

# HEDONIC PRICES AND QUALITY ADJUSTED PRICE INDICES POWERED BY AI

P. BAJARI\*, Z. CEN\*, V. CHERNOZHUKOV\*, M. MANUKONDA\*, S. VIJAYKUMAR\*, J. WANG\* AND R. HUERTA, J. LI, L. LENG, G. MONOKROUSSOS, S. WANG

**ABSTRACT.** We develop empirical models that efficiently process large amounts of unstructured product data (text, images, prices, quantities) to produce accurate hedonic price estimates and derived indices. To achieve this, we generate abstract product attributes (or “features”) from descriptions and images using deep neural networks. These attributes are then used to estimate the hedonic price function. To demonstrate the effectiveness of this approach, we apply the models to Amazon’s data for first-party apparel sales, and estimate hedonic prices. The resulting models have a very high out-of-sample predictive accuracy, with  $R^2$  ranging from 80% to 90%. Finally, we construct the AI-based hedonic Fisher price index, chained at the year-over-year frequency, and contrast it with the CPI and other electronic indices.

**Key Words:** Hedonic Prices, Price Index, Transformers, Deep Learning, Artificial Intelligence

---

*Date:* March, 2019. This version: December 30, 2025.

\*P. Bajari, Z. Cen, V. Chernozhukov, M. Manukonda, S. Vijaykumar, and J. Wang (in alphabetical order) are the principal authors who contributed to the current version of this paper. R. Huerta and G. Monokrousos were also principal contributors to an earlier version of the paper. For helpful comments, we would like to thank the participants at the 2018 Federal Economic Statistics Advisory Committee meeting, the 2019 Allied Social Science meetings, the 2019 Federal Reserve Board conference on "Nontraditional Data, Machine Learning, and Natural Language Processing in Macroeconomics", the 2019 Brookings Conference "Can big data improve economic measurement?" and the 2021 CEMMAP conference "Measuring prices and welfare," and seminars at Berkeley, European Bank of Reconstruction and Development, MIT, UCL, and York. We are grateful to Andrew Chesher, Greg Duncan, Kevin Fox, John Haltinwager, James Heckman, Daniel Miller, Mathew Shapiro, Bernhard Schölkopf, James Stock, and Weining Wang for helpful comments during the various stages of this project. The paper was first circulated in 2021 as a CEMMAP Working Paper (CWP04/21).

## 1. INTRODUCTION

Economists and policy-makers rely on price indices, such as the Consumer Price Index (CPI), to measure inflation, consumer welfare, and the cost of living. As a result, the methods used to construct price indices warrant considerable attention. Two standard methods are the Laspeyres and Paasche indices, which measure changes in the cost of a standardized basket of products between two periods, where the basket is chosen to represent aggregate demand in either the initial period (Laspeyres) or the final period (Paasche). The two are often combined into what is known as the Fisher Price Index (FPI).<sup>1</sup> The FPI has been shown to accurately approximate the relative cost of living between two periods when the difference in prices is small.<sup>2</sup>

A common problem with the aforementioned indices is product entry and exit, where the previous period's prices are not available for new products, and vice-versa. Consequently, economists often produce so-called 'matched indices': indices restricted to the set of products that are bought and sold in both periods. This introduces selection bias because products exiting the marketplace may not resemble those that remain, especially when products turn over quickly. To mitigate this bias, economists often use high-frequency chaining combined with compounding: for example, computing price indices at a monthly frequency and then compounding monthly inflation rates to get yearly rates. This approach does mitigate the turnover problem if the rate of turnover is not high from month to month. Still, it can suffer from chain-drift bias—bias in estimated price levels due to geometric compounding of measurement errors.

Hedonic price models were introduced by Court (1939) and Griliches (1961); these models postulate that the prices of differentiated products are determined by the market value of each product's underlying characteristics. Court and Griliches suggested measuring inflation or deflation by modeling how hedonic prices change while holding product characteristics fixed. Importantly, one can compute hedonic prices for a

---

<sup>1</sup>These quantities bound other measures of inflation based on the expenditure function under certain assumptions; see Diewert (1998) and also Diewert and Fox (2022) for a state-of-art review.

<sup>2</sup>The FPI captures the cost of living for a representative consumer with a quadratic utility or expenditure function, and provides a second-order approximation for any smooth utility or expenditure function under small price changes (see Diewert, 1976). Price indices with this property are called superlative indices, with another prominent example being the Tornqvist index (see, e.g. Office of National Statistics, 2020). For both matched and hedonic indices, the Tornqvist indices we obtain are numerically very similar to the corresponding (matched or hedonic) Fisher indices.

basket of goods at any time point—despite entry and exit—because such prices are determined only by product characteristics. These prices can then be used in the Fisher price index and other price index formulas, giving rise to hedonic price indices. Hedonic price indices are regularly used both in academic research and by statistical agencies (e.g., Wasshausen and Moulton, 2006; Office of National Statistics, 2020).

The resulting hedonic price indices are sometimes called *quality-adjusted* price indices because we implicitly fix the set of characteristics (“qualities”) of a basket in a reference period and compute the ratio of costs of the basket in the comparison period and in the reference period. The ability to compute hedonic prices at any time point for any product allows us to address the entry/exit problem and also reduce chain-drift biases by making “long” or “low-frequency” comparisons (year-to-year, for example).<sup>3</sup> Success of this approach depends both on the ability of hedonic price models to approximate real-world prices and on our ability to accurately estimate the hedonic price function.

Provided that hedonic models provide a good approximation of real world supply and demand, we can hope to model equilibrium prices by regressing observed prices on product characteristics  $X_j$ . In traditional empirical hedonic models, the construction of  $X_j$  is performed using human expertise, and statistical agencies conduct extensive field surveys and interviews to this end.

In this paper, we develop an alternative approach to the construction of price indices. Our method uses electronic data and artificial intelligence to automatically generate product characteristics,  $X_j$ , which are used to accurately predict prices. The resulting approach is highly scalable and cost-effective compared to the traditional approach. The potential to use electronic records to improve the construction of price indices, as we do here, has received substantial attention from both academic researchers and statistical agencies (see e.g. Groves and Harris-Kojetin, 2017; Ehrlich et al., 2019; Jarmin, 2019; Lebow, 2023). Our approach thus offers a valuable complement to traditional methods for measuring price-level changes.

---

<sup>3</sup>Year-to-year chaining is recommended in the CPI manual to deal with chain drift (ILO et al., 2004). This approach was used, for example, by Handbury et al. (2013) to construct non-hedonic indices using electronic data from Japan. See Diewert and Fox (2022) for pertinent discussion and other methods used to address chain drift. One prominent class of alternative methods are multilateral indices, such as the GEKS index, which take the average of matched indices with varying base periods (Gini, 1931; Eltetö and Köves, 1964)



FIGURE 1. An example of product characteristics for a product sold in the Amazon store.

**1.1. Deep Learning and Structured Sparsity.** Let us illustrate the problem we are solving. Figure 1 shows the product characteristics visible to customers, including the product title, description, and images. Our goal is to represent this information using a numerical vector  $X_j$  of moderately high dimension (in our case it has roughly 2000 entries), which can be used to predict prices accurately. Moreover, the representation needs to be algorithmic and scalable.

The success of this approach depends on the existence of parsimonious structures behind images and text. Traditionally, analysts relied on human experts to summarize the salient features of products using a low dimensional vector,  $X_j$ . Experts produced low-dimensional representations for certain groups of products, which proved to be successful in building hedonic models (e.g., Pakes, 2003 reports very high accuracy for predicting computer prices). However, this approach does not scale well to many types of products and is prone to judgment biases. These issues raise important questions: when human experts do succeed, what is the underlying reason?; can we replicate this success with artificial intelligence, and can these methods deliver scalable inference?

Arguably, humans can easily summarize the red dress depicted in Figure 1 and its accompanying text, even though the original representation of this information lives in an extremely high-dimensional space. Indeed, the image consists of nearly one million

pixels (three layers of 640 x 480 pixels encoding the blue, red, and green color channels), while words belong to a dictionary whose dimension is in the tens of thousands and sentences live in much higher-dimensional space (Milton and Treffers-Daller, 2013). However, we believe that information in images and sentences can be effectively represented in a much lower-dimensional space, a phenomenon we call “structured sparsity.” Human intelligence can exploit this structured sparsity to process information effectively—perhaps using the geometry of shapes in images, the relative simplicity of color schemes and shade patterns, the similarity of many words in the dictionary, and the context-specific meaning of words. The field of artificial intelligence (AI) developed neural networks to mimic human intelligence in many information-processing tasks. These models do create parsimonious structures from high-dimensional inputs and often surpass human ability in such tasks. Going forward, we will employ state-of-the-art solutions from AI to construct price indices.

In this paper we use deep learning models to extract relevant product attributes, or “features,” from text and images. We then use these features to estimate a non-parametric model for prices. Finally, we use the model to perform quality adjustment, mirroring the traditional construction of hedonic price indices.

Our features, which are denoted by  $X_j$  (of moderate dimension) and  $V_j$  (of low dimension), are analogous to traditional hedonic features tabulated by human experts. Like the vector of covariates in a traditional hedonic regression,  $V_j$  is a concise summary of product attributes that linearly explains variation in prices. Both  $X_j$  and  $V_j$  may be viewed as representations of the product (through its image and description) in a latent characteristic space. However, there are also important differences. For one, the entries of  $X_j$  and  $V_j$  lack a straightforward explanation in terms of product characteristics. Moreover, the mapping from images and descriptions to  $V_j$  is not canonical, and may change when the model is updated (e.g., to incorporate new data). Nevertheless, for the remainder of the paper we will refer to the resulting models as “hedonic,” and will treat the vectors  $X_j$  as we would hedonic characteristics. We also refer to the vector  $V_j$  as a *value embedding*, to reflect the idea that it concisely summarizes a product’s valuable characteristics.

To generate  $X_j$  and  $V_j$ , we convert text information about the product to numeric features (embeddings) using the BERT model of Devlin et al. (2018), a transformer-based

large language model; our version of the model has been fine-tuned on Amazon’s product descriptions and prices. We similarly use a pre-trained ResNet-50 model, a convolutional neural network used to understand images, to produce embeddings for product images (He et al., 2016). The BERT and ResNet-50 models are trained to comprehend text and images in tasks unrelated to predicting prices (e.g., image classification or prediction of a missing word in a sentence). We then take the internal numeric representation generated by each model as an information-rich, parsimonious *embedding* of the input.<sup>4</sup> With these embeddings, we then estimate the hedonic price function using a deep neural network. In particular, we use *multi-task* networks, which predict a complete time-series of hedonic prices  $(\hat{H}_{it})_{t \in T}$  for a given product  $i$ , where  $t \in T$  covers all time periods in the study.<sup>5</sup>

**1.2. Deriving Indices using Amazon Data.** We apply these models to Amazon’s data for apparel sales to estimate hedonic prices.<sup>6</sup> The resulting models have high predictive accuracy, with  $R^2$  in the holdout sample ranging from 80 to 90%. Therefore, our approach can attribute up to 90% of the variation in prices to variation in embeddings that encode product attributes. We find this performance remarkable for two reasons:

- (1) The price imputation is completely automatic and scalable, without relying on any human-based feature extraction.
- (2) The performance suggests that characteristics-based price models provide a good first-order approximation of real-world prices.

