FYS-STK4155

Project 3

Alf Sommer Landsend

Diederik van Duuren

Abstract

The aim of this project was to predict healthcare costs. We did this by means of both a feed forward neural network and linear regression. The R^2-value was higher and the mean squared error was lower for the neural network method than for linear regression. The β-values for age, BMI and smoking were further away from zero than the other β-values (for sex, number of children and region), and they were all positive. This may suggest that these factors are important for the development of disease.

Introduction

In western countries large sums of money are spent on health services. It is important to be able to estimate these costs. In this project we used a dataset with the insurance premium and other information on 1338 persons. The data were split into train and test data. In order to predict the premium of the test data, we used a feed forward neural network and linear regression, including Ridge and lasso regression and included a RandomForrest regression at the end

Methods

The data we used in this project were the medical cost, the premium, of 1338 persons and their age, sex, body mass index, number of children, whether they smoke or not, and the region where they live. The data were converted from a csv file into a pandas data frame.

```
df = pd.read_csv('C:.\Data_project_3\insurance.csv')
```

The categories in the data frame were replaced with numbers.

```
df['sex'].replace('female',0,inplace=True)

df['sex'].replace('male',1,inplace=True)

df['smoker'].replace(['no','yes'], [0,1],inplace=True)

df['region'].replace(['northwest','northeast','southeast','southwest'],
[1,2,3,4],inplace=True)
```

We then scaled the features so that all values were between 0 and 1.

```
scaler = MinMaxScaler()

df_scaled = pd.DataFrame(scaler.fit_transform(df), columns=df.columns)
```

A column of numbers with value one was added in order to get a constant beta term. The numbers for the premium were converted into a numpy array. The rest of the data were also converted into a numpy array. This array was used as a design matrix. The data were then split into train and test data.

To try to predict the premium from the test data, we first used a feed forward neural network. For this we used the MLPregressor from Scikit-learn. We ended up with three hidden layers. We also used ordinary least squares regression, Ridge regression and lasso regression.

Results

To explore the data a bit we first made a heatmap to show correlation between the different parameters. This can be seen in figure 1 below

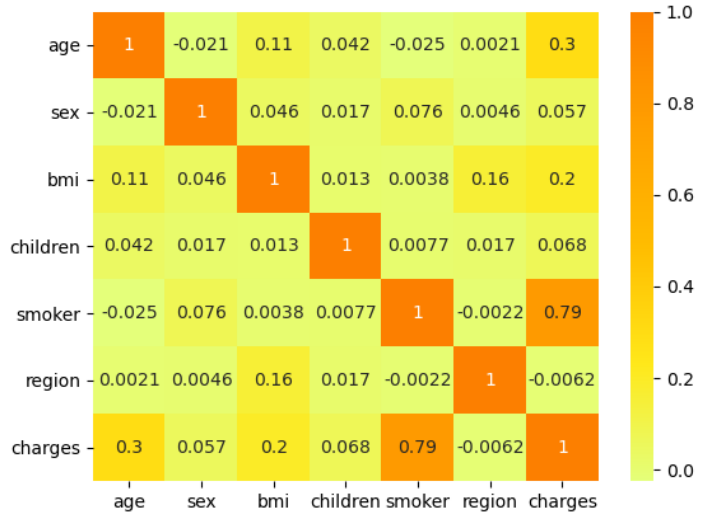

figure 1: correlation heatmap

One can see a quite a strong correlation between charges en smoking, as one could expect.

The neural network gave the highest R^2-value. It was about 0.80 to 0.85. The MSE was approximately 0.006.

Ordinary least squares regression gave an R^2-value of about 0.75 and an MSE about 0.01. The β-value that constitutes the constant term, was -0.05201767. The rest of the β-values were:

β-values with corresponding category

Category	β-value
Age	0.18859674
Sex	-0.00072959
BMI	0.19664388
Children	0.04797373
Smoker	0.37916967

Region	-0.01724261

As far as Ridge and lasso regression were concerned, we found that the lower the λ was, the higher the R^2-value became. This is shown in the figures below.

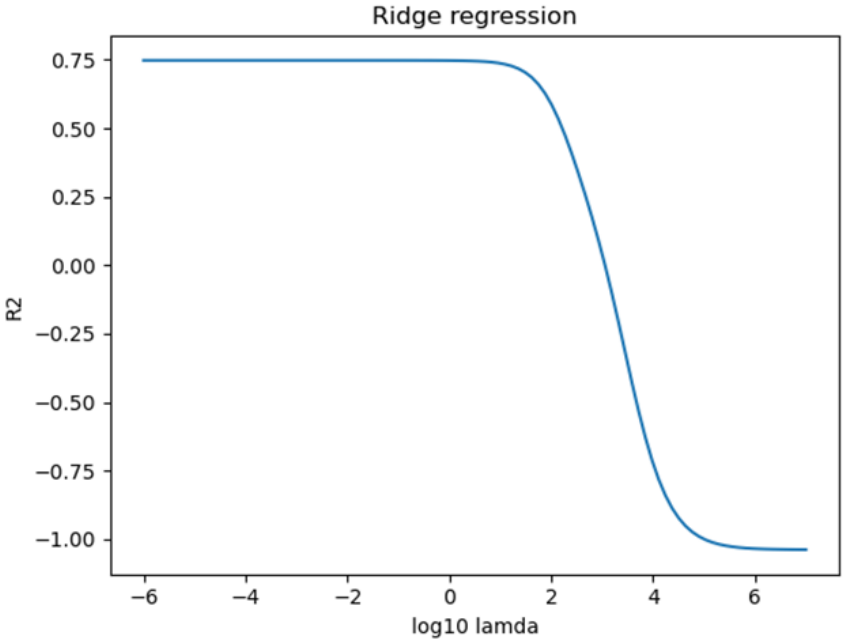

Figure 2. R2-value as a function of the logarithm of lambda. Ridge regression.

