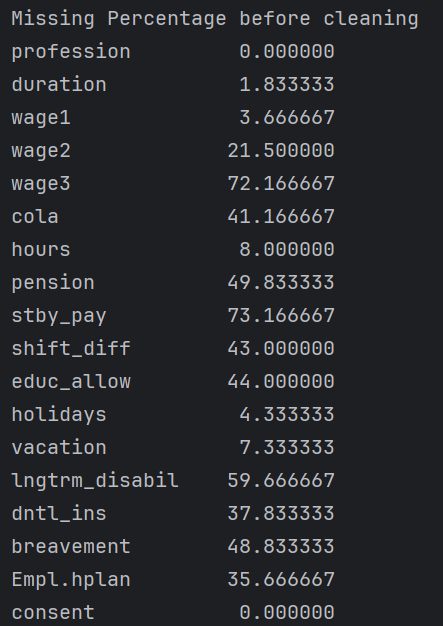
**Data Mining Exercise 3**

Magdalena König

01455794

1. Labor negotiation

* Task 1: Data cleaning & exploration

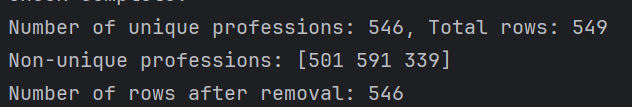


* Ein Bild, das Text, Screenshot, Schrift enthält.

  Automatisch generierte Beschreibungall columns that have a value of more than 40% missing values will be deleted
* The still missing values will be filled with the medium (numerical columns) and most frequent ones (categorical columns)

Rows with unexpected values are deleted in the categorical columns (and consent).

Also those rows are deleted that don’t have a unique professions in order to identify the workers clearly:



Ein Bild, das Diagramm, Kreis enthält.

Automatisch generierte BeschreibungEin Bild, das Text, Screenshot, Diagramm, Reihe enthält.

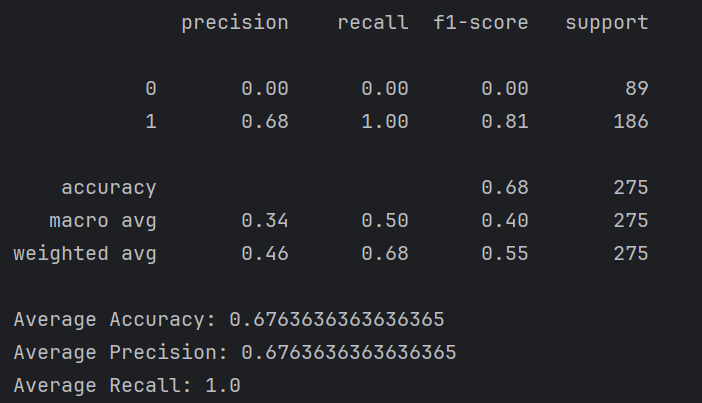
Automatisch generierte Beschreibung

* Task 2: Classification

Logical Regression was chosen here 🡪 statistical method to determine the properbility of the occurance of a dependent variable (consents) in combination with one or more independet factors (vacation, hours, wage,…).

* The training data was unbalanced (consent 0 (bad/unaccepted) was way less frequent than 1 (good/accepted).

When unbalanced there was the following result:



Results with balancing strategies:

* SMOTE (SMOT creates synthetic training data for the underrepresented groups)

Ein Bild, das Text, Screenshot, Schrift enthält.

Automatisch generierte Beschreibung

* Oversampling (Oversample the bad consent rows so its more balanced)

Ein Bild, das Text, Screenshot, Schrift enthält.

Automatisch generierte Beschreibung

* Undersampling (undersample the overrepresented data – good consent)

Ein Bild, das Text, Screenshot, Schrift enthält.

Automatisch generierte Beschreibung

* To see a good effect every model was ran 10 times (the classification report was computed at the last run) – and the average accuracy, precision and recall was taken.
* The balanced models have a worse accuracy than the unbalanced one, however the model using SMOTE showed the best precision – a balanced sample would surely make sense here, even if the accuracy is lower.

