Quality-Adjusted Price Indices Powered by ML and AI

Amazon Core AI

Science-Engineering Team:

P. Bajari, V. Chernozhukov (+MIT), R. Huerta (+UCSD), G. Monokrousos, M. Manukonda, A. Mishra, B. Schoelkopf (+ Max Plank)

Motivation

- Inflation indices are important inputs into measuring aggregate productivity and cost of living, and conducting monetary and economic policy.
- We want to contribute to the science of inflation measurement based on quality-adjusted prices.
- Main challenges today:
  1. millions of products (global trade environment);
  2. prices change quite often (often algorithmically by sellers);
  3. extremely high turnover for some products (e.g., apparel, electronics).
- Our teams addressed these challenges to produce a method that utilizes scalable ML and AI tools to predict quality-adjusted prices using text and image embeddings

- We want to share our findings:
  - 1/Deep learning embeddings work as input features for hedonic price models.
  - 2/ Random Forest and other Machine Learning models lead to superior price prediction.
  - 3/ Fusion of engineers and scientists in teams lead to faster experimentation and deployment of models.

Outline

1) Price Indices
2) Quality-Adjusted (Hedonic) Price Indices
3) Hedonic Prices Indices Using ML and AI
    1) Feature Engineering from Text
    2) Feature Engineering from Images
    3) Nonlinear Price Prediction using Random Forest
4) Conclusion

Transaction-Price Quantity Index (TPQI)

- Price P_{jt} and quantity Q_{jt} for product *j* in period *t*
- Transaction-Price Quantity Indices are based on **matching**:

Paasche Index: $$R_t^{P,M} = \frac{\sum P_{jt}Q_{jt}}{\sum P_{j(t-1)}Q_{jt}}$$

Laspeyres Index: $$R_t^{L,M} = \frac{\sum P_{jt}Q_{j(t-1)}}{\sum P_{j(t-1)}Q_{j(t-1)}}$$

Fisher Index: $$R_t^{F,M} = \sqrt{R_t^{P,M} R_t^{L,M}}$$

where the summation in the denominator/numerator is over the matching set (largest common set).

- Missing products **create biases in the matching set**.

Need for Hedonics (Quality-Adjusted Pricing)

- To avoid biases in the **matching set**, we can predict prices of missing products in period-to-period comparisons.
- This is especially relevant for product categories with high turn-over.
- In product groups like apparel, **about 50% of products get replaced** with new products **every month**.
- **Use predicted prices**, **using product attributes or qualities**, instead of the observed prices

Hedonic Price Quantity Index

- **Replace prices by quality-adjusted prices**

Paasche Index: $$\hat{R}_t^{P,M} = \frac{\sum \hat{P}_{jt} Q_{jt}}{\sum \hat{P}_{j(t-1)} Q_{jt}}$$

Laspeyres Index: $$\hat{R}_t^{P,M} = \frac{\sum \hat{P}_{jt} Q_{j(t-1)}}{\sum \hat{P}_{j(t-1)} Q_{j(t-1)}}$$

Fisher Index: $$\hat{R}_t^{F,M} = \sqrt{\hat{R}_t^{P,M} \hat{R}_t^{L,M}}$$

The Hedonic Price Model

The model we want to fit is

$$\underbrace{\log(P_{it})}_{Y_{it}} = f_t(\underbrace{W_{it}, I_{it}}_{X_{it}}) + \epsilon_{it},$$

where P_{it} is the price of product i at time t, and the pricing function f_t can change from period to period. Most of the product characteristics X_{it} will remain time-invariant, but some may change over time. We will use the data from period t to estimate the function f_t. We will estimate f_t by the nonlinear regression method called the Random Forest (RF), which is known to be well-designed for parallel implementations.

What are the features?

