Predicting company default with webscraping data

Josep Domenech
Universitat Politècnica de València, Spain
jdomenech@upvnet.upv.es

Predicting company default

- Business credits depend on the ability of partner companies to return the credit
- Economic literature usually relies on:
  - Age
  - Size
  - Sector
  - Debt
  - Financial ratios
- It is specially challenging to predict SMEs default

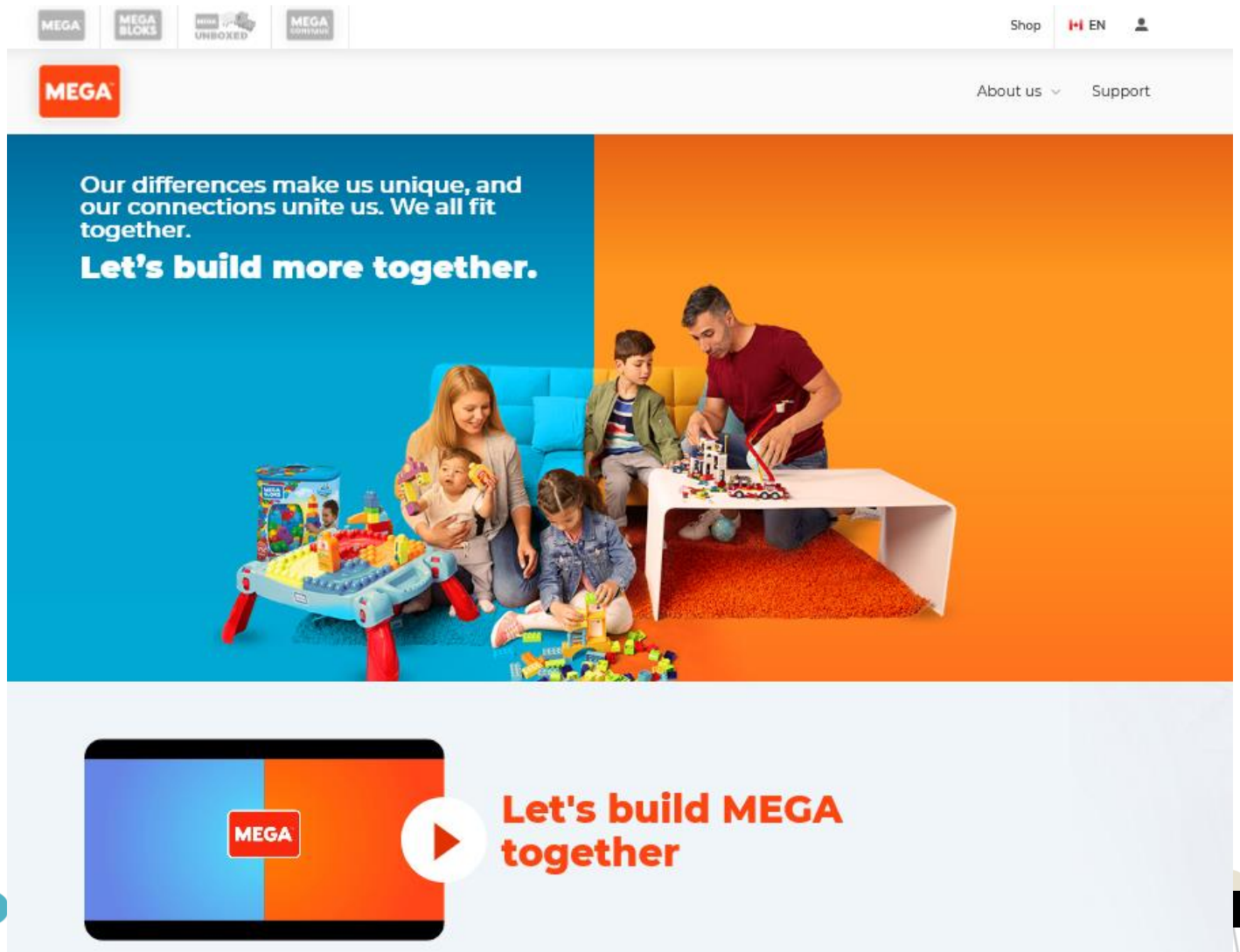
Are websites informative?

- Intuition:
  - Company websites show company activities
  - Lower economic activity and increased default risk when:
    - Lack of website activity
    - Using stale technologies

Are websites informative?

