Scalable Machine Learning Pipeline for Metadata Discovery from eBay Listings

Qing Zhang, Rui Li

eBay

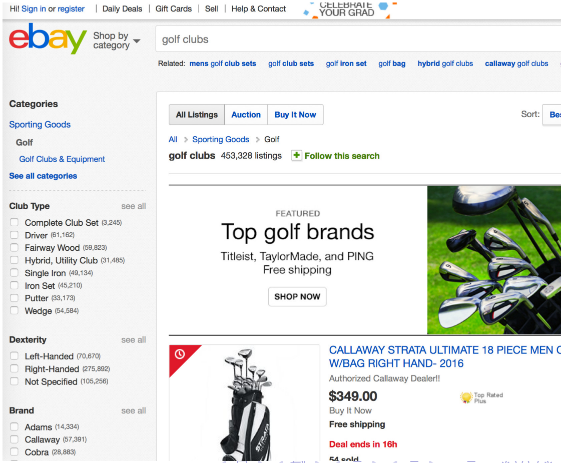
Spark Summit 2016, June 6-8 San Francisco

Table of Contents

1. Who We Are
2. Metadata Discovery and Challenges
3. Spark Solution
4. Summary

eBay Structured Data

- Metadata discovery and management
- Listing classification
- Catalog and mapping listing to product
- Inventory insights



Hi! Sign in or register | Daily Deals | Gift Cards | Sell | Help & Contact

CELEBRATE YOUR GRAD

Shop by category

golf clubs

Related: mens golf club sets golf club sets golf iron set golf bag hybrid golf clubs callaway golf clubs

Categories

Sporting Goods

Golf

Golf Clubs & Equipment

See all categories

Club Type see all

- □ Complete Club Set (3,245)
- □ Driver (61,162)
- □ Fairway Wood (59,823)
- □ Hybrid, Utility Club (31,485)
- □ Single Iron (49,134)
- □ Iron Set (45,210)
- □ Putter (33,173)
- □ Wedge (54,584)

Dexterity see all

- □ Left-Handed (70,670)
- □ Right-Handed (275,892)
- □ Not Specified (105,256)

Brand see all

- □ Adams (14,334)
- □ Callaway (57,391)
- □ Cobra (28,883)

All Listings | Auction | Buy It Now

Sort: Bes

All > Sporting Goods > Golf

golf clubs 453,328 listings ✚ Follow this search

FEATURED

Top golf brands

Titleist, TaylorMade, and PING

Free shipping

SHOP NOW

CALLAWAY STRATA ULTIMATE 18 PIECE MEN C W/BAG RIGHT HAND- 2016

Authorized Callaway Dealer!!

$349.00

Buy It Now

Free shipping

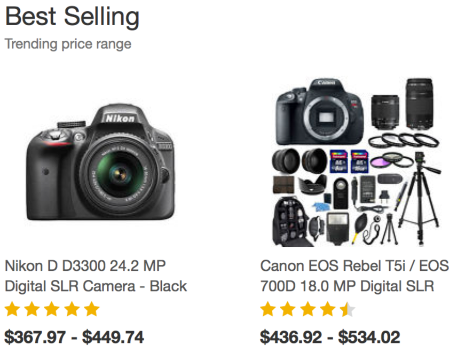
Top Rated Plus

Deal ends in 16h

54 sold

eBay Structured Data

- Metadata discovery and management
- Listing classification
- Catalog and mapping listing to product
- Inventory insights



Best Selling

Trending price range

Nikon D D3300 24.2 MP Digital SLR Camera - Black

★★★★★

$367.97 - $449.74

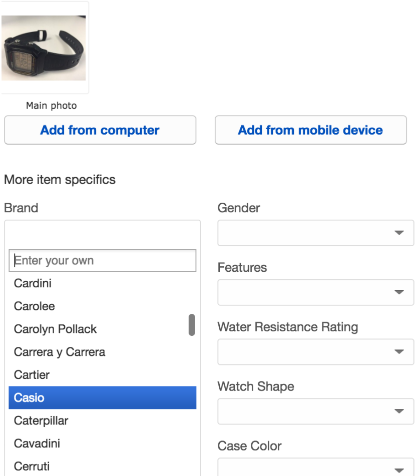
Canon EOS Rebel T5i / EOS 700D 18.0 MP Digital SLR

★★★★½

$436.92 - $534.02

eBay Structured Data

- Metadata discovery and management
- Listing classification
- Catalog and mapping listing to product
- Inventory insights



Metadata Zoom In

- Important name-value pairs: *brand* - Dell
- Selling flow item specifics
- Search navigation
- Powers internal applications

Metadata Discovery

memory box	weiss	yamaha	fisher-price	generic
modway	duluth trading	mek usa dnm	other	sahara club
gokey	longhorn	outdoor gear	trax	wolverine
sk	spiderman	vintage	mixed	orchard corset

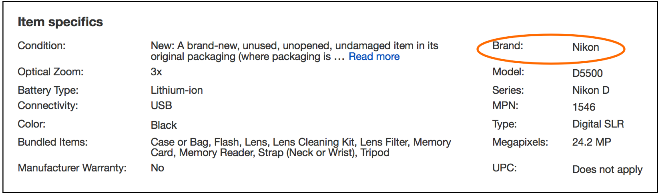
- Highly rely on manual review
- Unfamiliar candidates
- The same candidate appears in multiple categories