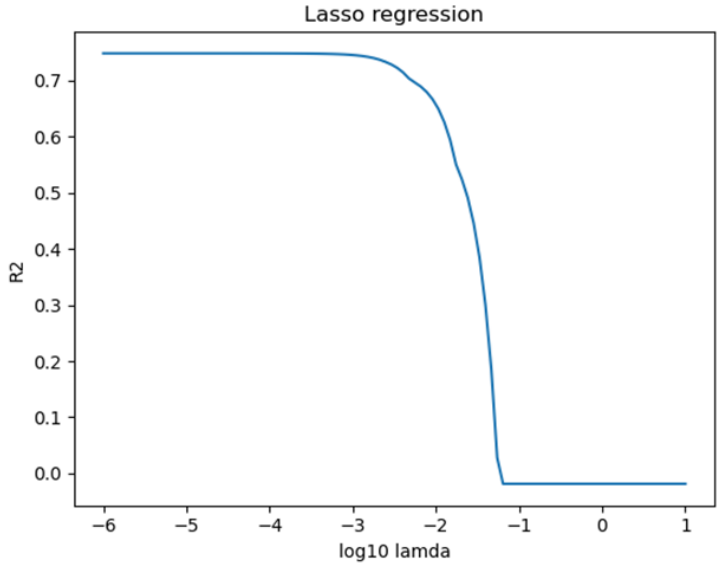

Figure 3. R2-value as a function of the logarithm of lambda. Lasso regression.

We too did a Randomforrestregression with sklearn that gave us the following results:

	MSE	R2
test data	19241950.109	0.878

table: RandomForrest scores

to visualise these results we made a plot on the predicted values for the test and train data in figure 4.

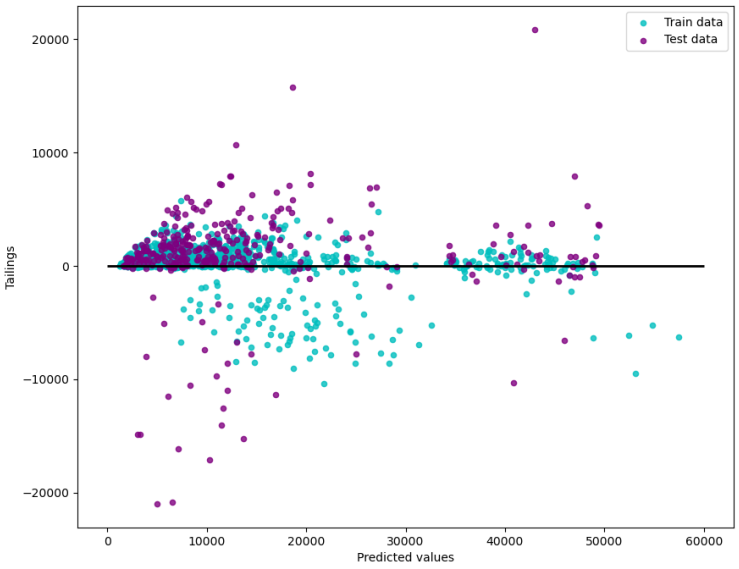

Figure 4: predicted values for RandomForrest

Discussion

When λ is low, the results from Ridge and lasso regression will be similar to the results from ordinary least squares regression. The fact that the R^2-value is higher the lower λ is, shows that ordinary least squares regression in this case gives better results than the other two types of regression. The RandomForrest however got the best result from all of the methods, although it had a MSE that was off the charts, that we could not explain. it might be a stacked value, but i couldn't find out why it was so high

The β-values that are most different from zero, are those for age, BMI and smoking. All of them are positive. It is well known that the risk of disease is higher among old people than young people. It is also well known that smoking increases the risk of disease. Many studies show that obesity increases the risk of disease and health care spending.[1] This is consistent with the values we have found for the age, BMI and smoker βs.

In this case the neural network gave better results than linear regression. One reason for this may be that the linear regression result would have been better if we had used functions of age, BMI and so on instead of only the values of age and BMI. Another reason may be that there can be interactions between the different covariates. The effect of for instance BMI on health costs may depend on the value of other covariates. This was not taken into account when we did linear regression.

Conclusion

We tried to predict the insurance premium by means of both a neural network, linear regression RandomForrest. We got the highest R^2-value when we used the RandomForrest followed by the neural network. The R^2-value of the linear regression could perhaps have been higher if the linear regression model had been better fitted. The β-values for age, BMI and smoking were further away from zero than the other β-values, and these three were all positive. This was as we would expect.

References

[1] Ana Aizcorbe (editor), Colin Baker (editor), Ernst R. Berndt (editor), David M. Cutler (editor), Measuring and Modeling Health Care Costs, University of Chicago Press, Chicago 2018, page 211