- Information on company websites:
  - Content
    - Description of the activity:
      - Sector
      - Main products / services
      - Contact information
  - HTML Code
    - HTML
      - Organization of the content
      - Links to other pages/sites
      - Contact methods
    - Metadata
      - Technology (generator)
      - User orientation (search engines)
  - Server response
    - Server technology
    - Content freshness (age or last modified)

Content


MEGA
MEGA BLOKS
UNBOXED
Shop
EN
About us
Support
Our differences make us unique, and our connections unite us. We all fit together.
Let's build more together.
Let's build MEGA together

HTML Code

```
<!DOCTYPE html>
<html>
    <head>
        <meta name="google-site-verification" content="x-AAwX5KCaF-U4WZf_21P76oRwxsy2lMVjK4zgMn-0M">
        <meta name="msvalidate.01" content="610FD5B0342477063B48B429B4D57F42">
        <meta name="viewport" content="width=device-width, initial-scale=1">
        <meta http-equiv="X-UA-Compatible" content="IE=edge">
        <meta charset="utf-8">

            <meta property="og:title" content="Mega Brands">
    <meta property="og:url" content="https://www.megabrands.com/en-ca/">
    <meta property="og:image" content="https://s4.megabrandsmedia.com/2019/05/03/10/00/19/MOZ10U2es51556892019
    <meta property="og:description" content="Mega Brands">

        <link rel="shortcut icon" href="/favicon.ico" type="image/vnd.microsoft.icon"/>
        <title>
            Mega        </title>

                    <link rel="stylesheet" href="https://proxy.megabrands.com/css/mega-4cbf196fea.min.css" typ

        <link rel="stylesheet" href="https://proxy.megabrands.com/fonts/mega/mega-icons/v1/css/mega-icons.min.
        <link rel="stylesheet" href="https://proxy.megaconstrux.com/fonts/construx/construx-icons/css/construx

                    <link rel="stylesheet" href="https://www.megabrands.com/css/vendor-2b1e89ad08.min.css" typ
            <link rel="stylesheet" href="https://www.megabrands.com/css/main-334b5a1843.min.css" type="text/cs

                            <script src="https://www.megabrands.com/js/vendor-98f2422960.min.js"></script>

            <link rel="canonical" href="https://www.megabrands.com/en-us/" />

        <script type="application/ld+json">
    {
        "@context": "https://schema.org",
        "@type": "WebSite",
        "url": "https://www.megabrands.com/",
        "potentialAction": {
```

Server response

GET https://www.megabrands.com/en-ca/

Status	200 OK
Version	HTTP/2
Transferred	223.11 KB (354.82 KB size)

Response Headers (904 B)

```
age: 10633
cache-control: public, s-maxage=14400, public, must-revalidate
content-encoding: gzip
content-type: text/html; charset=UTF-8
date: Sat, 08 May 2021 10:09:45 GMT
server: Apache
set-cookie: lc=en-ca; expires=Tue, 06-May-2031 10:09:45 GMT; Max-Age=315360000; path=/; domain=.megabrands.com; secur
set-cookie: incap_ses_885_926694=2FOud4+xa3XVYsvSHidIDGhjImAAAAAAShapX/6I4f+F1d+dLamF6w==; path=/; Domain=.me
via: 1.1 92974644c95de2a8e1e1b0062afcb761.cloudfront.net (CloudFront)
x-amz-cf-id: -tJ5URvQc-95SmtCVfXDTgyIQavpYGrVkstnJY3ozBc-E0f2MGhLvA==
x-amz-cf-pop: MAD51-C1
x-cache: Miss from cloudfront
x-cdn: Imperva
x-content-digest: enff25e53e25a843e7f8afe5e0374be2d2375bb6507c870a9ebcedbe314210b0fc
x-content-type-options: nosniff
X-Firefox-Spdy: h2
```

Machine learning approach

- Supervised learning problem:

  $p(default_i) = f(website_i)$

  – Any classification method could be applied:
    - Logistic regression
    - Classification trees
    - Random forest
    - SVM
    - …
- Performance evaluation with cross validation
  – Train set
  – Test set

