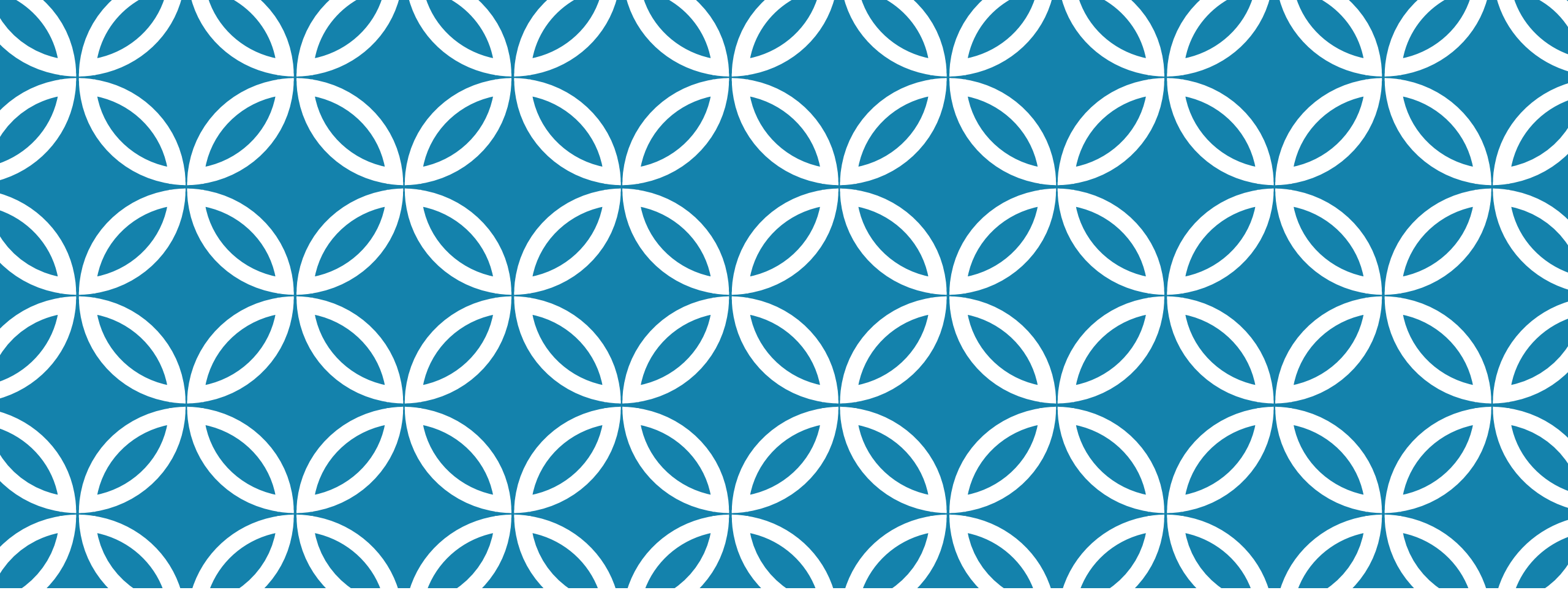
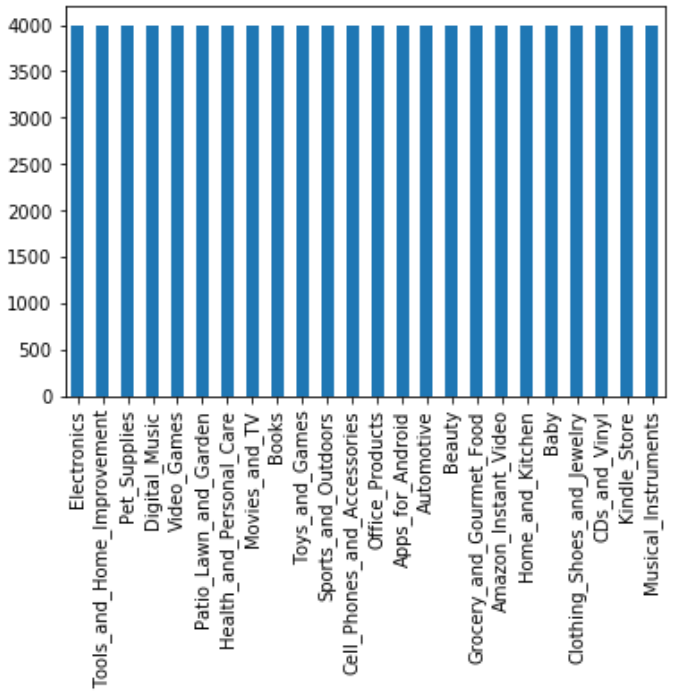


# DATA SCIENCE FOR BUSINESS: PROJECT

Rodrigues Alves Danny
Robin Genolet

# STEP 1, PREPROCESSING

Data is uniformly distributed




Features that we dropped:

