184.702 Machine Learning (VU 3,0) 2022W

Exercise 0

Ardit Luzi (12226089), Dante Godolja (11929150), Lukas Burtscher (11925939)

Exercise 0 Dow Jones Dataset Analysis.

The chosen dataset from our group it is the Dow Jones dataset which contains relevant data about the stock market prices, and it has a couple of attributes which we can apply data preprocessing and filtering for it to be adequate to generate a Machine Learning model, or perform data analysis on a given requirement. We want to start by describing the dataset which contains a total sample of 751, and it has 14 attributes. This dataset has a mix of datatypes, for example we have the stock attribute which is nominal, and the other 13 are continuous numerical data.

An interesting data which we have made some analysis would be the closing price which has the following information: a mean of 53.33$, mode = 20$, min= 10$ and max = 170$, these values were abstracted from the closing column to better understand what is the data range, learning what values are most repeated on the column, better visualizing the mean value of the data which we are handling, visualizing what price parameters have in its extremums like the minimum value and the maximum, in this way by taking these pieces of information apart from the whole column makes the attribute closer to human perception and readability so you can apply the given logic based on the requirement on in. The preprocessing steps that we have taken to properly work on this data set are:

- removing the US dollar Symbol out of the price values,
- converting the string price values into floats,
- dropping null values,
- turning the data attributes from string data types to Date data type for Pandas.

The tools that were used for the data analysis used to analyze this data set its Jupyter notebook and pandas library.

As a conclusion for this paper, we can summarize that the dataset that we are working on has in total 14 attributes, a mix of nominal and continuous, we analyzed an attribute and described the properties of its distribution, and explained the data processing that took place

Speed Dating

The dataset consists of 8378 instances containing 121 features. In the approach we had to deal with 18372 missing values.

The data is from speed dating events in 2002-2004. During the date the participant awnsers questions about hemselves and rate their partner. They were also asked to rate their date on six attributes: Attractiveness, Sincerity, Intelligence, Fun, Ambition, and Shared Interests.

During the preprocessing process we worked with pandas and numpy as Python libraries.

Missing Values

in order to deal with the missing values, we had to replace the Strings "?" with np.nan, so we were able to count the missing values per column.

Importance of Race

Participants were asked how important they rate the fact, that their partner has the same race as they have.

To fill in the missing values we decided to fill by the mode of the race which has these missing values. We assume that all rows have the race specified so that we can fill the Importance of Race. There were some rows which had no specified race and no importance of race so we dropped these.

From further observation, only Europeans/Caucasian-Americans had empty values in this dataset so we can just fill them with the mode of importance of race for this specific race.

Age

For age values we filled the NaN values with the median of all collected ages through the dataset.

Met

met column tells if the people from speed dating has met before. Since it is not very common to meet people at speed dating more than one time we assume that cells with NaN can be filled

with 0 that stands for "have not met before".

Drop Rows

After the assumtions we had a few NaN columns left and after a short check we found out that rows containing that NaN value in one column, also contained NaN values in other column. Participants producing those values skiped a whole question section, so we decided to drop the rest of the rows containing the NaN values, bacause we had not enough data do fill a whole question section by simulated values.

Value Casting

Numeric values in the data set are only rounded numbers, so we hat to cast float values into int values. Columns in the data set containing the Substring "pref" contain values about how the important the partner rated the specific category. Unfortunately we had casting errors and some rows contained decimal places. Thus we rounded those numbers with *.round()* and casted the float into int.

Duplicate in Field

For the coulumn Field we observed a lot of duplicates that were a result of lower/upper-case differences, as well as specific fields which could be generalized in one. For Example we took the all the fields that had "MBA-Busniess" , "Business and Economics" or anything related to business and put under the same field name. From such transformations we reduced the number of fields from 219 to 86

Intervall Description

The dataset contains fixed intervalls to recognize if a person rated something low, medium or high. The categorical values are ordinary and represent the given Interval in a String, for example [0-1] for low [2-5] for medium and [6-10] for high. Instead of using the specific Interval values we decided to replace the Strings by the according categorical values (not important, important, very impirtant).