SAKI SS 2021 Homework 1 (Transaction Classification)

Silatsa Tchatchum Jovial(jovial.t.silatsa@fau.de)
University of Erlangen-Nuremberg
(Friedrich-Alexander-Universität Erlangen-Nürnberg)

May 11, 2021

Program code:`https://github.com/aj52izov/homework-1/releases/tag/homework-1`

1 Problem definition and solution approach

As the title says, the problem here is to classify clients transactions of a bank. So for a given client's transaction, we want to assign it to one of six categories(Income, Private, Living, StandardOfliving, Finance, Traffic, Leisure). Since the available dataset is labeled and with assumption that each new transaction is assigned to one and only one category, this problem is a supervised multi-class classification problem, and our goal is to build a classifier using a Naive Bayes algorithm to solve it.

2 Training and testing data-set

The data-set used consists of 209 bank transactions and each transaction below to one category. The data is splited in training (80%) and testing (20%) data-sets.

One of our main concerns when developing a classification model is whether the different classes are balanced. This means that the data set contains an approximately equal portion of each class. When the data set is imbalanced, a naive application of a model may focus on learning the characteristics of the abundant observations only, neglecting the examples from the minority class. The Figure 1 below show the number of transactions per category in our datasets.

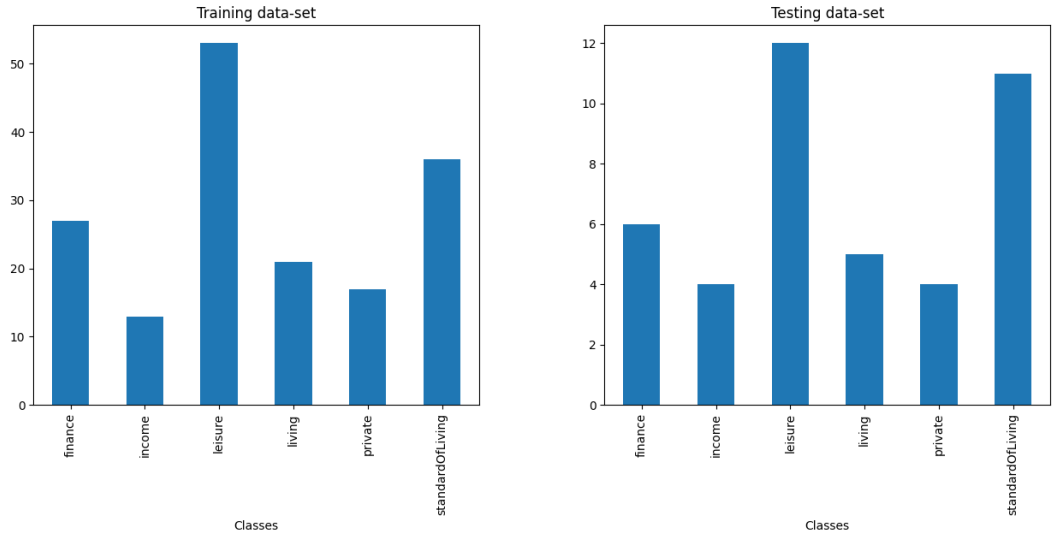

Figure 1: Datasets

looking at the figure 1 our training data-set is imbalanced. Therefore, the classifiers can have a difficult time classifying less common categories such as "Income" and will be more accurate for well-represented categories like Leisure and Standard-Ofliving. There are several ways of dealing with imbalanced datasets. However, the most commonly used approaches consist of undersample the majority class or oversample the minority one, so as to obtain a more balanced dataset. In this project, in oder to have an balanced training data-set, the minorities categories are reused multiple time. For example, transactions belonging to the category "Income" are used four times in the final training dataset. the Table 1 shows the number of times each transaction category has been used for the final training data-set and the Figure 4 shows the final training data-set.

3 Model Pipeline

The transactions of our dataset have characteristics and the proposed model uses three of them ("Buchungstext" , "Verwendungszweck", "Beguenstigter/Zahlungspflichtiger"). The output of the pipeline is the average of the outputs of four modules

consisting of a vectorizer and a classifier. Three modules take each of these characteristics and the last takes their concatenation as input. Each modul applies first the vectorizer then its classifier. The modules use the same vectorizer but different classifiers of the same architecture. The Figure 2 shows the model pipeline.

