SAKI SS 2021 Homework 1

Author: Stephanie Mehltretter

Program code: https://github.com/defaultUser3214/saki-21-homework

Summary

Introduction:

The fully automated classification of financial transactions may enable an analysis on the buying preferences of the account owner and thus also indirectly assign the account owner to certain marketing target groups. This report elaborates a way to classify the kind of financial transactions using a machine learning model based on a Naïve Bayes classifier. The data set for this task consisting of 209 labelled transaction information was provided by adorsys GmbH & CO KG and the chair for Open Source Software at FAU.

Development process:

The development process implements the first four process steps of the CRISP-ML standard [1]:

Business & Data Understanding > **Data Preparation** > **Modeling** > **Evaluation**

Business & Data Understanding: After loading the data set with a Pandas data frame, it was evaluated, which columns of the data set provide meaningful information for the training of the machine learning (ML) model:

Column	**Description**	**Relevance for model building**
Waehrung	Contains only one value	Not relevant
Unamed 0	Increasing enumeration of the rows	Not relevant
Auftragskonto	Contains *missing values* and two values *89990210.0* and *89990201.0*	*89990210.0* shows a significant correlation the label *leisure* → relevant
Buchungstag & Valutadatum	Are different only in one single row. Column *Buchungstag* has a correlation with the label, e.g., transactions with label *leisure* are more frequently booked on Mondays or Tuesdays	Sufficient to use only one column like *Buchungstag*
Kontonummer & BLZ	There are no entries of the column *Kontonummer* that belong to more than one *BLZ*.	Sufficient to use column *Kontonummer* and ignore *BLZ*
Betrag	Has a big variety of entries. Per label the number of different entries of Betrag reaches from 21 (label private) to 65 (label leisure)	Column Betrag may have impact on model performance
Buchungstext, Verwendungszweck	Contains text information indicating the label, e.g., keyword *Lohn* indicated belonging to label *income*	Relevant
Beguenstigter/ Zahlungspflichtiger	Contains text information indicating the label, e.g., keyword *Adorsys GmbH & Co. KG* indicates belonging to label *income*	Relevant

Data Preparation:

Remove irrelevant features: The feature columns *Waehrung, Unamed 0, Valutadatum* and *BLZ* were removed from the data frame as they do not provide relevant information for the training.

Feature extraction on date information: The ISO standardized weekday was extracted from the column *Buchungstag*, as it is more relevant for the training to know on which weekday the booking was performed than the date (which has a bug variety).

Reformatting numerical feature: The column entries of *Betrag* were reformatted to the English decimal format.

Text feature preparation: The text columns *Buchungstext, Verwendungszweck, Beguenstigter/Zahlungspflichtigter* were transferred into a matrix representation of token counts using the CountVectorizer from sklearn. Also, every word was converted to lowercase and typical words without a high information relevance, i.e. stop words, were removed. Therefore, lists for the languages English and German provided by the plugin stop-words were used.

Split into test and training data: A split into a training and test data set with a ratio of 80:20 is performed.

Rebalancing to minimize class imbalance: An analysis on the distribution of the labels showed a significant class imbalance, whereas, e.g., the classes *leisure* and *standardOfLiving* are frequent, whereas the classes *private* and *income* is barely present in the data set. To minimize the class imbalance by overfitting the underrepresented class, the package imblearn.over_sampling.SVMSMOTE is used.

Modeling:

Three GaussianNB classifiers from the package sklearn.naive_bayes are trained. One with the imbalanced training data set and the other with the rebalanced data set. The third classifier deals with the rebalanced data and ignores the column *Betrag*.

Evaluation

Metrics: To evaluate the model's performance, three metrices are considered:

1) The mean accuracy score from scikit-learn [2]: Describes how often the predictions match the labels in the test data set. Over 50 iterations the mean is calculated.
2) Confusion matrix: Visualizes the average precision, the recall and the f1-score as well the precision, the recall and the f1-score per class, which are elaborated in the following:
    a. The precision of a class is also known as the positive predictive value and describes the fraction of the correctly predicted members of the class among all the elements that were assigned to the class (= set of the correctly predicted class members and the ones that are incorrectly predicted as class member) by the model.
    b. The recall of a class points out the fraction of all correctly predicted class members among all the entities of the class (= set of the correctly predicted class members and the ones that are incorrectly predicted as not class members). Recall is also known as relevance.
    c. The f1-score combines precision and recall in the following way: $f1 - score = \frac{2*precision*recall}{precision+recall}$

Results: On the test data, **a mean accuracy score** of **~ 0.90** is archived by the Gaussian Naïve Bayes classifiers trained on the imbalanced, rebalanced and the rebalanced without the column *Betrag*. The confusion matrices of all three classifiers are equal. The **mean weighted f1-score** is **0.98,** the **mean weighted precision** is **0.98** and the **mean weighted recall** is **0.98**.

