**INVESTIGATE CUSTOMER BUYING PATTERNS**

**Appendix**

**Preprocessing**

The dataset was import into Rapidminer v9.5. The data type of variable “in.store” changed from numeric to binary, also variable “region” changed from numeric to polynomial. After exploratory analysis, the variable “age” was discretized for the final machine learning and decision tree analyzes. The age groups are: young people between 18 and 28 years old, adults between 29 and 42 years old, adults between 43 and 55 years old, elderly between 56 and 74 years old and seniors over 75 years old. The variable “amount” was either log10 transformed or discretized into 3 bins. The correlation matrix was also analyzed.

A decision tree was built to verify the variation in the amount spent and the ages of customers between regions. For a better understanding of the factors that may influence a customer’s decision to buy online or in one of our stores, a machine learning was built with different classification algorithms: decision tree, random forest, gradient boosting trees, id3, and logistic regression. The F-metric was used to identify the best model. Another machine learning was built in order to verify whether amount spent is related to the number of item purchased.

**Topic 1: regional variations in amount spent and age of customers**

Results show that the amount spent per transaction varies across regions. Checking the median, customers spent more money in the central region and spent less in west region (Figure 1A). They also show that customers’ age varies slightly between regions. There are more old people in the West region and young people in the Central region.

The decision tree (Figure 2) shows that customers who bought our products in the central region tend to spend more in our stores (> $1000) than online (> $502.34). In the western region, people older than 29 years old spent online up to $500. In the southern region, people between 56 and 74 preferred to visit our stores. The eastern region, young people spent online up to $70.47, but overall people bought online up to $501.

If the data is right, we could improve our business by expanding to the west region, as we did not get good economic results. We could provide offers to older people as a target audience.

**B**

**A**

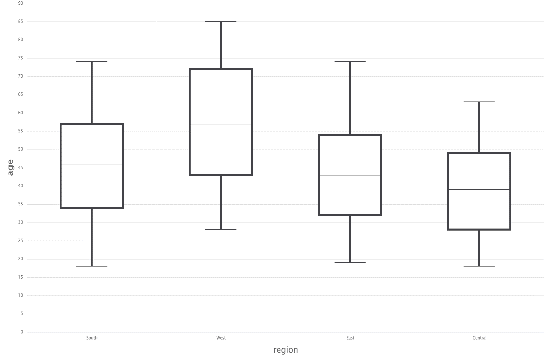
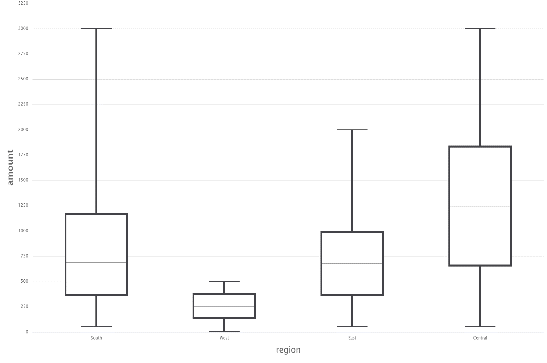


Figure 1. (A) Region categories are on the x-axis (South, West, East and Central, respectively) and the amount spent per transaction on the y-axis. (B) Region categories are on the x-axis (South, West, East and Central, respectively) and the age on the y-axis.