Challenges in Metadata Discovery

Term	Site	Categories
scott james	US	Men's Clothing: Blazers & Sport Coats
		Men's Clothing: Pants
		Men's Clothing: Casual Shirts
		Men's Clothing: Dress Shirts
tiella	US	Chandeliers & Ceiling Fixtures
		Lighting Parts & Accessories
turf	US	Sports Mem, Cards & Fan Shop: Cards: Football
turf	UK	Collectables: Cigarette/Tea/Gum Cards: Cigarette Cards: Other Cigarette Cards

Data Driven Approach for Brand Discovery

- Utilize seller input item specifics
- Utilize supply demand signals from sellers and buyers
- Training data available from previously reviewed candidates



Item specifics

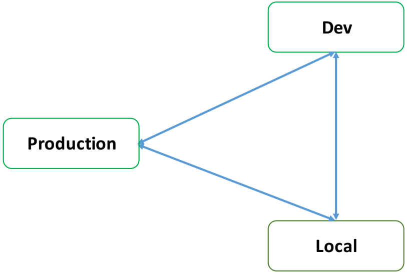
Condition:	New: A brand-new, unused, unopened, undamaged item in its original packaging (where packaging is ... Read more	Brand:	Nikon
Optical Zoom:	3x	Model:	D5500
Battery Type:	Lithium-ion	Series:	Nikon D
Connectivity:	USB	MPN:	1546
Color:	Black	Type:	Digital SLR
Bundled Items:	Case or Bag, Flash, Lens, Lens Cleaning Kit, Lens Filter, Memory Card, Memory Reader, Strap (Neck or Wrist), Tripod	Megapixels:	24.2 MP
Manufacturer Warranty:	No	UPC:	Does not apply

Supervised Machine Learning Approach

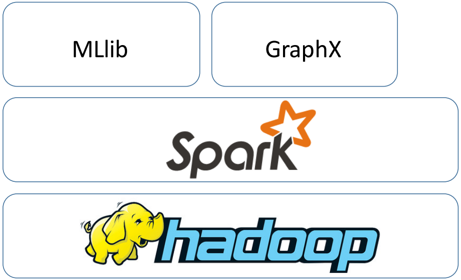
- Data : 35,000 previously human reviewed metadata candidates
- Feature : supply and demand signals
- Prototypes with Python Scikit
- Logistic regression, gradient boosting trees, random forest etc
- Random forest F1 0.878

Production Development

- In the past, train offline and implement prediction component on production
- File transferring and configurations are time-consuming



Spark and Spark MLlib

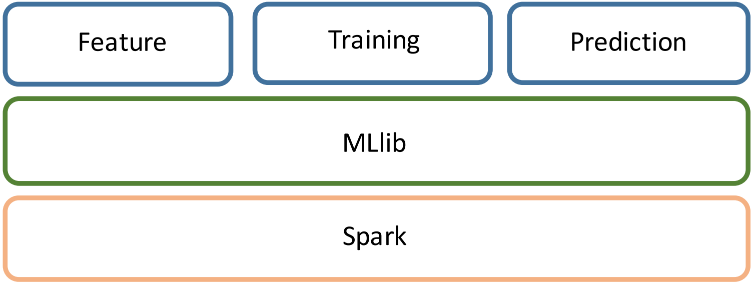


MLlib

GraphX

- Spark provides powerful data processing APIs
- MLlib is a comprehensive machine learning package powered by Spark
- Regression, classification, clustering, dimensionality reduction etc
- Efficient development with local model, and flexible file access

The Machine Learning System



Model Training with MLlib

```
val pipeline = new Pipeline()
  .setStages(Array(labelIndexer, featureIndexer,
                   rf, labelConverter))

val evaluator = new MulticlassClassificationEvaluator()
  .setLabelCol("indexedLabel")
  .setPredictionCol("prediction")

val cv = new CrossValidator()
  .setEstimator(pipeline)
  .setEvaluator(evaluator)
  .setEstimatorParamMaps(paramGrid)
  .setNumFolds(5)
```

Evaluations

Model	F1
Python Scikit prototype	0.878
MLlib local	0.865
MLlib Hadoop (200 executors, production)	0.862
MLlib Hadoop (400 executors)	0.861
MLlib Hadoop (50 executors)	0.857
MLlib Hadoop (2 executors)	0.862

- The performance variations among implementations are acceptable

Speed

Stage	Data Size	Time
Feature Generation	1.73 Billion	6 min
Train	33,000	8 min
Prediction Input	650,000	4 min

- Capable of running the job daily
- Speed up the metadata discovery process, from months to days

Newly Discovered Brand

Brand	Probability
Milkies	0.84
BEABA	0.85
OXO	0.83
Lorex	0.87
Plan Toys	0.85
Safety 1st	0.82
Blabla	0.81
Combi	0.88
Graco	0.88
TotsBots	0.85
Realtree	0.85

Summary

- Spark enables fast iterations of ML application development
- MLlib is comprehensive, and well integrated with Spark framework
- Dev and test locally, straightforward production deployment
- Compact code : 600 lines
- Need better understanding of the ML algorithm implementations in MLlib

Acknowledgement

Thejas Durgam
Anu Mandalam
Meital Tahar Zahav & eBay SDO Team
Jean-David Ruvini

Thank You!

Qing Zhang, qzhang12@ebay.com
Rui Li, ruili1@ebay.com