We then proceed to construct Fisher hedonic price indices (FHPI) over 2013-2017, constructing monthly-chained, yearly-chained, and the geometrically combined FHPI

---

<sup>4</sup>This strategy follows a paradigm called “transfer learning,” which has proved a very successful way to use neural networks in new domains (Ng, 2016).

<sup>5</sup>This requires training the network using data for all time periods, and re-training the full model when additional data become available. See Section 2 for a comparison to other models which do not share this constraint, and Appendix C for further discussion.

<sup>6</sup>Our data cover “first party” sales: those in which Amazon first buys the product from a trusted seller and then sells it directly to the consumer.

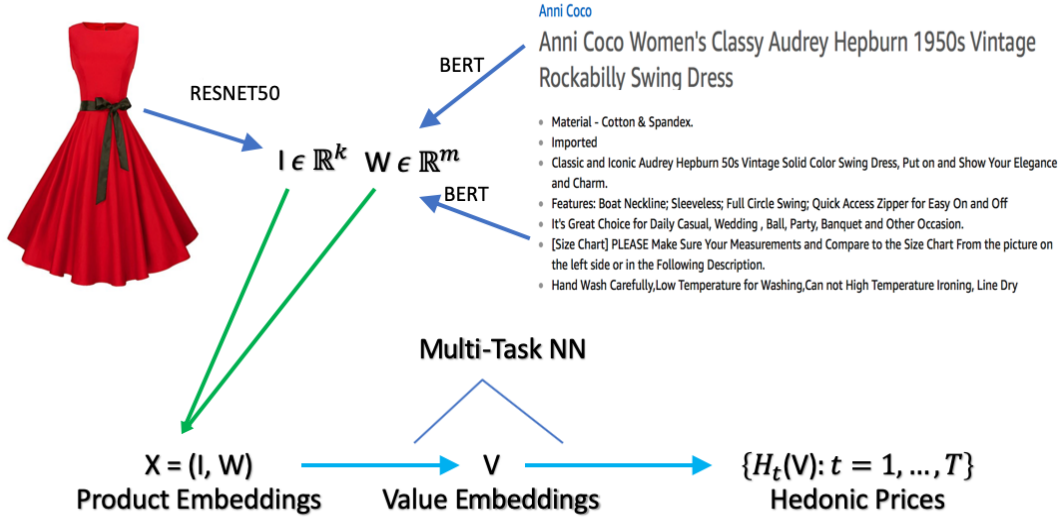


FIGURE 2. Our method for generating hedonic price: The input consists of images and unstructured text data. The first step of the process creates the moderately high-dimensional numerical embeddings  $I$  and  $W$  for images and text data via state-of-the-art neural networks: ResNet-50 and BERT. The second step takes input  $X = (I, W)$  and creates predictions for hedonic prices  $H_t(X)$  using a multi-task neural network. Our multi-task model creates an intermediate lower dimensional embedding  $V = V(X)$ , called a *value embedding*, and then predicts the final prices in all periods  $\{H_t(V), t = 1, \dots, T\}$  using linear functional forms; this makes it easy to perform inference on the last step using hold-out data.

(GFHPI).<sup>7,8</sup> We focus discussions on yearly chained FHPI and refer to it as the Fisher hedonic price index unless stated otherwise. We compare this index with

- the matched (repeated sale) Fisher index,
- the posted price Jevons index, which is the geometric mean of relative prices, chained at a daily frequency,

<sup>7</sup>The GFHPI gives equal geometric weight to yearly and monthly chained indices. This brings in some seasonality, allowing the GFHPI to reflect within-year price changes. This index is not recommended for use over long horizons, however, due to chain drift.

<sup>8</sup>This approach was inspired by the GEKS index, which computes the geometric average over all chains; see e.g., Diewert and Fox (2022) for a precise definition. The GEKS index reduces chain drift caused by measurement error in the product shares used to construct the monthly index. We are currently exploring the use of hedonic GEKS indices as follow-up work.

- the BLS Urban CPI index for apparel (CPI), constructed by the Bureau of Labor Statistics,
- the Adobe Digital Price Index (DPI) for apparel, constructed by Goolsbee and Klenow (2018) using the Adobe Analytics data.

All indices suggest the apparel price level declines over this period. The annual rate of inflation estimated by FHPI is  $-0.98\%$ , by the Jevons index  $-3.01\%$ , and by the matched Fisher index  $-3.12\%$ . In comparison, the annual rate of inflation in apparel prices estimated by the Adobe DPI is  $-2.02\%$ , and by the CPI is  $-0.31\%$ .

The yearly chained FHPI, our preferred index, has the smallest discrepancy with the CPI. Part of the remaining difference with the CPI could be attributed to the limited ability of the CPI to address quality change and substitution (as pointed out by Moulton, 2018, and others), to specifics of product categorization in the apparel segment, or to differences in the composition of baskets purchased by Amazon’s consumers. The difference may also reflect Amazon-specific cost improvements and supply-chain logistics.

In comparison to the FHPI, neither the matched Fisher index nor Adobe’s DPI adjusts for change in the quality of goods sold. These indices potentially also suffer from chain drift—the systematic accumulation of errors due to frequent compounding. Compared to the FHPI, the Jevons index does not incorporate quantity weighting, and it is also subject to chain drift.<sup>9</sup> Both of these issues produce a large discrepancy between the Jevons index and the FHPI.

**1.3. Contributions to the literature.** We view our paper as an original contribution towards the modernization of hedonic price models and their application to large-scale data. In this way, we also contribute to the empirical literature dedicated to hedonic price indices and their uses in measuring inflation.

**1.3.1. AI-based hedonic models.** At a high level, our work stems from the observation that hedonic modeling is fundamentally a prediction problem: “how much demand will there be for characteristics  $X_j$  at time  $t$ , based on the demand for similar products?”

---

<sup>9</sup>Chain drift due to measurement error is thought to be less severe for the Jevons index because the index is computed using very many prices, so that within-period errors are small. The Jevons index is a special case of the Tornqvist index with revenue shares for products set to be equal. This reduces another important source of chain drift: measurement error in quantity weights (Ivancic et al., 2011).



As such, it presents an excellent opportunity for AI to help us measure policy-relevant economic variables, such as the rate of inflation. This is especially significant given the massive amount of unstructured data available in the form of electronic records, which are not easily incorporated into conventional models.

We are not aware of any prior work that develops large-scale price prediction models from unstructured product text and image descriptions. In addition, our data are unique in that they cover the universe first party transactions in Amazon’s U.S. stores from 2013 to 2019. From these models and data, we generate interesting findings: we demonstrate the power of hedonic models to characterize prices and document the decline in the quality-adjusted Fisher price index for apparel. To the best of our knowledge, there are very few related studies in economics: An independent and contemporaneous work by Zeng (2020) develops a related approach to hedonic prices using scanner data, but uses random forest methods and low-dimensional data instead of neural network embeddings of unstructured text and images. An independent and contemporaneous work by Han et al. (2021) explores the use of image embeddings to characterize typesetting fonts as products, and analyzes the effect of mergers on product differentiation via font producers’ design choices. In the coming years, we expect to see a much wider use of AI-based embeddings for text and images to power empirical research in economics.

*1.3.2. Inflation, e-commerce, and electronic data.* This paper contributes to a rapidly growing literature on the use of electronic data to measure inflation and other aggregate quantities. This is motivated by the significant growth of e-commerce, along with evidence that online prices may fluctuate differently than their offline counterparts.

The MIT Billion Prices Project (BPP) constructs Jevons indices using web-scraped retail price data and directly-provided retailer data. The advantages of such data include real-time availability at daily frequencies, low collection costs, large product counts, and uncensored price spells. Consequently, BPP indices constructed using modern real-time data can serve as useful benchmarks for official government statistics. For instance, Cavallo and Rigobon (2011, 2016) study several countries and establish that inflation measures constructed using such online data can differ substantially from official government statistics on consumer prices (the most extreme example being the case of Argentina, see Cavallo, 2013).

A recognized limitation of this approach is that it does not incorporate quantity information. To this end, Goolsbee and Klenow (2018) constructed matched Törnqvist indices, including the Digital Price Index (DPI), using Adobe Analytics data from e-commerce clients of Adobe, which notably include quantities as well as prices. This approach overcomes the quantity limitation of the BPP index and thus can account for substitution resulting from consumers' cost minimization. They find that the DPI for U.S. online transactions measures inflation to be substantially lower than the official CPI's measurement for the categories they study in the period 2014-2019. This is analogous to our findings for apparel, albeit the discrepancy between our preferred index and the CPI is much smaller. The difference with our approach could possibly occur due to chain drift in the DPI due to monthly chaining.

There is also a body of research on using scanner data to survey demand and to construct measurements of inflation; of particular note are Handbury et al. (2013), Ivancic et al. (2011), Leicester (2014), Diewert and Fox (2022), and Office of National Statistics (2020). We contribute to this development by providing AI-based hedonic prices that economists can employ alongside the approaches in these studies. This provides a valuable complement, as scanner data are generally tilted towards food and beverages sold in grocery stores (Kaplan and Schulhofer-Wohl, 2017).

In related work using online price data, Cavallo (2017) finds that online prices are identical to their offline counterparts about 70% of the time (on average across countries). Gorodnichenko and Talavera (2017), Cavallo (2018) and Gorodnichenko et al. (2018) find that online prices change more frequently than offline prices, thereby responding to competition more promptly. These papers also find that online prices exhibit stronger pass-through in response to nominal exchange-rate movements than prices found in official CPI data; such results have important implications for the price stickiness literature and the law of one price. These results also highlight the potential benefit of real-time, electronic data and derived price indices like ours.

Since the circulation of an earlier version of this manuscript as a CEMMAP Working Paper (Bajari et al., 2019), there has been substantial and very promising research that adapts our methods to other, publicly available data sources, with the ultimate aim of complimenting or improving national statistics. Of particular note are the working papers of Cafarella et al. (2023) and Ehrlich et al. (2023), both of which discuss the use of ML-based hedonic indices with point-of-sale scanner data.

