Feature Engineering in Decision Trees

Estimated time: 5 minutes

Unlike linear models, which require feature scaling as part of the data preprocessing step, decision trees are invariant to feature scaling. This means that their performance is not affected by scaling or not scaling feature variables. They are not sensitive to outliers or the variance in the data. The splitting algorithm itself is indifferent to scale. Recall that it uses information gain and impurity measures to determine the optimal splits to make. These decisions are based on comparing a particular feature to some value, and it doesn't matter what scale the feature is on.

Categorical variables are handled differently in different models and different programming languages. With scikit-learn, categorical variables have to be encoded for ML algorithms. The problem is that, for decision trees, one-hot encoding categorical features is very inefficient, because it creates a sparse matrix in the dataset where there are many more columns full of zeros.

Another approach is label encoding:

This approach converts categorical variables to integers. For example, it might convert the lifestyle column to numbers, but this implies a ranking order where "weight trainer" is greater than or more important than "athlete" or "cardio enthusiast", and this is not intended. If categorical data is not truly ordinal the tree will end up making splits that do not make sense, because the model will interpret these numbers as ranked.

	lifestyle
1	Cardio Enthusiast
2	Athlete
3	Weight Trainer
4	Sedentary

lifestyle
0
1
2
3