

Preparation

```
In [6]: import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

Introduction

The purpose of this project is to implement a logistic model from scratch. We begin with preparing the data for analysis, then we implement 3 optimization algorithms for training the model, and finally we assess the result using statistical measures.

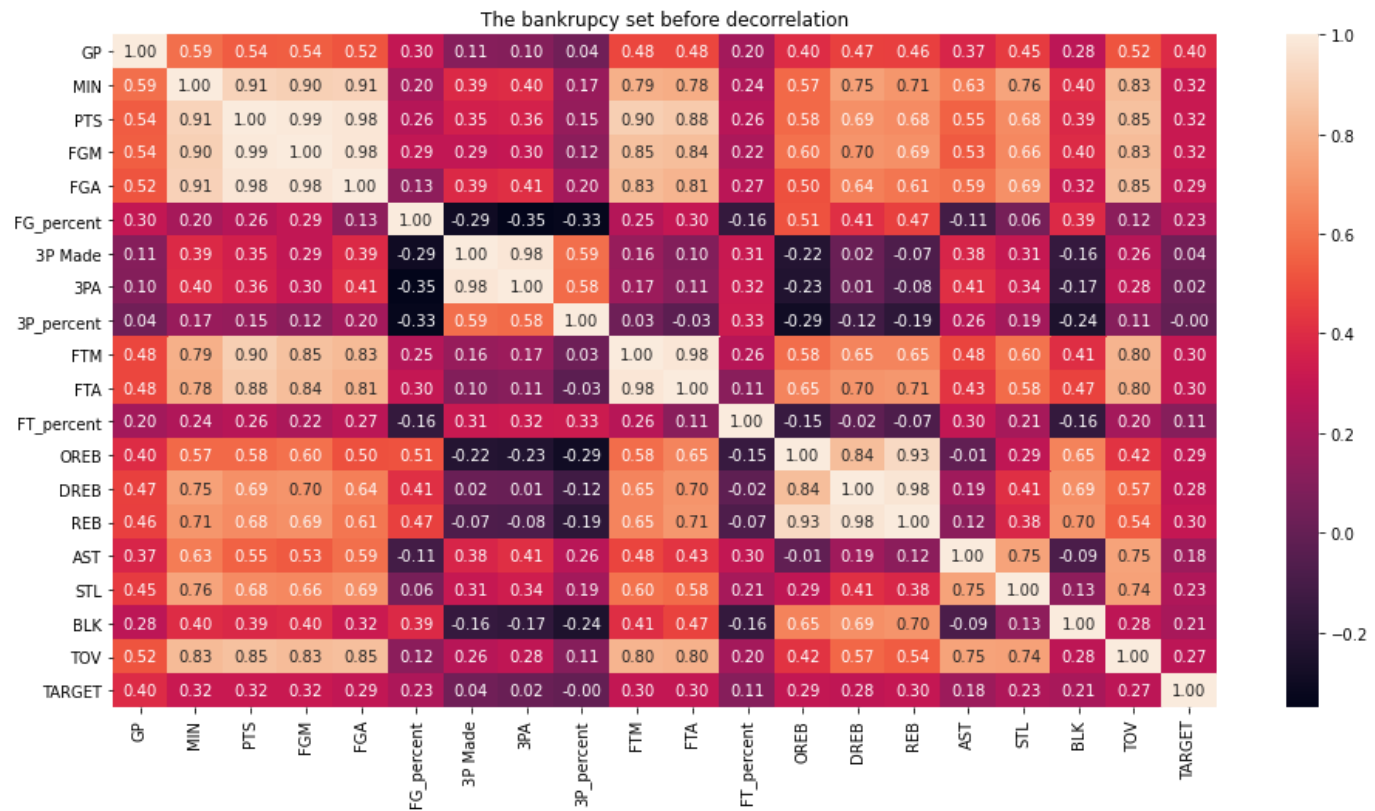
Datasets

Since the model needs to be trained and tested, there is a need to prepare some data. We collected 5 datasets:

- **candies** - dataset concerning production of candies. Sample questions addressed by the variables: *Does it contain chocolate?, Is it fruit flavored?, Is there caramel in the candy?, Does it contain peanuts, peanut butter or almonds?, Does it contain nougat?*. The target variable describes percentage of people who liked this candy type.
- **breast cancer** - dataset describing health state of a patients suffering from breast cancer. The target variable describes if the cancer is malignant.
- **bankruptcy** - this is a dataset from a recruiting task from a consulting firm. Variables include information of each candidating company, for example income, marketing statistics, ratings, sector number of employees, etc. The target variable represents information about bankruptcy. This dataset is significantly larger than other.