**1.4. Organization of the Paper.** We organize the rest of the paper as follows. Section 2 defines the hedonic price models and price indices. Section 3 discusses modeling and estimating the hedonic price functions via Neural Networks (NNs). Section 4 provides a non-technical description of the process of obtaining product embeddings via AI tools. Section 5 examines the empirical performance of the AI-based hedonic price functions and constructs the AI-based hedonic price indices. Appendix A provides a technical description of BERT, a large language model used to generate embeddings for the product description. Appendix B gives a technical description of ResNet-50, a tool used to generate embeddings of the product’s image. Appendix C gives a technical description of various models we’ve considered and tested, of which the best-performing methods are described in the main text.

**1.5. Notation.** We use capital letters as  $W$  as random vectors and  $w$  the values they take; we use  $W$  to denote matrices. Functions are denoted by arrows  $w \mapsto f(w)$  or simply  $f$ . Greek symbols denote parameter values, with the exception of  $\epsilon$  which denotes the regression error (see below).

## 2. HEDONIC PRICES AND HEDONIC PRICE INDICES

**2.1. The Hedonic Price Model.** We denote the product by its index  $i$  and the time period (month) by  $t$ . An empirical hedonic model is a predictive model for the price given product features:

$$P_{it} = H_{it} + \epsilon_{it} = h_t(X_i) + \epsilon_{it}, \quad E[\epsilon_{it} | X_i] = 0, \quad (1)$$

where  $P_{it}$  is the price of product  $i$  at time  $t$ ,  $X_i$  are the product features. The price function  $x \mapsto h_t(x)$  can change from period to period, reflecting the fact that product features may be valued differently in different periods. For our purposes, the advantage of these models is that they allow us to compare new goods to old rather directly; we simply compare the value consumers attach to the characteristics of the old good to those of the new.

We will use the data from time period  $t$  to estimate the function  $h_t$  using modern non-linear regression methods, such as deep neural networks. We contrast this approach with classical linear regression methods as well as other modern regression methods, such as a random forest. The key component of our approach is the generation of product features  $X_i$  using neural network embeddings of text and image information about

the product. Thus,  $X_i$  consists of text embedding features  $W_i$ , constructed by converting the title and product description into numeric vectors, and image embedding features  $I_i$ , constructed by converting the product image into numeric vectors:

$$X_i = (W_i', I_i')'. \quad (2)$$

These embedding features are generated respectively by applying the BERT and ResNet-50 mappings detailed in the next section.

There is a substantial body of economic research on hedonic price models. On the theory side, economists have developed a theory of supply and demand in terms of product characteristics (Lancaster, 1966; McFadden, 1974; Rosen, 1974; Gorman, 1980); they have also established existence of the hedonic price function and characterized its behavior under various assumptions (Berry et al., 1995, 2004; Ekeland et al., 2004; Benkard and Bajari, 2005; Chiappori et al., 2010; Chernozhukov et al., 2020); they have also developed the use of hedonic prices for measuring changes in consumer surplus and welfare (Bajari and Benkard, 2005). On the empirical side, economists have estimated a variety of hedonic price models and linked them to the consumer's utility and their marginal willingness to pay for certain characteristics, see Nesheim (2006) and Greenstone (2017); they have also used hedonic theory to measure the value of non-tradable goods—for example, to measure the effects of air quality or hazardous waste cleanup on housing prices, (e.g., Chay and Greenstone, 2005; Stock, 1991), or to study equalizing differences in the labor market, (e.g., Brown, 1980). Our main use of hedonic prices is to estimate the rates of deflation (or inflation) for apparel products purchased at Amazon.com. This follows prior work on using hedonic models to construct official price statistics (see e.g., Griliches, 1961; Pakes, 2003; Wasshausen and Moulton, 2006; Office of National Statistics, 2020), but with a major deviation: we use product features generated using deep learning instead of using human experts to tabulate product characteristics, and we also estimate the hedonic price function using deep learning rather than using classical regression methods.

The literature typically specifies three building blocks of theoretical hedonic models: consumer utility functions defined over products' characteristics (rather than products themselves); producer cost functions likewise defined over characteristics of the product; and an equilibrium assumption (or existence of the equilibrium is shown as a part

of the analysis); see e.g. Pakes (2003) and Rosen (1974).<sup>10</sup> This determines prices and quantities given demand and input costs, and establishes the existence of the hedonic price function

$$(x, u) \mapsto H^*(x, u)$$

as a function of product attributes  $(x, u)$ . Here, both  $x$  and  $u$  are observed by the consumer, while only  $x$  (and not  $u$ ) is observed by the modeler.

Price functions give us information about customer preferences. For example, when the customer's utility is given by:

$$V(x, u, p, m) = V_0(x, u) + m - p$$

where  $p$  is the price of the product to be paid by the customer and  $m$  is their income, the first order conditions for the utility maximization problem  $\max_{(x, u)} V_0(x, u) + m - H(x, u)$  is given by:

$$\partial_{x_k} V_0(X, U) = \partial_{x_k} H^*(X, U),$$

where  $\partial_{x_k} = \partial/\partial x_k$ , where  $x_k$  refers to the  $k$ -th component of the vector  $x$  (for continuously varying attributes). Therefore, under suitable assumptions, a standard argument for identification of the average derivative of a structural function gives

$$E[\partial_{x_k} H^*(X_j, U_j) | X_j] = \partial_{x_k} h_t(X_j).^{11}$$

In other words, the average marginal willingness to pay for a given characteristic is equal to the average derivative of the hedonic price map, and is identified by the derivative of the hedonic regression function. We remark here that the expectation is taken over the random variables' distribution at equilibrium; we can not say anything about features of the hedonic price map away from equilibrium.

Given a parametric consumer utility function, preference parameters can be recovered from first-order conditions provided that either (i)  $(x, u) \mapsto H^*(x, u)$  is additively separable in  $(x, u)$ , so that  $\partial_{x_k} H^*(x, u) = \partial_{x_k} h_t(x)$ , does not depend on  $u$ , or (ii) we can identify  $H^*$  and the unobservable  $U$  by other means (for example, by making quantile or multivariate-quantile type assumptions on the way  $U$  appears in  $H^*$ ). For example, if the utility is Cobb-Douglas over observed characteristics,  $V_0(x, u, p) = \sum_{k=1}^K \alpha_k \log(x_{jk}^*) +$

<sup>10</sup>Typically the utility and cost functions will also depend upon characteristics of the consumer and producer, respectively.

<sup>11</sup>This holds, for example, under independence of observable and unobservable characteristics at equilibrium, see Appendix D.1.

$\beta g(u) - p$ , then under additive separability,  $H^*(X, U) = H_0^*(X) + U$ , we recover  $\alpha_k = \partial_{x_k} h_t(X_j) X_{jk}$  for a consumer who has purchased product  $j$ . The distributions of consumer taste parameters  $\alpha$  can be recovered under this type of modeling approach; see Bajari and Benkard (2005) for further relevant discussion. Here we have a different goal: we use hedonic prices to construct indices and thereby measure price level changes, following accepted practice in applied economics and in the work of statistical agencies (e.g., Wasshausen and Moulton, 2006; Pakes, 2003; Office of National Statistics, 2020).

**2.2. Price Indices: Hedonic vs Matched.** We focus on hedonic price indices and compare them to matched (repeated sale) price indices. The matched price index tracks changes in the price of a basket of products sold in both the base period and in subsequent time periods. While the matched price method is subject to selection bias (due to product entry and exit), it ensures the index tracks goods from a common pool of products. A major shortcoming of this method is that the common pool of products across time can be small and non-representative; this will be apparent in our data.

The hedonic price index we develop replaces transaction prices with predicted values using a rich set of product characteristics (obtained using a combination of AI and machine learning methods). In principle, hedonic models measure changes in the value consumers assign to product attributes.<sup>12</sup> They are especially helpful for predicting the prices of new goods and correcting for selection bias due to product entry and exit, where transaction prices are not observed. This is especially relevant in our case, where we observe a very high turnover of products.

We consider three broad classes of price index:

- The Laspeyres (L) type, which uses base period quantities for weighting the prices;
- The Paasche (P) type, which uses current period quantities for weighting the prices;
- The Fisher (F) type, which is the geometric mean of the L- and P-type indices.

---

<sup>12</sup>We refer the reader to Aizcorbe (2014) for further discussion of hedonic price indices and their application.

One defines the L- and P-type matched indices as measures of the total rate of price change of a basket of matching products from the current period  $t$  with a previous period  $t - \ell$ :

$$R_{t,\ell}^{P,M} = \frac{\sum_{i \in \mathcal{C}_t \cap \mathcal{C}_{t-\ell}} P_{it} Q_{it}}{\sum_{i \in \mathcal{C}_t \cap \mathcal{C}_{t-\ell}} P_{i(t-\ell)} Q_{it}}; \quad R_{t,\ell}^{L,M} = \frac{\sum_{i \in \mathcal{C}_t \cap \mathcal{C}_{t-\ell}} P_{it} Q_{i(t-\ell)}}{\sum_{i \in \mathcal{C}_t \cap \mathcal{C}_{t-\ell}} P_{i(t-\ell)} Q_{i(t-\ell)}};$$

and the F-type index takes the form

$$R_{t,\ell}^{F,M} = \sqrt{R_{t,\ell}^{P,M} \cdot R_{t,\ell}^{L,M}},$$

where  $Q_{it}$  is the quantity of the product  $i$  sold in month  $t$ ,  $P_{it}$  is the average sales price for product  $i$  at time  $t$ ,  $\mathcal{C}_t$  is the set of all products with transactions at time  $t$ ,  $\mathcal{C}_t \cap \mathcal{C}_{t-\ell}$  is the match set, the set of all products with transactions both at time  $t$  and at time  $t - \ell$ .

In a standard representative consumer model, matched indices should obey the order restriction  $R_{t,\ell}^{L,M} \geq R_{t,\ell}^{P,M}$ . Even when aggregate quantities are not described by a representative consumer model, the relation often holds empirically. Similar intuition applies to hedonic indices. The two indices may be combined using Fisher's ideal index, which is a "superlative" index: it measures the exact cost of living when the utility function is quadratic and provides a second-order approximation to the cost of living at the given prices when the utility function is smooth (Diewert, 1976).

We define the L-, P-, and F-type hedonic indices similarly, as measures of the total rate of hedonic price change of a basket of product attributes from the current period  $t$  with the previous period  $t - \ell$ :

$$R_{t,\ell}^{P,H} = \frac{\sum_{i \in \mathcal{C}_t} H_{it} Q_{it}}{\sum_{i \in \mathcal{C}_t} H_{i(t-\ell)} Q_{it}}; \quad R_{t,\ell}^{L,H} = \frac{\sum_{i \in \mathcal{C}_{t-\ell}} H_{it} Q_{i(t-\ell)}}{\sum_{i \in \mathcal{C}_{t-\ell}} H_{i(t-\ell)} Q_{i(t-\ell)}}; \quad R_{t,\ell}^{F,H} = \sqrt{R_{t,\ell}^{P,H} \cdot R_{t,\ell}^{L,H}}. \quad (3)$$

We note that the P-type index is defined over sets of products  $\mathcal{C}_t$  and the L-type index is defined over the set of products  $\mathcal{C}_{t-\ell}$ , which are supersets of the matching set  $\mathcal{C}_t \cap \mathcal{C}_{t-\ell}$ . When hedonic prices follow a linear model  $H_{it} = \theta'_t V_i$ , where  $V_i$  are product attributes, we have e.g.  $\sum_{i \in \mathcal{C}_t} H_{is} Q_{it} = \theta'_s (\sum_{i \in \mathcal{C}_t} V_i Q_{it})$ . Thus, these indices may be viewed as measuring the relative price of a quantity-weighted basket of product attributes.

