Kiss Dániel Márk

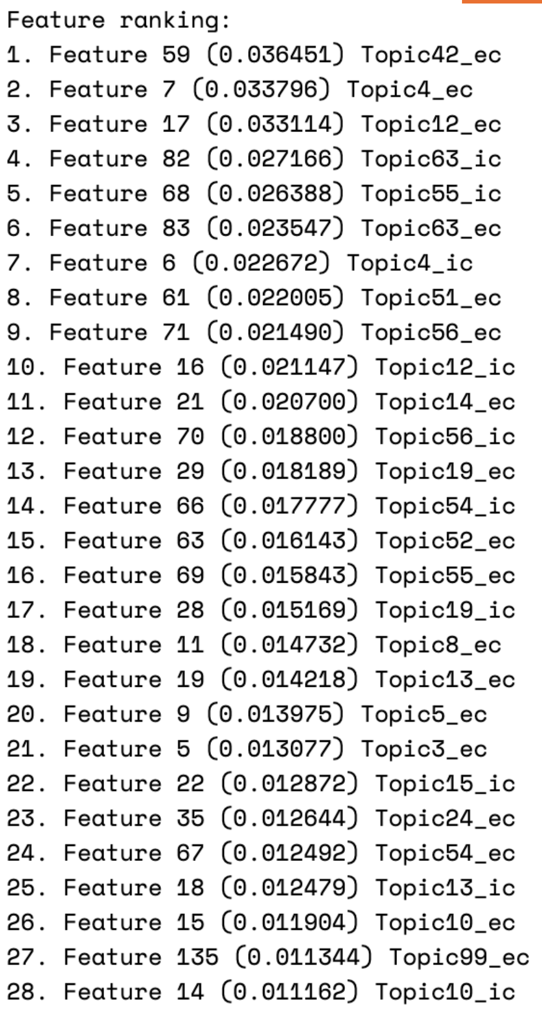
WP871Q

Customer Analytics

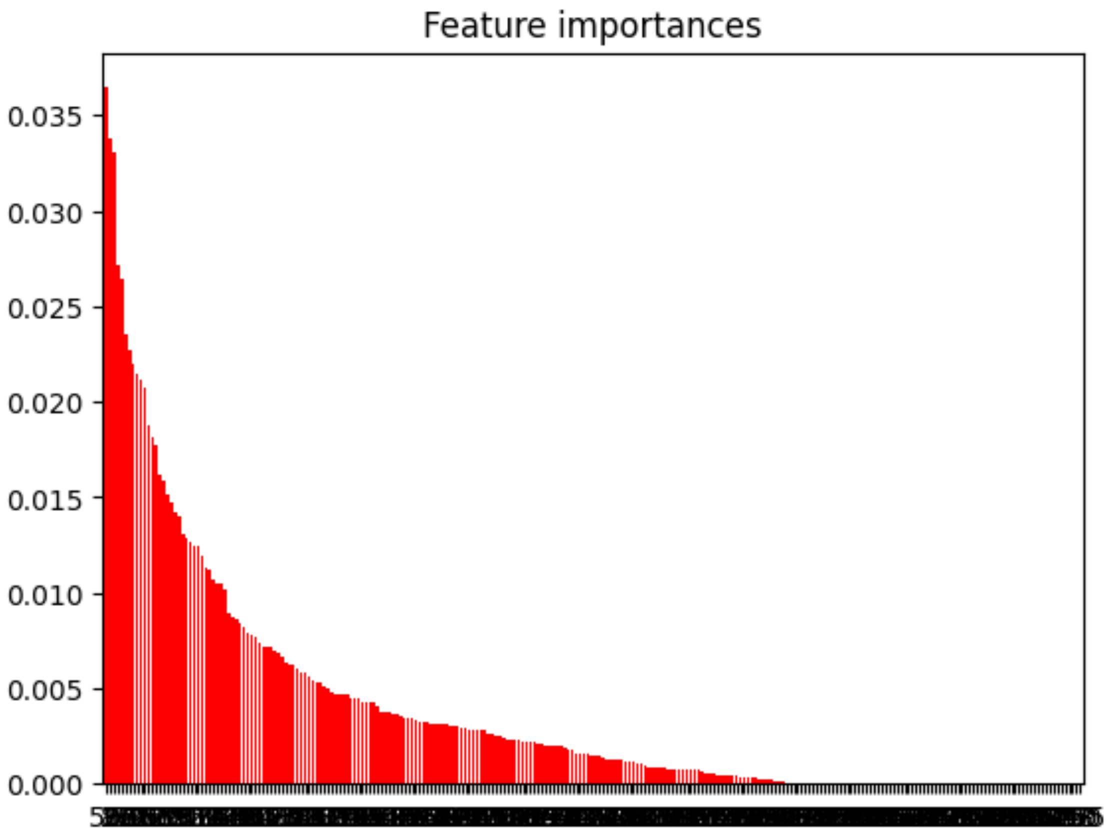
Exploratory data analysis

- .columns() – 258 unique features
- .describe() – statistical informations
- Selecting target value
- .dropna() – droping rows with nan value
- Sorting values by variance – keeping only top 100

