UNIVERSITÄT
BIELEFELD

Report third project

Yasmin Bouhlada, Annalisa Dettori, Andrea Schinoppi

17^th January 2024

2 Introduction

3 Data Set

3.1 First data set

The first data set was made of ten columns X for the independent variable and one column for the dependent variable y. The data set had 1000 numeric samples and it has no NaNs. We didn't know what the variables represented, but we could resume the main information of the variables with this table:

	mean	std dev	min	max
X	0.01	4.89	-9.99	9.98
y	0.5	0.5	0	1

3.2 Second data set

The first data set was made of thirteen columns X for the independent variable and one column for the dependent variable y. The data set had 110 numeric samples and it has no NaNs. We didn't know what the variables represented, but we could resume the main information of the variables with this table:

	mean	std dev	min	max
X	0.01	4.89	-9.99	9.98
y	0.5	0.5	0	1

FARETABELLAAA

3.3 Third data set

The third data set (named *real world data*) was a real data set regarding the prices of some houses. It was made of ten columns $X_1, \ldots, X_{10}$ for the independent variables and one column for the dependent variable y. The data set had 1095 numeric samples and it had no NaNs.

The labels of the variables were: "LotArea", "TotalBsmtSF", "1stFlrSF", "2ndFlrSF", "GrLivArea", "WoodDeckSF", "OpenPorchSF", "3SsnPoarch", "ScreenPorch" and "PoolArea".

We could resume the main information of the variables with this table:

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
mean	10722.41	1159.84	0	338.71	1505.13	91, 06	47.26	2.78	15.09	2.14
std dev	11054.40	376.46	34900	432.04	514.24	120.64	66.79	25.18	56.55	35.79
min	1300	0	343	0	334	0	0	0	0	0
max	215245	3206	3228	1872	4676	670	547	407	480	738

4 Foundations

4.1 Methods

4.1.1 Filter methods, wrapper methods and unbedded methods

For variables selection what we decided to do was to apply three methods: filter methods, wrapper methods and unbedded methods.

As filter method, we used the *F measure*. We selected both 2 and 6 features:

4.2 Evaluation

4.2.1 Filter method

With the *F measure* we selected the following features

- Two features: `['TotalBsmtSF' 'GrLivArea']`
- Six features: `['2ndFlrSF' 'OpenPorchSF' 'WoodDeckSF' '1stFlrSF' 'TotalBsmtSF' 'GrLivArea']`

that provided the respective results for the R^2:

	R^2
2 features	0.22
6 features	0.47