The above indices exclusively use hedonic prices  $H_{it}$ , even when true prices  $P_{it}$  are available; this is known as "full imputation." Alternative approaches, particularly "single imputation" and "double imputation," make use of observed prices  $P_{it}$  for products in the matching set  $\mathcal{C}_t \cap \mathcal{C}_{t-\ell}$  (Aizcorbe, 2014, Sec. 3.2). In subsequent work on ML-based hedonic indices, Ehrlich et al. (2023) finds that full imputation tends to reduce chain

drift, as estimated hedonic prices are less volatile than true prices. For our preferred index, both strategies produce nearly indistinguishable results.

For an arbitrary index  $R_{t,\ell}^{\bullet,\bullet}$ , for positive integers  $t, \ell$ , we measure the price changes up to time  $t \geq t_0$  as follows. Let  $m = \lfloor (t - t_0)/\ell \rfloor$  denote the largest integer no greater than  $(t - t_0)/\ell$ , and write  $t - t_0 = m\ell + r$ . We construct a chained index by taking the product

$$R_{t,\ell}^{\bullet,\bullet,C} = R_{t,r}^{\bullet,\bullet} \prod_{\bar{m}=1}^m R_{t_0+\bar{m}\ell,\ell}^{\bullet,\bullet} \quad \text{where} \quad R_{t,0}^{\bullet,\bullet,C} = 1.^{13}$$

For the hedonic index we shall use month-over-month chaining with  $\ell = 1$  and year-over-year chaining with  $\ell = 12$ , getting two types of indices:

$$R_{t,1}^{F,H,C} \text{ and } R_{t,12}^{F,H,C},$$

where the first index captures month-over-month changes in prices, especially for non-seasonal apparel, and the second index measures year-over-year changes in prices for products sold in the same month as  $t_0$ , including seasonal apparel. The second index is also less susceptible to the chain-drift problem that arises from an accumulation of errors due to repeated compounding. For this reason, we choose the yearly-chained index to be our preferred index, and refer to it as the Fisher Hedonic Price Index (FHPI). To capture seasonality while mitigating chain drift, we also consider the geometric mean of the monthly-chained and yearly-chained index (GFHPI):  $R_t^{GF,H} = \sqrt{R_{t,1}^{F,H,C} R_{t,12}^{F,H,C}}$ .

### 3. PRICE PREDICTION AND INFERENCE WITH DEEP NEURAL NETWORKS

**3.1. The Multi-Price Prediction Network.** Our model takes in high-dimensional text and image features as inputs, converts them into a lower-dimensional vector of value embeddings using state-of-the-art deep learning methods, and then outputs simultaneous predictions of price in all periods.

Our regression model is a composition of several non-linear vector-valued functions, called *layers*. It takes the form

$$Z_i = \begin{bmatrix} \text{Text}_i \\ \text{Image}_i \end{bmatrix} \xrightarrow{e} X_i \xrightarrow{g_1} E_i^{(1)} \dots \xrightarrow{g_m} E_i^{(m)} =: V_i \xrightarrow{\theta'} \{H_{it}\}_{t=1}^T := \{\theta'_t V_i\}_{t=1}^T. \quad (4)$$

<sup>13</sup>For long-chained indices with  $\ell > 1$ , this approach compares a given period's prices to the most recent "pivot"  $\{t_0, t_0 + \ell, t_0 + 2\ell, \dots\}$ . This is a standard practice, see e.g. Statistics Bureau of Japan (2022, Appendix 4). Resulting indices are nearly identical to those given by alternative long-chaining methods.



Here  $Z_i$  is the original input, which lies in a very high-dimensional space.  $Z_i$  is non-linearly mapped, via the embedding layer  $e$ , to an embedding vector  $X_i$  which is of moderately high dimension (up to 5120 dimensions). This embedding is again non-linearly mapped to a lower dimension vector  $E_i^{(1)}$  by the first hidden layer  $g_1$ , and so on, until the final hidden layer  $g_m$ . The output of the final hidden layer  $g_m$ , given by  $V_i := E_i^{(m)}$ , is then *linearly* mapped to the final output, consisting of hedonic price  $H_{it}$  for product  $i$  in all time periods  $t = 1, \dots, T$ .

The output of the final hidden layer,  $V_i = E_i^{(m)}$ , is called the *value embedding* in our context. It is a moderately high-dimensional summary of the product (256 dimensions in our primary model). It is derived from product attributes and directly determines the predicted hedonic price of the product. Note that the embeddings  $V_i$  do not depend on time and thus represent intrinsic, potentially valuable attributes of the product.<sup>14</sup> However, the predicted price does depend on time  $t$  via the coefficient  $\theta_t$ , reflecting the fact that these intrinsic attributes are valued differently across time.

The network mapping (4) makes use of the repeated composition of nonlinear mappings of the form

$$g_\ell : v \mapsto \{E_{k,\ell}(v)\}_{k=1}^{K_\ell} := \{\sigma_{k,\ell}(v' \alpha_{k,\ell})\}_{k=1}^{K_\ell}, \quad (5)$$

where the  $E_{k,\ell}$ 's are called neurons, and  $\sigma_{k,\ell}$  is the activation function that can vary with the layer  $\ell$  and can vary with  $k$ , from one neuron to another.<sup>15</sup> Standard examples include the sigmoid function:  $\sigma(v) = 1/(1 + e^{-v})$ , the rectified linear unit function (ReLU),  $\sigma(v) = \max(0, v)$ , or the linear function  $\sigma(v) = v$ . Individual neurons' activations can be linear or non-linear. The use of a non-linear activation function has been shown to be an extremely powerful tool for generating flexible functional forms, both yielding successful approximations in a wide range of empirical problems and backed by approximation theory. Good approximations can be achieved by considering sufficiently many neurons and layers (e.g., Chen and White 1999; Yarotsky 2017; Kidger and Lyons 2020).

Our empirical model uses up to  $m = 3$  hidden layers, not counting the input. The dimensions of each layer are described in Appendix C. The first layer is a sophisticated

<sup>14</sup>The mapping from  $Z_i$  to  $V_i$  is learned by the neural network, so it may change when the model is re-trained to include additional data.

<sup>15</sup>The standard architecture has an activation function that does not vary with  $k$ , but some architectures such as ResNet-50 (discussed in Section 4) can be viewed as having an activation function that depends on  $k$ , with some neurons linearly activated and others non-linearly activated.

embedding of the input, trained using auxiliary text and image comprehension tasks using a separate dataset; further details are provided in Section 4.

The model is trained by minimizing the loss function

$$\min_{\eta \in \mathcal{N}, \{\theta_t\}_{t=1}^T} \sum_t \sum_i (P_{it} - \theta'_t V_i(\eta))^2 Q_{it}, \quad (6)$$

where  $\eta = (g_1, \dots, g_m)$  denotes all of the parameters of the mapping  $X_i \mapsto V_i =: V_i(\eta)$ . Here we are weighting by the quantity  $Q_{it}$ . Regularization can be used to limit fluctuations of predicted prices across time. This is done by adding a penalty to the loss function:

$$\lambda \sum_i \sum_{t=1}^{T-1} |\theta'_{t+1} V_i(\eta) - \theta'_t V_i(\eta)|, \quad (7)$$

where the penalty level  $\lambda$  is chosen to yield good performance in the validation sample. Such regularization can bias price changes towards zero; however, validation ensures that overall price prediction accuracy is improved.<sup>16</sup> We discuss the sample splitting procedure below in our description of model training.

Next, we give an overview of how the initial embedding is generated. A multilingual BERT model is used to convert text information into a sub-vector  $W_i$  of  $E_i^{(1)}$ , and likewise a ResNet-50 model is used to convert images into another sub-vector  $I_i$  of  $E_i^{(1)}$  (both models are publicly available; see Section 4 below). These models are trained on auxiliary prediction tasks with auxiliary output  $A_{T_i}$  for text and  $A_{I_i}$  for image, which can be illustrated diagrammatically as:

$$Z_i = \begin{bmatrix} \text{Text}_i \\ \text{Image}_i \end{bmatrix} \xrightarrow{e} X_i := \begin{bmatrix} W_i \\ I_i \end{bmatrix} \begin{array}{c} \uparrow A_{T_i} \\ \downarrow A_{I_i} \end{array} \mapsto E_i^{(1)} \mapsto \dots \mapsto E_i^{(m)} := V_i \xrightarrow{\theta'} \{H_{it}\}_{t=1}^T. \quad (8)$$

The text and image embeddings  $W_i$  and  $I_i$ , which form  $X_i$ , are obtained by a procedure known as transfer learning. They are extracted from separate neural networks which were trained to map text or images to auxiliary outputs  $A_{T_i}$  or  $A_{I_i}$ ; in training, these auxiliary outputs were scored on natural language processing tasks and image classification tasks, respectively. This first step was performed by other researchers and used

<sup>16</sup>Following Pakes (2003, p. 1587), we take the view that predictive accuracy should be the primary objective of the hedonic model when applied to price index construction.

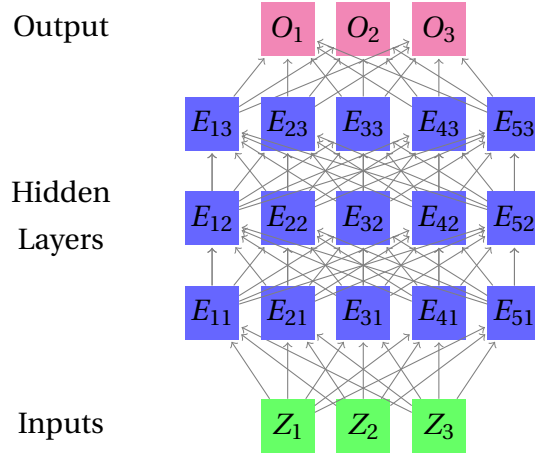


FIGURE 3. Standard architecture of a Deep Neural Network. In the hedonic price prediction network, the penultimate layer is interpreted as an embedding of the product’s hedonic value and the output layer contains predicted hedonic prices in all time periods. In comparison, the networks used for text and image processing have very high-dimensional inputs and outputs, with intermediate hidden layers composed of neural sub-networks. The dense embeddings typically result from taking the last hidden layer of the network.

data unrelated to prices—its aim was to produce versatile, high-quality embeddings  $W_i$  and  $I_i$  for general text and image data, as we elaborate upon in Section 4. Finally, we further fine-tuned parameters of the mapping that generates  $W_i$  for price prediction tasks, yielding some improvements. We did not attempt to fine-tune image embeddings  $I_i$ .

