University of St.Gallen

School of Management, Economics, Law, Social Sciences and International Affairs

Data Analytics II: PC3

University of St. Gallen

Jonas Husmann | 16-610-917

Niklas Leander Kampe | 16-611-618

Prof. Dr. Michael Lechner

March 14, 2022

Part 1: Trees and Forests

a. Load Data Set + Descriptive Statistics

```
      Y     X
0  4.30  0.86
1  0.02 -0.70
2 -1.06 -0.50
3  0.44 -0.89
4  6.84  1.60
```

```
Descriptive Statistics:
-----------------------------------------------
   mean   var   std    max   min   na  unique
Y  1.97  8.76  2.96  14.54 -2.51  0.0  1800.0
X  0.01  0.97  0.98   3.17 -3.17  0.0  1800.0
-----------------------------------------------
```

b. Description of Regression Trees + Outcome with only the Root Node + SSE/MSE Function

A general regression tree π consists of a root node (= initial node), terminal nodes and leaf nodes (final node for every unique tree structure), whereby every parent node consists of two child nodes, their splits in form of functions of the covariate vector X, a predetermined stopping rule and an error/accuracy measure. The prediction algorithm starts at the root node with the entire training sample of the outcome variable and covariates. From there, for each predictor X_i and cut-point s, two different, exclusive sub-sets are defined, according to $R_1 = \{X|X_i < s\}$ and $R_2 = \{X|X_i \geq s\}$. Based on this initial split, the mean values $\bar{Y}_1$ and $\bar{Y}_2$ of the outcome variable in the two "filtered" sub-sets based on the split condition are calculated. Afterwards, the final split of the data set is determined based on the covariate X_{ij}^* and cut-point s_j^* that minimize the sum of squared residuals. This procedure is continued for all following parent nodes in every unique tree path until a predefined stopping rule is reached, and hence every unique tree path reaches a leaf/final node. The optimal stopping rules could be determined as the maximum tree size (depth of the tree) or the minimum gain on the SEE/MSE calculation through an additional split, next to many others. Furthermore, these stopping rules can also be determined through hyperparameter tuning, e.g., based on a cross-validation approach. Hence, the regression tree follows a top-down approach in which the entire data set gets split through its depth and paths, which is called "recursive partitioning". The algorithm is also greedy as it adds a split at each step of the tree to improve the prediction power in terms of the prediction error minimization. Thereby, it needs to be recognized that a higher complexity, in form of an increased depth or the number of terminal leaves, could be first considered as an optimization in the training set as a lower prediction error, but could lead to overfitting of the model on the training data which is shown by an increased variance and hence an increased sensitivity to the data sample, i.e., the prediction on the testing sample. (Mareckova, 2021, S. 2-11)

Based on the previous theory about regression trees, a regression tree-based prediction of the outcome variable Y only on the node rood means that the prediction is just equivalent to the mean value of the outcome variable Y in the training data set. In this application, the full data set is split on an 70-30-split, whereby 70% of the data I used as training and 30% is used as testing data for final validation of the model. From the results below, the prediction of the outcome variable Y based on the covariate X only on the root node results in a value of 1.93, which is very close to the mean value of the full data set as stated in task 1a. Furthermore, the accuracy measures SSE and hence also the MSE are very high compared to the scale of the outcome variable Y, which indicates a bad performance of the model when only using the root node. The result does not seem surprising as the mean value of the training set does usually not lead an accurate prediction for the testing sample.

```
Predictive Regression Tree: Root Node Only
-----------------------------------------------
Predicted Y: 1.93
SSE: 3514.19
MSE: 9.76
-----------------------------------------------
```

c. Regression Tree Prediction of Outcome Variable Y

```
SSE Optimizing Regression Tree
--------------------------------------------------------------------------------
Best splitting value of covariate X: 0.5008
Row Index of X: 143
SSE: 4970.5531
MSE: 13.8071
--------------------------------------------------------------------------------
```

Literature & Sources

Mareckova, J. (2021). Data Analytics I: Regression Trees and Forests. St. Gallen, Switzerland.