HarvardX - Data Science

Marco Schicker

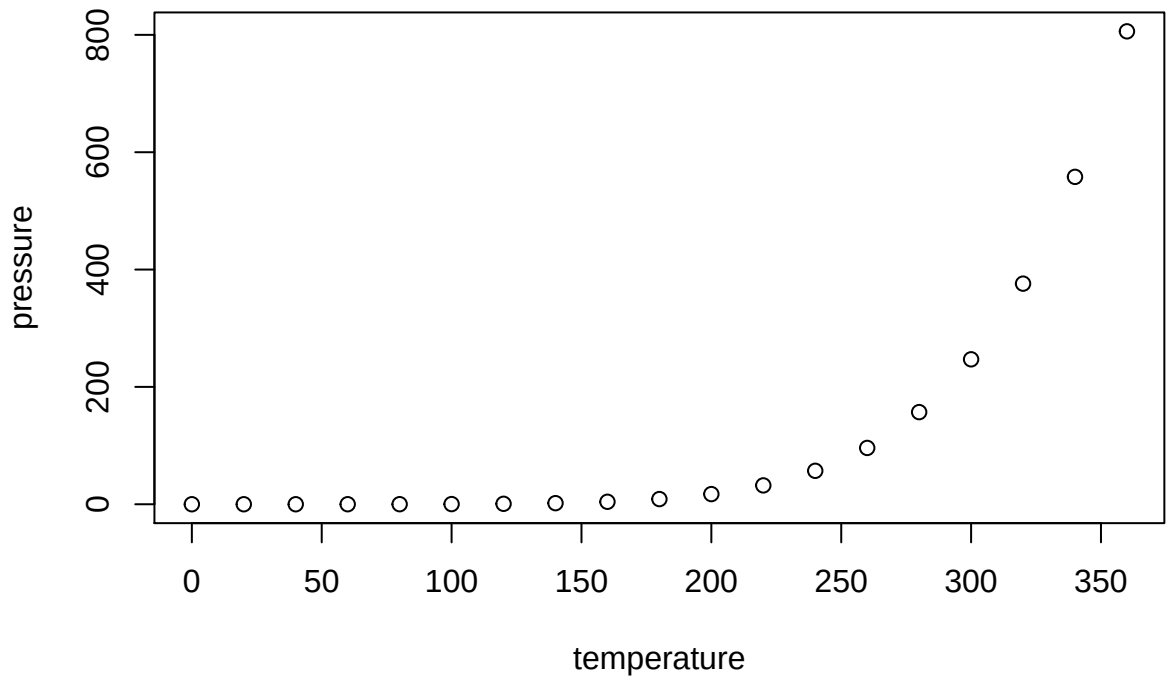
2021

```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

Including Plots

You can also embed plots, for example:



Executive summary

The goal of this project is to predict movie ratings with a high accuracy. The means to do so are by analyzing a data set provided and training an algorithm to make predictions for a validation set of data.

the data set provided consists

- describe data set
- summarize goal of project
- key steps performed

Methods & Analysis

The process to train the algorithm consists of the following basic steps: * data import * data cleaning * data exploration and visualization * creating training and test set *

data cleaning

- find predictors that are highly correlated and remove some

```
correlation matrix
```

- remove predictors with near zero variation

```
List of NZ-predictors and values of STD-deviation
```

data exploration and visualization

looking at the predictor´s influence on the rating we can see that...

```
grid Y=rating, X=predictors
```

some data about single predictors * histogram

insights gained

We can see from the analyzed data that

modeling approach

identify predictors

identify prediction model

Results

Following the approach mentioned above we can see that we reach the best result by combining the predictor set X (a, b, c, d) with the prediction method Y (ensemble)

```
matrix predictor sets vs prediction method
```

comparison of model performance

Conclusion

To sum up the results of this report it was possible to train an algorithm and reach an *RMSE of* . The best performance was reached by using the following predictors: * * *

The best results were reached using an ensemble of the following prediction methods: * naivebayes * knn * random forest * ...

The limitations of this report lie in... Next steps to improve performance even more would be to increase training data size, keep on learning or use neural networks to find new predictors in deep learning algorithms.