Gender Prediction from SuperMarket Sales Data

Utpalraj Kemprai, Aryan Joshi

March 2024

1 Introduction

The SuperMarket Sales dataset contains data for a supermarket with three branches. We attempt to build two classifiers to predict Gender, one using a decision tree and one using a random forest.

2 Exploratory Data Analysis

We carried out our analysis in the following manner:

1. We first checked the datatypes of all columns in the data.
2. We checked for missing values and found there were no missing values in the data
3. We plotted the correlation Heat map for numeric columns and found that Tax and Total had perfect correlation.
4. We checked for outliers by standardising all the numeric columns and then plotting boxplot diagrams for them. We found a few outliers in Tax and Total (some values lied outside outlier fences). Since these two are perfectly correlated, the outliers corresponded to the same rows in the dataset.
5. We checked if the data is imbalanced. Our dataset had 496 males and 495 females. So the data was not imbalanced.

3 Data Preprocessing

- We encoded the non-numeric columns using One Hot Encoding.
- We dropped rows that had Total value outside the upper bound of the box-plot. There were 9 such rows. There were no values below the lower bound.

- We dropped the InvoiceID in the training data as it was clearly independent of Gender.
- We also dropped the Tax column as it had perfect correlation with Total and hence, dropping it would not result in any loss of information.
- We split the data into train and test data (80% for Training and 20% for testing).

4 Training the models

4.1 Decision Tree Classifier

4.1.1 Approach

We performed Hyperparametric Tuning for our Decision Tree classifier on the criterion(function to measure quality of split), max depth, Cost complexity pruning parameter(ccp_alpha) and max_features(maximum number of features to consider) with 3 fold cross validation.

4.1.2 Results

1. The best parameters found were ccp_alpha=0, criterion='gini', max_depth=5 and max_features = 'log2', no. of estimators= 175
2. The accuracy on the training data was 0.607
3. The accuracy on the test data was 0.517
4. The confusion matrix for the test data is $\begin{pmatrix} TP & FP \\ FN & TN \end{pmatrix} = \begin{pmatrix} 56 & 27 \\ 69 & 47 \end{pmatrix}$ where N = Male and P = Female

4.2 Random Forest Classifier

4.2.1 Approach

- We performed Hyperparametric tuning for the Random Forest classifier on the number of decision trees, max_features,max_depth.
- We use Gini index as our criterion.

4.2.2 Results

1. The best parameters found were max_depth = 3, max_features = 'log2', n_estimators = 200.
2. The accuracy on training data was 0.636
3. The accuracy on test data was 0.527

4. The confusion matrix for the test data is $\begin{pmatrix} TP & FP \\ FN & TN \end{pmatrix} = \begin{pmatrix} 66 & 17 \\ 77 & 39 \end{pmatrix}$ where N = Male and P = Female

5 Comparing results for the two classifiers

In our case of Gender classification we are typically interested in knowing whether the buyer is male or female. Our interest typically is not finding all male buyers of female buyers. Also as the test data is not imbalanced (proportion of male and female instances about the same), we care equally about both True Positives and True Negatives. So accuracy is a suitable metric for comparing their performance.

Decision Tree had an accuracy of 0.517 on test data while Random Forest had 0.527. So the Random Forest fares a bit better than the Decision Tree although ever so slightly.