Query: red dress

Image

Customer behavior data

3,989 customer reviews | 259 answered questions

Description

- Material - Cotton & Spandex.
- Imported
- Classic and Iconic Audrey Hepburn 50s Vintage Solid Color Swing Dress, Put on and Show Your Elegance and Charm.
- Features: Boat Neckline; Sleeveless; Full Circle Swing; Quick Access Zipper for Easy On and Off
- It's Great Choice for Daily Casual, Wedding , Ball, Party, Banquet and Other Occasion.
- [Size Chart] PLEASE Make Sure Your Measurements and Compare to the Size Chart From the picture on the left side or in the Following Description.
- Hand Wash Carefully,Low Temperature for Washing,Can not High Temperature Ironing, Line Dry

$X \in \mathbb{R}^d$

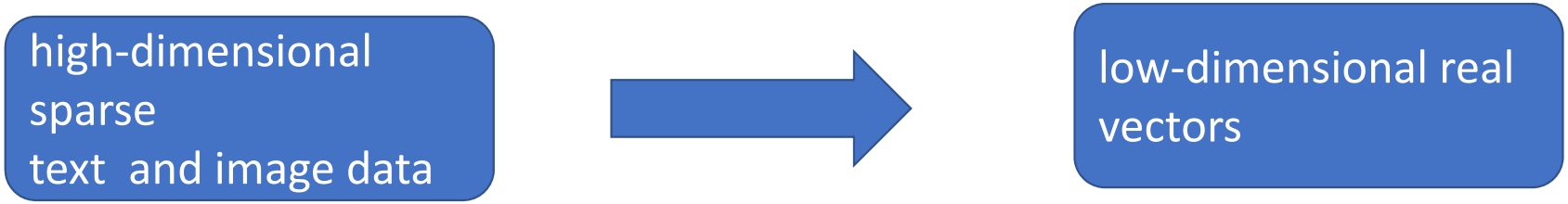
Title

Anni Coco

Anni Coco Women's Classy Audrey Hepburn 1950s Vintage Rockabilly Swing Dress

On Deep Learning Features

- Think of them as produced by dimensionality reduction:



- Open Source State-of-the-art Deep Learning methods:

a) Text: **Word2Vec**

b) Images: **GoogLenet, Resnet, Alexnet**

The Benefits of Text and Image Features in Hedonic Regression

- Using only conventional features in linear regression gives R^2 **for predicting log-price lower than 10%.**
- Using ***W*** features in linear regression gives R^2 **of 30%.**
- Using ***I*** features in linear regression gives R^2 **of 25%.**
- Using ***W*** and ***I*** features in linear regression gives R^2 **of 36%.**
- Using ***W* and *I* features** plus **Random Forest** brings R^2 **of about 45-50% (up to 70% for very deep forests).**

Performance of the predictive model

Year/Month	RMSE	R^2
1607	0.46	0.52
1608	0.46	0.51
1609	0.48	0.50
1610	0.48	0.51
1611	0.48	0.52
1612	0.48	0.50
1701	0.49	0.50
1702	0.48	0.49
1703	0.48	0.49
1704	0.47	0.49
1705	0.46	0.48
1706	0.46	0.47
1707	0.46	0.47
1708	0.47	0.45
1709	0.47	0.47
1710	0.47	0.45
1711	0.48	0.44
1712	0.47	0.44

TABLE 2. The Out-of-Sample Performance of the Empirical Hedonic Pricing Function obtained Using Random Forest every month since July 2016.

Details of Feature Engineering

Query: red dress

Image

Customer behavior data

3,989 customer reviews | 259 answered questions

Description

- Material - Cotton & Spandex.
- Imported
- Classic and Iconic Audrey Hepburn 50s Vintage Solid Color Swing Dress, Put on and Show Your Elegance and Charm.
- Features: Boat Neckline; Sleeveless; Full Circle Swing; Quick Access Zipper for Easy On and Off
- It's Great Choice for Daily Casual, Wedding , Ball, Party, Banquet and Other Occasion.
- [Size Chart] PLEASE Make Sure Your Measurements and Compare to the Size Chart From the picture on the left side or in the Following Description.
- Hand Wash Carefully,Low Temperature for Washing,Can not High Temperature Ironing, Line Dry

