Email sorting with Natural Language Processing and Conformal Prediction

Patrizio Giovannotti and Daljit Rehal | Computer Learning Research Centre, Royal Holloway University of London | Centrica

Problem

- Automatically forward customer email to the right agents
- Based on message content
- Control the number of errors

Data

We undersampled the largest 7 classes to have 5,000 examples per class

Text pre-processing

I've already paid £150 on 10/03/2019. Why the new bill?

already paid __money__ __date__ new bill

Sparse Text Vectors

$d_1 =$ "my bill is too high"

$d_2 =$ "last bill was a high bill"

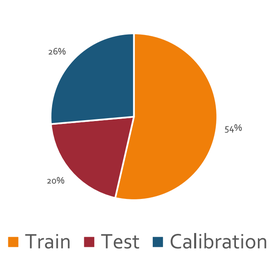
	a	bill	high	is	last	my	too	was
w_1	0	1	1	1	0	1	1	0
w_2	1	2	1	0	1	0	0	1

Feature scaling

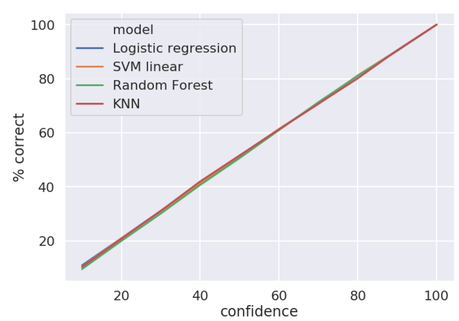
- Term t Frequency $tf(t, d_i) = \#t \in d_i$
- Document Frequency $df(t) = \#d : t \in d$
- **TF-IDF**$(\boldsymbol{t}, \boldsymbol{d_i})$ $:= tf(t, d_i) \cdot \log\left(\frac{D}{df(t)}\right)$
- TF-IDF score penalizes terms that appear in too many documents

Final dataset: matrix $\boldsymbol{M} \in \mathbb{R}^{40{,}000 \times 7{,}500}$

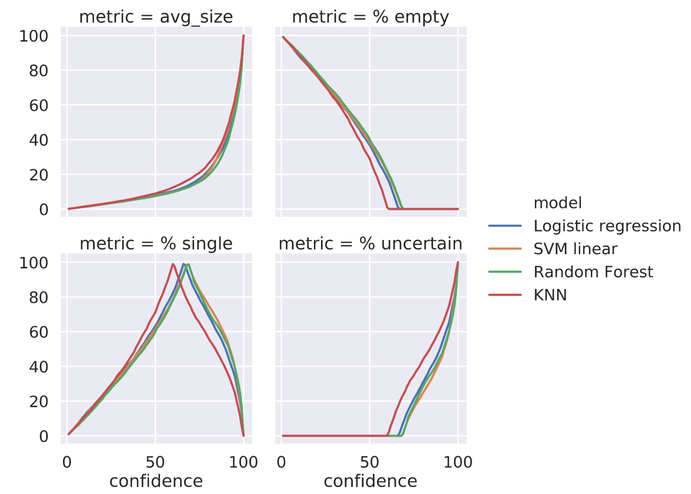
Dataset split



Validity



Performance



- Random Forest's prediction region is the most **efficient**
- Apart from KNN, models produce correct **single predictions 70%** of times
- **Slowest**: SVM (6+ hours) and KNN
- Recommended: **Random Forest** (1000 trees)

Conclusion & Future work

- **Good result** given the limitations:
  - Undersampled dataset
  - Several wrongly labelled examples
- We can decide if a human intervention is needed in each case
- Will use **Mondrian** predictors for imbalanced classes
- Will use **dense embeddings** and deep neural network as underlying algorithms

References

- Manning, C. D. *et al.* Introduction to Information Retrieval (2008)
- Vovk, V. *et al.* Algorithmic Learning in a Random World (2005)
- Eliades, C. *et al.* Detecting seizures in EEG recording using conformal prediction (2018)
- Linusson, H. `nonconformist` Python module

ROYAL HOLLOWAY UNIVERSITY OF LONDON

centrica