Homework_2

Damiano Orlandi

Introduction

Prostate cancer remains one of the most prevalent forms of cancer among men worldwide, posing significant challenges to public health systems and individuals. Understanding the intricacies of this disease and its associated factors is crucial for devising effective treatment strategies and improving patient outcomes. In this study, i investigate the complex interplay between prostate-specific antigen (PSA) levels, clinical measures, and prostate cancer progression. To accomplish so, i employ decision regression trees, random forests and boosting decision trees as analytical methodologies in order to deepen the relationship among the response and predicted variables given by a dataset of 97 men.

Data Exploration

The dataset under investigation comprises data collected from 97 men who were on the brink of undergoing radical prostatectomy, a common surgical procedure for treating localized prostate cancer. Among the variables examined, the level of prostate-specific antigen (lpsa) serves as a pivotal indicator, measured in nanograms per milliliter (ng/ml) and log-scaled to facilitate analysis. Additionally, several clinical measures have been considered, each offering unique insights into disease's progression and severity:

- `lcavol`: log-transformed cancer volume in cubic centimeters (cm^3).
- `lweight`: log-transformed prostate weight in grams (g).
- `age`: age patients in years.
- `lbph`: log-transformed amount of benign prostatic hyperplasia in square centimeters (cm^2).
- `svi`: a binary variable denoting the presence (1) or absence (0) of seminal vesicle invasion.
- `lcp`: log-transformed capsular penetration in centimeters (cm).

- `gleason`: the Gleason score, a grading system reflecting the aggressiveness of prostate cancer, ranging from 6 to 9.
- `pgg45`: the percentage of Gleason scores 4 or 5, recorded over the patients' visit history before their final current Gleason score.

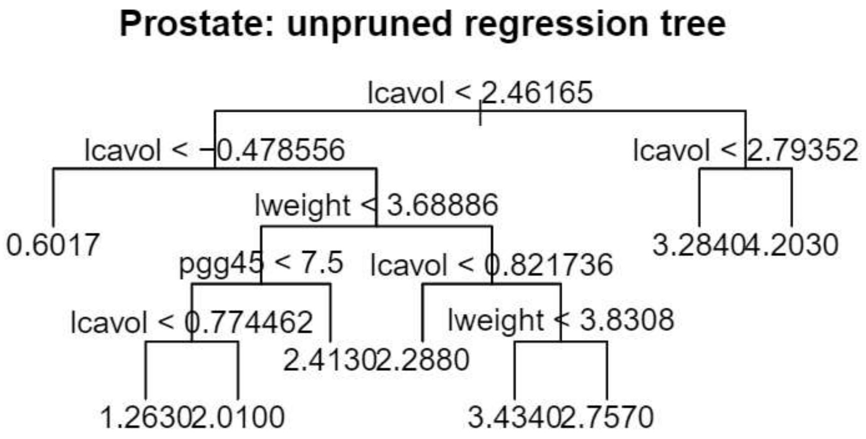
Notably, the dataset is devoid of any missing values (NAs), ensuring the integrity and reliability of our analyses. Moreover, the dataset contains only numeric values, so all the analysisi wiil be oriented to regression methods.

Methods

For consistency in results, I've set the random seed to 18 before computing every model, ensuring reproducibility across multiple runs. The analysis relies on key R libraries: tree, randomForest, gbm, Metrics, and caret, providing essential tools for model building and evaluation.

