# Webscraping prices to estimate hedonic models and extensions to other predictive methods

Ottawa Group
Rio de Janeiro
May 2019

Cliquez pour ajouter un titre

Insee
Mesurer pour comprendre

Mesurer pour comprendre



# 01 Context

# Price collection for electronic goods in France

- Data collection is performed manually, in physical stores and on the Internet
  - Representativity of different types of stores matters: geographic stratification involves 1pt difference on the CPI from december to april
- We use a fixed basket
- Webscraping is a very interesting source of data, especially for this type of goods:
  - More and more products are sold on the Internet
  - Detailed information on the products, lower cost of collection
- Experiments have been lead on webscraping, mostly for transports
  - Automatic data collection is in production for airplane tickets and maritime transportation
  - We still have to set a general organisation and infrastructure for further use of webscraping in production → we will first use **manually collected data** with models estimated with webscraped data

# Hedonic models and innovative goods

- Importance of taking innovation into account, because it is a major driver of prices
  - New types of products
  - New technical characteristics
  - Improvement of technical characteristics
  - Impact on the price of products already on the market

- Hedonic models can help us measure the technical improvement of electronic goods

- Hedonic re-pricing is preferred in France, and currently in use for household products, because:
  - Easier to check the robustness of the model
  - Fewer statistical analysis are needed
  - For webscraping, no need to have a production infrastructure, only need to gather data at the base month

# Goal: using webscraped data efficiently to estimate the difference in quality

- This involves :
  - Getting data from the websites
  - Cleaning data: from raw data found on the description pages to data sets which can be used for statistical needs
  - Extracting relevant information from these data sets
  - Estimate the price of the good from its technical characteristics → hedonic models, or any type of predictive method
- Hedonic repricing will be used to adjust the base month price:
  - $P_0'(X_k) = P_0(X_k).f(Y_k)/f(X_k)$, where f is the quality function linking the technical characteristics to the price
  - log(f) is linear in the hedonic regression case, but we could use other prediction methods

# 02

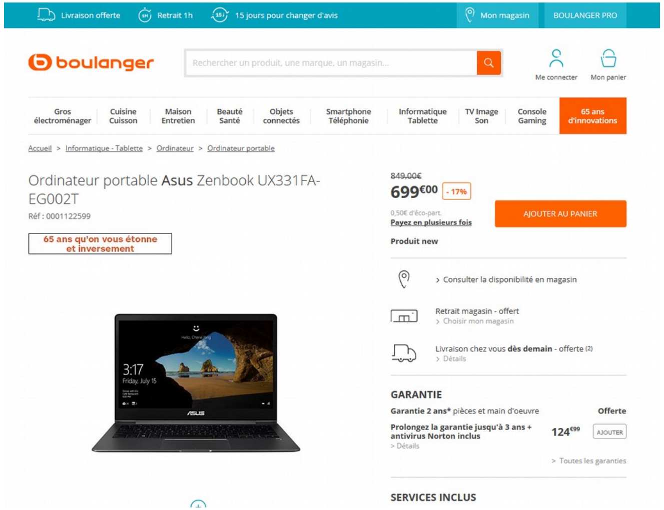
# Data collection and treatment

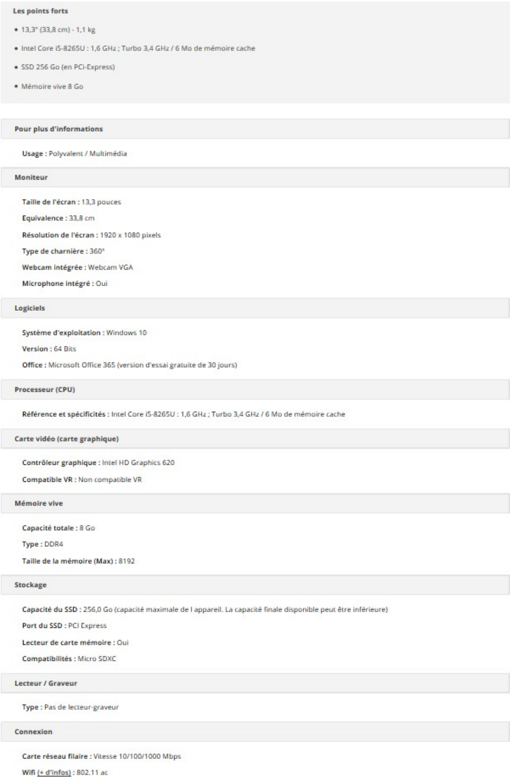
# Data collection

- Data was webscraped from four ecommerce websites; the detailed product page provides information on the technical characteristics

