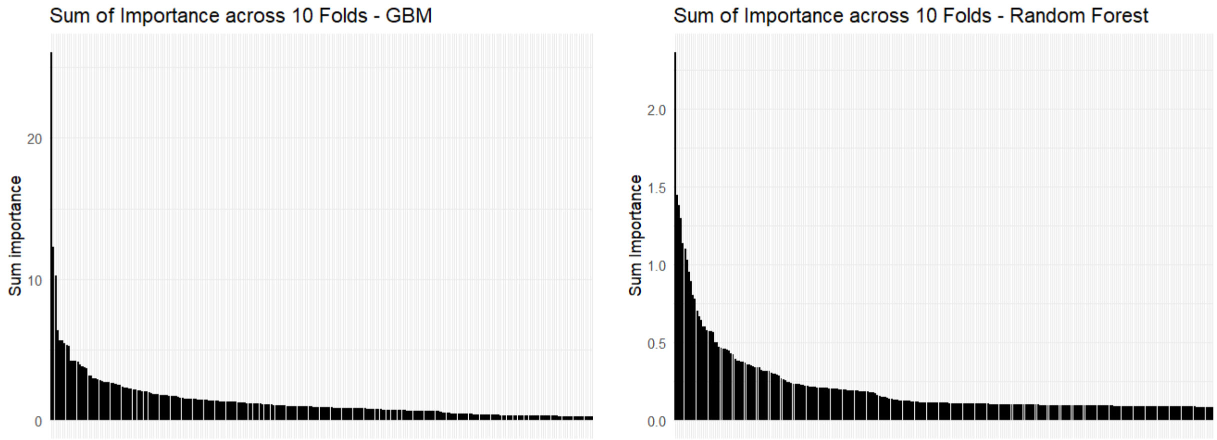

Figure 2. Sum of importance of variables across all folds for GBM and Random Forest.

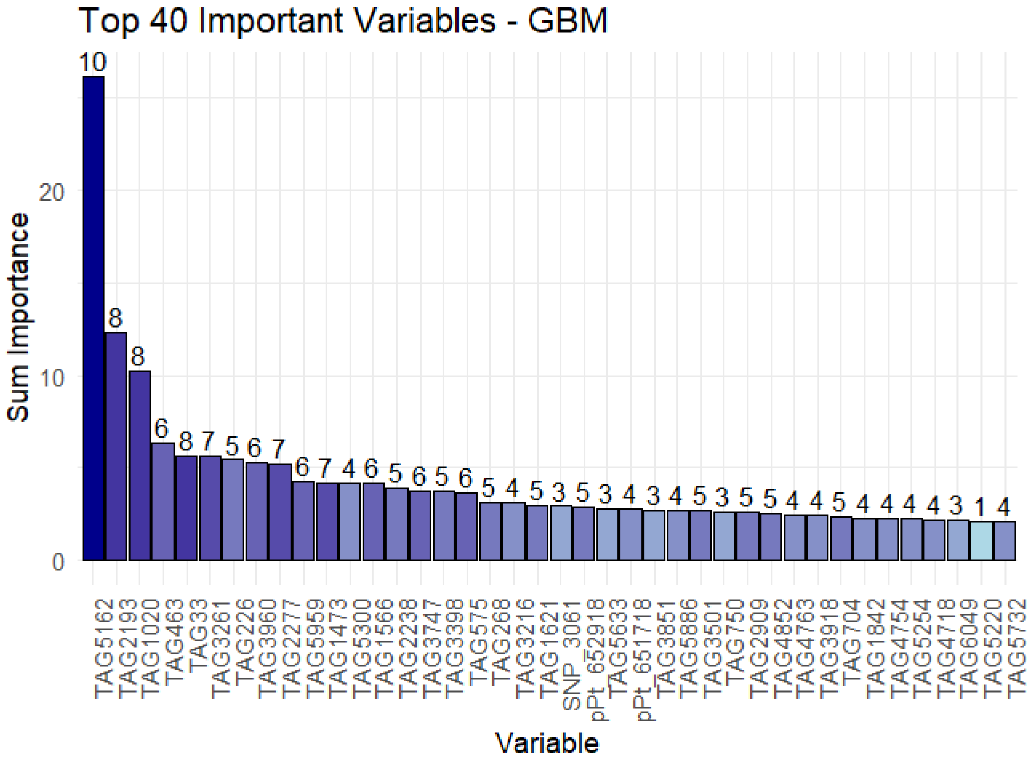

Figure 3. The top 40 important variables for GBM model.

