Decision Tree - Insights

Information related to code and method

A few points in respect of the Python code are given below:

(1) Each part of the code includes comments mentioning the activity step being performed.

(2) For creating decision tree, class 'DecisionTreeClassifier' uses criterion “gini” for Gini impurity (by default), and considers a binary split for each attribute.

(3) Postpruning approach of ‘minimal cost complexity pruning’ has been used to create the pruned decision tree.

(4) The pruned decision tree is created by using the value of effective 'alpha' for which accuracy of the decision tree model is maximum on testing set.

(5) The root node is for criterion ‘Contract’ as ‘Month-to-month’.

Some insights/rules to identify which customers are likely to churn (not churn)

(1) For the root node, the pruned decision tree has:

Contract_Month-to-month <= 0.5

gini = 0.391

samples = 4930

value = [3614, 1316]

class = Churning_No

This indicates that:

(a) The split criterion is “Contract_Month-to-month” and the split value is 0.5.

(b) The gini score for this split is 0.391.

(c) There are 4930 samples in this node.

(d) There are:

- 3614 samples belonging to “Churning_No” class.
- 1316 samples belonging to “Churning_Yes “class.

(e) The predicted class for this node is “Churning_No”.

The variable ‘Contract_Month-to-month’ has two values (binary) after one-hot encoding, as:

(a) 0 (indicating absence of ‘Contract’ as ‘Month-to-month’).

(b) 1 (indicating presence of ‘Contract’ as ‘Month-to-month’).

As this criterion is “Contract_Month-to-month <= 0.5”, customers who do not have ‘Contract’ as ‘Month-to-month’ (i.e. “Contract_Month-to-month = 0”) are predicted to belong to class “Churning_No”.

Thus, customers who do not have ‘Contract’ as ‘Month-to-month’ are predicted to not churn.

(2) Based on the pruned decision tree, the customers that are predicted as likely to churn are with the following attributes and values:

(a) tenure <= 13.5

(b) TotalCharges <= 120.0

(c) MultipleLines_Yes <= 0.5

(d) MonthlyCharges <= 76

(3) Based on the pruned decision tree, the customers that are predicted as not likely to churn are with the following attributes and values:

(a) Contract_Month-to-month <= 0.5

(b) MonthlyCharges <= 93.75 (excluding MonthlyCharges <= 76, as mentioned above)

(c) InternetService_Fiber optic <= 0.5

(d) OnlineBackup_No <= 0.5

(e) Contract_Two year <= 0.5

(f) tenure <= 3.5

(g) TechSupport_No <= 0.5

(h) MultipleLines_No phone service <= 0.5

(i) PaymentMethod_Electronic check <= 0.5

(j) tenure <= 22.5 (excluding tenure <= 13.5, as mentioned above)

(k) tenure <= 43.5

(l) TotalCharges <= 3083.85 (excluding TotalCharges <= 120.0, as mentioned above)

(4) It is noteworthy that demographic criteria such as 'gender', 'SeniorCitizen', 'Partner' and 'Dependents' do not appear as a node in the pruned decision tree.

(5) Criterion 'tenure' appears the most (four times) as a node in the decision tree. This may indicate that tenure is an important factor that can have an impact on whether a customer is likely to churn or not churn.

(6) Criteria 'TotalCharges', 'MonthlyCharges' and 'MultipleLines' also appear twice each. Hence, these three may also be important factors for determining whether a customer is likely to churn or not churn.

<u>Comparison of these findings with the exploratory data analysis performed on this data set earlier</u>

(1) In the exploratory data analysis, the demographic criteria such as 'gender', 'SeniorCitizen', 'Partner' and 'Dependents' appeared to play a role in determining whether a customer is likely to churn or not churn. This is in contrast to the demographic criteria not appearing as a node in the pruned decision tree.

(2) The exploratory data analysis indicated that for attributes related to services availed, customers having multiple lines with fiber optic are most likely to churn. The nodes in the pruned decision tree appear to agree with it.

(3) Exploratory data analysis had indicated that for payment related attributes customers having payment method as electronic check are most likely to churn. The related node in the pruned decision tree appears to agree with it.

(4) The exploratory data analysis indicated that for billing related attributes customers having contract as month-to-month are most likely to churn. The related node in the pruned decision tree appears to agree with it.

(5) Exploratory data analysis had indicated that for technology support related attribute customers with no tech support are not likely to churn. The related node in the pruned decision tree appears to agree with it.

(6) The exploratory data analysis indicated that in respect of the attributes related to data protection, online security and online backup, the customers that were most likely to churn are those that have no internet service. The related node in the pruned decision tree appears to agree with it.

(7) Attributes related to consumption (entertainment) patten such as 'StreamingTV' and 'StreamingMovies' appeared to affect whether a customer is likely to churn or not churn. This is in contrast to such criteria not appearing as a node in the pruned decision tree.

==========