HOST
Hochschule Stralsund

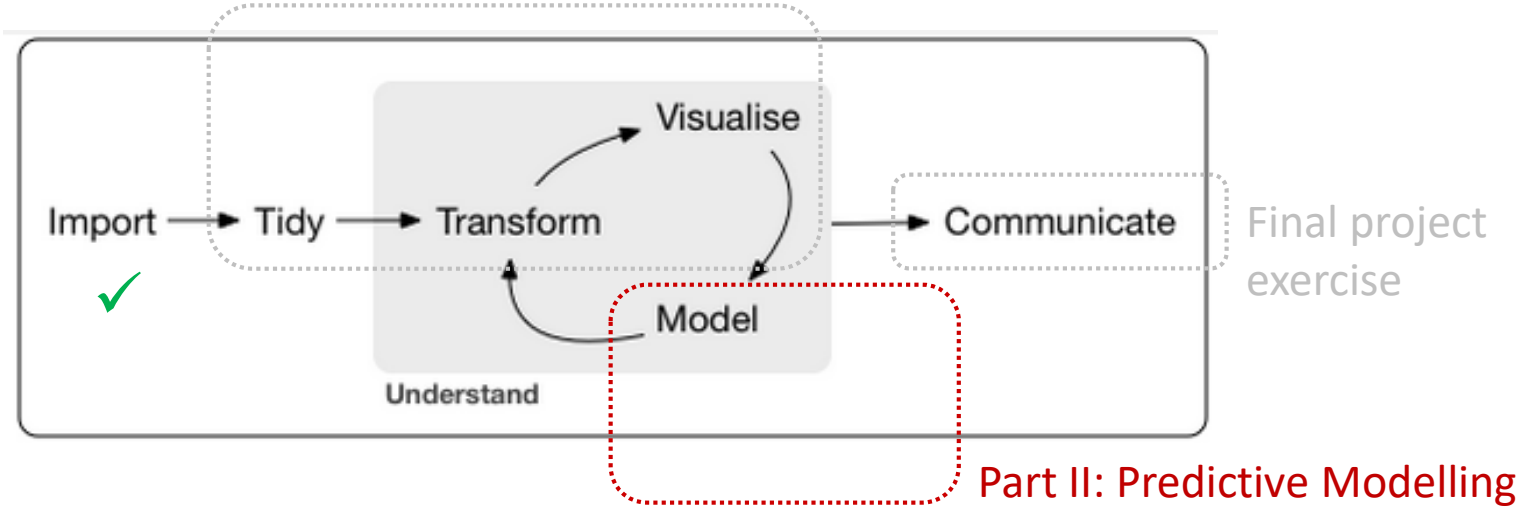
Data Science

Analyzing Facebook Likes

Gero Szepannek

Data Science Process

Part I: Data Engineering & Exploration ✓



Final project exercise

Part II: Predictive Modelling


WDR
UNTERSUCHUNGEN ANGEKÜNDIGT
Konzern gerät durch
Datenaffäre unter Druck
00:00:02
00:03:45

Facebook Data

1. Import the data `fb_likes.csv` as an R object of name `fb`.
2. What do columns represent?
3. How many rows do the data have? What does a row correspond to?
4. What does the entries contain?
5. What is the last column?
6. Create a histogram of the last variable!

Kosinski, M., Wang, Y., Lakkaraju, H., & Leskovec, J. (2016). Mining big data to extract patterns and predict real-life outcomes. *Psychological Methods, 21*(4), 493-506. http://dx.doi.org/10.1037/met0000105

Some Preliminary Text Mining…

Run the following code in order to get an overview on the most frequent likes:

```
library(wordcloud2)

# count '1's for each column (i.e. word)
wordcount <- colSums(MM[,-182])

?wordcloud2
wordcloud2(data.frame(word = names(wordcount), freq = wordcount), size = 0.5)
```

Creating a prediction model for your personality

1. Run the following code to keep only users with at least ten likes:

```
nmin <- 10
fb10 <- fb[rowSums(fb[,-182]) >= nmin,]
```

2. How many users (what percentage) do remain in the sample?

3. Create a **decision tree** personality predition model using the following code!

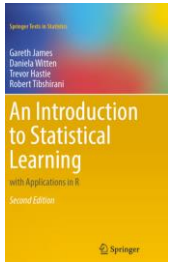
```
library(rpart)
rpmod <- rpart(openness ~ ., fb10, cp = 0.005)
```