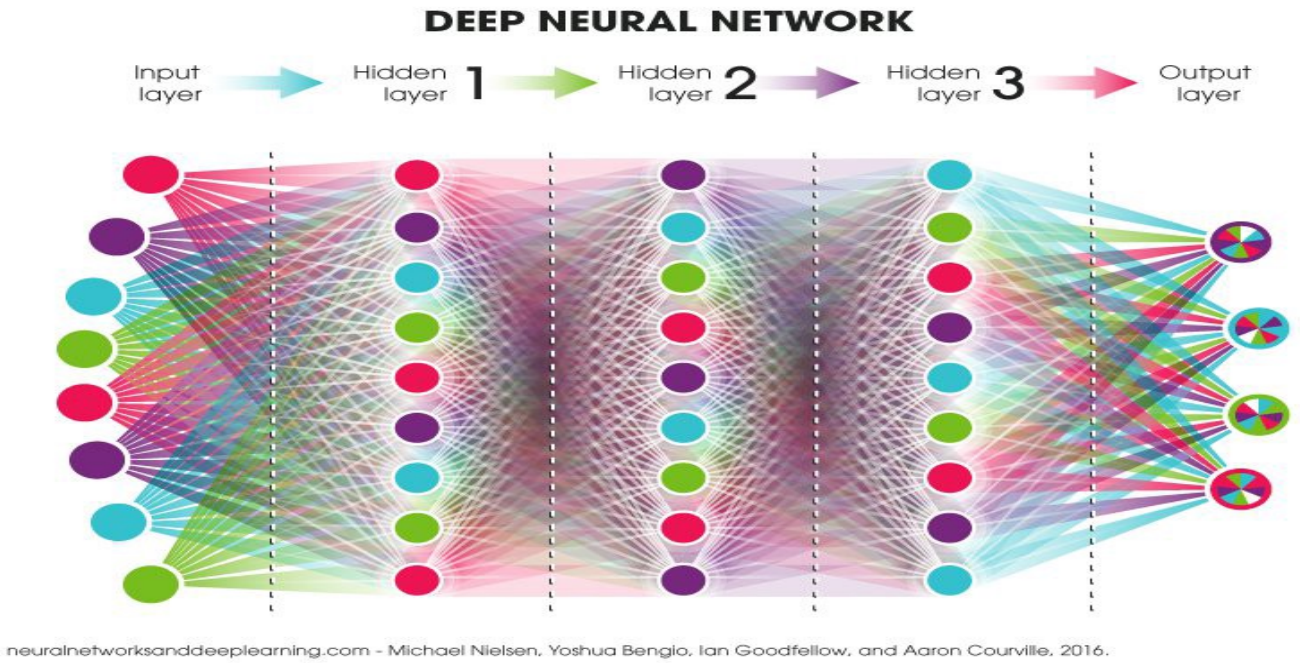
$X \in \mathbb{R}^d$

Title

Anni Coco

Anni Coco Women's Classy Audrey Hepburn 1950s Vintage Rockabilly Swing Dress

Features are created by (Deep) Neural Nets



Word2vec

- From sentence of words we predict the middle one using the left and the right words. Training is unrelated to prices.
- Words, V, are coordinate (sparse) vectors in $\mathbb{R}^p$, are mapped into X:

$$\mathrm{V} \longmapsto X := wV,$$

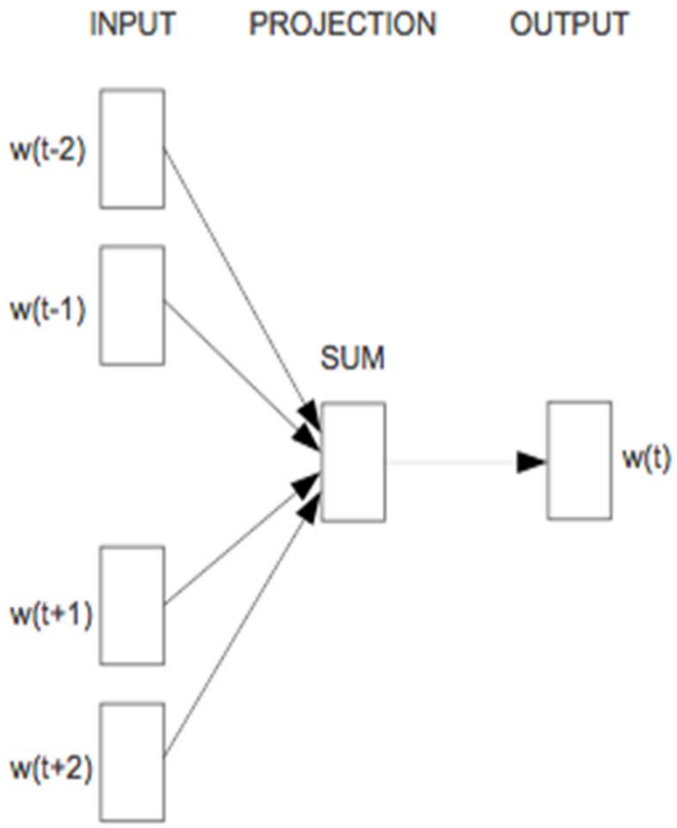
which is composed with the logistic map to classify the middle word:

