UNIVERSITÄT
BIELEFELD

Report third project

Yasmin Bouhlada, Annalisa Dettori, Andrea Schinoppi

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:

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
mean	0.66	-0.06	0.02	-0.2	0.04	-0.1	0.12	0.06	0.08	0.06
std dev	2.60	5.80	3.50	5.76	2.69	5.75	5.78	5.90	3.31	5.79
min	-7.32	-9.98	-8.44	-9.95	-9.13	-9.98	-9.98	-9.99	-7.91	-9.98
max	7.4	9.97	7.43	9.97	8.92	9.98	9.98	9.97	8.26	9.97

	y
mean	0.5
std dev	0.5
min	0
max	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:

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
mean	-0.28	-0.33	-0.25	0.02	-0.03	-0.31	-0.48	0.08	-0.19	-0.31
std dev	0.92	0.79	0.99	1.06	1.26	1.13	1.00	1.13	1.15	0.8
min	-1.49	-1.23	-3.31	-2.42	-1.95	-2.62	-2.61	-1.84	-2.5	-1.51
max	2.13	2.17	1.79	3.31	4.05	2.44	1.71	2.76	2.84	2.23

	X_{11}	X_{12}	X_{13}	y
mean	0.36	-0.39	-0.34	1.82
std dev	0.90	1.16	0.85	0.38
min	-1.6	-2.88	-1.46	1
max	2.33	2.04	1.96	2

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 $X_1,\ .\ .\ .\ ,X_{10}$ 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

	y
mean	179984.82
std dev	77610.06
min	34900
max	755000

4 Foundations

4.1 Methods

4.1.1 Filter methods, wrapper methods and embedded 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*. SPIEGAZIONE

As wrapper method we used sequential feature selection both forward and backward. SPIEGAZIONE

As embedded method we used Lasso. SPIEGAZIONE

4.1.2 Random forest classifier

SPIEGAZIONE

4.1.3 Naive Bayes classifier

SPIEGAZIONE

4.1.4 Logistic regression

SPIEGAZIONE

4.1.5 k nearest neighbours

SPIEGAZIONE

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 on the test set:

numer of features	R^2
2	0.22
6	0.47

Both with two variables such as with six variables, the R^2 seems to be very low, for two variables is even less than 0.5, so we can assume the models are both not reliable.

4.2.2 Wrapper method

With the *forward feature selection* we selected the following features

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

that provided the respective results for the R^2 on the test set:

numer of features	R^2
2	0.22
6	0.46

As we can see *forward feature selection* chose the same two variables as the *F measure* did, but others for six variables. In any cases these models are not acceptable, because the values for the R^2 are too small.

With the *backward feature selection* we selected the following features

- Two features: `['TotalBsmtSF' '2ndFlrSF']`
- Six features: `['LotArea' 'TotalBsmtSF' '1stFlrSF' '2ndFlrSF' 'OpenPorchSF' '3SsnPorch']`

that provided the respective results for the R^2 on the test set:

numer of features	R^2
2	0.20
6	0.44

Now the second variable has change for the *backward feature selection* from the *F measure* method and *forward feature selection*. We can still say that the values of the R^2 are still very small.

4.2.3 Embedded method

With *Lasso* we selected the following features

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

that provided the respective results for the R^2 on the test set:

numer of features	R^2
2	0.18
6	0.44

The variables selected are a little bit different from the previous ones, but as for the other methods, there are no significant emprovements in the values of the R^2

5 Bonus task

eheh

A Appendix

banana