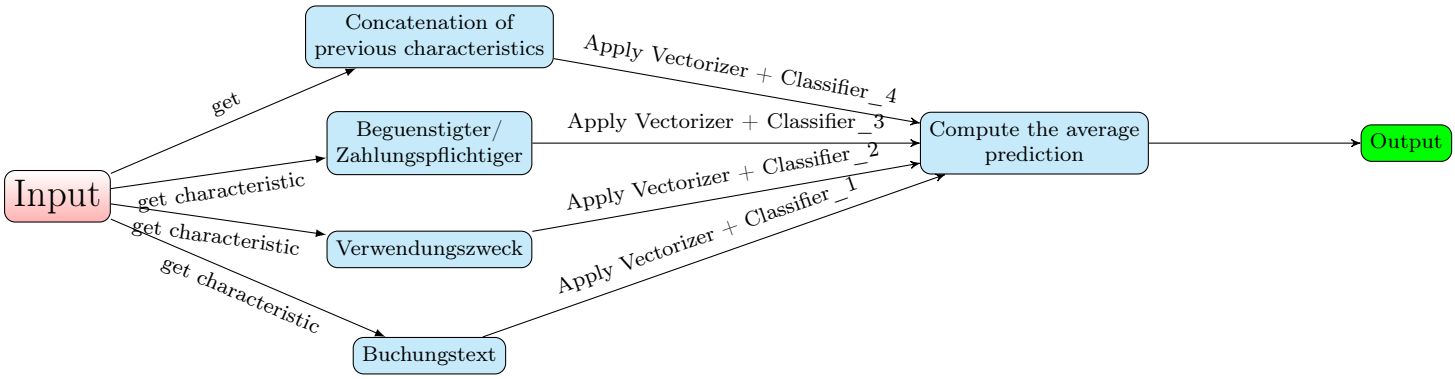

Figure 2: Model Pipeline

3.1 Pre-processing and Vectorization

The Pre-processing step here is essentially text cleaning that help to reduce the noise present in text data in the form of stop words, punctuations marks, suffix variations and the vectorization step, also called the feature engineering step, consists of transforming raw text data into feature vectors. The vectorization algorithm used is the TF-IDF (Term Frequency-Inverse Document Frequency) algorithm.

3.2 Classifier_{1,2,3,4}

since the problem is a multi-class classification problem that is a subset of mult-label classification, a problem transformation method is proposed to solve the task. The problem transformation method such as Binary Relevance transforms the classification problem into several single label problems. So, the one-hot-encoding algorithm is applyed to the input's label then a Classifier Chains compute the output. The Classifier Chains technique transforms a multi-label classification problem with L labels into L separate single-label binary classification problems. In addition, it takes label correlation into account. This approach uses a chain of classifiers where each classifier predicts the membership of one class and uses the predictions of all the previous classifiers as input. The total number of classifiers is equal to the number of classes. The base classifier used for the Classifier Chains is the Naive Bayes Classifiers.

4 Model evaluation

For the evaluation of the proposed model, its Accuracy is calculated. Informally, accuracy is the fraction of predictions our model got right. Furthermore, the accuracy of each class can be observed on the diagonal of the normalized confusion matrix. The Figure 3 hows the normalized confusion matrix.

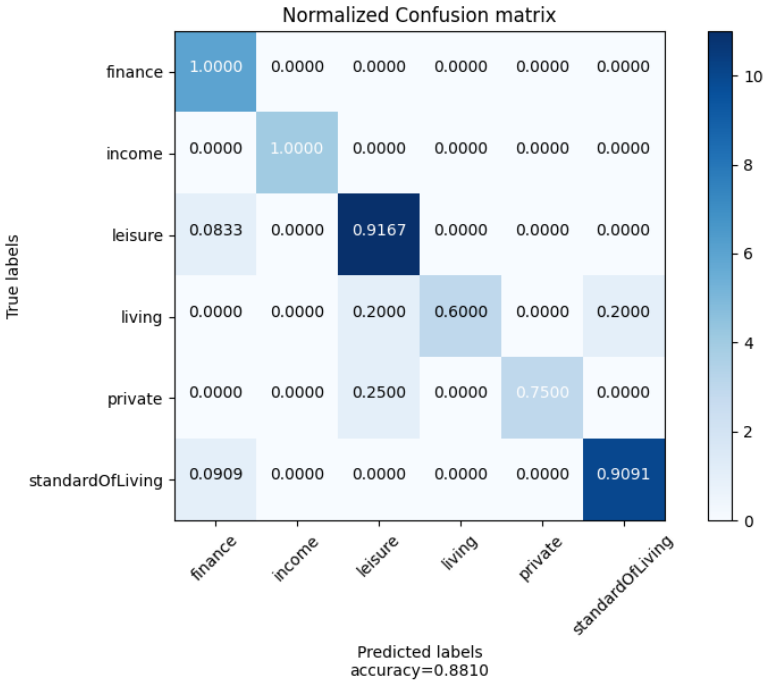

Figure 3: Classification Metrics

Cotegory	**Used**
Income	4 times
Finance	2 times
Living	2.5 time
Leisure	1 time
Private	3 times
StandardOfliving	1,4 times

Note: Used 0.4 time mind that 40% of the data was taked.

Table 1: Number of times the transactions of each category have been used

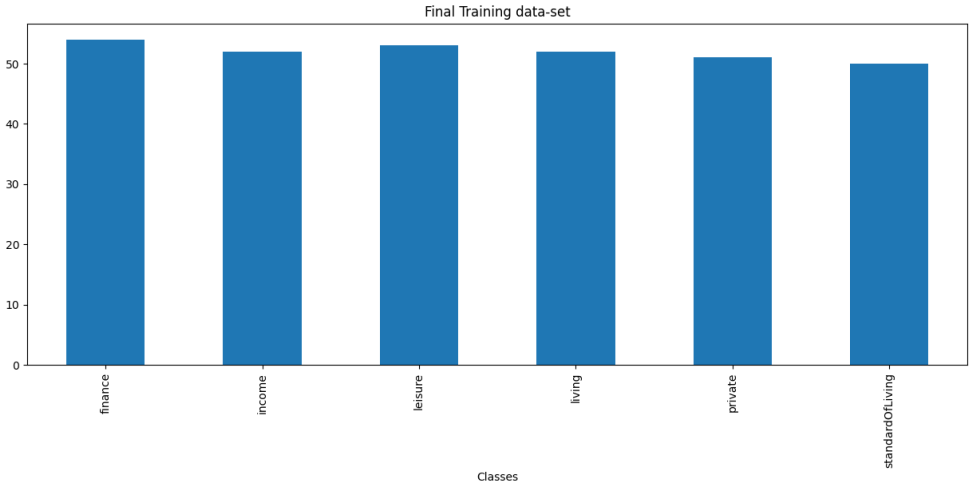

Figure 4: Final training Dataset