Machine learning approach

- Challenge: website feature extraction
  - Text content: Natural Language Processing
    - Bag of words
    - Lemmatization
    - …
  - HTML: Occurrences
    - Tags
    - Content of some specific tag attributes:
      - href in A tags
      - href in LINK tags
      - name in META tags
    - Parts of attribute contents
      - extension in hrefs (e.g., asp, php, html, pdf)
      - some specific words in hrefs (e.g. mailto, tel, twitter)
  - Server response: Occurrences
    - Headers
    - Specific header contents

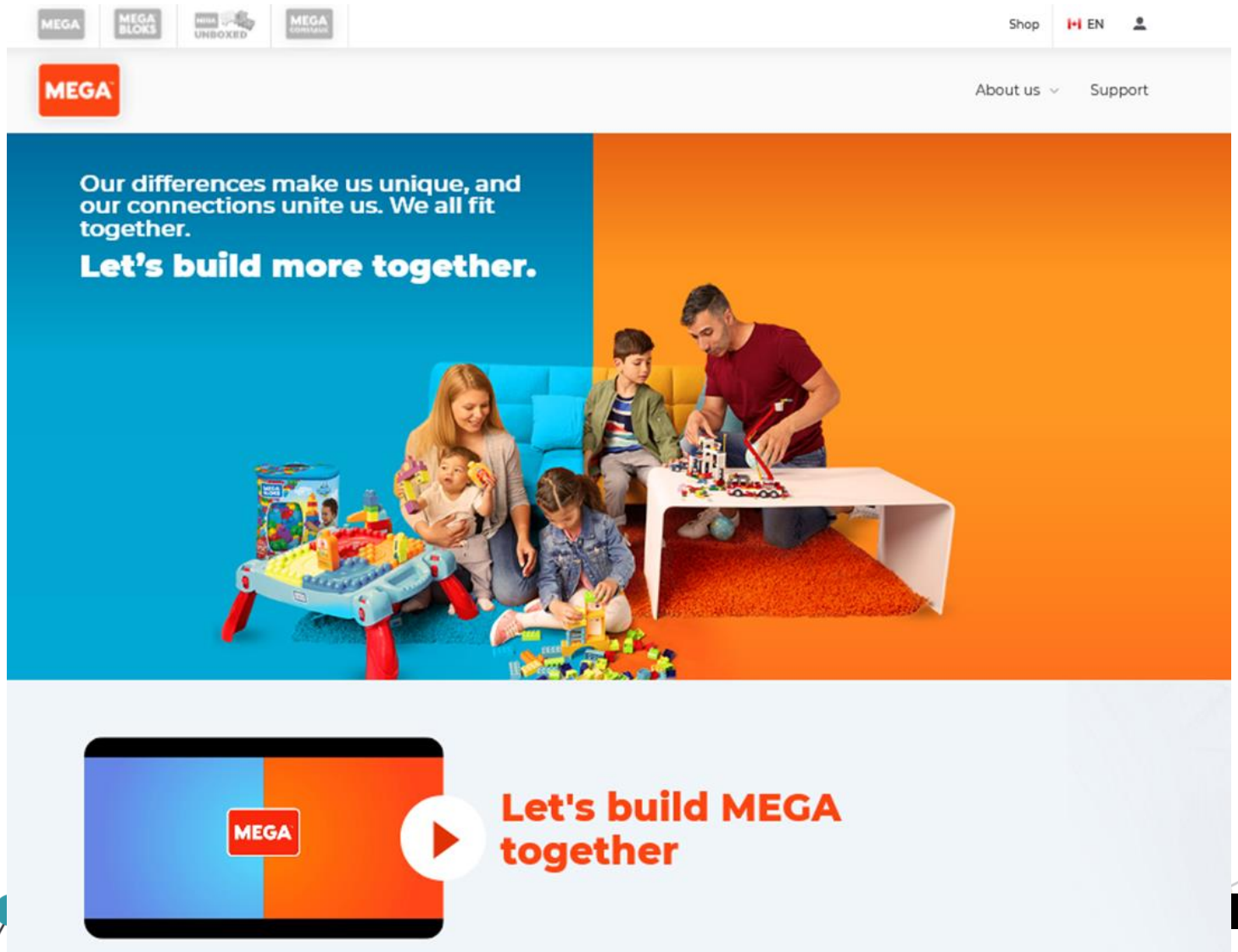
Training the model

- Putting data together:
  - Default information from company databases (e.g., Bureau Van Dijk)
  - Website scrapping
    - Home page or whole site?
  - Temporal consistency:
    - Do default and website information refer to the same time period?
    - Prediction horizon

