Exercise 0: Dataset description

Course: 184.702 Machine Learning WS2020
Submission deadline: 11:59 pm, 25.10.2020

Group 8:

- Alexander Leitner, 01525882
- Peter Holzner, 01426733
- Mario Hiti, 01327428

Dataset overview

dataset	type	samples	dimensions	missing [# / cols][1]	nominal cols	numerical cols
Polish companies	classification	~41k (6-10k/a)	64	~1,46%/ 64	-	64
Moneyball	regression	1232	14	3600 / 4	2	12

Tab. 1 Overview of chosen datasets.
[1]: Total number of missing values in dataset / affected columns.

We chose two datasets of wildly different sizes and dimensionalities, that both contained missing values. However, the missing values are distributed differently among all instances and columns in both sets.
We describe both datasets and how they compare to each other in more detail on the following two pages.

Polish companies bankruptcy (Classification)

UCI ML: https://archive.ics.uci.edu/ml/datasets/Polish+companies+bankruptcy+data

Predicting the success of an enterprise is one of the most important topics in economics. However, companies are complex entities that cannot be easily described by a mathematical level.
The “polish companies bankruptcy” dataset describes financial characteristics of various polish companies between 2007 and 2013. Each of the datasets contains 64 columns for attributes (labeled “Attr1”, ”Attr2”, …) and 1 column for classification (labeled “class)”. The classifier is either 0 if the company is solvent in the year 2013 or 1 if the company went bankrupt between 2007 and 2013. The attributes contain numerical information such as: Working capital, net profit / total assets, total costs / sales, total assets, etc.

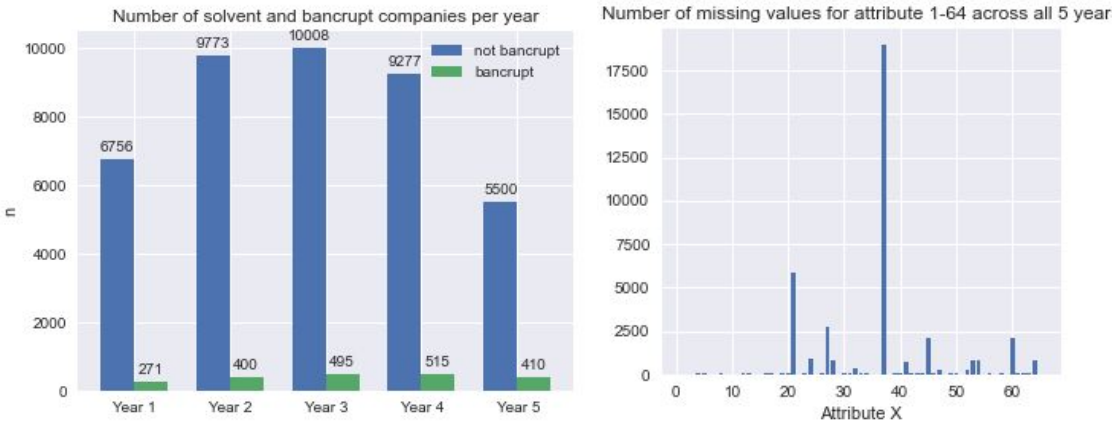

Fig. 1 Comparison of classified companies (left) and total number of missing values for Attr1-Attr64 across all years(right)

As seen in Fig. 1 right missing values are distributed unevenly with Attr21 and Attr37 lacking the most entries. The attributes also have wildly varying ranges, which will also have to be taken care of during preprocessing.

The companies do not have a unique identifier assigned. Therefore, it is not possible to determine if a company that did not go bankrupt in year 1, is also represented in year 2. For this reason, we will only focus on one of the five datasets for further analysis. Good candidates are either Year 4 (most bankruptcies total) or year 5 (highest ratio of bankrupt companies). In general, the classes are very unevenly distributed. Guessing ‘not bankrupt’ for every instance in Year 5 would yield a success rate of ~93%.

