Data Presentation

Objectives

A company develops a Product Discovery services using Machine Learning solutions. We describe the dataset for highlighting important behaviors of the product and explain them to the clients.

- How perform the tags discovery according to the gender ?
- What is the more important tag ? Other tags depend on it ?
- How discovery model performs for each tag prediction ? Is it reliable for every category of tags ?
- Do we have equivalent results for the same item ?

Dataset - General Descriptions and Assumptions

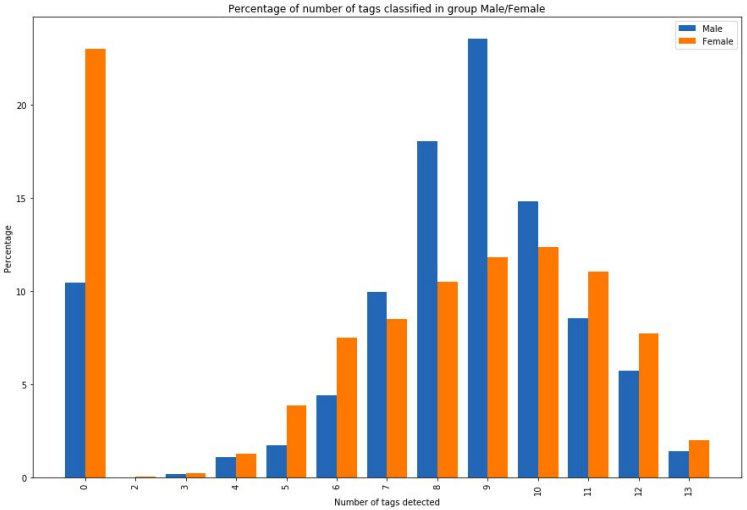
- The Dataset is composed of 50k items and 28 variables:
  - **Title** of the item.
  - **Gender** associated to the item.
    - Female gender represents **85%** of the dataset.
  - **Tags** information which represents category tags and their probability to be tagged.
    - an item can be tagged **13 times at maximum**.
    - the same item can have **different tags combinations.**
    - 'Cat', 'Type', 'Look', 'Color', 'Texture', 'Style', 'Pattern', 'Detail', 'Embellishments', 'Length', 'Sleeve', 'Neckline', 'SleeveStyle'.

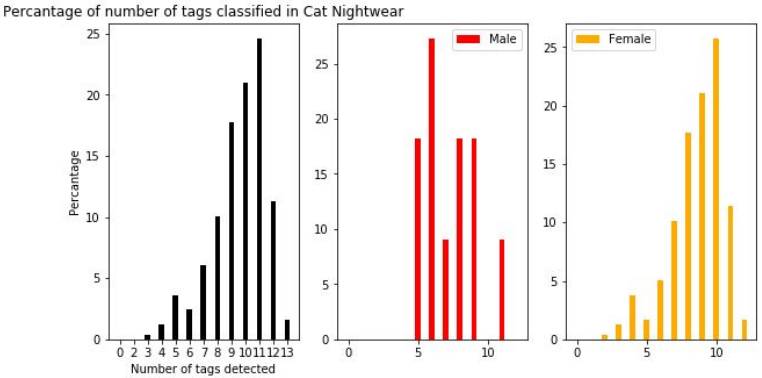
{'title': 'Woman Script Chunky Trainers',
 'gender': 'Female',
 'tags': {'Cat': 'SportShoes', 'Cat_prob': 0.99,
          'Type': 'Trainers', 'Type_prob': 0.71,
          'Look': 'Sportive', 'Look_prob': 0.66,
          'Color': 'Natural', 'Color_prob': 0.28,
          'Texture': 'Leather', 'Texture_prob': 0.098,
          'Style': 'SlipOn','Style_prob': 0.016}}

- Each row is considered to be an event (a sale for example) and the data stores the title of the item, the gender associated and a list of tags likely classified (by an item discovery model).

Model Discovery of Tags

- The Tags Discovery Model has a different behavior according to the gender.
  - It classifies **more tags on Male** items than Female items.
- 21% of all items have been not tagged.
  - When the tag **'Cat' is not classify** then no other tags is found.
  - Moreover, for each Tags (variables) **46% are not tagged in average.**

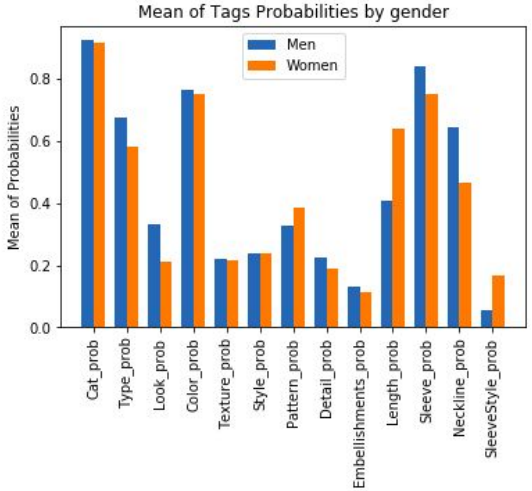




- Tags discovery model detect more tags in average for male items than female items.
  - Male's items are maybe more specific and Female's items are widespread.
- The number of tags classified is correlate with the tag 'Cat' found and the gender.

Probabilities of Tags

- Each Tags recognition model has its own reliability.
  - Detection of tags ‘Cat’, ‘Type’, ‘Color’, ‘Length’, ‘Sleeve’ and ‘Neckline’ gives higher probabilities in average.
- Tags probabilities can have different values for the same tag on the same item.
  - Tags probabilities are not dependent on each other but the tags are. **If a tag is misclassified it can affect the others.**



{'title': 'Plus T-Shirt & Cycle Short Co-ord',
'gender': 'Female',
'tags': {'Color': 'White', 'Color_prob': 0.50,
'Cat': 'Shorts', 'Cat_prob': 0.19,
'Type': 'Tights', 'Type_prob': 0.12,
'Length': 'Longline', 'Length_prob': 0.04,
'Texture': 'Leather', 'Texture_prob': 0.02,
'Style': 'Biker', 'Style_prob': 0.01,
'Look': 'Casual', 'Look_prob': 0.01}}

{'title': 'Plus T-Shirt & Cycle Short Co-ord',
'gender': 'Female',
'tags': {**'Cat': 'Shorts', 'Cat_prob': 0.98**,
'Type': 'Tights', 'Type_prob': 0.92,
'Color': 'Gray', 'Color_prob': 0.79,
'Length': 'Longline', 'Length_prob': 0.76,
'Look': 'Sportive', 'Look_prob': 0.59,
'Style': 'Biker', 'Style_prob': 0.53,
'Texture': 'Cotton', 'Texture_prob': 0.10}}

Summary

- The dataset is composed of 50k items, represented by their title, gender and a list of tags and their probabilities.
- Each item has its own tags combination but the same item can be tagged differently.
  - Male’s items are prone to have more tags than female’s items.
  - The tag ‘Cat’ is the most important tag, without it other tag can not be detected.
  - Is rare to have 13 tags - the item ‘Jacket’ is the most likely item for getting 13 tags.
- The tags probabilities’ reliability is varying according the tags category and the results can change for the same item.