Mar 27, 2023

Machine Learning Methods and Applications

Week 5. Decision trees - I

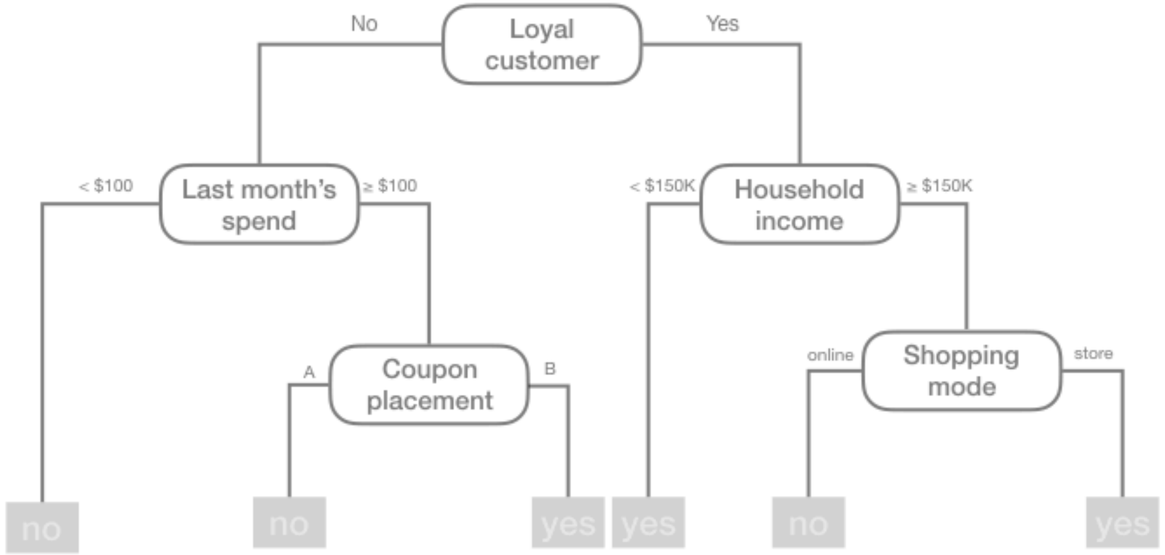
Remember

- Model validation ~ model performance check
- Holdout (splitting as train and test) provides only an estimate of model performance.
- Resampling based validation methods provides more than one estimation.
- Handling with missing data: remove or impute according to the missing data mechanism
- Transformations (scaling) may be needed to improve model performance