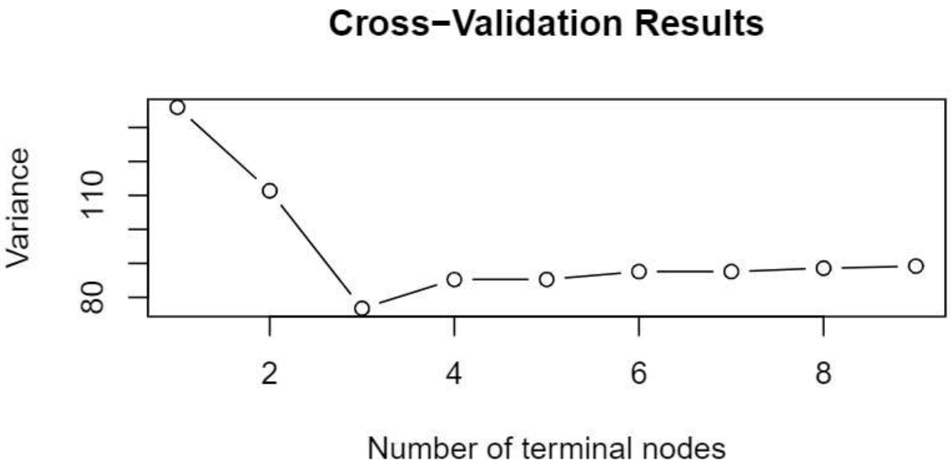
Decision Tree

The decision tree structure visually illustrates how the dataset is divided into distinct regions or segments based on predictor variables. At each node, a decision is made to split the data further until terminal nodes, or leaves, are reached.

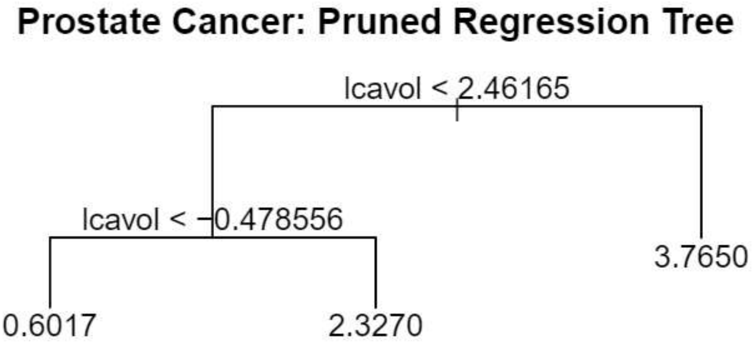


The diagram illustrates that cancer volume, whether below or above 2.46165, is the primary predictor in the model. The following divisions involve cancer volume and age variables, underling their reducted impact on the variance.

Considering the regression decision tree peculiarities of being easly understandable, but difficulty able to generalize predictions (due to the tipical high variance), i computed cross-validation in order to identify the best level of complexity.



As shown, the number of nodes that minimizes the error is denoted as 3, indicating the needing of attempt pruning and evaluating better performances.



The pruning process strongly optimize the prediction: it confirmes the already identified importance of lcavol (with threshold of 2.46165), and intorduce another split still given by lcavol (with the threshold of -0.478556). The output underline the pivotal role that the cancer volume has on the prediction.

Random Forest

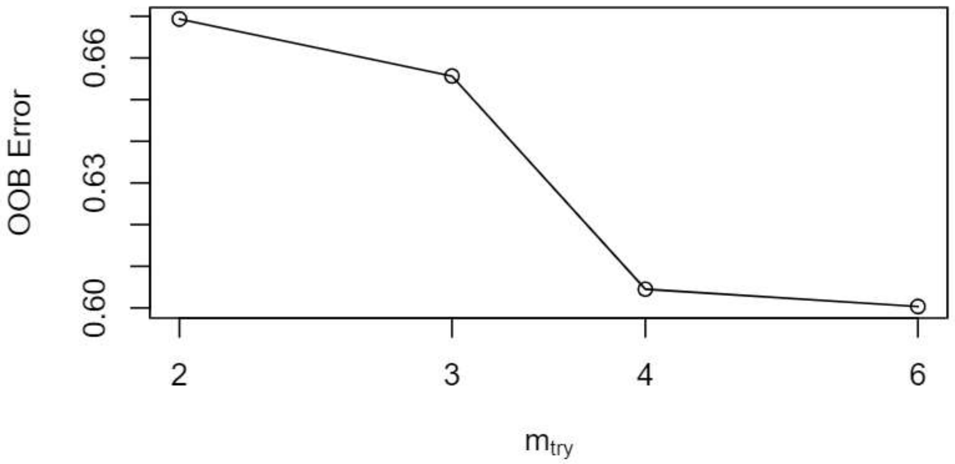
Random Forest is an ensemble learning method that combines multiple decision trees to predict continuous outcomes. It operates by generating bootstrap samples from the dataset and training individual decision trees on these samples. In regression tasks, each tree predicts a continuous value, and the final prediction is obtained by averaging the predictions of all trees. Leaving the default number of trees set at 500 let me having an overview of the MSR and the explained variance, and focusing on the number of variables tried at each split. Firstly i set **mtry** as the number of features in the dataset (8) and i obtained:

```
Call:
 randomForest(formula = lpsa ~ ., data = prostate_data, mtry = default)
              Type of random forest: regression
                    Number of trees: 500
No. of variables tried at each split: 8

          Mean of squared residuals: 0.6285734
                    % Var explained: 52.34
```