The estimates are computed using a stochastic gradient descent algorithm; care must be taken because the optimization problem is generally not convex. For the price prediction task, we used the Adam algorithm (Kingma, 2014). To train, we randomly split the set of products into three subsets: training (60%), testing (20%), and validation (20%). This split occurs exclusively across products  $i$ , while preserving each product’s time-series of prices and quantities. The training and validation sets are used by the Adam algorithm to minimize the penalized loss given in Equations (6) and (7), while the testing set is used to measure predictive performance in Section 5. The overall process has many tuning parameters; in practice, we chose them by cross-validation. The most important choices concerned the number of neurons and the number of neuron layers.

We depict the process conceptually in Figure 3, where we have a regression problem, and the network illustrates the process of taking raw regressors and transforming them into outputs, the predicted values. In the first row we see the inputs, and in the second row we see the first layer of neurons. The neurons are connected to the inputs, and the connections represent coefficients. Finally, the last layers of neurons are combined to produce a vector of outputs. The coefficients  $\theta_m$  are shown by the connections between the last hidden layer of neurons and the multivariate output. Networks with vector outputs are called multi-task networks.

Prediction methods based on neural networks with many layers of neurons are called deep learning methods. Neural networks recently emerged as a powerful and all-purpose method for a wide range of problems—ranging from prediction and classification analysis to natural language processing tasks. Using many neurons and multiple layers gives rise to deeper networks that are very flexible and can approximate the best prediction or classification rules very well in settings where data are plentiful. In Section 4 and Appendix, we overview the ideas and details of the neural networks for dealing with text and images.

**3.2. Assessing Statistical Significance and Confidence Intervals.** In many settings, researchers may wish to construct standard errors for predicted prices or for model coefficients. We give a brief explanation of how this can be carried out, and refer to Bajari et al. (2023, Section 5) for an empirical illustration.

To this end, the last layer of the neural network provides us with what we interpret as “value embeddings,”

$$V_i = (V_1, \dots, V_p)',$$

where we have chosen  $p = 256$  in our primary specification. We can condition on the training and validation data so that  $V_i$  is considered fixed for all products  $i$ . Then we can use the test data to estimate the following linear regression model:

$$P_{it} = V_i' \tilde{\theta}_t + v_{it}, \quad \tilde{\theta}_t = (\tilde{\theta}_{1t}, \dots, \tilde{\theta}_{pt})', \quad E[v_{it}|V_i] = 0.$$

This is a low-dimensional linear regression model for which we can apply standard inference tools. Applying linear regression to the test data, we obtain a new estimate  $\hat{\theta}_t$  and a new estimated hedonic price

$$\hat{H}_{it} = V_i' \hat{\theta}_t. \tag{9}$$

These additional estimates  $\hat{H}_{it}$  and  $\hat{\theta}_t$  will generally differ from  $H_{it}$  and  $\theta_{it}$  estimated by our primary model. Their main purpose is to allow the construction of confidence intervals for the hedonic price.<sup>17</sup> In particular, due to random sampling of the test data,  $\hat{\theta}_t$  will be approximately Gaussian.

Treating  $V_i$  as fixed, the statistical significance of entries of  $\hat{\theta}_t$  can then be assessed by testing whether the regression coefficients  $\tilde{\theta}_t$  are equal to zero.<sup>18</sup> We may also construct confidence intervals for individual coefficients, as well as level  $1 - \alpha$  confidence intervals for the predicted hedonic price:

$$[L_{it}, U_{it}] = [\hat{H}_{it} \pm \Phi^{-1}(1 - \alpha/2) \text{SE}_{it}], \quad \text{SE}_{it} = \sqrt{V'_{it} \widehat{\text{Cov}}(\hat{\theta}_t) V_{it}}. \quad (10)$$

The same ideas can be used to assess confidence in the downstream estimates of indices and inflation. As we noted in the discussion following the definition of hedonic indices in (3), the estimated indices may be expressed in terms of quantities  $\hat{\theta}'_s \sum_{i \in \mathcal{C}_r} Q_{it} V_i$  for some time periods  $(r, s, t)$ , and these quantities are linear in  $\hat{\theta}$ . Since the L- and P-type indices are (products of) ratios of these quantities, while the F-type is in turn a geometric mean of L- and P-type indices, inference may be established for these indices using the delta method.

**Remark 1.** An advantage of this approach is its simplicity, while a disadvantage is that predictive confidence intervals do not account for uncertainty in estimating value embeddings themselves (indeed, we consider them to be frozen conditional on the training and validation samples). Chernozhukov et al. (2018) offer a way to account for this variability. They propose constructing multiple random splits,  $s = 1, \dots, S$ , of the data into test, training and validation subsets (stratified by month), and taking the median of central values across splits. We do not pursue this approach in the present paper as it requires training  $S$  models instead of only one.

<sup>17</sup>When the training and test datasets are many orders of magnitude larger than the dimension of  $V_{it}$ , as is the case in our data, the differences between  $\hat{H}_{it}$  and  $H_{it}$  are typically small. In some settings,  $\hat{H}_{it}$  may be preferable, e.g., because they lie at the center of the resulting confidence intervals.

<sup>18</sup>It should be noted that since  $V_i$  is determined by our model, and  $\tilde{\theta}'_t V_i = (A' \tilde{\theta}_t)' (A^{-1} V_i)$  for any invertible matrix  $A$ , the coefficients of  $\tilde{\theta}_t$  are only identified if the embeddings  $V_i$  are considered fixed.

#### 4. IMAGE AND TEXT EMBEDDINGS VIA DEEP LEARNING

Typically, customers view products as depicted in Figure 1 of the introduction, where we present details such as the product’s title, its description, and accompanying images. Our task is to convert these product characteristics into numerical vectors, which can be used for estimating hedonic prices, as shown in Figure 2. In what follows, we give a high-level description of how these text and image features are generated.

**4.1. Text Embeddings from the Title and Product Description.** We begin with text embeddings and give a non-technical description of the main ideas. We will mostly focus on the transformer-based large language model BERT, which is one of the most successful text embedding algorithms. We give a more technical review of BERT and its predecessors (Word2Vec and ELMO) in Appendix A.

**4.1.1. High-Level Objectives.** First, we would like to stress that the high-level objective of the text-embedding algorithms is to construct a concise (low-dimensional) numeric representation of sequences of words, such as product titles and descriptions.

Conceptually, the  $j$ -th word in the product description can be treated as a categorical variable and represented with a very large number  $d$  of dummy variables—where  $d$  is the size of the dictionary of words. Still, this representation is not very useful: it is not able to explore word similarity to compactly approximate the dictionary. Indeed each distinct word has a distance of 1 every other word.

Instead, we aim to represent words by vectors of much lower dimension  $r \ll d$ , such that the distance between similar words is small. Denote such potential representation of  $j$ -th word by  $u_j$ , then the dictionary is  $r \times d$  matrix

$$\omega = \{u_j\}_{j=1}^d,$$

where  $r$  is the reduced dimensionality of the dictionary, then each word  $t_j$  in a human-readable dictionary can be represented by the word  $u_j$ .

Text embedding algorithms aim to find an effective representation of dimension  $r$ , where  $r$  is much smaller than  $d$ . This is achieved by treating the entries of  $\omega$  as parameters and estimating them so that the model performs well in certain natural language processing tasks, such as predicting an omitted word in a sentence using surrounding

words, or detecting when two sentences are presented in reverse order. Examples are drawn from corpora of published text (ranging from “small” data, such the entirety of Wikipedia, to a large fraction of all digitized text in the case of GPT-3). These tasks are not related to hedonic prices. However, precisely because they are not related, one can generate extremely large data sets of examples on which the model can be trained.

Once embeddings for individual words have been obtained, we can generate the embedding for the title or description of product  $i$ , containing the embedded words  $\{U_{j,i}\}_{j=1}^J$  by taking averages:

$$W_i = \frac{1}{J} \sum_{j=1}^J \lambda_j U_{j,i}, \quad (11)$$

where  $\lambda_j$ ’s are weights given to the  $j$ -th word. A simple choice is  $\lambda_j = 1/J$ , but one can also use data-dependent weights (see the Appendix for details). One may also simply concatenate the embeddings:  $W_i = \text{vec}(\{U_{j,i}\}_{j=1}^J)$ . This leads to  $rJ$ -dimensional embedding, but as long as  $rJ$  is not overly large, it remains practical.

Once we have obtained the embeddings, how do we judge whether the text embedding is successful? The most obvious check we can do, in our context, is to see if the embeddings are useful for hedonic price prediction. We find that they are extremely useful, as we report later in the paper. One can also check qualitatively to see if words that have similar meanings have similar embeddings. Ordinarily, this is done through the correlation or cosine-similarity:

$$\text{sim}(T_k, T_l) = U_k' U_l / (\|U_k\| \|U_l\|).$$

Here, similar pairs of words—based upon our human understanding of the words’ meaning in context—should attain a higher formal similarity measure, up to a maximum value of 1. Even the first generation of text embedding algorithms was able to achieve remarkable qualitative performance at this task. We give such examples in Appendix A; the latest generation of the algorithms have only improved in their ability to understand language. In what follows, we present a non-technical discussion of BERT, a state-of-art language understanding model which we have used extensively; in the Appendix, we present a more technical discussion.

**4.1.2. An Overview of BERT.** BERT (Bi-directional Encoder Representations from Transformers, Devlin et al., 2018; Vaswani et al., 2017) is a transformer-based model initially

developed by Google. The model is trained using a transfer learning approach, meaning auxiliary prediction tasks—predicting a masked word in a sentence or predicting the order of two sentences—that are not connected to the final tasks, such as predicting hedonic prices in our case. The auxiliary tasks are selected such that the amount of examples is very large (e.g., all of Wikipedia), and such that performing the tasks accurately reflects a high level of comprehension. This initial training with auxiliary tasks is only performed once, so that users may apply the model to smaller datasets without incurring the computational burden of the initial training. Fine-tuning, or further optimization of the network’s weights—which is not computationally expensive—is performed using Amazon’s product descriptions for apparel, as well as price-prediction data.

BERT is particularly good at understanding the meaning of words in context, which is generally attributed to the transformer blocks present in the neural network (Vaswani et al., 2017). The transformer blocks in BERT allow the model to understand the context of a word by looking at the words that come before and after it, rather than just relying on the individual word. The transformer blocks comprise two main components: a self-attention mechanism and a standard feed-forward neural network. The self-attention mechanism allows the model to weigh the importance of different words in a sentence when making predictions, producing so-called “attention weights.” BERT uses several self-attention mechanisms in parallel, thus allowing the model to “attend to” different parts of the input simultaneously. The feed-forward neural network, also known as a fully connected layer, takes the output from the self-attention mechanism and applies a series of non-linear transformations. Both components are applied to the input in parallel, and then the outputs from both are linearly summarized. This summary is then used as the input for the next transformer block.

