Using R to automate the classification of e-commerce products

Aidan Boland

4 November 2017

CLAVIS INSIGHT

EARL
ENTERPRISE APPLICATIONS OF THE R LANGUAGE

Who are Clavis Insight?

- Global leader in action-ready insights for product manufacturers.
- We monitor online retailers to provide insights for manufacturers on their online performance.

Video link

Brands

Grocery, Personal, Home, & Pet

P&G, Unilever, ConAgra Foods, Dr Pepper Snapple, essity, Meyer, Canidae, Henkel, Nestlé, KraftHeinz, Kimberly-Clark, Colgate-Palmolive, Pinnacle, Britvic, Mondelēz, Ferrero, Hershey's, Bimbo, Campbell's, Ocean Spray, Perdue, bel, Keurig

Health & Wellness

Abbott, gsk, Actavis, Johnson & Johnson, rb (Health · Hygiene · Home), Perrigo, Darlie, Pfizer, Bayer, Nature's Bounty

Beauty, Luxury, Fragrance

L'ORÉAL, LVMH, Beiersdorf, PUIG, COTY, The Body Shop

Toys & Games

Mattel, Hasbro, Nintendo

Hardlines

3M, Delta, Bosch

Data & Standards

GS1, World Sync

Alcoholic Beverages

Bacardi Limited, Diageo, SAB Miller, Molson Coors, Beam Suntory, William Grant & Sons

Consumer Elec & Home Appliances

hp, Philips, LG, Duracell, dyson, Miele

B2B

3M, 3663

Sporting & Apparel

Amer Sports

Retailers

450 retailers across x countries

Volume of Data Handled

- Clavis processes a large amount of data monthly.
  - $\approx$ 530,000,000 rows.
  - From 450 stores.
- $\approx$ 40,000 new products added to the database monthly.
- One particular bottleneck is classifying new products, adding tags to help with further analysis.

Classification of Products

- eCommerce sites are fast moving and no two are exactly alike.
  - Different stores have different category names and structures.
- There is a need to provide a standardized category view across multiple sites.
  - Each manufacturer has a different view of categories.
  - Need to be flexible when assigning a product to a category.

Classification of Products

- Amazon
  - Grocery & Gourmet Food > Candy & Chocolate > Bars
- Walmart
  - Food > Candy & Gum > Chocolate

Classification of Products

	Portfolio Availability	Price and Promotions	Ratings	Reviews	Share of Search	Menu	Total Content	Core Content	Extended Content
All	44 %	90 %	61 %	47 %	36 %	22 %	54 %	61 %	92 %
Breath Fresheners	55 %	100 %	72 %	60 %	16 %	0 %	31 %	49 %	88 %
Candy - Chocolate	39 %	93 %	62 %	45 %	33 %	23 %	66 %	73 %	93 %
Candy - Non-Chocolate	59 %	83 %	59 %	41 %	17 %	20 %	32 %	41 %	88 %
Grocery	48 %	88 %	69 %	50 %	20 %	41 %	68 %	75 %	94 %

Classification of Products

- When a new product is found on a retailer, the information is collected.
- Each new product is then classified into it's correct category.
- As a manual task this is very time consuming, error prone and requires a lot of manual intervention by users.
  - $\approx$ 5 hours per customer end to end.
  - Source of complaints from customers.

- The difficulty of this process was a major indication for the need to automate.

Automating Classification

Roadmap for Automating Classification

- Research
  - Find suitable algorithm.
- Beta testing
  - Test the algorithm within the current process.
- Integrating code
  - Build the code into the current company platform.

Research Phase

- Task
  - Classify new products into their correct categories and subcategories (up to 8 levels).
    - Efficient.
    - Accurate.
- Solution
  - Supervised classification.

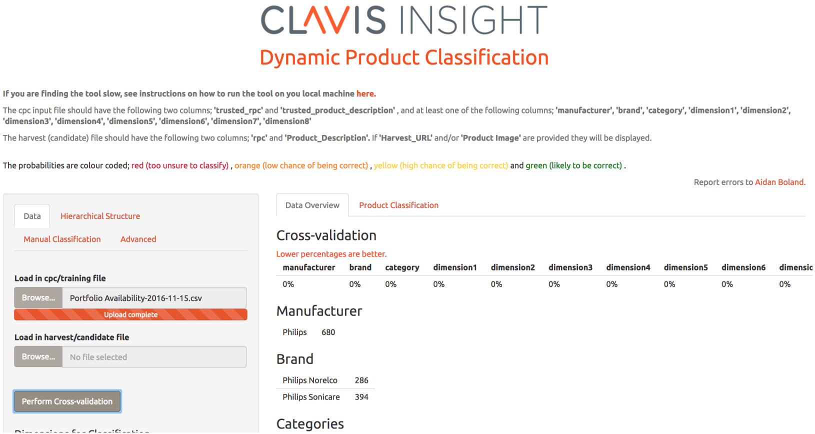
Supervised Classification.

- Learn from the current products in order to classify new products.
- Many methods were considered (maxent, random forests, k-nn).
- Final algorithm was personalized to this particular task.
  - Classify's all levels of hierarchy simultaneously.
  - Relatively efficient ($<$30 seconds for 1000 products)
  - Ability to handle all languages.