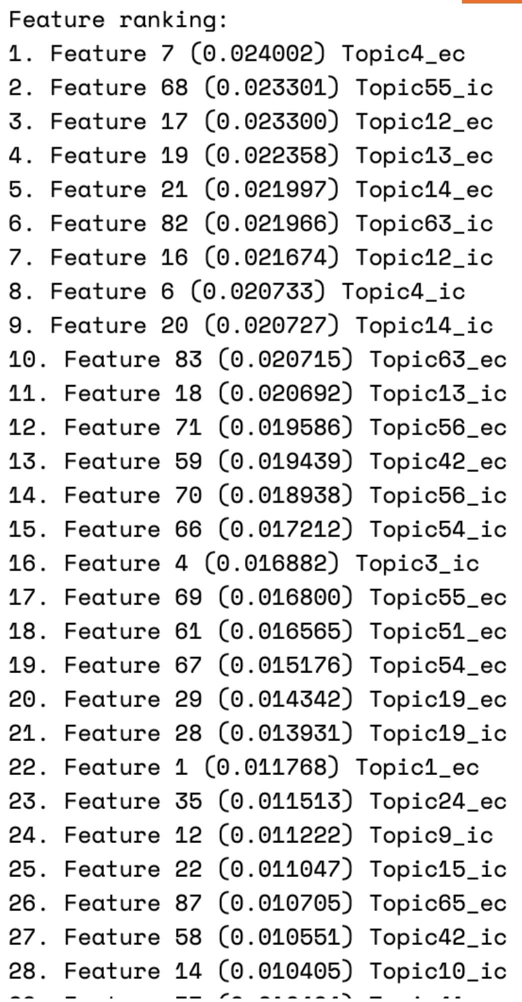
Feature importance – Decision tree



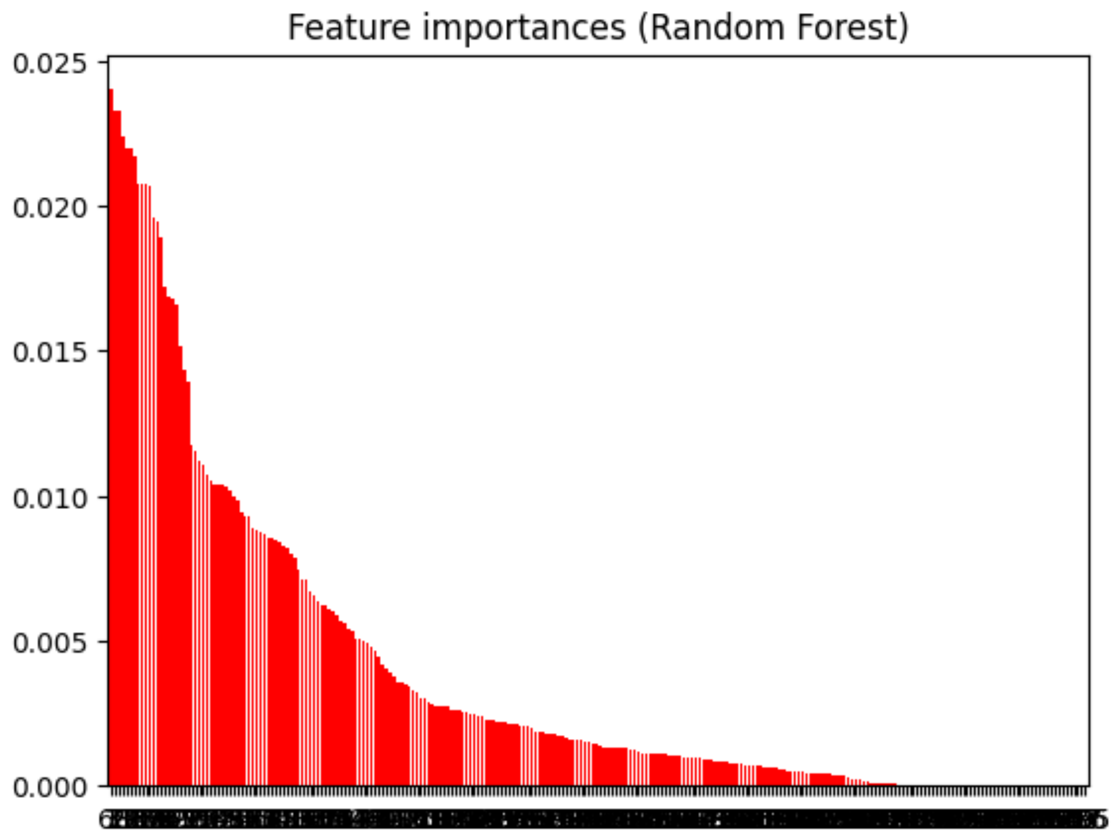
```
Feature ranking:
1. Feature 59 (0.036451) Topic42_ec
2. Feature 7 (0.033796) Topic4_ec
3. Feature 17 (0.033114) Topic12_ec
4. Feature 82 (0.027166) Topic63_ic
5. Feature 68 (0.026388) Topic55_ic
6. Feature 83 (0.023547) Topic63_ec
7. Feature 6 (0.022672) Topic4_ic
8. Feature 61 (0.022005) Topic51_ec
9. Feature 71 (0.021490) Topic56_ec
10. Feature 16 (0.021147) Topic12_ic
11. Feature 21 (0.020700) Topic14_ec
12. Feature 70 (0.018800) Topic56_ic
13. Feature 29 (0.018189) Topic19_ec
14. Feature 66 (0.017777) Topic54_ic
15. Feature 63 (0.016143) Topic52_ec
16. Feature 69 (0.015843) Topic55_ec
17. Feature 28 (0.015169) Topic19_ic
18. Feature 11 (0.014732) Topic8_ec
19. Feature 19 (0.014218) Topic13_ec
20. Feature 9 (0.013975) Topic5_ec
21. Feature 5 (0.013077) Topic3_ec
22. Feature 22 (0.012872) Topic15_ic
23. Feature 35 (0.012644) Topic24_ec
24. Feature 67 (0.012492) Topic54_ec
25. Feature 18 (0.012479) Topic13_ic
26. Feature 15 (0.011904) Topic10_ec
27. Feature 135 (0.011344) Topic99_ec