Decision Trees

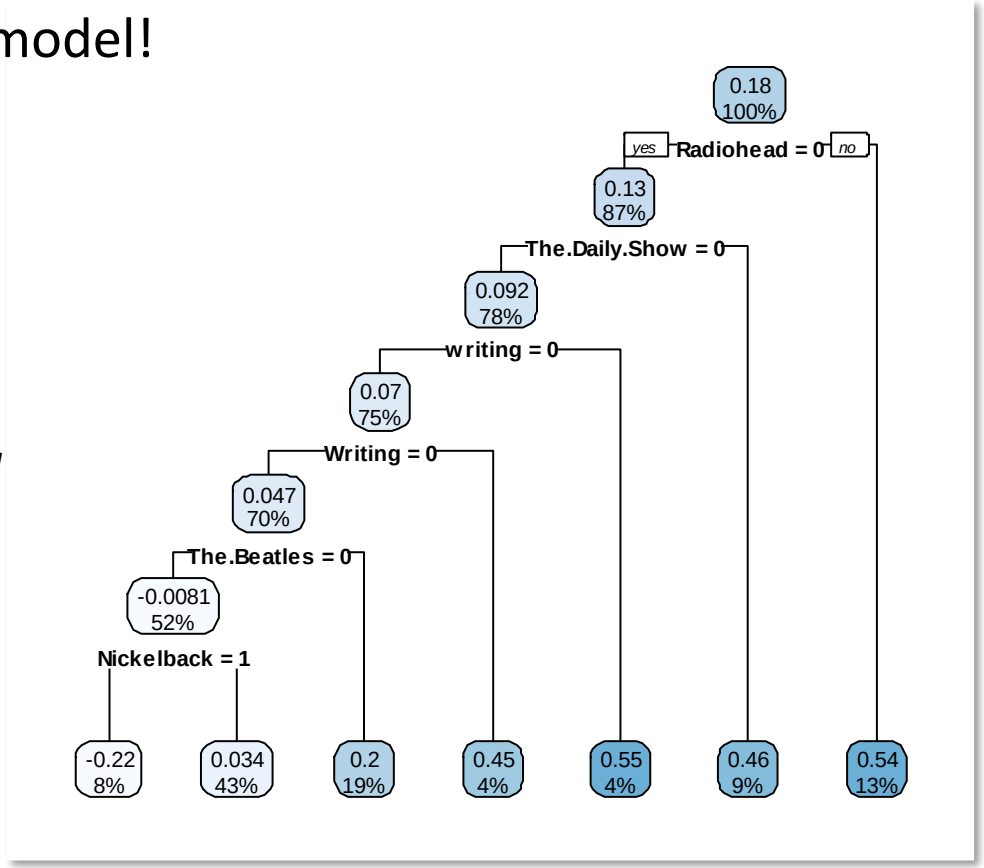
1. Run the following code in order to understand the model!

```
library(rpart.plot)
rpart.plot(rpmod)
rpmod
```

2. Explain the two numbers within the boxes!
3. …**Inverted classrom**: Read ISLR, chp. 8.1
   1. Find the answer to the question: *„Why is liking Radiohead selected first?“*
   2. …Prepare questions!



https://www.statlearning.com/



Quality of the model

1. Run the following code to create predictions $\hat{y}_i = \hat{f}(x_i)$ of the data!

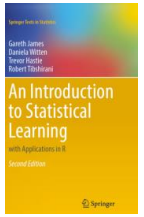
```
# predictions
yhat <- predict(rpmod, fb10)
```

2. …Compare predicted openness and the true values!

How can we assess the quality of our model?

1. Brain storming: Do you remember any measure from STATS that could be used to compare the fit of predictions and true data?
2. Read p.29 of the ISLR book: What measure is proposed to assess the goodness of fit?
3. Compute this measure in R!