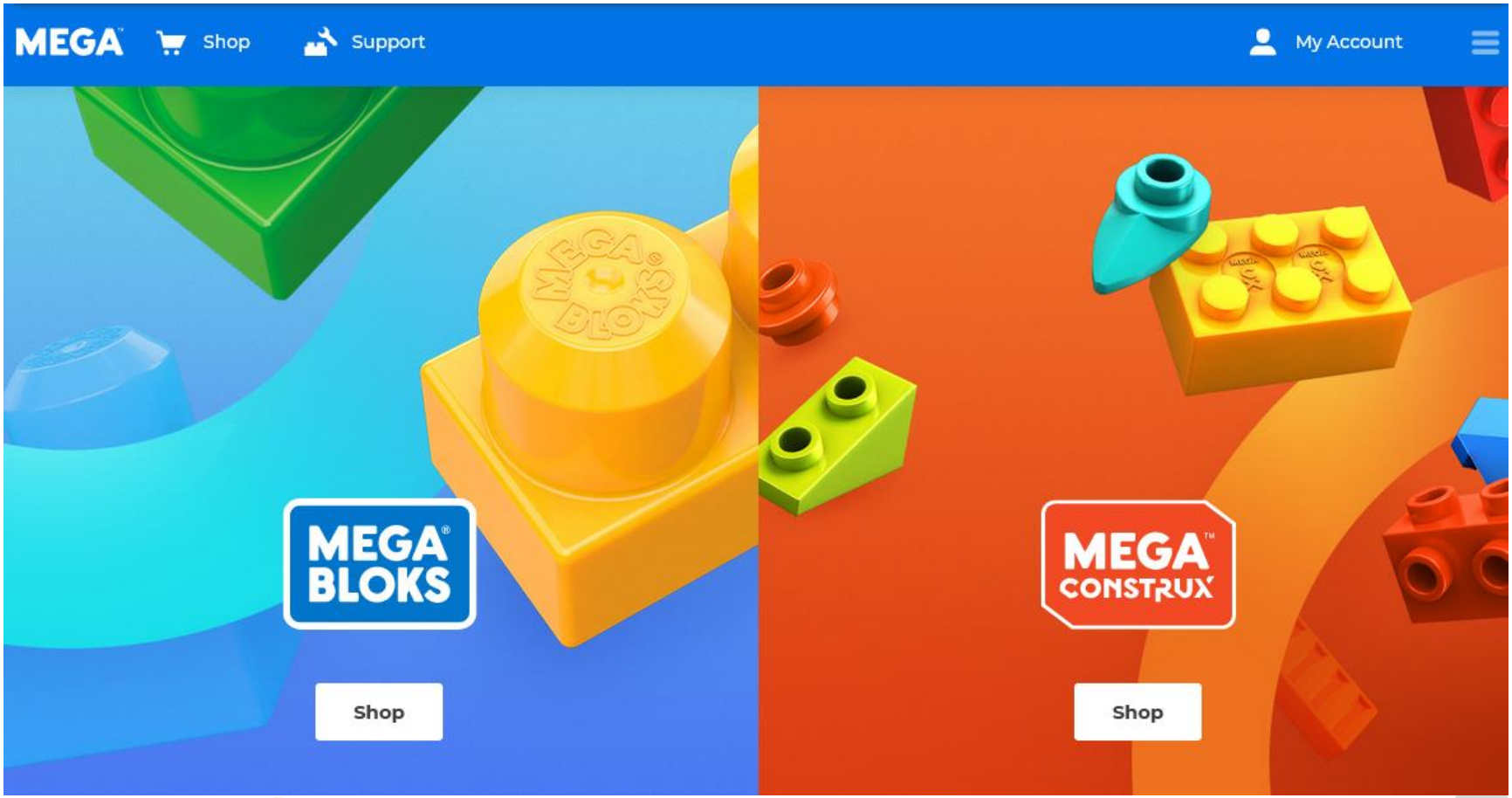
Training the model

- Accessing websites from the past:
  - Wayback Machine of the Internet Archive
- Limitations
  - Incomplete coverage, both in terms of websites and pages
  - Server response headers are not stored