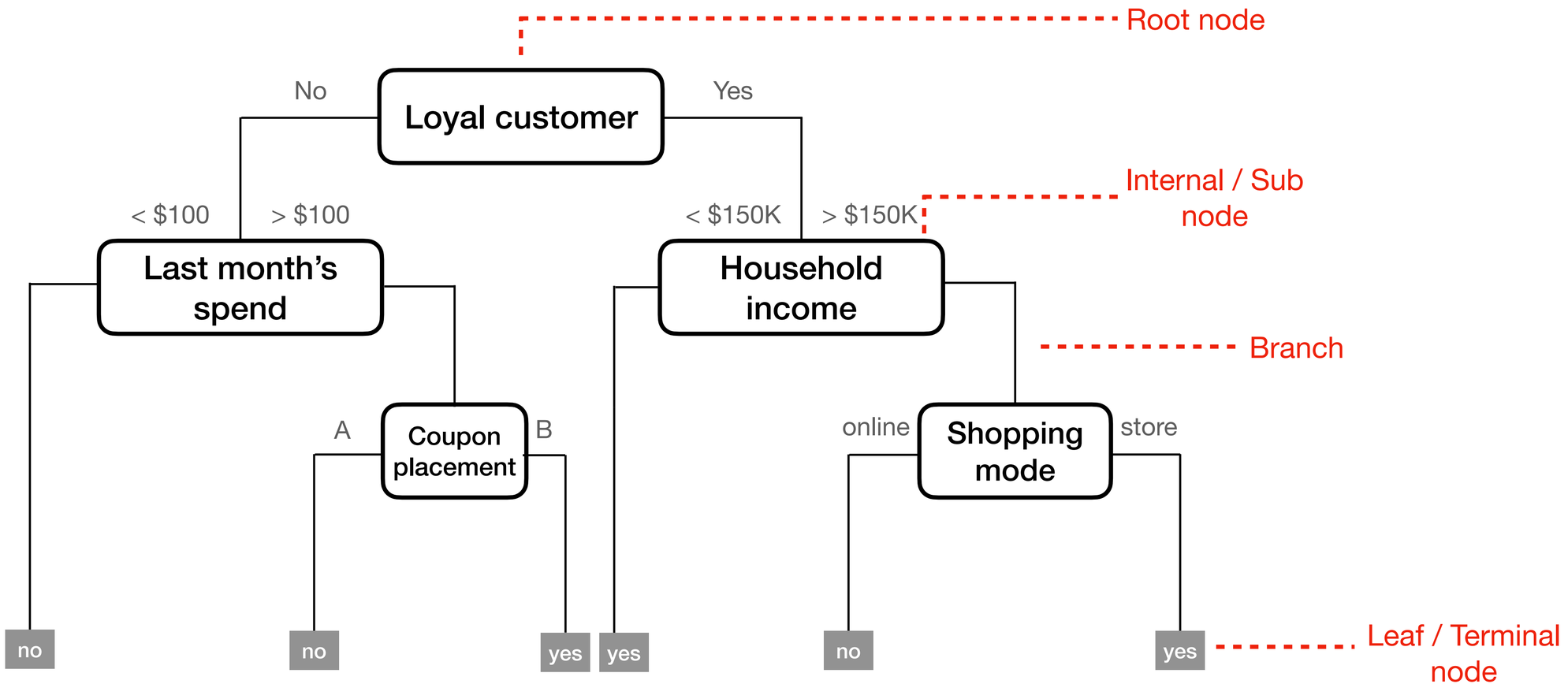
Decision trees

Decision tree

- Non-parametric algorithm
- Work by partitioning the feature space into a number of smaller regions
- Using splitting rules
- Produce simple rules that are easy to interpret
- Visualize with tree diagrams



Decision tree structure and terminology



Decision tree

Most well-known method for constructing decision trees is the classification and regression trees (CART) proposed in Breiman (1984).

1. A basic decision tree partitions the training data into homogeneous subgroups.
2. The subgroups are formed recursively using binary partitions.
3. This is done a number of times until a stopping criteria is satisfied.
4. The model predicts the output based on:
    - The average target/response values for all observations that fall in that subgroup
    - The class that has majority representation

Partitioning

Partitioning

CART uses binary recursive partitioning. The objective at each node is to find the best feature to partition the remaining data into one of two regions such that the overall error between the actual response and the predicted constant is minimized.

For regression problems, the objective function to minimize is the total SSE as defined below:

$$SSE = \sum_{i \in R_1} (y_i - c_1)^2 + \sum_{i \in R_2} (y_i - c_2)^2$$

Gini index

The Gini index is calculated as follows:

$$G = 1 - \sum_i p_i^2$$

where p_i is the probability of class i. The Gini impurity measures the frequency at which any observation in the dataset is misslabelled. It takes the values in the interval $[0,0.5]$.

Entropy

The Entropy is calculated as follows:

$$E = -\sum_{i} p_i log(p_i)$$

where p_i is the probability of class i. It takes the values in the interval $[0,1]$.

Gini index vs. Entropy

- Gini index takes the values in the interval $[0,0.5]$ whereas the interval of Entropy is $[0,1]$.
- Entropy is computationally expensive because it consists the logarithm step.
- Maximum purity of the Gini index and Entropy is 0.
- Minimum impurity of the Gini index is 0.5, where 1 for Entropy.

Hyperparameters

Hyperparameters of decision trees

- minimum number of observations required to attempt a split (**minsplit**)
- complexity parameter (**cp**)
- maximum depth of a decision tree (**maxdepth**)

Problems in decision trees

- **Deep tree** (tree with too many splits)
  - Too complex, danger of overfit

- **Shallow tree** (tree with too few splits)
  - Predictions too coarse-gained

Pros and cons

- Interpretable
- Easy to understand
- Can be used both of regression and classification tasks
- Non-linear concept
- No need to transformation
- Can handle missing values
- Robust to outliers

- Large trees may be hard to interpret
- High variance causes the poor model performance
- May overfit easily

Application

See the R codes on the course GitHub repository!

The video recording of today's lecture will be available on YouTube, and slides on **GitHub**.
Feel free to contact me via e-mail: mustafacavus@eskisehir.edu.tr