- Scraping was done in Python by statisticians; we still need to figure out how to organise the work between stat and software teams if we do webscraping in production

- Each website has a common page structure for all types of goods, which limits the development costs



# Cleaning data

- We get raw data, we must first transform them to make them useable:
  - Remove technical characteristics with too many missing values
  - Harmonise the format, the levels of discrete variables, the units for continuous ones, transform text into numbers, etc.
  - Detect anomalies
  - Imputate missing values

→ webscraping can make it easier to get data for many types of products, but
  - We have to set a general canvas of treatment
  - We must adapt it to the specificities of each product → modularity

- We have to be careful because some websites mix their products with other sellers, or mix new and repackaged products
  - Some categories even contain a few products which have nothing to do with the other ones!

# Many information on the website… what is relevant ?

- There are many technical characteristics available, but our price collectors will not be able to collect them all

- Even in the case of webscraping in production, we want to limit the number of variables in the case of hedonic models,

- → We have to select the ones which can predict the price
  - Using sectorial expertise/intuition can be a first step
  - Automatic selection through statistical analysis can be more efficient, especially if we want to apply it to many products

# 03 Variable selection and price prediction methods

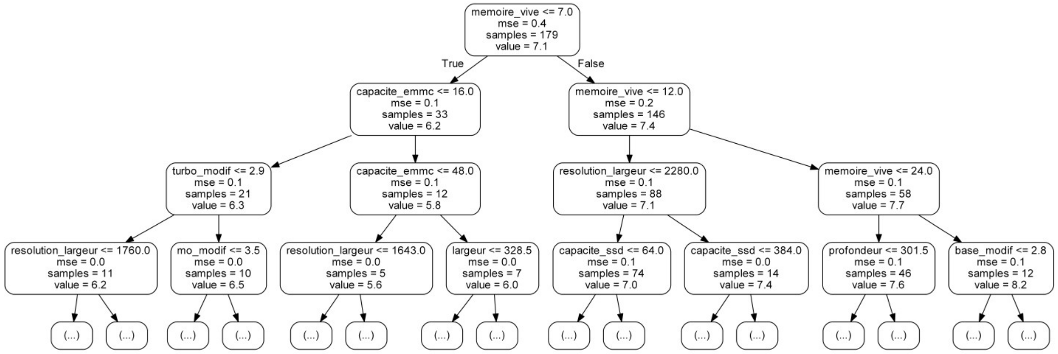
# Variable selection and price prediction



- We want to select a subset of our variables…
- We also want to predict prices with the technical features of the good

→ some statistical learning tools do both!

# Random forests




- Decision trees split the data set at each node, into subsets minimizing the intra-classes variances
- The most relevant features are at the top
- At the bottom of the tree, each cell makes a prediction for observations satisfying conditions of each of the upper nodes (*e.g.* screen size < *15 inch)*
- Among tree-based methods, random forests average several estimators, each one coming from a sample of the original data (sampling observations and variables)
- We can cut the tree at a defined level to get only the most influential nodes
- The trees can be used for prediction, variable selection and variable transformation

# LASSO regression

- LASSO (least absolute shrinkage method) is a regression with penalization of the coefficients

$$\min_{\alpha_1,\ldots,\alpha_p} \sum_{i=1}^{n} (y_i - \alpha_0 - \sum_{j=1}^{p} \alpha_j x_{i,j})^2 + \lambda \sum_{j=1}^{p} |\alpha_j|$$

- The L1 penalization cancels some coefficients, as opposed to the ridge regression (L2)
- We can choose to have more or less non-zero coefficients by making λ vary
  - For prediction purpose, we prefer to use **cross-validation**

