Data analyst assignment:

Question 2:

Naïve Bayes classifier:

**Accuracy**: 0.316

**ROC AUC**: 0.529

**Confusion matrix**:

[[ 938 9887 956 0]

[ 1275 15148 1668 0]

[ 1256 13743 1717 0]

[ 796 8109 862 0]]

**FP** = [ 3327 31739 3486 0]

**FN** = [10843 2943 14999 9767]

TP = [ 938 15148 1717 0]

TN = [41247 6525 36153 46588]

Decision Tree classifier:

**Accuracy**: 0.288

**ROC AUC**: 0.508

**Confusion matrix**:

[[3887 5569 1634 691]

[6128 5702 5238 1023]

[4033 5924 4446 2313]

[2448 2505 2634 2180]]

**FP** = [12609 13998 9506 4027]

**FN** = [ 7894 12389 12270 7587]

TP = [3887 5702 4446 2180]

TN = [31965 24266 30133 42561]

The NB classifier did not predict any entry to belong to class 4.

Both classifiers have a very low accuracy score. While the classifiers' accuracy improves slightly when using only the original vars (Snapshot Date, Checking Date, DayDiff, Hotel Name, WeekDay), the NB classifier loses the predictions of class 1 on top on class 4.

The classifiers shown here are based on the original vars, in addition to the date vars separated to their components (year, month, day).

The tree classifier was also tried with the addition of converted dates to a cyclical format (month and day) to consider the cyclical nature of theses vars, but the scores did not improve.

Each var was checked on its own to understand its contribution. Most classifiers predicted only class 2, which was taken to assume this variable is not a good predictor.

In the tree classifier, Hotel Name predicted all classes, while Checkin Date and WeekDay predicted classes 1-3, and so are considered the most important features accordingly. The removal of WeekDay and Checkin Date indeed improved the accuracy to 0.43 and the ROC AUC to 0.74. By looking at the feature importance, Hotel Name was indeed the most important feature, though it was followed by Snapshot Date and DayDiff, which did not appear to predict the class well.

In the NB classifier, Snapshot month, Checkin year, and Checkin month were able to predict more than one class (classes 1 and 2). No combination of vars resulted in predictions of class 4 and removing parameters did not improve scores. This aligns with the features being treated independently by the algorithm, and the class chosen is the one with the highest probability.

As some of the vars are clearly dependent (e.g. Checkin Date, Checkin year, Checkin month, and Checkin Day), it explains the inability to improve this classifier.

As the NB classifier works with categorical features, an improvement can be made using further preprocessing and grouping of some of the parameters. For example, DayDiff can be group by percentiles, and Hotel Name can be grouped by location (e.g. by neighborhood).

It seems to me that although the tree-based algorithm can treat the numeric variables as numeric and give them relatively more weight, in the end they seem to not be good predictors for the class, which explains the low scores when using them.

A clear advantage of the tree-based classifier is its ability to predict all classes, and its ability to improve when picking the more important features.

Question 3:

By looking at the clusters (image below), we can see 5 clusters with a similar pricing policy. When choosing a higher threshold (at threshold 838), the two top clusters (brown and purple) combine to a single cluster, and the two bottom clusters (green and orange) combine to a single cluster (at threshold 839). This achieves three similarly sized major clusters of hotels with similar pricing.

From this we can conclude that the hotels' pricing policy divide to three major groups, while two of these groups can be further divided each to two groups by more closely related pricing policies, with these differences being similar between the major groups.

Better understanding of this clustering can be achieved by possibly getting more information about the hotels, such as location or chain. My assumption is that hotels belonging to the same chain will have similar pricing, while also hotels located in the same areas/near the same landmarks.

Diagram, schematic

Description automatically generated