Coffee data analysis

Wojciech Sarnowski

1. Task

In this task you will analyze coffee brewing data.

Attached document coffee_data.csv contains data gathered from cupping sessions which contains such parameters as:

```
brewing time,
brewing temperature,
grinding level,
TDS,
water pH,
processing method,
region and plantation height.
```

Each coffee was rated from 1-5. Your task is to tell which parameters affects brewing process the most.

Interpretation

By affecting the brewing process we understand affecting the coffee rating, a target variable, represented in data set by column "mark". Using only above-mentioned parameters we want to produce a ranking (known as variable importance) which reflects their predictive power. Variable importance score will be extracted from a machine learning model. The model will be constructed with the help of the xgboost algorithm. Taking into account that "mark" is a multiclass target (5 levels) the softmax objective will be chosen.

Problem will be solved using the R package.

2. Preliminary data analysis

Data available in the coffee_data.csv file consist of 500 records and 11 columns (10 predictors plus target variable). However only 8 predictors will be used to build a model and assess the variable importance. Let's look at top 10 records of the data.

```
##    brewing_time brewing_temp grinding_level  TDS water_ph
## 1           189         91.1              7   NA     6.67
## 2            95         94.0              2 0.14     3.00
## 3           147         91.6              4 0.13     5.51
## 4           158         92.3              4 0.13     7.11
## 5           178         98.9              3 0.11     6.06
## 6           169         88.9              5 0.15     8.93
## 7           166         88.8              5 0.09     5.12
## 8           119         85.2              2   NA     3.99
## 9           151         92.8              5 0.14     5.73
## 10          142         92.0              4 0.14     6.29
##    processing_method   region plantation_height mark
## 1              Honey   Panama              1710    3
## 2              Honey   Panama              1530    3
## 3              Honey Colombia              1370    3
## 4              Honey   Rwanda              1630    3
## 5             Washed   Panama              1400    3
## 6              Honey    Kenya              1540    5
## 7            Natural Ethiopia              1470    5
## 8             Washed   Panama              1080    4
## 9              Honey   Panama              1540    5
## 10             Honey    Kenya              1780    3
```

It is obvious that "region" and "processing_method" are categorical variables. We can see also that "TDS" has some missing values.

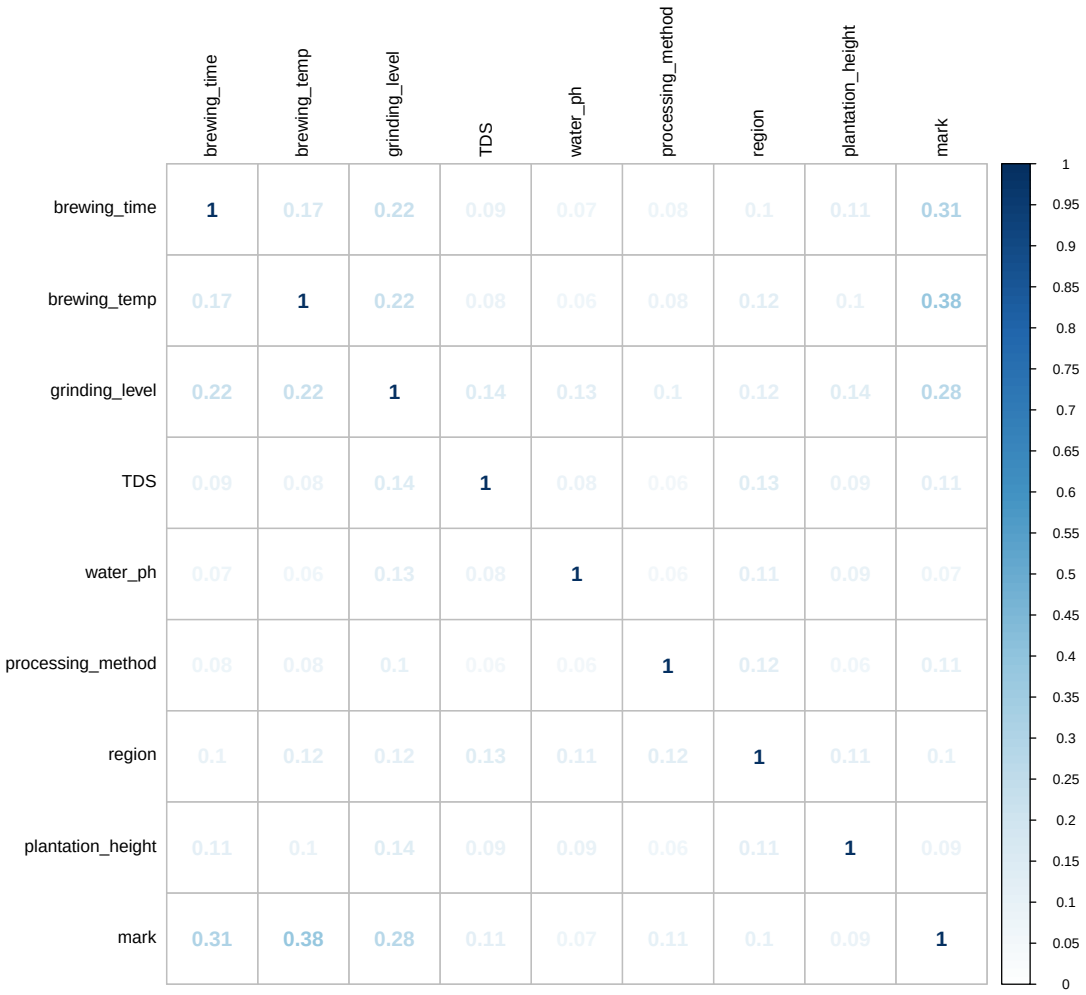
To get more intuition about the data we present the table with basic statistics for each variable. It turns out that only "TDS" contains missing values (almost 20%). Probably "grinding_level" is an ordinal variable (9 scores 1, 2, ..., 9 assigned to levels of grinding) as well as target variable "mark" (but this was known in advance). The other columns are numerical variables. These facts are crucial since R's implementations of xgboost algorithm cannot work directly with categorical data. Hence, before modelling, some feature engineering should be done. One-hot encoding will be applied to categorical predictors. Special treatment of missing values is not required.