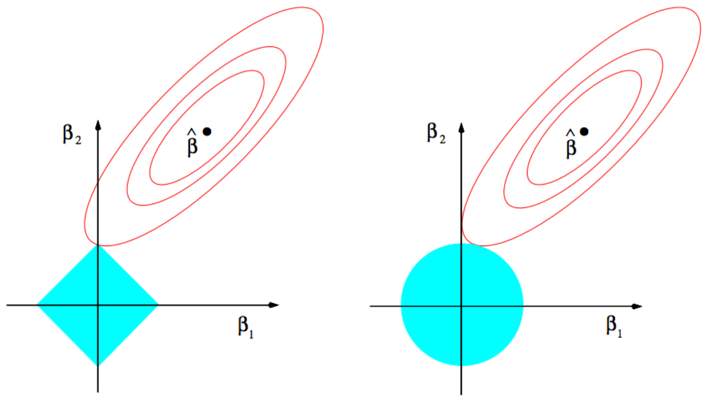



**FIGURE 3.11.** *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \leq t$ *and* $\beta_1^2 + \beta_2^2 \leq t^2$*, respectively, while the red ellipses are the contours of the least squares error function.*

Source: Hastie, Tibshirano, Friedman, *Elements of Statistical Learning*, Springer

# Results for variable selection

- Random forests show that for laptops, the RAM is the most relevant feature (importance around 73% in the trees), followed by the frequency (base and boost) of the processor
  - Other variables selected: dimensions, cache size, weight, resolution, screen size, brand (Apple)

- Random forests provide more stable results (with respect to the website and the collection date) than LASSO

- AIC or BIC stepwise selection could also be used

# Prediction

- Predictive approach :
  - Define a training set, to be split in to subsets in the case of cross-validation
  - Compare models using :
    - **mean squared error**
    - **mean average error**
    - **accuracy = 1 – MAPE** (mean average percentage error) → easier to use

- RF can predict the price with an accuracy around **85%**
  - Up to **89%** if we include the model of graphics card, but difficult to use when new models appear

- LASSO has lower accuracies

- These models usually perform better on the log of price (but we always evaluate the prediction on the price)

# 04 Hedonic replacements

## Use of hedonic models

- We want to reprice our product at base month, using :

$$P_{j,t=0} \cdot \frac{f(X_k)}{f(X_j)}$$

where *j* denotes the replaced product, and *k* the replacing product, and *f* is the function:

$$\exp(a + b_{1.}\, x_1 + b_2 . x_2 + \ldots + b_n . x_n)$$

coming from the model :

$$\log(P) = a + b_{1.}\, x_1 + b_2 . x_2 + \ldots + b_n . x_n + \varepsilon$$

- We plan to apply the model to data collected in stores
  - → we assume that the difference in quality has a comparable effect on the price of different stores, even if the products are priced differently in the different stores

## Use of hedonic models

- Linear models provide good accuracies, around 84%
- Accuracy drop when we combine data from different websites
- R-squared > 0.9
- On the webscraped data, price change estimates were computed using bridged overlap and these hedonic models → results are close, work to be continued throughout the year!

| | March/January with basis month = January | April/March with basis month = March |
|---|---|---|
| Bridged overlap | 95.8 | 98.3 |
| Hedonic model | 96.3 | 98.5 |

# 05 Conclusions and further work

# Conclusion



- Webscraping can provide detailed information about technical characteristics, and help us estimate hedonic models

- Statistical learning models, such as random forests, can be useful for selecting relevant variables quickly
  - → quickly expand the scope of hedonic models to many products, without too much analysis

- They can perform better than traditional hedonic models and could be interesting to use, combined with webscraping in production
  - However, the gain in accuracy is not very important
  - **Possible bias**?

# Some ongoing developments include:

- Testing our models over a longer period
- Using random forests and LASSO predictor to adjust for quality, and compare them to the log-linear model and to (bridged) overlap
- Testing generalized additive models with cubic splines
- A better treatment of missing values (stratified hotdeck)
- Outlier detection → estimating the models without aberrant observations, and selecting replacing products more carefully
- Extension to other electronic goods

# Thanks for your attention !

insee.fr

INSEE
Price Consumer Index Division

Mesurer pour comprendre