BERT produces text embeddings by breaking the input text up into standardized words or sub-words known as “tokens.” These tokens are then passed through the BERT model. The transformer blocks process the tokens in parallel and learn to represent each token in a vector space; this representation is called a word embedding. Indeed, the last hidden layers of the network comprise the embeddings used in our study. During training, BERT learns a set of parameters that can be used to generate embeddings for any input text. The embeddings are then fine-tuned to perform a specific natural language understanding task; in our case, we fine-tune them for price prediction. The embeddings produced by BERT are contextual, meaning that a word’s embedding depends both on



both that individual word and its context. This is in contrast to non-contextual embeddings, such as Word2Vec, which assign the same embedding to a word regardless of surrounding words. We present more technical details in Appendix A.

It is helpful to compare BERT to GPT (3 and 4), which has received widespread media attention. GPT also uses transformer blocks, but it does so in an autoregressive fashion. It uses supervised learning, which means it is trained on specific tasks to generate human-like text. GPT is pre-trained on a massive amount of text data. It can then be fine-tuned for various natural language generation tasks such as text summarization, text completion, and translation. While GPT is a tool for generating human-like language, BERT is a tool for understanding language. However, since GPT also must understand language in order to generate conversation, future work may investigate whether GPT can provide embeddings that outperform those generated by BERT.

**4.2. Image Embeddings using ResNet-50.** ResNet-50 is a convolutional neural network designed for image classification (He et al., 2016). Convolutional neural networks work by pooling information extracted from sub-images (say, from all  $16 \times 16$  and  $4 \times 4$  pixel fragments of a  $128 \times 128$  pixel image). The key innovation in ResNet is the use of residual connections: shortcut connections that act as the identity map, bypassing one or more layers. Residual connections allow for a very deep network structure (i.e., a highly flexible model) without suffering from vanishing gradients, a problem that typically limits the ability to learn from data. Consequently, it becomes possible to train much deeper networks while maintaining good performance. ResNet-50 is trained on the ImageNet dataset (with over 14 million images and more than 22,000 different labels). The model has been trained to classify object images into 1000 groups with high accuracy. At the time of its release, the ResNet-50 model achieved the best results in image classification, particularly for the ImageNet and COCO datasets. There are recent advances in computer vision—vision transformers—that import the transformer architecture from large language models such as BERT. They also achieve a near state-of-art performance, but we have not yet explored their use in our context.

Just like with text embeddings, we are not interested in the final predictions of these networks but rather in the last hidden layer, which is taken to be a meaningful summary or “embedding” of the image. The reason is that the last hidden layer is used for object classification using a simple logistic model, so it should represent the object type accurately. This information is clearly useful for our purposes.

## 5. AI-BASED HEDONIC PRICE MODELS AND PRICE INDICES FOR APPAREL

Having constructed numeric embeddings which capture price-relevant characteristics of the product as expressed in its image and text description, we may set about estimating the hedonic price function. In this section, we compare the performance various estimation strategies. We then report derived indices for apparel under our preferred method and contrast it with other major indices.

**5.1. Data.** We use Amazon’s proprietary data on daily average transaction prices and quantities for the first party sales from the entire population of apparel products, with tens of millions of products sold in the Amazon marketplace.<sup>19</sup> Our study covers the period from 01/2013 to 12/2018. The transaction prices of a product  $i$  in month  $t$  are defined as the ratio of total sales ( $S_{it}$ ) over the quantity sold ( $Q_{it}$ ),

$$P_{it} = S_{it}/Q_{it},$$

where the price is treated as missing for the case of no sales.

To estimate the hedonic price function, we collected the most recent description ( $\text{Text}_i$ ) and image ( $\text{Image}_i$ ) available for each product as of July 2019—shortly after the end of the study. Descriptions include the product’s title, brand name, a list of high-level bullet points, and a description provided by the seller.

These images and descriptions are occasionally updated. We remark that changes to images and descriptions do not reflect changes to the underlying product, which would require creation of a new product identifier (see Section 5.1.1 below). Thus, the captured descriptions and images represent each product throughout the study’s duration. Changes to the description or image do not reflect changes to a product’s underlying valuable attributes, hence they should not change its hedonic price.

One key characteristic of our data is the very high turnover of products from month to month, as shown in Figure 4. Moreover, Figure 5 shows that there is considerable growth

---

<sup>19</sup>The quantity information is proprietary, but we note that there are methods of approximating quantity weights based upon product ranking data; see, e.g., Chessa and Griffioen (2019) and Office of National Statistics (2020). Therefore, hedonic prices and hedonic price indices derived using quantity weight can be approximated to various extents by publicly available price information, product information and images, and product rank information.

in the selection of products. As discussed in the introduction, these properties motivate the use of hedonic indices.

The size of the resulting dataset is approximately 20 terabytes, owing to the vast number of products and the length of the time period covered. Cloud computing tools based upon Apache’s Hadoop and Spark were nonetheless highly effective for holding and processing the data (both in terms of cost and computing time).<sup>20</sup> In this case, these cloud computing resources were provided by Amazon Web Services (AWS). Computation of the product embeddings and estimation of the hedonic price function were carried out on a GPU cluster, also provided by AWS.

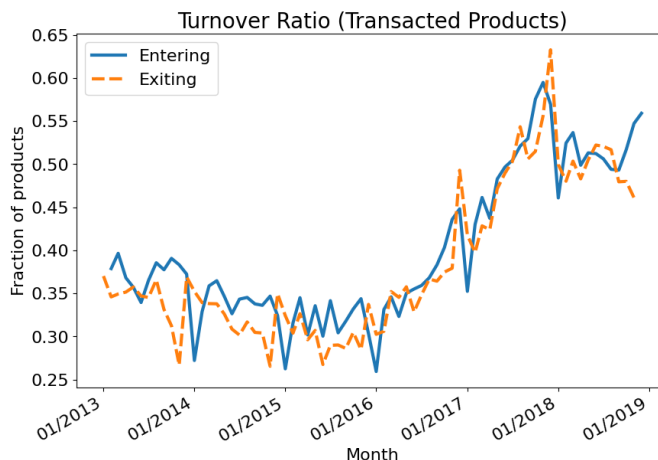


FIGURE 4. Turnover Rate for Products. The Figure shows the share of products with transactions in a given month and no transactions in the previous month (blue line), as well as the share with transactions in a given month and no transactions in the next month (orange line).

5.1.1. *Defining products.* In conducting our study, we have chosen to define “products” at a very granular level. In particular, different sizes, colors, and versions of an article of clothing are treated as distinct products, even if they share the same brand and item name. This gives the neural network model flexibility to group products and adapt to differences in demand between, say, more- or less-popular colors of a given item, and

<sup>20</sup>For example, training our model takes under 24 hours on a cloud computer with 8 Nvidia V100 GPUs, corresponding to less than \$500 in cloud computing fees using the SageMaker environment provided by Amazon Web Services. The largest expense was data storage, which costed roughly \$460 per month.

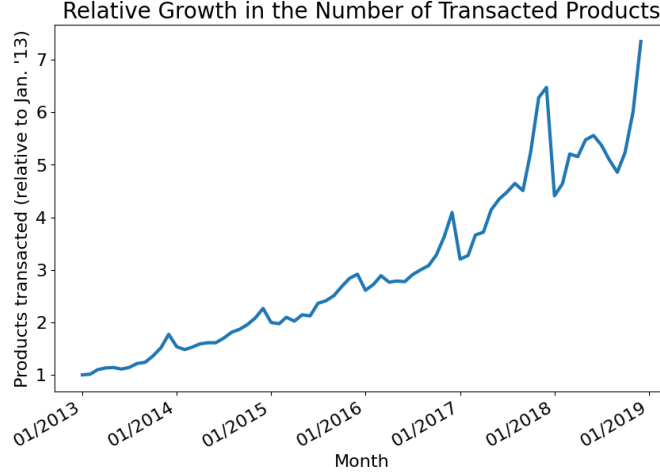


FIGURE 5. Products transacted per month relative to the number of products transacted in January, 2013.

limits the role of human judgement.<sup>21</sup> This level of granularity is particularly natural in our context, since product variations are typically listed at different prices in Amazon’s marketplace. It does, however, influence the interpretation of Figures 4 and 5 as it affects the measurement of product variety and product turnover in our data. Using an auxiliary catalogue of all product variations, we verify in Appendix D.2 that the high out-of-sample accuracy reported later in this section is not driven by variations of essentially similar products: we observe similar accuracy when restricting to products from the hold-out sample which are not a variation of any product in the training data.

**5.2. Out-of-sample Performance for Predicting Prices.** In Table 1, we first examine the predictive performance, recording the  $R^2$  for predicting prices in the hold-out sample of products. In addition to the multi-task price prediction neural network using text and image embeddings  $W$  (from BERT) and  $I$  (from ResNet-50), which is discussed extensively in Sections 3 and 4, we consider the following alternative methods for estimating the hedonic price function:

- linear regression using product category and sub-category indicators;
- linear regression using text and image embeddings  $W$  and  $I$  generated by BERT and ResNet-50 as discussed in Sec. 4;

<sup>21</sup>Other choices are possible. See Aizcorbe (2014, Sec. 4.1) for further discussion on how researchers may choose to define the product.

Method	$R^2$
Linear Model with basic catalogue features	$\approx 30 - 45\%$
Linear Model with embeddings $W$ and $I$	$\approx 55 - 65\%$
Random Forest/Boosted Tree Models with embeddings $W$ and $I$	$\approx 70 - 80\%$
Single-Task Neural Network with embeddings $W$ and $I$	$\approx 75 - 85\%$
Multi-Task Neural Network with embeddings $W$ and $I$	$\approx 80 - 90\%$

TABLE 1. Summary of Out-of-Sample Performance of the Empirical Hedonic Price Function.

- gradient-boosted trees using the aforementioned features  $W$  and  $I$ ;<sup>22</sup>
- an alternative, “single-task” neural network with embeddings  $W$  and  $I$ .<sup>23</sup>

Results from the comparison are seen in Table 1. When we switch from the basic catalogue features to the embeddings  $W$  and  $I$ , we obtain a first major improvement—even using linear regression in the final step. We obtain a second major improvement when we switch from linear regression to a scalable implementation of tree-based methods (gradient-boosted trees). Finally, we obtain one more major improvement as we switch from tree-based methods to a multi-task neural network. The neural network model achieves substantially better predictive performance than other methods.