28. Feature 14 (0.011162) Topic10_ic
```



Feature importance – Random forest



```
Feature ranking:
1. Feature 7 (0.024002) Topic4_ec
2. Feature 68 (0.023301) Topic55_ic
3. Feature 17 (0.023300) Topic12_ec
4. Feature 19 (0.022358) Topic13_ec
5. Feature 21 (0.021997) Topic14_ec
6. Feature 82 (0.021966) Topic63_ic
7. Feature 16 (0.021674) Topic12_ic
8. Feature 6 (0.020733) Topic4_ic
9. Feature 20 (0.020727) Topic14_ic
10. Feature 83 (0.020715) Topic63_ec
11. Feature 18 (0.020692) Topic13_ic
12. Feature 71 (0.019586) Topic56_ec
13. Feature 59 (0.019439) Topic42_ec
14. Feature 70 (0.018938) Topic56_ic
15. Feature 66 (0.017212) Topic54_ic
16. Feature 4 (0.016882) Topic3_ic
17. Feature 69 (0.016800) Topic55_ec
18. Feature 61 (0.016565) Topic51_ec
19. Feature 67 (0.015176) Topic54_ec
20. Feature 29 (0.014342) Topic19_ec
21. Feature 28 (0.013931) Topic19_ic
22. Feature 1 (0.011768) Topic1_ec
23. Feature 35 (0.011513) Topic24_ec
24. Feature 12 (0.011222) Topic9_ic
25. Feature 22 (0.011047) Topic15_ic
26. Feature 87 (0.010705) Topic65_ec
27. Feature 58 (0.010551) Topic42_ic
28. Feature 14 (0.010405) Topic10_ic
```