$$X \longmapsto \pi = \exp(X)/(1^T \exp(X))$$

- Parameters w are trained by maximizing the logistic likelihood function applied to text data

$$\{Y(t), C(t)\}, \quad t = 1, \ldots, T;$$
$$Y(t) := V(t);$$

C(t) := (V(t−2), V(t−1), V(t+1), V(t+2))



CBOW

Word Embeddings: Examples

womens	0.387542	0.03051	-0.19703	0.179724	-0.222901	-0.606905	0.306091	-0.597467
mens	0.758868	0.372418	0.370116	0.706623	-0.124954	0.5088	0.106177	0.208935
clothing	0.149283	0.5161	-0.027684	0.218484	-0.851416	-0.409885	0.386088	0.170605
shoes	1.323812	-0.358704	-0.007683	-0.552144	0.011261	0.365239	0.228273	-0.565655
women	0.601477	-0.045845	-0.099481	0.010576	-0.096852	-0.605281	0.25606	-0.550759
girls	0.417473	-0.005265	-0.40939	-0.531189	-1.31938	-0.034746	-0.940507	-0.361215
men	0.778298	0.406613	0.426292	0.534272	-0.056103	0.51756	0.107846	0.245275
boys	0.896637	-0.016821	-0.001602	-0.181901	-1.313441	0.449006	-0.828408	0.52121
accessories	0.86825	-0.378385	-1.247708	1.541265	0.323952	0.282909	-0.491176	0.081314
socks	0.27636	0.354296	0.185734	0.301311	-0.643142	-0.021945	0.320751	0.240676
luggage	0.796763	1.749548	-2.30671	-0.559585	0.03054	0.921458	0.417333	0.313436
dress	0.282053	0.233192	0.043318	0.174759	-0.50114	-0.381047	0.297995	-0.026033
baby	0.346065	-0.550016	-1.136202	-0.043899	-2.004979	0.689747	-1.091575	0.009901
jewelry	-0.315784	0.347808	-0.308736	0.878713	-0.766016	1.124318	-0.079883	-2.039485
black	0.427496	0.030204	-0.019082	0.224096	-0.162242	-0.325359	0.170407	-0.172714
boots	1.009074	-0.30359	0.03197	-0.334004	-0.095679	0.111328	0.11769	-0.51878
shirts	0.444152	0.452918	0.393656	0.517929	-0.531462	0.099621	0.146202	0.204338
shirt	0.328998	0.421561	0.226565	0.455649	-0.700352	0.067224	0.106364	0.233862
underwear	0.230821	0.490978	0.226338	0.202376	-0.774363	0.004693	0.228712	0.310215

Embeddings have interesting properties

Word2Vec("handbag")+ Word2Vec("men")- Word2Vec("woman")
≈Word2Vec("briefcase")

Word2Vec("tie")+ Word2Vec("woman")- Word2Vec("men")
≈ Word2Vec("pashmina") , Word2Vec("scarf")

- Distance is the Cosine Distance = Euclidian distance after normalizing vector norms to unit

ResNet50 Image Embedding

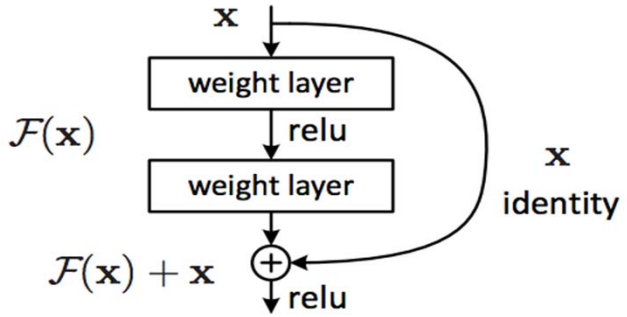


Figure 2. Residual learning: a building block.

a repeated composition of the partially linear score with the rectified linear unit.

