Case Study

Binary classification NLP problem for BRAIN ONE

Ronny Hein

19 Oktober 2019

Overview

Facts:

- 11 features
- 248252 observations
- 70 % test/train split
- 192426 observations in train set

Insights:

- overall is left-skewed
- target is imbalanced with 89.19 % positive

Visualisation

Postive

Visualisation

Negative

Model

Bayes' theorem: $p(C_k \mid \mathbf{x}) = \frac{p(C_k)\ p(\mathbf{x}|C_k)}{p(\mathbf{x})}$

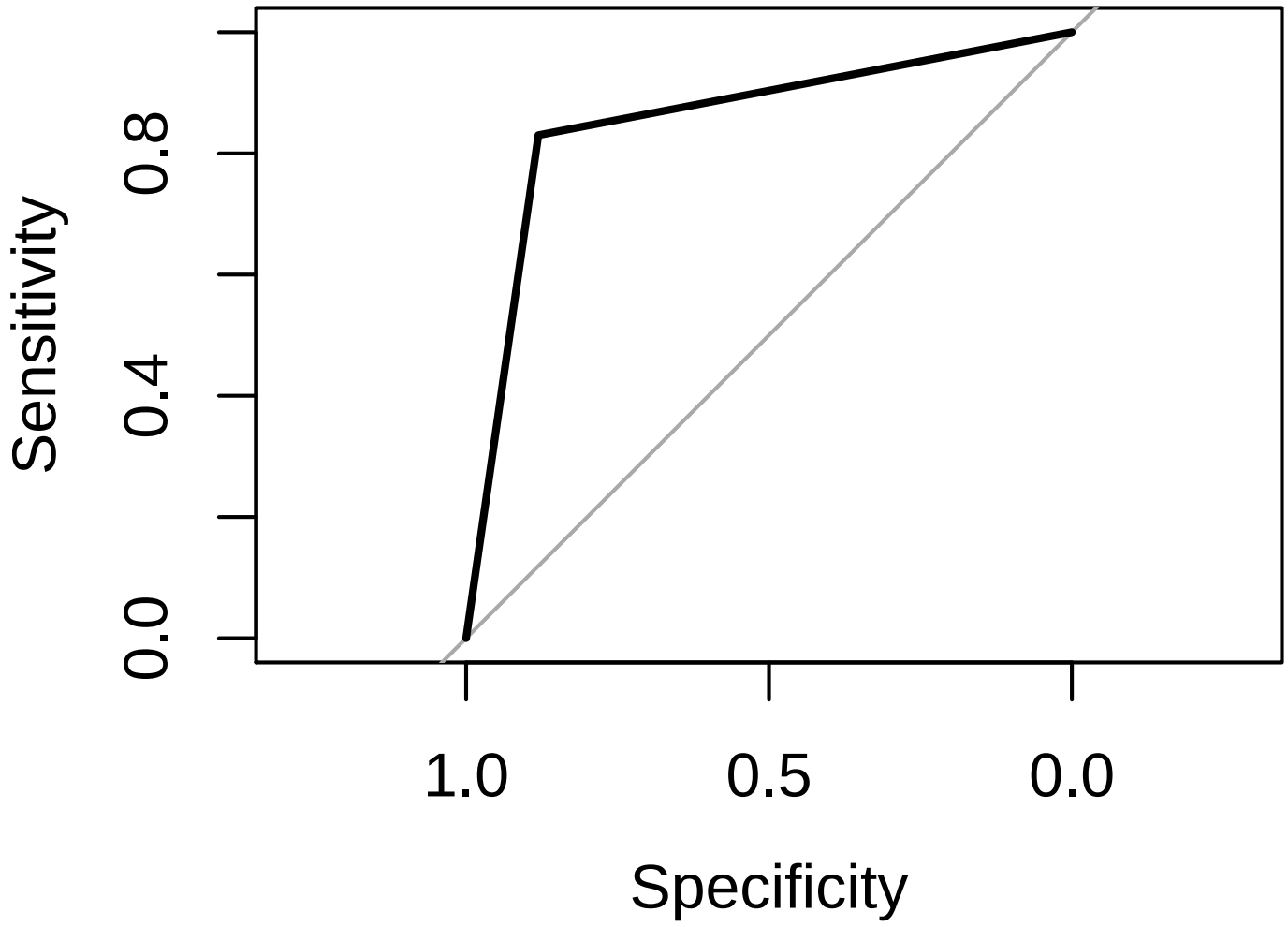
Results on train set:

Table 1: actual (cols) versus predicted (rows)

	-1	1
-1	78.97	1.13
1	10.22	9.67

Evaluation

- accuracy of classifier on test: 87.55 %
- true positive: 78.84 %
- false negative: 10.67 %
- area under the curve: 0.86



Considerations

1. more text preprocessing
2. use x-fold cross-valiation
3. apply more models and tune parameters
4. try more advanced NLP methods like BERT

End

Thank you