Figure 6 also presents the month-to-month performance of the various models. The out-of-sample  $R^2$  for the best multi-task neural network model ranges between 80% to 90%. Multi-task neural networks (NNs) uniformly dominate single-task NNs, which in turn uniformly dominate boosted tree models and linear models. We also contrast the performance of BERT-based embeddings of the product description,  $W$ , with an alternative embedding of the product description produced using the ELMO network<sup>24</sup> The BERT-based multi-task NN almost always outperforms the ELMO-based multi-task NN, although the difference in performance is relatively small.

<sup>22</sup>Gradient-boosted trees (see Chen and Guestrin, 2016) are a relative of random forests; they have been shown to perform comparably in general prediction tasks and are easier to estimate in very large datasets.

<sup>23</sup>Roughly, single-task learning corresponds to estimating a separate model in each time period. See Appendix C for a formal definition.

<sup>24</sup>ELMO is a large language model based upon a recursive neural network architecture. Although it does not use the modern transformer architecture, it performs nearly as well as BERT in our setting. We review details of ELMO in Appendix A

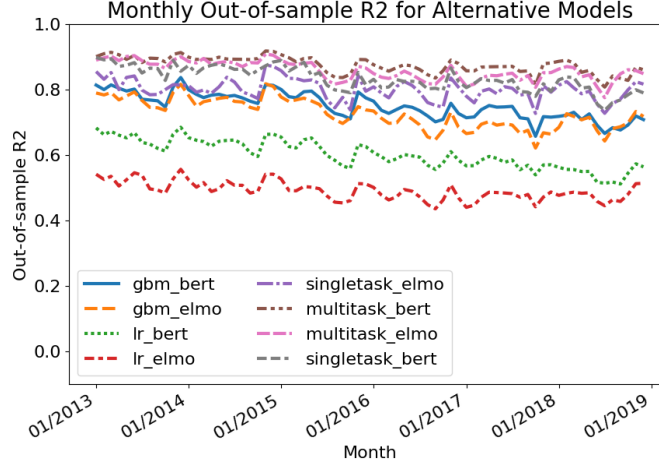


FIGURE 6. The out-of-sample performance of the empirical hedonic price function obtained using our neural network every month since January, 2013, in comparison to alternative models. Multi-task neural networks dominate single-task neural networks (discussed in Appendix C). Neural network models dominate boosted tree models (GBM), which in turn dominate linear models (LM).

One limitation the multi-task model in comparison to the single-task model is that the multi-task model must be retrained using the full dataset when new data become available. After retraining, the model will in general produce different embeddings  $V_i$  and different predictions for initial time periods. Due to the model's high out-of-sample accuracy, however, we expect that changes to resulting price indices would be minimal.

As a practical suggestion to limit retraining, a researcher could freeze the value embedding model  $(\text{Text}_i, \text{Image}_i) \mapsto V_i$  and resulting embeddings  $V_i$  for some time, so that only  $\hat{\theta}_t$  is estimated each time period, and monitor predictive accuracy to decide when it is necessary to retrain the full model. We recommend updating the value embedding model as frequently as is practical in order to account for drift in the distribution of product characteristics (due to factors such as technological change). Indeed, the performance of the best multi-task models is best at the beginning of the studied period and worsens over time, possibly due to the increasing variety and number of products being transacted (shown in Figure 5). It is worth mentioning that the out-of-sample  $R^2$  agrees with validation  $R^2$  we obtained in training (not reported), which suggests that our training approach successfully limits overfitting.

5.2.1. *Examples of hits and misses.* We can inspect the performance of the best predictive model using examples. In Figures 7 and 8, we present one example of accurate prediction and one of inaccurate prediction. For the first item, a sweater, the neural network model predicts a price of about 100. The time-averaged average price as shown on `camelcamel.com`—a third-party website which tracks prices on Amazon—is \$97, with the price ranging from \$39 (corresponding to liquidation events) to \$120. For the second example, a designer dress, the neural network predicts the price of this item at about \$300, but the most recent offer prices for this item were around \$2400. While this seems to be a serious inaccuracy, the price history for this item, again per `camelcamel.com`, suggests that there were periods when the price ranged between \$206 and \$2800; the average listed price was \$464, which is not as far off from the prediction.

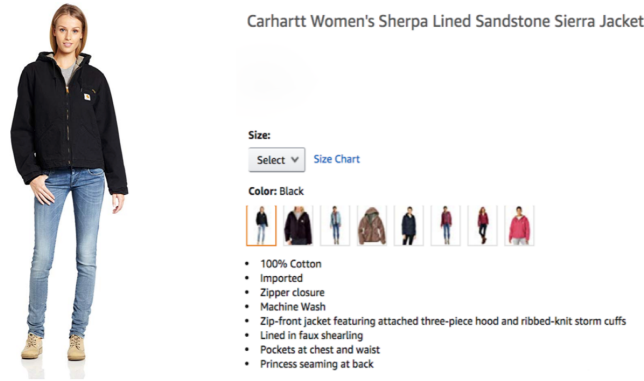


FIGURE 7. An example of accurate price prediction for a women's jacket (black, size medium): The neural network predicts the price of this item at about \$100. The average price on the price aggregator `camelcamel.com` is \$97, with the offer price ranging from \$39 to \$120.

5.3. **Hedonic Price Indices for Apparel.** We will now use our price predictions to construct hedonic indices, following the definitions introduced in Section 2. For this exercise, we use the best predictive model from the preceding comparison, which we have described in detail in Sections 3 and 4.

In Table 2, we present our main results—estimates of the average annual rate of inflation in apparel from 2014 to 2019, via Fisher hedonic price indices (yearly-chained, monthly chained, and the geometric mean of the two), the Jevons posted-price index, the Adobe DPI (Adobe Inc., 2024), and CPI (Bureau of Labor Statistics, 2021):

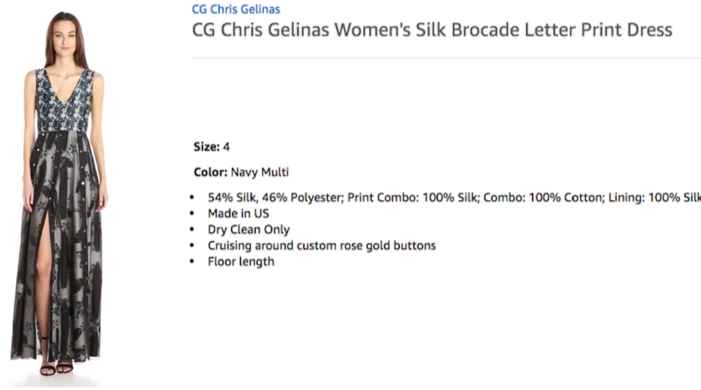


FIGURE 8. An example of inaccurate prediction: The neural network predicts the price of this item at about \$300, but the true price is roughly \$2400. However, the price history for this item as recorded by the price aggregator `camelcamel.com`, shows that its price ranged between \$206 and \$2800, with an average price of \$464.

<i>Apparel Indices</i>	<i>Change in Price Index</i>
<b>Fisher Hedonic, Yearly Chaining (FY)</b>	-0.98%
Fisher Hedonic, Monthly Chaining (FM)	-5.27%
Fisher Hedonic, Geometric Mean ( $\sqrt{FY \cdot FM}$ )	-2.28%
Fisher Matched, Monthly Chaining (FI)	-3.12%
Jevons Posted Price Index, Daily Chained (JPI)	-3.01%
Adobe Digital Price Index, Monthly Chained (DPI)	-2.02%
<b>U.S. Urban Apparel (BLS)</b>	-0.31%

TABLE 2. Estimates of Average Annual Rate of Inflation in Apparel over five years, 2014-2019: Fisher Hedonic Index, Fisher Matched Index, Jevons Posted Price Index, Adobe DPI, and the BLS urban CPI.

The main conclusions we draw from these results are the following.

- E.1 The main index – the yearly-chained FHPI – suggests an average price decline in 2014-2019 of .98%. In contrast, the BLS CPI for apparel suggests that there was a somewhat smaller decline in the price level.



E.2 The yearly chained FHPI declines at a slower rate than the matched Fisher index, Adobe DPI, and the monthly-chained FHPI, highlighting the importance of using hedonic adjustment and long chaining.

The matched index potentially contains selection bias induced by the rapid turnover of products, illustrated in Figure 4. Moreover, as we emphasized in the introduction, hedonic prices allow us to perform “long” chaining, mitigating chain drift.

Differences from the CPI could be attributed to a number of sources: There are methodological differences: the BLS uses a hybrid index where price levels are first measured within narrow subgroups of products, without quantity weighting, and then aggregated using a Tornqvist index with expenditure shares for subgroups. Moreover, the BLS also uses their own hedonic models to perform quality adjustments.<sup>25</sup> Moulton (2018) estimates that the upward bias in the overall (non-apparel specific) CPI index may be in the range [4%, 1.3%], which could potentially reconcile some of the difference. Other sources of the difference could be the Amazon-specific productivity improvements leading to lower prices (e.g., improvements in supply-chain productivity) and different shares of products in customers’ baskets.

We also performed similar calculations using a linear model instead of the neural network (not reported), and conclusions are qualitatively very similar. However, the neural network models—which do exhibit superior predictive ability—result in a more pronounced quantitative drop in the price index level. The differences between the two are relatively small, likely due to the fact that quantity-weighted averaging eliminates noise present in the linear model’s predictions.

In Figure 9, we demonstrate the dynamics of the year-chained FHPI and compare it with the Jevons index and the monthly chained FHPI. The Jevons index is a geometric mean of the posted price relatives and does not incorporate quantity weighting. It underlies the Billion Price Project (Cavallo, 2018) and is also used by statistical agencies as a complementary index (Office of National Statistics, 2020). It is convenient because

---

<sup>25</sup>Note that our goal here is also not to replicate the BLS CPI index but to construct a modern hedonic version of a classical Fisher index. We also note that the replication is simply precluded in our setting because we can not assign extremely many products manually to the subcategories used by BLS.

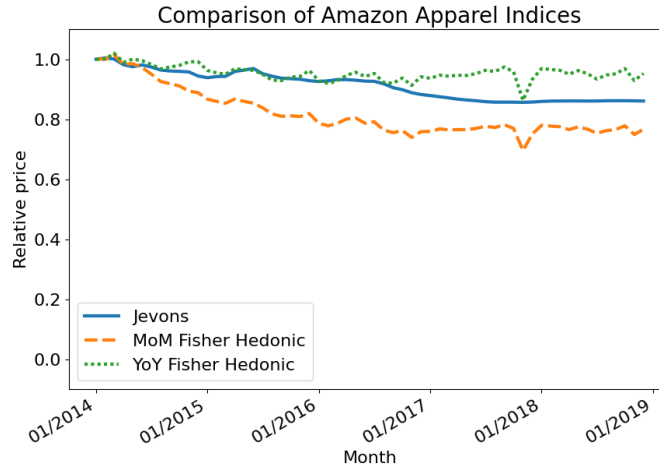


FIGURE 9. Dynamics of the yearly-chained Fisher Hedonic Price Index (dotted line), the Jevons Price Index (solid line), and the monthly-chained Fisher Hedonic Price Index (dashed line).

it only relies on publicly posted prices and does not use quantity information, measuring within-product price level changes that are averaged over a much larger universe of products (including those without transactions). In contrast, the Fisher index reflects both within-product price changes and the as well as changes due to substitution arising from utility-maximizing behavior by customers with budget constraints. Our main empirical finding from this comparison is the following.