- Researching and creating the algorithm is only one step.

Beta Testing

Shiny Application



Beta Testing

- Shiny application
  - Hosted on an internal server.
- Allows the ability to manually load in training data and new data.
- Final downloaded is in the correct format to load into the production database.

- Still manual touch.

Full Integration

Integration into Platform

- The backend is coded in Java.
- Algorithm is coded in R.

- Needed to integrate the R code into the backend code.
- Must follow the companies standard practices.
  - Version control (bitbucket).
  - Change testing (unit tests).

Solution

- R library (including unit tests).
- Stored in a repository on Bitcucket.
  - This allows version control and updates which are in line with company practices.

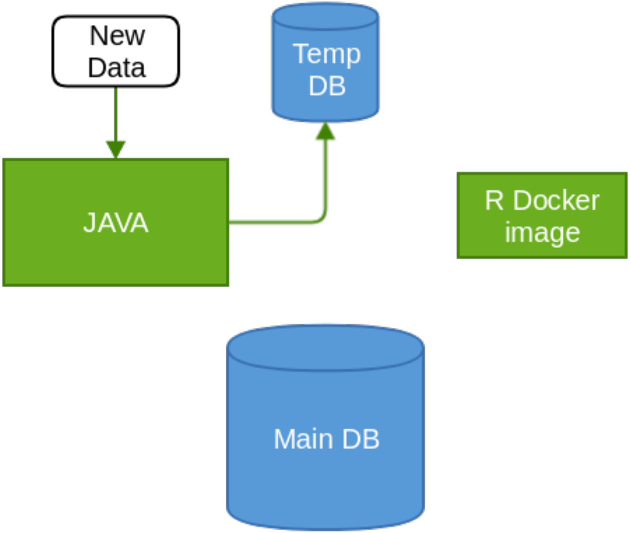
- The **plumber** library is used to allow the code to be called using an API.
  - Arguments in the API point the R code to the database containing the new data.
  - R code classify's the new data and updates the database with the new categories.

Solution

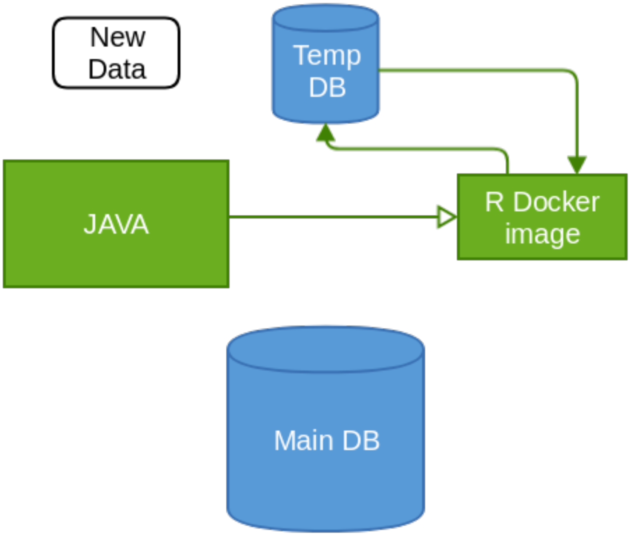
- A Docker image containing the R library and any requirements is stored.
  - The Docker image is automatically rebuilt when code is updated or a new production branch is created.
  - R library is recompiled and unit tests are run when the Docker image is rebuilt.

- Java code can spin up a new server and run the Docker image which will contain the code and will allow calls to be made to the API.
  - Multiple Docker images can run at any one time.

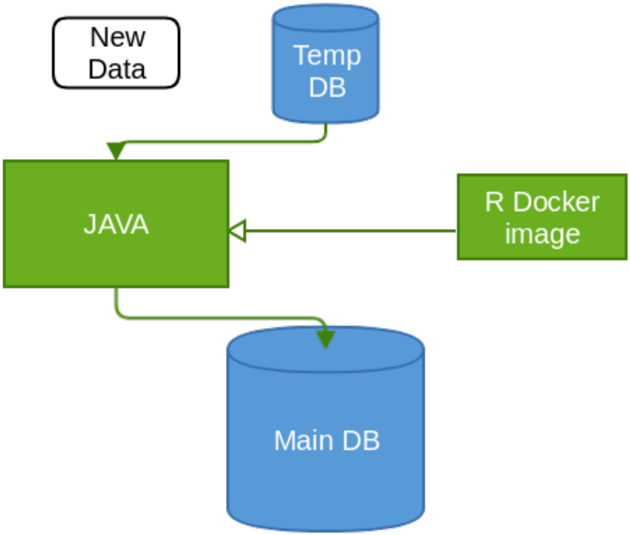
Final Flow of Data



Final Flow of Data



Final Flow of Data



CLAVIS INSIGHT

EARL
ENTERPRISE APPLICATIONS OF THE R LANGUAGE

Improvement in Time to Add New Products

- Before automated classification
  - $\approx$ 5 hours per customer end to end.
  - 76 customers processed monthly.
  - Major source of complaint from customers.
- After automated classification
  - $\approx$ 2 hours per customer end to end.
  - 159 customers processed monthly.
  - No ‘red’ complaints since March.