Comparison to other dataset

Compared to the second dataset (Moneyball) this dataset has:

- A large sample size (5k – 10k per year, ~41k in total)
- a high number of dimensions (64 attributes + 1 classifier)
- missing values (1.46% of the data is missing) distributed unevenly among almost all attributes

Moneyball (Regression)

OpenML: https://www.openml.org/d/41021

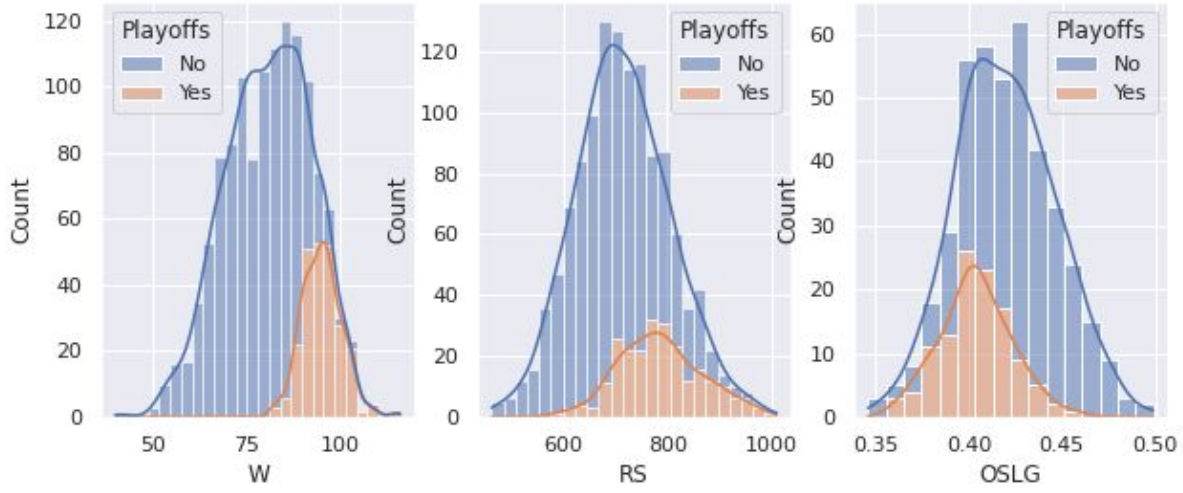

Fig. 2 Histograms for target (*W*) and two attributes (*RS*, *OSLG*).

The dataset is similar to one that two statisticians used in the early 2000s to gain better insight into what actually makes a baseball team more successful. The dataset contains per season performance statistics for various NA baseball teams from 1962 to 2012. Our regression target is the number of wins *W* which is contextualized by additional attributes such as the *Playoff* status (yes/no - could be used for classification) and what ranking they achieved in the season and the playoffs, which all could serve as prediction targets as well.

The dataset contains mostly numerical data such as the number of runs scored and allowed (*RS* and *RA*), batting averages (*BA*), slugging percentages (*SLG*, *OOBP*, *OSLG*). Contextual information is also given via nominal features such as Team, League and Year and the number of games played during the season. For the regression, the numerical features seem to be of more interest than the rest. We split the columns into three categories: ID (*Team*, *Year*, *League*), Targets (*W*, *Rankings*, *Playoff*) and the remaining eight as proper Features for the regression.

Fig. 2 shows histograms for our regression target *W* and two attributes, that are representative for the rest of the features in terms of their distributions and value ranges. Scaling will have to be used in the preprocessing step due to their wildly different ranges (hundreds vs percentages). There are also a lot of missing values, specifically in the *OSLG* and *OOBP* columns, that will have to be taken care of during the preprocessing.

Comparison to other dataset

We chose this dataset because it is a very small dataset compared to the Bankruptcy set, both in terms of dimensionality and number of instances. Both of them contain missing values, but they are distributed in different ways across the columns. In the Moneyball set the missing values are concentrated on four columns - the two ranks, *OOBP* and *OSLG*. For the last two, only 420 instances actually have non-missing entries which will pose a challenge for how to deal with them.