Limitations and future work: To further increase the prediction performance of the model, the amount of training data should be increased, and a harmonic class balance should be archived. Other data augmentation tequniques than SVMSMOTE may be evaluated. Also, further experiments should be conducted using other classifiers and different preprocessing methods, e. g., it may be promising to manually adapt the stop-word lists.

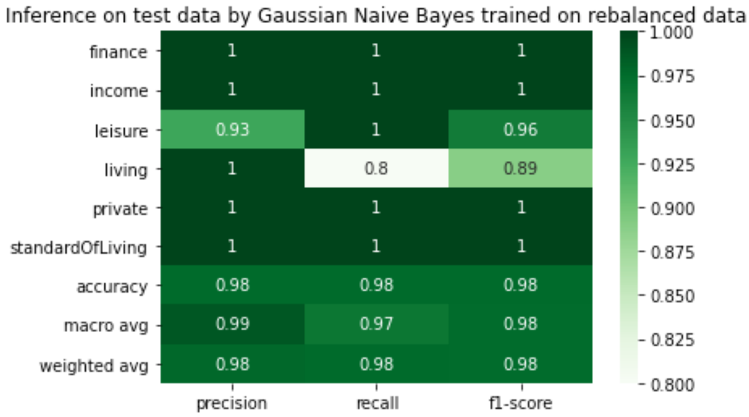
Sources:

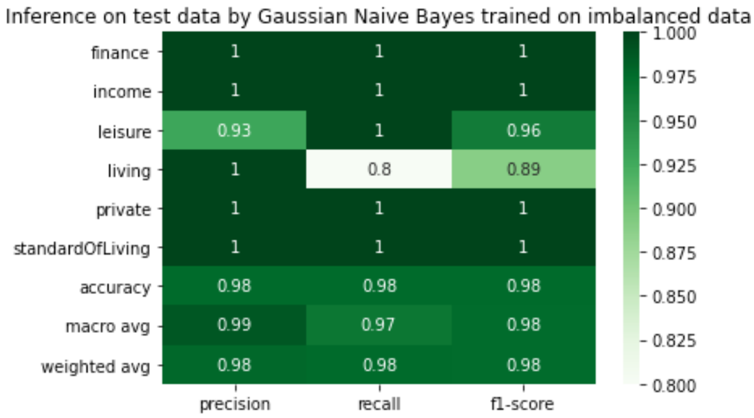
[1] Studer, S., Bui, T. B., Drescher, C., Hanuschkin, A., Winkler, L., Peters, S., & Müller, K. R. (2021). Towards CRISP-ML (Q): a machine learning process model with quality assurance methodology. *Machine Learning and Knowledge Extraction*, *3*(2), 392-413.

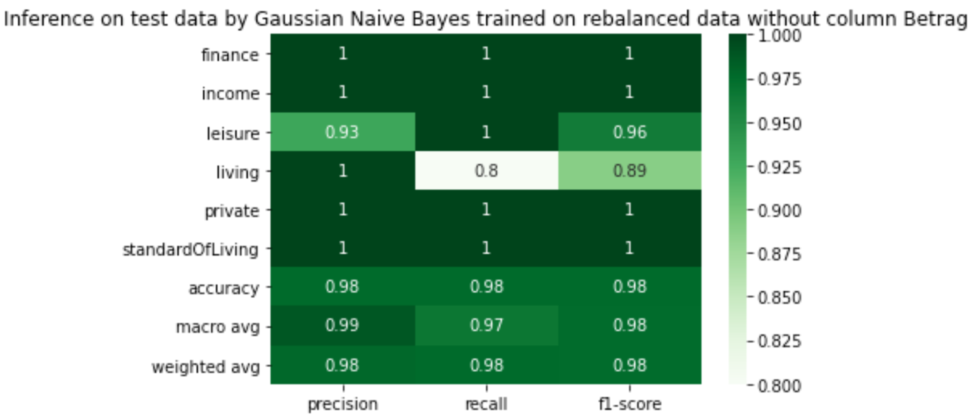
[2] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html, 09.05.2021

[3] https://towardsdatascience.com/beyond-accuracy-precision-and-recall-3da06bea9f6c#:~:text=The%20precise%20definition%20of%20recall,the%20number%20of%20false%20negatives.&text=Recall%20can%20be%20thought%20as,of%20interest%20in%20a%20dataset., 09.05.2021

Screenshots of the evaluation results








```
mean accuracy score (rebalanced training data): 0.8990476190476191
mean accuracy score (imbalanced training data): 0.8980952380952381
mean accuracy score (rebalanced training data without column Betrag):   0.8995238095238095
```