Classifying products by activity automatically

OCTOBER 2022

NAJWA MOULINE

Presentation Outline

1. Objectives
2. Dataset Preparation
3. Pre-processing and Modeling
4. Conclusions

Objectives

Context

- Retail company has launched an e-commerce marketplace
- Sellers offer items to buyers by providing a description of their products
- Assignment of product by family (or activity) done manually by sellers – unreliable and tedious

Business Problem – Multiclassification problem

- Improving user experience for sellers and buyers, and improving business activity overview for the retail company through reliable product classification
- Automating the assignment of the family (or activity) of products for sale

Mission

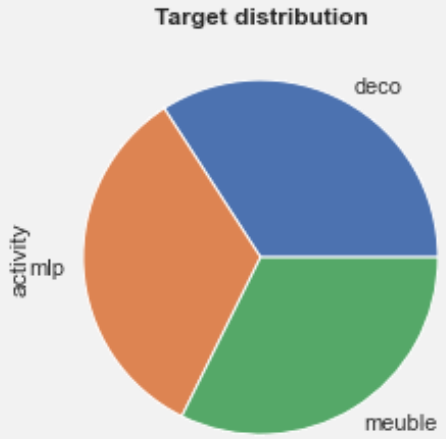
- Build a classification model to determine the family of a product based on product features. Steps involved are:
  - Dataset analysis
  - Modeling
  - Results

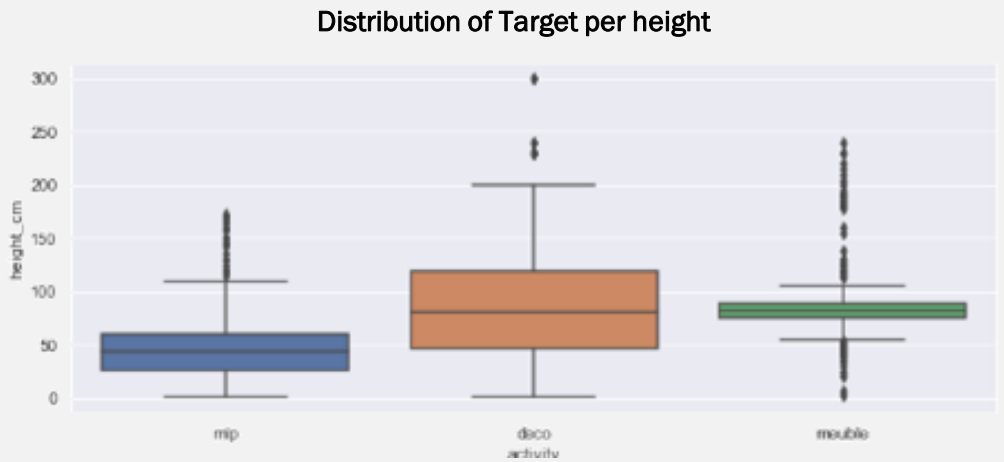
Data Preparation

Dataset

Dataset - structured data

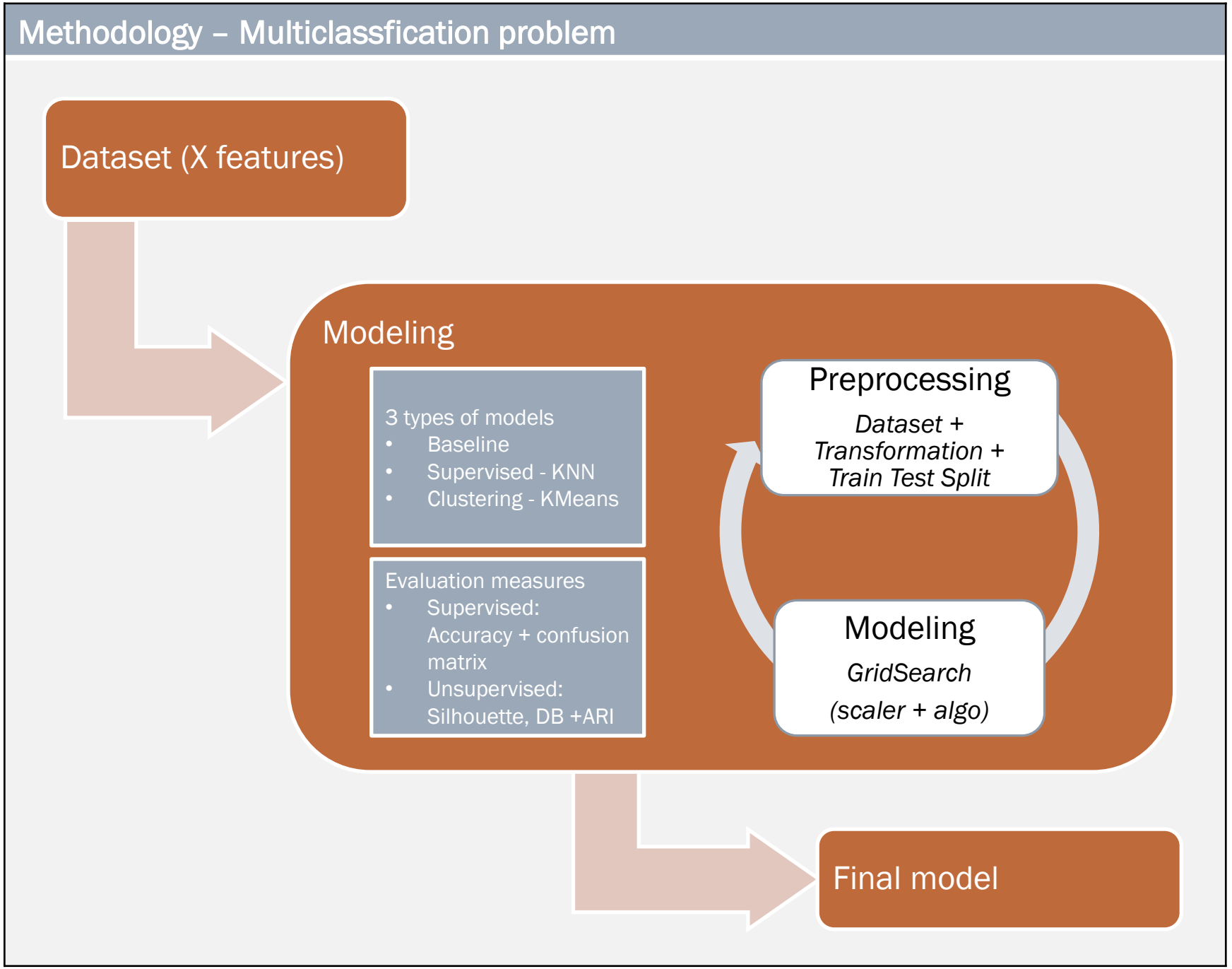
Observations	900 products
Variables	5 variables – 4 quantitative and 1 qualitative (target)
Duplicates	18 removed
NaNs	None
Outliers	3 removed
Features	Product features: height, width, depth and weight Added features: volume, density
Target	Target “activity” has 3 activities The dataset is small but well distributed by activity