- **nba** - dataset on all the NBA finals results and NBA Regular Seasons MVP.
- **wine** - dataset related to red and white variants of the Portuguese "Vinho Verde" wine.

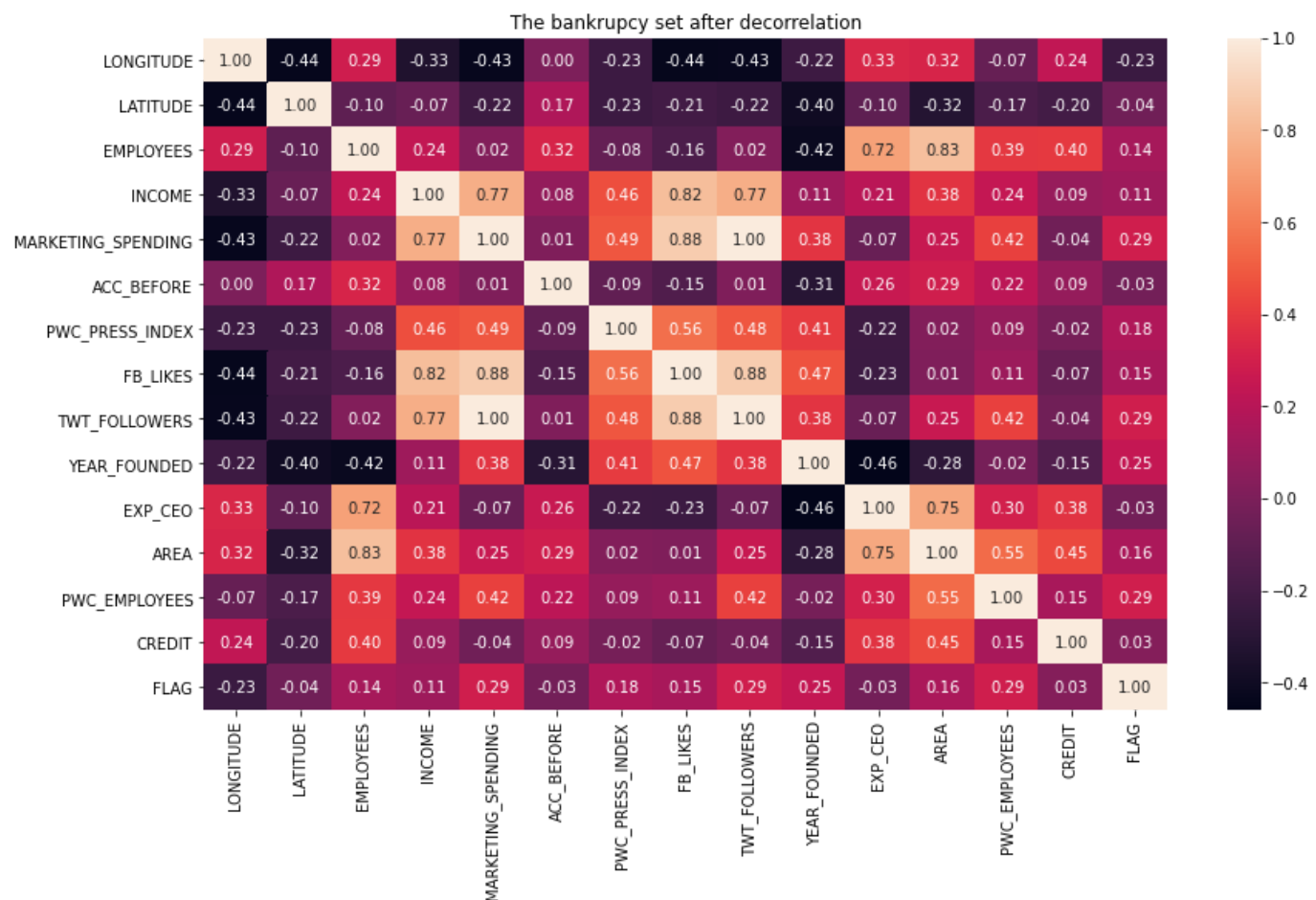
The first step was downloading the data and conversion of types. We saved the initially preprocessed data under directory `data/before preprocessing`. Next step was to replace missing data with mean value of each variable. Then we analyzed correlation matrix for each dataset. Let's consider the **nba** dataset

```
In [13]: nba = pd.read_csv('data/before preprocessing/nba.csv')
plt.figure(figsize=(16, 8))
plt.title('The bankruptcy set before decorrelation')
_ = sns.heatmap(nba.corr(), annot=True, fmt='.2f')
```



We dropped the most correlated variables. Below we present the correlation plot after decorrelation.

```
In [15]: nba = pd.read_csv('data/after preprocessing/nba.csv')
plt.figure(figsize=(14, 8))
sns.heatmap(bankruptcy.corr(), annot=True, fmt='.2f')
_ = plt.title('The bankruptcy set after decorrelation')
```



Measures

We introduce 3 measures of fit: *accuracy*, *precision* and *recall*.

$$accuracy(y, \hat{y}) = \frac{1}{n} \sum_i I_{\{y_i = \hat{y}_i\}}$$

$$precision(y, \hat{y}) = \frac{\sum_i I_{\{y_i = \hat{y}_i = 1\}}}{\sum_i I_{\{\hat{y}_i = 1\}}}$$

$$recall(y, \hat{y}) = \frac{\sum_i I_{\{y_i = \hat{y}_i = 1\}}}{\sum_i I_{\{y_i = 1\}}}$$

Model