2021


MEGA
MEGA BLOKS
UNBOXED
Shop
EN
About us
Support
Our differences make us unique, and our connections unite us. We all fit together.
Let's build more together.
MEGA
Let's build MEGA together

2019


MEGA
Shop
Support
My Account
MEGA BLOKS
Shop
MEGA CONSTRUX
Shop

2011


Français | Español
Search
MEGA
Home
Brands
Kids Zone
Corporate
Family Club
Videos
BLOK SQUAD
VISIT THE WEBSITE
MEGA BLOKS
RoseArt
MEGA PUZZLES
The BOARD DUDES

2008

MEGA™
Creativity to the Rescue.™

This page requires Adobe Flash Player 8 or above and javascript enabled.

Click Here to download Flash Player

Home Page

About Us | Shop | Kid Zone | Family Fun | Customer Service

Contact Us
Legal Notice
Privacy Policy

Investor Relations
Investor Info
Financial Reports
Press Releases
Donations
Careers

Shop
- Construction Blocks
- Magnets
- Arts and Crafts
- Games and Activities
- Stationery
- Tech Toys

Award-Winners

Toys
- Magnetix
- Dragons
- Pirates Of The Caribbean 2
- Pirates Of The Caribbean 3
- Pyrates
- Plasmaverse
- Neoshifters
- B'Chic

Movie/Animation
- Magnetix
- Magnaman
- Magwarriors
- Dragons
- Dragons Metal Ages
- Pirates Of The Caribbean 3
- Spiderman 3

Family Fun
Activity Center
Learning Benefits
MEGA Club
Parent Panel
Contest

Request Center
Instructions
F.A.Q
Recall Information

Training the model

- Classification methods perform better with balanced samples
- Balancing the sample
  - Oversampling default firms
  - SMOTE
    - Adds synthetically generated observations

Training the model

- Too much variability?
  - Transformation of website features into binary variables
- Too many variables?
  - LASSO regression
    - Variable selection
    - Regularization
  - Reducing dimensionality
    - Multiple Correspondence Analysis
    - Kernel Discriminant Analysis

Evaluation

- Evaluation on the test set
  - Multiple repetitions
- Metrics:
  - Accuracy
    - Which proportion of companies were correctly classified as defaulters/non-defaulters?
  - Sensitivity
    - True positive rate
    - Which proportion of defaults was detected?
  - Specificity
    - True negative rate
    - Which proportion of non-defaulters was detected?

Results

After 100 repetitions:

Model		Sensitivity	Specificity	Accuracy
		Online data		
RBF	mean *(sd)*	0.706 *(0.207)*	0.708 *(0.073)*	0.708 *(0.056)*
	median *(mad)*	0.778 *(0.165)*	0.733 *(0.066)*	0.747 *(0.052)*

https://doi.org/10.1016/j.econlet.2021.109888

Conclusions

- Websites are a rich source of information yet to be exploited
- Specific techniques must be used to transform them into usable data sources

Predicting company default with webscraping data

Josep Domenech
Universitat Politècnica de València, Spain
jdomenech@upvnet.upv.es

Crosato, L., Domenech, J., & Liberati, C. (2021). Predicting SME's default: Are their websites informative?. *Economics Letters*, 109888. 10.1016/j.econlet.2021.109888
Blazquez, D., Domenech, J., & Debón, A. (2018). Do corporate websites' changes reflect firms' survival?. *Online Information Review*. 10.1108/OIR-11-2016-0321

Main discriminant indicators

Indicator	Defaulted	Survived	Total
htmltags.div	0.28	0.65	0.63
htmltags.a	0.31	0.67	0.66
hrefwords.www	0.21	0.53	0.52
htmltags.script	0.34	0.67	0.66
htmltags.br	0.17	0.49	0.48
htmltags.meta	0.45	0.76	0.75
htmltags.title	0.48	0.77	0.76
htmltags.h	0.14	0.41	0.40
htmltags.link	0.55	0.81	0.80
linkhref.ext.css	0.55	0.81	0.80
htmltags.html	0.55	0.80	0.79
hrefwords.index	0.07	0.30	0.30
hrefwords.de	0.03	0.24	0.23
htmltags.style	0.14	0.34	0.33
hrefwords.web	0.03	0.22	0.21
htmltags.meta.L	0.00	0.13	0.13
hrefwords.mailto	0.03	0.16	0.16