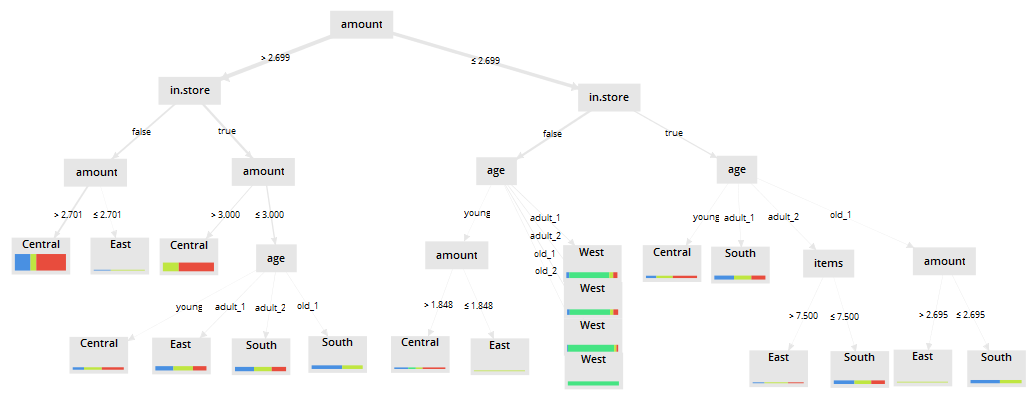


Figure 2. Decision tree for region as label and other variables (amount spent log10-transformed, customers’ age discretized, number of item purchased and place of purchase (online/store).

**Topic 2: correlation of age of customers and place of purchase (online/store)**

The correlation between age of customers and place of purchase shows that there is not a real correlation (ρ = -0.176), but the Figure 3 shows that old people (>= 57 years old) prefer to buy online and young people (<= 28 years old) prefer to buy in store.

The best model of machine learning was with decision tree algorithm (F = 84.97). You can check the performances of the other algorithms in table 1. The decision tree shows that customers’ age could help in identifying the pattern of consumption in the central region (Figure 4). For instance, people between 18 and 55 years old preferred to visit our stores, while a portion of older people prefer to shop online. In the other regions, the amount spent is a better classifier.

Figure 3. Place of purchase (store or online) is on the x-axis and ages are on the y axis.

Table 1. Performance of different algorithms.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Metric (%) | Decision tree | Random forest | Gradient Boosting trees | id3 | Logistic regression |
| Precision | 73.88 | 74.15 | 74.35 | 74.37 | 74.54 |
| Recall | 99.98 | 99.41 | 96.20 | 96.84 | 82.24 |
| F | **84.97** | 84.94 | 83.87 | 84.13 | 78.20 |

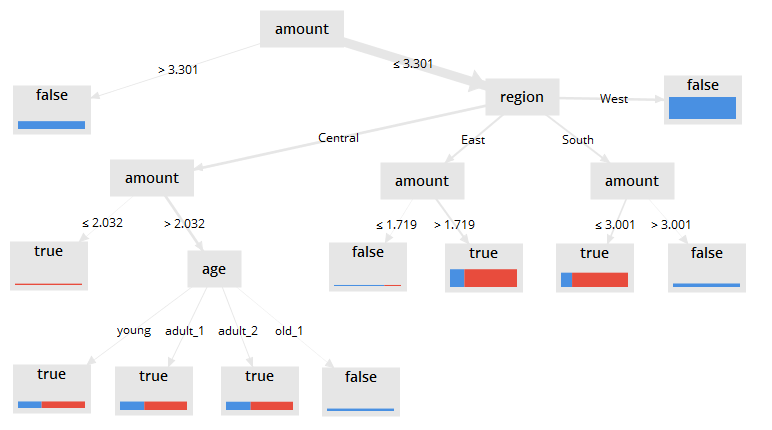


Figure 4. Decision tree for place of purchase (online/store) as label and other variables (amount spent log10-transformed, customers’ age discretized, number of item purchased and region.

**Topic 3: relationship between number of items purchased and amount spent**

The correlation between number of items purchased and amount spent shows (ρ = -0.002) that there is not relationship between them. For this reason, we cannot use one of this variable to predict another variable.

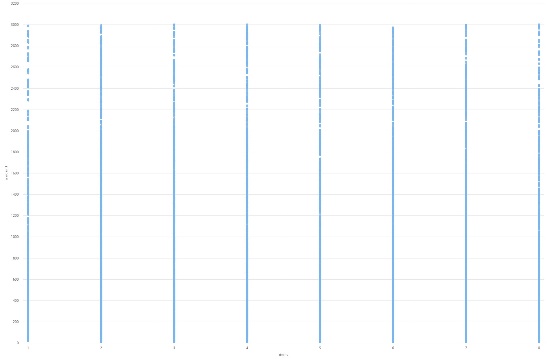


Figure 5. Relationship between number of items purchased (x-axis) and amount spend (y-axis).