Then, i opted to compute the Out-Of-Bag as an internal validation metric used to estimate model prediction error. As shown in the plot, the lowest error value (0.6003244) is matched where the number of features is 6, our updated best value.

```
mtry = 2  OOB error = 0.6693536
Searching left ...
Searching right ...
mtry = 3    OOB error = 0.6556615
0.02045573 0.01
mtry = 4    OOB error = 0.6044948
0.07803823 0.01
mtry = 6    OOB error = 0.6003244
0.006898928 0.01
```



Then, i compute a new random forest with the updated value of mtry keeping the previous parameters and then, in order to have a better understanding of the role played by all the variables in the variance, i performed the variance importance.

```
var_importance <- importance(rf_otm)
print(var_importance)
```

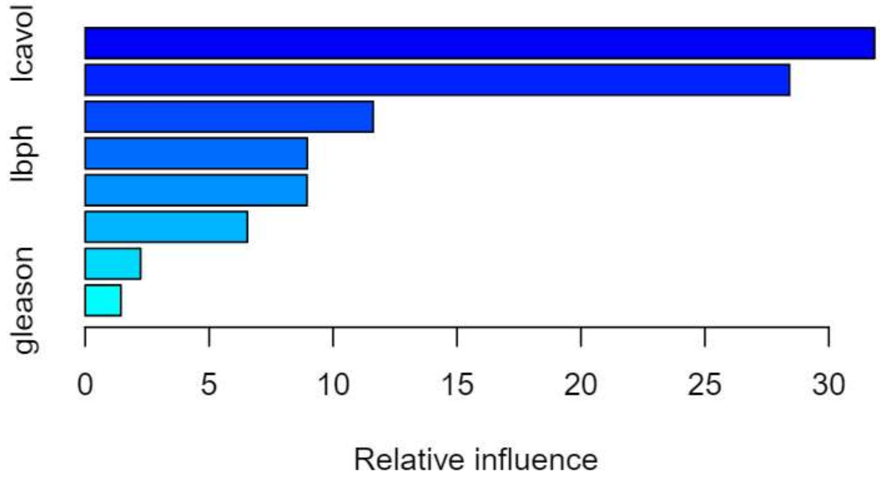
```
        IncNodePurity
lcavol      63.753754
lweight     17.885406
age          6.141768
lbph         6.131388
svi         11.993369
lcp          6.867909
gleason      1.277111
pgg45        6.641365
```

lcavol emerges as the most influential predictor, displaying a substantial increase in node purity of 63.753754. This suggests that variations in lcavol significantly aid in segregating observations into distinct classes, making it a pivotal feature for predicting the outcome.Following closely are lweight, svi, and lcp, each demonstrating some contributions to node purity, with values ranged from 6 to 17. Conversely, age, lbph, pgg45, and gleason exhibit comparatively lower increases in node purity. While still contributing to the model, their impact appears less pronounced when compared to the aforementioned predictors.

Boosted Regression Trees

Boosted Regression Trees (BRT) melds regression trees with boosting for superior predictive accuracy. It sequentially refines predictions by iteratively adding trees that correct errors made by previous models. Employing gradient boosting, BRT minimizes the loss function to optimize model fit. The final prediction aggregates contributions from all trees, weighted by their performance in reducing the loss.