Table 1: Basic statistics for variables (continued below)

variable	distinct_values	fraction_missing_values	min	mean
brewing_time	135	0	60	164.8
brewing_temp	117	0	83	91.39
grinding_level	9	0	1	4.744
TDS	15	0.198	0.05	0.1192
water_ph	328	0	3	6.304
processing_method	4	0	NA	NA
region	8	0	NA	NA
plantation_height	101	0	700	1406
mark	5	0	1	3.194

median	max	variance
166	240	1126
91.3	100	8.389
5	9	2.195
0.12	0.2	0.0005153
6.285	10	1.808
NA	NA	NA
NA	NA	NA
1400	2000	49420
3	5	1.58

At the stage of preliminary analysis we also explore the correlation matrix of the data set. Since we deal with different types of variables the v-Cramer measure of association will be used. The figure below shows the correlation structure. As we can see there is no substantial correlation between predictors. However there are some associations within the triple "brewing_time", "brewing_temp", "grinding_level". It is also reasonable to investigate correlation between predictors and "mark" variable. Such the correlation reflects the simple importance of each predictor (interaction between predictors is not taken into account). In this approach the strongest variable is "brewing_temp", next "brewing_time" and "grinding_level". Rest of predictors is significantly weaker. The smallest impact on target has "water_ph".