https://www.statlearning.com/

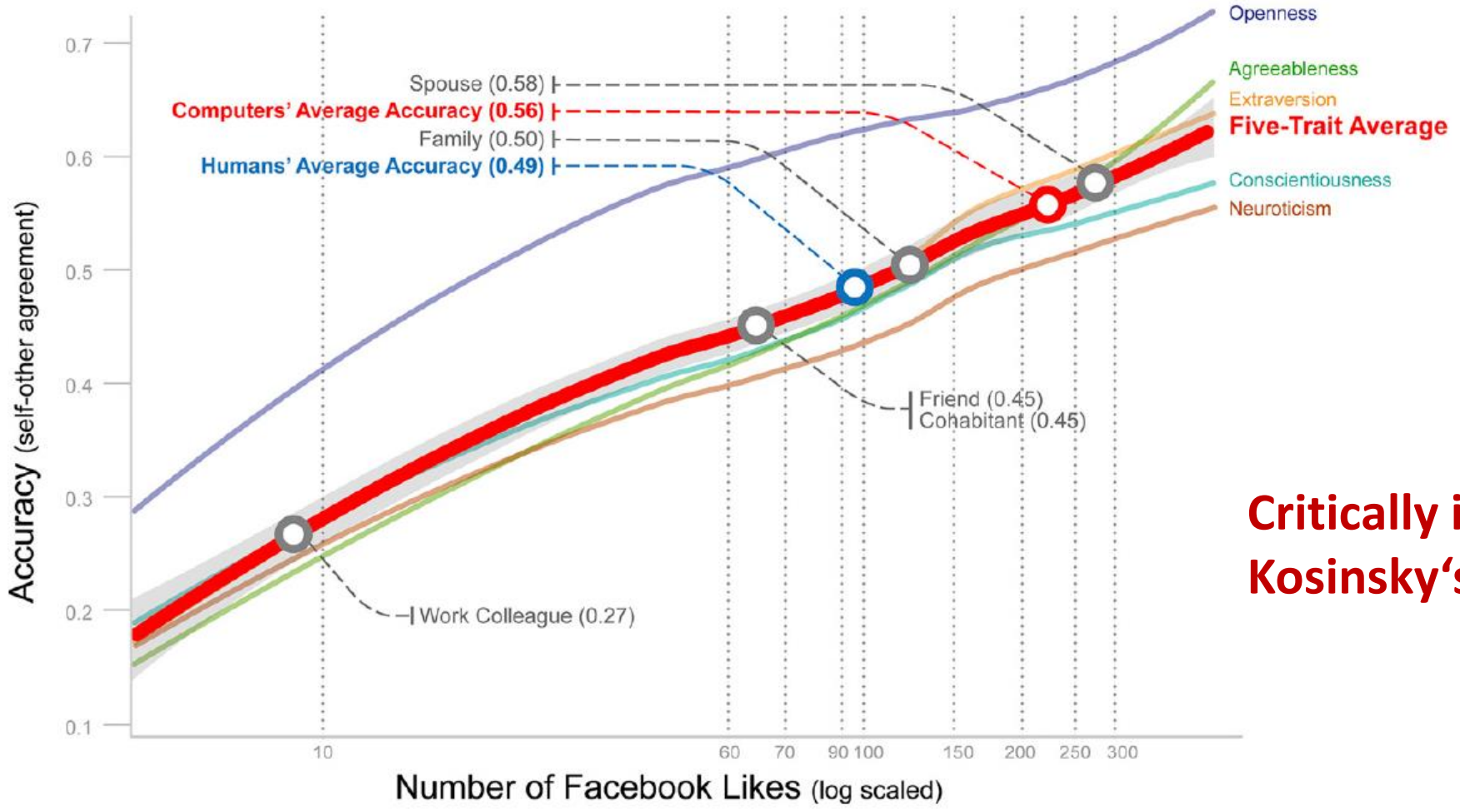


MAE

- **How do we have to modify the formula of the MSE in order to compute the mean absolute error (MAE)?**

- **What are pros and cons of the MAE compared to the MSE?**

Kosinsky‘s Results



Critically interpret Kosinsky‘s results!

Wu Youyou, Michal Kosinski, David Stillwell (2015): Computers judge personalities better than humans, PNAS 112 (4) 1036-1040; DOI: 10.1073/pnas.1418680112

Tuning: How deep is the tree?


```
rpmod <- rpart(openness ~ ., fb10, cp = 0.005)
cor(predict(rpmod,fb10), fb10$openness)
rpart.plot(rpmod)
```

1. The argument `cp` is called complexity parameter!
2. Re-run the tree and set `cp` to different values!
   1. Compute the correlation between predicted and true openness!
   2. Plot the tree!
3. Answer the questions: What happens if we increase (/decrease) `cp`? …cf. also ISLR, p. 309
4. Try to **find a value for `cp` that maximizes the predictive quality of the model!**