- ❑ asin
- ❑ reviewTime
- ❑ unixReviewTime
- ❑ reviewerName

For this step, we use only 2 features

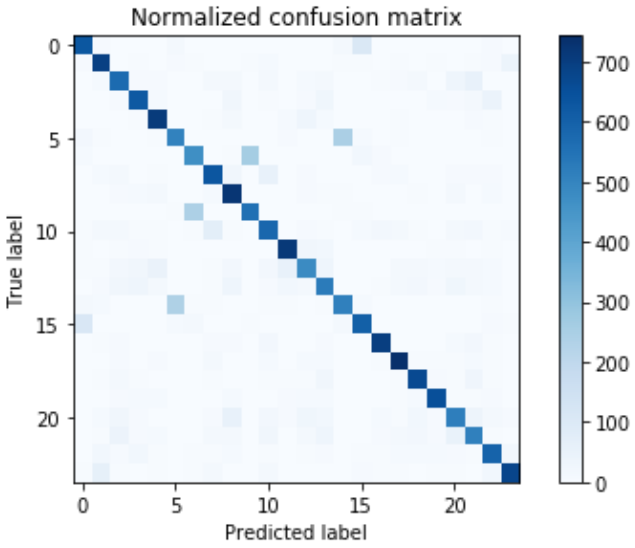
- ❑ reviewText tf-idf
- ❑ summary tf-idf

Why?

- ➢ Models yield better results without the features that we create in steps 2 & 3

# STEP 1, BEST MODEL

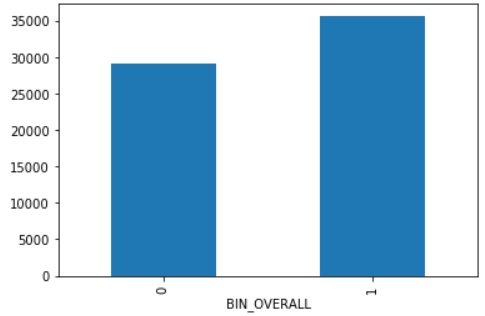
And the winner is:
Linear Support Vector Classification




| Final ranking: | Precision (%) |
|---|---|
| 1. Linear Support Vector Classification | 76.1 |
| 2. Linear SVM (with SDG) | 73.5 |
| 3. Random Forest Classifier | 60 |
| 4. KNN Classifier | 47.75 |
| 5. Dummy classifier | 4.5 |
| | $\frac{1}{24}$ : 4.16 |

# STEP 2, PREPROCESSING

- Skewed data




- We don't merge features reviewText and Summary
  - Avoid loss of information (e.g. summary has specific keywords

- We could have added features that guess if user is good/random/bad worker
  - Avoids trolls

- helpful [a, b] becomes helpfulPercentage:

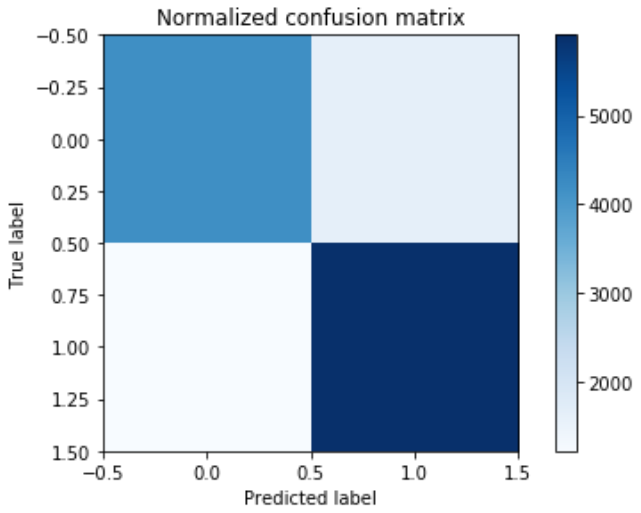
  $\frac{a}{b}$ if b is not zero, 1 otherwise

- from reviewText we create:
  - reviewTextLength
  - reviewTextCountPoints
  - reviewTextCountExcl
  - reviewTextCountInterr
  - reviewTextCountComas
  - reviewTextPositiveSmiley
  - reviewTextNegativeSmiley
  - reviewTextAllCAPS
  - reviewTextPositiveSmiley, reviewTextNegativeSmiley

- positiveRatio (uses positive & negative word lists)

# STEP 2, BEST MODEL

And the winner is (again):
Linear Support Vector Classification



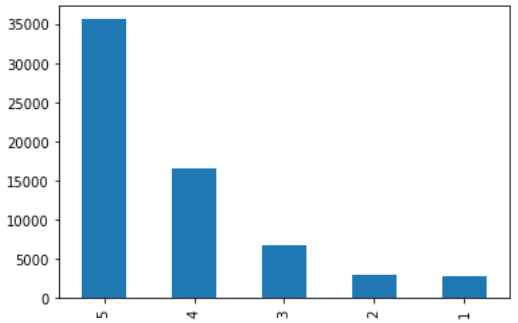

Final ranking:

| | | Precision (%) |
|---|---|---|
| 1. | Linear Support Vector Classification | 77.9 |
| 2. | Linear SVM (with SDG) | 77.3 |
| 3. | Random Forest Classifier | 76.1 |
| 4. | KNN Classifier | 63.3 |
| 5. | Dummy classifier | 4.5 |

$\frac{1}{2} : 50$

# STEP 3

- Skewed data:




- Use all created features

Final ranking:

| | | Precision (%) |
|---|---|---|
| 1. | Linear Support Vector Classification | 57 |
| 2. | Linear SVM (with SDG) | 50.93 |
| 3. | Random Forest Classifier | 57.1 |
| 4. | KNN Classifier | 44 |
| 5. | Dummy classifier | 20.2 |

❖ Same order in each step!

$\frac{1}{5} : 20$