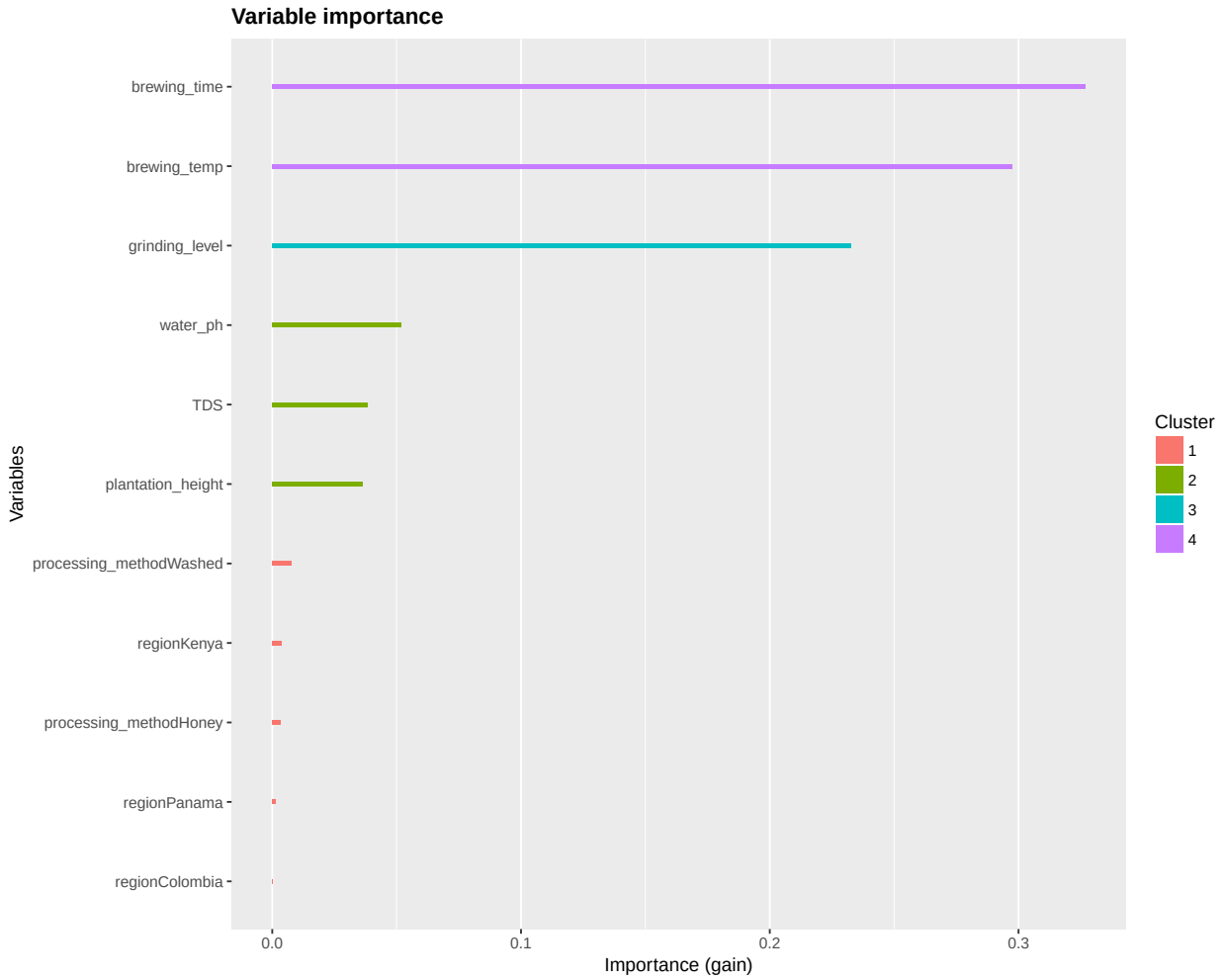
3. Model-based approach for variable importance

We start with building the xgboost model to capture the relationship between predictors and the target variable. Because of small sample size the cross validation method for data splitting will be applied. The table presenting target structure shows that marks are imbalanced. We will assign weights to the training observations to get equally represented marks during learning process.

Table 3: Structure of the target variable

mark	count	percentage
1	52	0.104
2	99	0.198
3	149	0.298
4	100	0.2
5	100	0.2

Fitted model generates the following ranking of importance.



According to the figure variable "brewing_time" contributes most to explain the target. The second place occupies "brewing_temp". Next variable, "grinding_level", has a bit lower gain statistics than brewing time and temperature. These parameters have the biggest influence on "mark" variable (the same set of the best

predictors indicates v-Cramer statistics however with different order). Next three variables "water_ph", "TDS", "plantation_height" are significantly weaker. Remaining parameters, "processing_method" and "region" can be neglected. Notice that these variables are represented separately by their each attribute. This is because of using one-hot encoding approach to categorical variables. We can also see that not all attributes of "processing_method" and "region" are included in the model. Regularization imposed on the xgboost algorithm to decrease overfitting eliminates some of them.

REMARK. To prepare the importance ranking the gain statistics was used. However there are available other metrics. All metrics generated by R function are presented in the table below. Some difference in variables ordering can be found, but top 4 variables are the same regardless of the criterion used.

Table 4: Variable importance - different metrics

Feature	Gain	Cover	Frequency	Importance
brewing_time	0.3268	0.2753	0.2829	0.3268
brewing_temp	0.2974	0.2747	0.2432	0.2974
grinding_level	0.2326	0.1934	0.182	0.2326
water_ph	0.05187	0.1008	0.1009	0.05187
TDS	0.03848	0.04781	0.06667	0.03848
plantation_height	0.0362	0.07032	0.08468	0.0362
processing_methodWashed	0.00772	0.011	0.01441	0.00772
regionKenya	0.00388	0.00867	0.01081	0.00388
processing_methodHoney	0.00334	0.00803	0.00721	0.00334
regionPanama	0.00129	0.00784	0.00541	0.00129
regionColombia	0.00038	0.00211	0.0018	0.00038