2. Flight data

* Task 1: Data cleaning and preperation
  + Cancelled flights are deleted (rows)
  + The column cancelled are deleted
  + Each delay column was transferred from NaN to 0 (however I could have deleted those ones from the beginning)
  + Rows with missing values were deleted



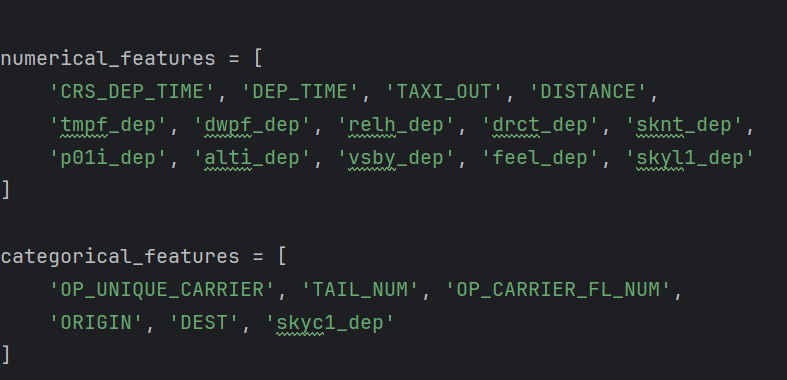
* + Later after the merge there was some more data cleaning necesarry (since the merging was a really time-consuming process it was done after merging in a second data cleaning step
  + Rows with the vlaue “M“ (for missing) were set to NaN
  + All Rows with NaN values were deleted as well as those columns deleted that had a precentage of more than 30% of missing values
  + Time and Date (date and wheels off) is converted to datetime and cleaned if necessarry + a new column for these values
  + Weather stations’s valid column is also converted to datetime
  + Unnecesarry weather column was dropped (lan and lat)
  + Wheels off times and the time from the weatherstation are rounded to 10 minutes in order to make them match better
* Ein Bild, das Text, Screenshot, Software enthält.

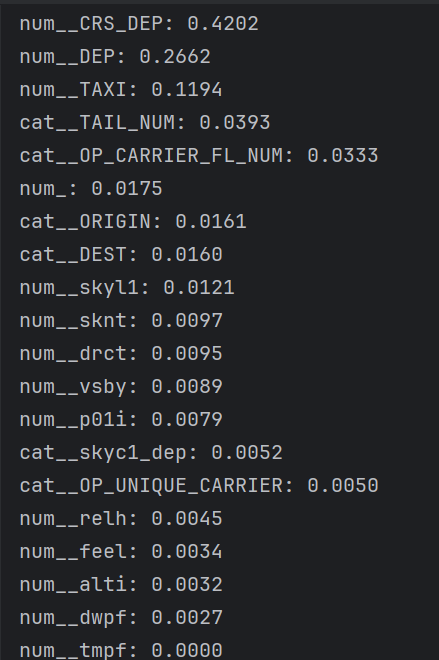
  Automatisch generierte BeschreibungTask 2: Merging

The merging was incredibly time consuming and took about 2h.

~5100 rows were in the merged dataset

* Task 3 – Classification
  + First step was to devide to determine which features make sense
  + Then they were divided into numerical and categorical features



After that the information gain of every feature was determinded with the following result:  


* After that a decision tree classification was created – first with all features, no matter which information gain the have:

Ein Bild, das Text, Screenshot, Schrift enthält.

Automatisch generierte Beschreibung

* All features with a information gain with more than 0.08:

Ein Bild, das Text, Screenshot, Software, Schrift enthält.

Automatisch generierte Beschreibung

* The best 5 features according to their information gain:

Ein Bild, das Text, Screenshot, Software, Multimedia-Software enthält.

Automatisch generierte Beschreibung

* The worst 5 features according to their information gain:

Ein Bild, das Text, Screenshot, Schrift enthält.

Automatisch generierte Beschreibung

When it comes to Accuracy the model with the worst features is showably lower than the other three. Accuracy- and precision-wise the model with only the top 5 features as parameters is the best, however only slightly. Between all features as parametes and only those over an information gain of 0.08 there are almost no differences. However what’s also worth meantioning is that the dataset is not balanced and with a balancing strategy the result could be different.