Data Preprocessing and Modeling

Methodology

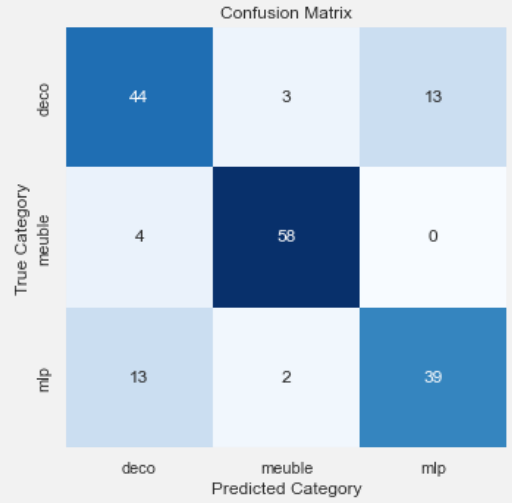


Data Preprocessing and Modeling

Results

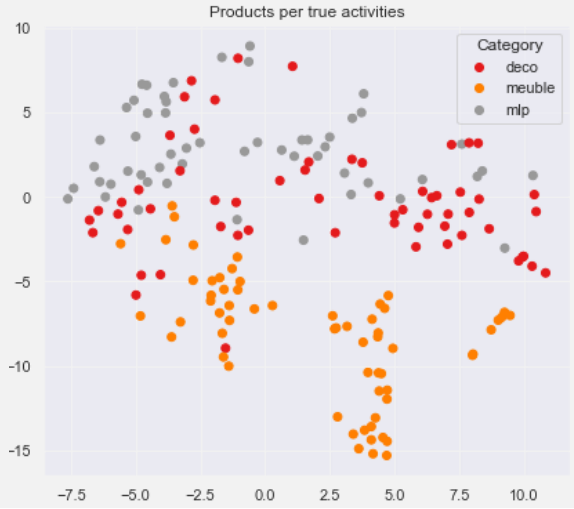
Modeling results

Confusion matrix using KNN

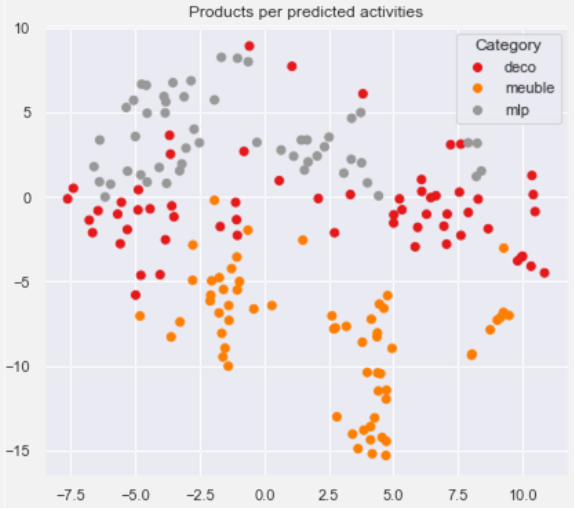


- Prediction of classification issued from KNN gives us best 80% accuracy, with almost right accuracy for « meuble »
- Confusions remain to predict « deco » and « mlp » as expected from EDA.
- Visualisation of products in 2D for both true and predicted activities show clusters that are almost even but not spaced enough.

Visualisation 2D TSNE – predicted activities with KNN considering all features



Accuracy
0.80



Conclusions

What's next?

Feasibility => possible with supervised classification

- Prediction of the classification issued from the KNN model trained on the 3 activities
- In 2D visualisation, true and predicted clusters are not spaced but are distributed equally
- Almost right accuracy for only one of the 3 categories : Meuble

Improvement => options/considerations to further improve the results

- Need a larger database – 300 products per activity is low
- Need more variables (picture, price, …) to better qualify a product to an activity
- Review assignment of activities to certain products – maybe initially miscategorised?
- Consideration of other pre-processing techniques specific to e-commerce
- Consideration of other algorithms and hyperparameters optimisation

Thank you for your attention!