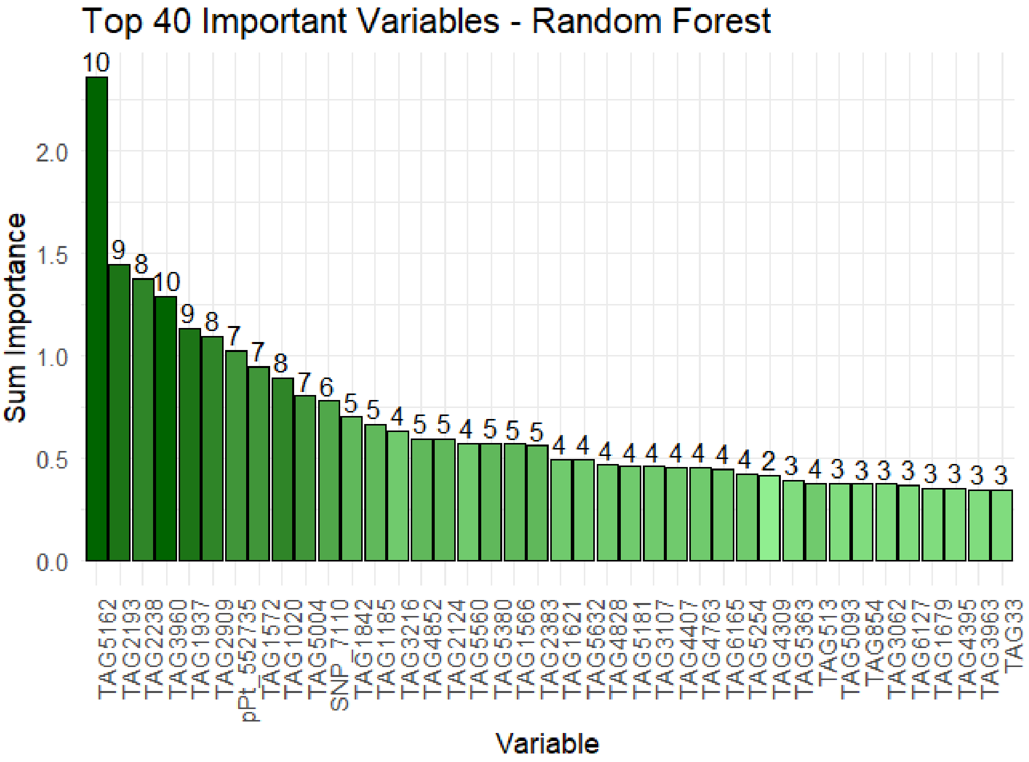

Figure 4. The top 40 important variables for Random Forest model.

In the predictive phase, the prediction is made for each test set using the fitted model. The performance of prediction can be assessed using confusion matrices to know whether the predicted outcomes match the observed class in the test set. The accuracy of the prediction of each fold is recorded and used to calculate the standard error across the ten folds.

Table 1. Confusion Matrix of GBM model

Prediction vs. True	1	2	3
1	4	4	0
2	32	174	36
3	1	1	1
Accuracy: 0.7075			
Standard error: 0.0148			

Table 2. Confusion Matrix of Random Forest Model

Prediction vs. True	1	2	3
1	0	1	0
2	37	178	37
3	0	0	0
Accuracy: 0.7036			
Standard error: 0.0074			

5.0 Conclusion

TAG5162 stands out as the most important variable in both models because it is at the top in all ten folds and has the highest sum importance value, according to Figures 3 and 4. Following closely is TAG2193, attaining the second highest importance score and frequently appearing among the top rankings in most folds in both models. Some other variables, such as TAG 2909 and TAG1020, are also in the top 40 for both models. Notably, the TAGs appear to be more influential than SNPs since the majority of the top variables are TAGs, with only a few SNPs listed. This difference suggests that gene expression profiles may have a more substantial impact on the potato tuber sprouting than DNA sequences.

Tables 1 and 2 show that both the GBM and Random Forest models proficiently identify the Tuber Sprout Index of value 2, evidenced by the significant number of true positives. This is in the context of the imbalanced dataset, with 37 observations for index values 1 and 3, and 179 for index value 2. The Random Forest model exhibits zero false positives in predicting index value 1 and 3, suggesting an inherent model bias towards the majority class. On the contrary, the GBM model shows a small number of false positives for the index value 1 and 3, indicating a relatively less biased approach. The model bias may be generated by the class imbalance of the original observations of the Tuber Sprout Index. Because the Tuber Sprout Index value of 2 significantly outnumbers index value of 1 and 3, the GBM and Random Forest models learn from the more abundant examples during the training phase, potentially leading to the better performance on the class of index value 2 at the expense of the other two classes.

The overall accuracy of both models is fairly similar, reflecting an adequate predictive capacity despite class imbalances. The Random Forest model has a lower standard error, indicating a better consistency of predictions across folds. However, the consistency may lead to overfitting to index value 2 and weakening the generalizability of the model. Although the GBM model captures several instances of index value 1 and 3, its performance on these minority classes remains limited. Therefore, while the GBM model better identifies the minority class compared to the Random Forest model, both still face challenges in attaining a balanced classification performance across classes within an imbalanced dataset.

For further analysis, balancing the classes of the Tuber Sprout Index is crucial to enhance the classification performance of GBM and Random Forest models. A possible approach could involve redefining the scoring criteria for the Tuber Sprout Index, aiming to achieve a more even distribution across the classes. Additionally, exploring other R packages like `XGBoost` may yield better performance with multi-class datasets. By continuously refining the analytical approaches, we can achieve more robust and reliable results and better understand and predict the factors influencing potato tuber sprouting.

Reference

Friedman, M. & Levin, C. E. (2016). Glycoalkaloids and Calystegine Alkaloids in Potatoes. *Advances in Potato Chemistry and Technology* (2nd ed.) Academic Press.

Hu, M. and Polyak, K. (2006). Serial analysis of gene expression. *Nature Protocols*, 1:1743-1760.

International Potato Center. (2024). Potato Facts and Figures. Retrieved from: https://cipotato.org/potato/potato-facts-and-figures/

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013). Bagging, Random Forests, Boosting. *An Introduction to Statistical Learning: with Applications in R*. Springer.

Liaw, A. & Wiener, M. (2022). *randomForest: Classification and Regression with Random Forest*. R package version 4.7-1.1.

Pinhero, R. G., Coffin, R. & Yada, R. Y. (2009). Post-harvest Storage of Potatoes. *Advances in Potato Chemistry and Technology*. Academic Press.