('Predicted:', [(u'n03450230', u'gown', 0.4549656), (u'n03534580', u'hoopskirt', 0.3363025), (u'n03866082', u'overskirt', 0.20369802)])

Final Step: Random Forest to Predict Prices

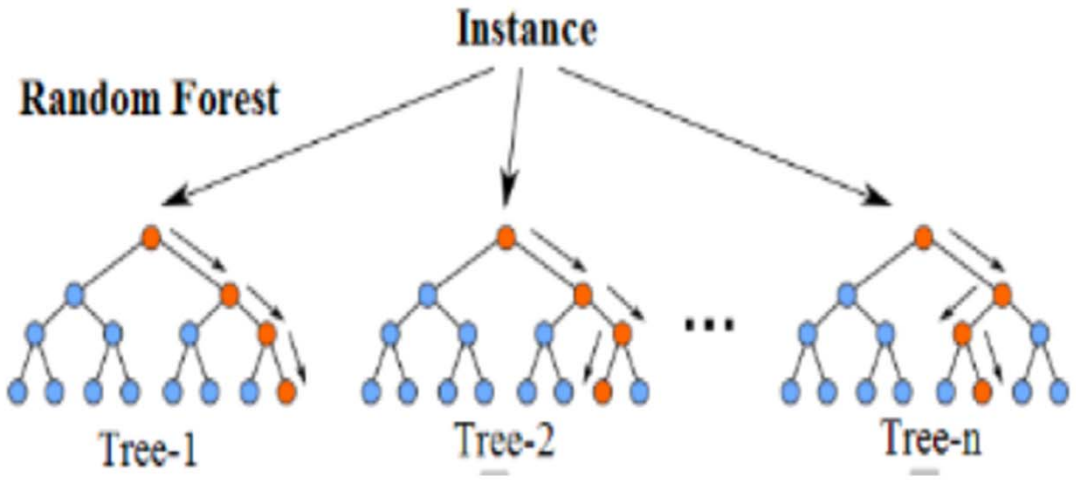


FIGURE 5. Example of a Random Forest Prediction from a Textbook. An instance $X = x$ gets mapped into predictions $g_1(x)$, $g_2(x)$, ..., $g_R(x)$, based on regression trees of depth 3 (the orange color describes how the prediction is generated at the particular instance $X = x$, by going from one decision node to another). Then the trees are averaged to produce the final prediction $\hat{f}_{random\ forest}(x)$.

Random Forest Continued

Let us explain in detail the idea of bootstrap aggregation. Each bootstrap sample is created by sampling from the our data on pairs (Y_i, X_i) randomly with replacement. So some observations get drawn multiple times and some don't get redrawn. Given a bootstrap sample, numbered by b, we build a tree-based prediction rule $\hat{f}_b(X)$. We repeat the procedure R times in total, and then average the prediction rules that result from each of the bootstrap samples:

$$\hat{f}_{random\ forest}(X) = \frac{1}{R}\sum_{b=1}^{R}\hat{f}_b(Z)$$

- Linear regression with text and image as features gives ***R^2*** **of about 36%.**
- Random forest brings the ***R^2*** **to 45-50% up to 70% if very deep**

Conclusions

- Inflation indices are important inputs into measuring aggregate productivity and cost of living, and monetary and economic policy
- We address the challenges in measuring inflation that arise due to
  - millions of products, with rapidly changing prices, and
  - extremely high turnover for some product groups.
- We do so by building quality-adjusted indices, which utilize
  - modern scalable computation that handles large amount of data
  - modern, open-source ML and AI tools to predict missing prices using product attributes.
- We would like to share our science and engineering expertise with U.S. statistical agencies.