PCA

REDUCING 258 TO 50 COMPONENTS

SCALING USING STANDARDSCALER

Evaluation – AdaBoost and Random forest with Voting

0,84538 public score

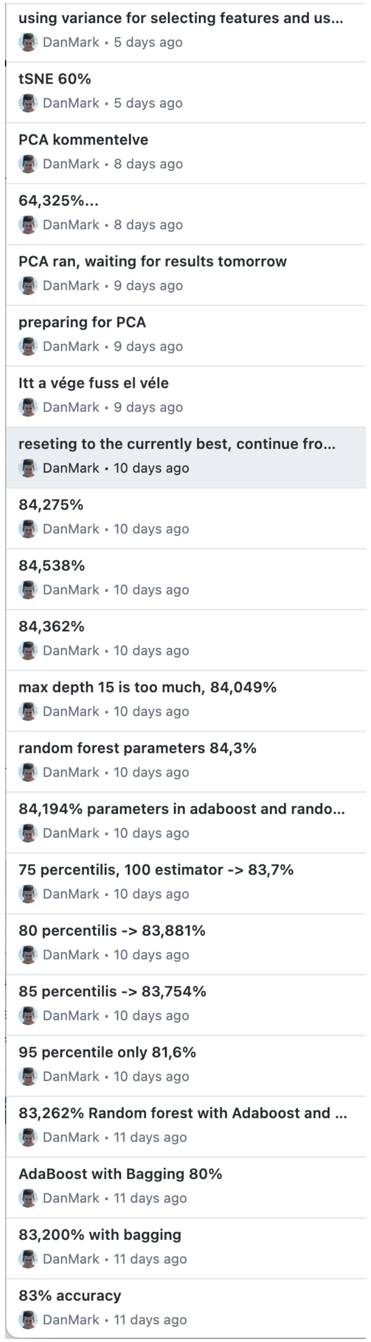
Parameter optimization

Random forest:

- 150 estimators
- Max depth: 12 (15 too much, 10 too few)
- Criterion for cutting: Entropy

AdaBoost

- 150 estimators
- Learning rate: 1,5



using variance for selecting features and us...
DanMark • 5 days ago

tSNE 60%
DanMark • 5 days ago

PCA kommentelve
DanMark • 8 days ago

64,325%...
DanMark • 8 days ago

PCA ran, waiting for results tomorrow
DanMark • 9 days ago

preparing for PCA
DanMark • 9 days ago

Itt a vége fuss el véle
DanMark • 9 days ago

reseting to the currently best, continue fro...
DanMark • 10 days ago

84,275%
DanMark • 10 days ago

84,538%
DanMark • 10 days ago

84,362%
DanMark • 10 days ago

max depth 15 is too much, 84,049%
DanMark • 10 days ago

random forest parameters 84,3%
DanMark • 10 days ago

84,194% parameters in adaboost and rando...
DanMark • 10 days ago

75 percentilis, 100 estimator -> 83,7%
DanMark • 10 days ago

80 percentilis -> 83,881%
DanMark • 10 days ago

85 percentilis -> 83,754%
DanMark • 10 days ago

95 percentile only 81,6%
DanMark • 10 days ago

83,262% Random forest with Adaboost and ...
DanMark • 11 days ago

AdaBoost with Bagging 80%
DanMark • 11 days ago

83,200% with bagging
DanMark • 11 days ago

83% accuracy
DanMark • 11 days ago

Hyper parameters & results

- Before variance – treshold optimization for feature selection
  - Percentile trashold: 0,8 was the best, also tried 0,75;0,85;0,95
- Random forest parameter optimization
  - Max_depth: 15 - overfitting
  - Estimators number: 100 - too few
  - Trying different criterions: “gini”, “entropy”,”log_loss”
    - Log_loss: emphasizes accuracy of probalistic prediction, penalizing confident but incorrect predictions
- AdaBoost parameter optimization
  - Learning rate: 1,7 – too much(0,841), 1 – too slow
  - SAMME algorithm is worse (will be default in future)
- Using GridSearch
  - Was too slow in Laboratory exercise

Evaluation - PCA

Using nearly the same parameters

0,779 best public score

Thank you for your attention!

Any questions?