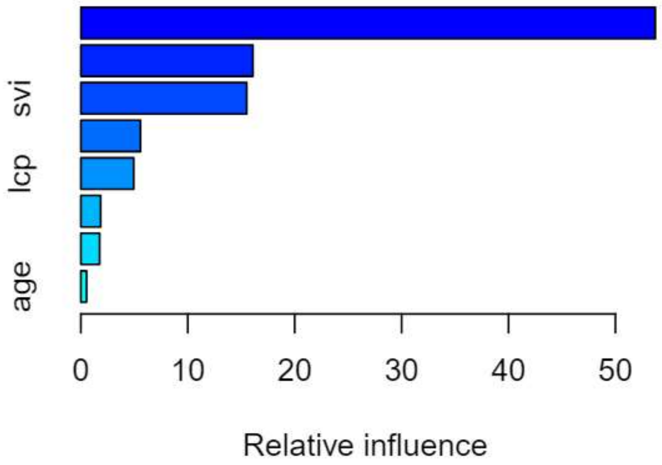
Firstly, i set 3000 as number of trees in order to have a sample big enough on which perform cross-validation to obtain the optimal value. Secondly, i set the k-folds of the cross validation at 5 considering the high values of trees. The analysis show that lcavol and lweight are responsable for the 31.840429 % and 28.415237% of the variance.



```
             var   rel.inf
lcavol   lcavol 31.840429
lweight lweight 28.415237
age          age 11.619323
lbph        lbph  8.960803
pgg45      pgg45  8.947434
lcp          lcp  6.544978
```

```
svi          svi  2.234651
gleason gleason  1.437144
```

Applying the cross-validation and optimizing the number of tree (now 29), i obtained a significant increase of importance for icavol feature, reaching 53.7179537%.



```
            var    rel.inf
lcavol   lcavol 53.7179537
lweight lweight 16.0970383
svi         svi 15.5149774
pgg45     pgg45  5.5787379
lcp         lcp  4.9434855
lbph       lbph  1.8545391
gleason gleason  1.7576638
age         age  0.5356043
```

Model Evaluation

In evaluating the models, I employed a systematic approach to prediction. Considering the reduced size of the dataset, i preferred assess the capability of the model in detecting the underlying patterns and relationships present in the entire dataset, rather than just on a subset (which would have been composed by 20-35 men, as i was initially considering for the split).

I choose MSE as indicator because of its relevance as a measure of prediction accuracy and its suitability for comparing the performance of different regression models.

Beginning with the Decision Tree model, I utilized the pruned regression tree to make predictions on the dataset, followed by the calculation of the Mean Squared Error (MSE) to quantify its performance. Similarly, for the Random Forest model, predictions were made using the optimized Random Forest model (rf_otm), and the corresponding MSE was computed. Lastly, leveraging the Boosted Regression Trees model (optimal_gbm), predictions were generated with a specified number of trees (best_trees), and the resulting MSE was determined. To compare the performance of these models comprehensively, I constructed a dataframe summarizing their MSE values. This structured approach facilitated a clear understanding of the models' predictive accuracy, with subsequent identification of the model with the lowest MSE for further analysis.

```
                   Model       MSE
1          Decision Tree 0.6174919
2          Random Forest 0.1116967
3 Boosted Regression Tree 0.5178462
```

Conclusion

The Random Forest model excelled in predicting prostate cancer progression with the lowest MSE of 0.1116967, leveraging ensemble learning to handle data complexity effectively and confirming variables like lcavol, prostate weight, and seminal vesicle invasion as key predictors. This robustness renders it highly suitable for clinical applications demanding high accuracy and reliability. On the other hand, the Boosted Regression Trees, despite a higher MSE of 0.5043923, provided insights into variable importance and potential for refining predictions with nuanced data interactions. Meanwhile, the Decision Tree model, while easy to interpret, exhibited the highest MSE of 0.6174919, indicating more susceptibility to overfitting and less reliability in predictions compared to the ensemble methods. Additionally, evaluating the model on the entire dataset can provide insights into its performance consistency and robustness. It ensures that the evaluation is not biased by the specific composition of the training and test sets and provides a more holistic view of the model's predictive accuracy.