E.3 The Jevons index exhibits steady declines over the studied period, with a total decline of more than 10%; the rate of decline is considerably larger and much more implausible than that of the yearly chained FHPI.

This may suggest that either the chain drift or the substitution effects are large enough to matter qualitatively. Without the ability to do long chaining and without any proxy for quantity information (such as sales rank data), it is doubtful that Jevons-type indices can accurately approximate inflation. Still, the Jevons index continues to be a very important, simple way to gauge large changes in inflation, as the BPP (based on the Jevons index) has amply demonstrated (Cavallo and Rigobon (2016)).<sup>26</sup>

<sup>26</sup>Note that given a reference series, such as the CPI, one can always use a post-processing or “now-casting” approach to modify Jevons or any other digital index  $R_g$  to track the reference index  $R_f$  as well as possible. For example, one can construct the best linear predictor  $\hat{a} + \hat{b} \log R_g$  of  $\log R_f$  using the historical data on  $(R_g, R_f)$  in a given time frame and then use  $R_g^* = \exp(\hat{a} + \hat{b} \log R_g)$  as the post-processed index.

Finally, when we compare yearly-chained FHPI to monthly-chained FHPI, we observe:

E.4 The monthly-chained FHPI exhibits a strong, very implausible 20% decline in the price level, in sharp contrast to the yearly-chained FHPI.

This highlights the importance of mitigating “chain drift,” which appears to be quite severe for frequently compounded series such as the monthly chained FHPI. For this reason, the monthly chained FHPI and other frequently chained indices (matched Fisher and daily-chained Jevons) are unlikely to measure inflation accurately, especially over longer horizons. Of course, as we stressed before for the Jevons index, these indices remain useful over short horizons to gauge large changes in the price level.<sup>27</sup>

## 6. CONCLUSION

We develop empirical, AI-based models for price prediction, and use them to derive quality adjusted price indices for measuring changes in consumer welfare. To achieve this, we generate product attributes (or ‘features’) from text descriptions and images using deep neural networks, namely BERT (for text) and ResNet-50 (for images). We then use these features to non-parametrically impute prices—again using deep learning—with a multi-task neural network that predicts a complete time-series of prices for each product. The method relies on publicly available, open-source software components. We apply the models to Amazon’s proprietary data on first-party sales in apparel. Resulting models have high predictive accuracy for several product categories, with the  $R^2$  ranging from 80% to 90%. Our main price index estimates the rate of inflation in apparel to be moderately negative, somewhat lower than the rate estimated by the CPI.

We find the performance of AI-based models remarkable for two reasons. First, their high predictive accuracy suggests that theoretical hedonic price models provide a good approximation of real-world equilibrium prices. Second, our methods are scalable to many products and avoid the significant manual effort required to construct more traditional matched or hedonic price indices. The good performance of these methods

---

This further highlights the potential usefulness of the Jevons and other digital indices, even though their raw versions exhibit large biases. In our discussions, we focus on raw indices without post-processing, with the understanding that post-processing can always be employed as needed.

<sup>27</sup>They can also be employed in a nowcasting fashion, as explained in the previous footnote.

suggests that AI-based methods, particularly embeddings, can be used to build real-time hedonic price indices using electronic data and facilitate price research in many settings. These methods can also be applied to economic research in other areas—for example, labor economics, where embeddings can be used to characterize workers' job-relevant characteristics, or industrial organization, where they can be used to characterize firms and their products. Since all of the main ingredients of our models are open source, future research in these directions can readily build upon our work.

There are natural directions for further research. First, we found that images add very little to prediction accuracy once the text embeddings are included in the model. Therefore, finding better ways to leverage image data is an important unresolved problem. Another important further research direction is model explainability. When the price of the product changes, we would like to explain the price change in terms of changing valuations of key attributes. Similarly, when comparing the prices of two products, we would like to attribute the price difference to the valuations of key attributes. Given the non-linearity of our models, explainability is not a simple problem. We hope that interested researchers will pursue these open problems.

## 7. ACKNOWLEDGEMENTS

We thank the participants at the 2018 Federal Economic Statistics Advisory Committee meeting, the 2019 Allied Social Science meetings, the 2019 Federal Reserve Board conference on "Nontraditional Data, Machine Learning, and Natural Language Processing in Macroeconomics", the 2019 Brookings Conference "Can big data improve economic measurement?" and the 2021 CEMMAP conference "Measuring prices and welfare," and seminars at Berkeley, European Bank of Reconstruction and Development, MIT, UCL, and York. We are grateful to Andrew Chesher, Greg Duncan, Kevin Fox, John Haltinwager, James Heckman, Daniel Miller, Mathew Shapiro, Bernhard Schölkopf, James Stock, and Weining Wang for helpful comments during the various stages of this project. All authors were fully or partially employed by Amazon.com while conducting the research.

## REFERENCES

- Adobe Inc. Adobe digital price index. <https://business.adobe.com/resources/digital-price-index.html>, 2024. [Accessed 16-10-2024].
- Ana Aizcorbe. *A practical guide to price index and hedonic techniques*. Oxford University Press, USA, 2014.
- Patrick Bajari and C Lanier Benkard. Demand estimation with heterogeneous consumers and unobserved product characteristics: A hedonic approach. *Journal of political economy*, 113(6):1239–1276, 2005. Publisher: The University of Chicago Press.
- Patrick Bajari, Zhihao Cen, Victor Chernozhukov, Manoj Manukonda, Jin Wang, Ramon Huerta, Junbo Li, Ling Leng, George Monokroussos, and Suhas Vijaykumar. Hedonic prices and quality adjusted price indices powered by ai. *CEMMAP, London*, 2019.
- Patrick Bajari, Zhihao Cen, Victor Chernozhukov, Manoj Manukonda, Suhas Vijaykumar, Jin Wang, Ramon Huerta, Junbo Li, Ling Leng, George Monokroussos, et al. Hedonic prices and quality adjusted price indices powered by ai. *arXiv preprint arXiv:2305.00044*, 2023.
- C. Lanier Benkard and Patrick Bajari. Hedonic price indexes with unobserved product characteristics, and application to personal computers. *Journal of Business & Economic Statistics*, 23(1):61–75, 2005. Publisher: Taylor & Francis.
- Steven Berry, James Levinsohn, and Ariel Pakes. Automobile prices in market equilibrium. *Econometrica: Journal of the Econometric Society*, pages 841–890, 1995. Publisher: JSTOR.
- Steven Berry, James Levinsohn, and Ariel Pakes. Differentiated products demand systems from a combination of micro and macro data: The new car market. *Journal of political Economy*, 112(1):68–105, 2004. Publisher: The University of Chicago Press.
- Charles Brown. Equalizing differences in the labor market. *The Quarterly Journal of Economics*, 94(1):113–134, 1980.
- U.S. Bureau of Labor Statistics. CPI for all urban consumers, 2014-2021 [data set]. <http://data.bls.gov>, 2021. [Accessed 16-10-2024].
- Michael Cafarella, Gabriel Ehrlich, Tian Gao, John C Haltiwanger, Matthew D Shapiro, and Laura Zhao. Using machine learning to construct hedonic price indices. Technical report, National Bureau of Economic Research, 2023.
- Alberto Cavallo. Online and official price indexes: Measuring Argentina’s inflation. *Journal of Monetary Economics*, 60(2):152–165, 2013.

- Alberto Cavallo. Are Online and Offline Prices Similar? Evidence from Large Multi-channel Retailers. *American Economic Review*, 107(1):283–303, January 2017. ISSN 0002-8282. doi: 10.1257/aer.20160542. URL <https://www.aeaweb.org/articles?id=10.1257/aer.20160542>.
- Alberto Cavallo. Scraped Data and Sticky Prices. *The Review of Economics and Statistics*, 100(1):105–119, March 2018. ISSN 0034-6535. URL [https://doi.org/10.1162/REST\\_a\\_00652](https://doi.org/10.1162/REST_a_00652).
- Alberto Cavallo and Roberto Rigobon. The distribution of the size of price changes. Technical report, National Bureau of Economic Research, 2011.
- Alberto Cavallo and Roberto Rigobon. The billion prices project: Using online prices for measurement and research. *Journal of Economic Perspectives*, 30(2):151–78, 2016.
- Kenneth Y Chay and Michael Greenstone. Does air quality matter? evidence from the housing market. *Journal of political Economy*, 113(2):376–424, 2005.
- Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.
- Xiaohong Chen and Halbert White. Improved rates and asymptotic normality for non-parametric neural network estimators. *IEEE Transactions on Information Theory*, 45(2):682–691, 1999. Publisher: IEEE.
- Victor Chernozhukov, Mert Demirer, Esther Duflo, and Ivan Fernandez-Val. Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in india. Technical report, National Bureau of Economic Research, 2018.
- Victor Chernozhukov, Alfred Galichon, Marc Henry, and Brendan Pass. Identification of hedonic equilibrium and nonseparable simultaneous equations. *Journal of Political Economy*, 2020.
- Antonio G Chessa and Robert Griffioen. Comparing price indices of clothing and footwear for scanner data and web scraped data. *Economie et Statistique*, 509(1):49–68, 2019. Publisher: Persée-Portail des revues scientifiques en SHS.
- Pierre-André Chiappori, Robert J McCann, and Lars P Nesheim. Hedonic price equilibria, stable matching, and optimal transport: equivalence, topology, and uniqueness. *Economic Theory*, 42(2):317–354, 2010. Publisher: Springer.
- AT Court. Hedonic price indexes with automotive examples. In *The dynamics of automobile demand*, pages 98–119. General Motors Corporation New York, 1939.

- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- W Erwin Diewert. Exact and superlative index numbers. *Journal of econometrics*, 4(2): 115–145, 1976. Publisher: Elsevier.
- W Erwin Diewert. Index number issues in the consumer price index. *Journal of Economic Perspectives*, 12(1):47–58, 1998.
- W Erwin Diewert and Kevin J Fox. Substitution bias in multilateral methods for cpi construction. *Journal of Business & Economic Statistics*, 40(1):355–369, 2022.
- Gabriel Ehrlich, John Haltiwanger, Ron Jarmin, David Johnson, and Matthew D Shapiro. Minding your ps and qs: Going from micro to macro in measuring prices and quantities. In *AEA Papers and Proceedings*, volume 109, pages 438–443. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, 2019.
- Gabriel Ehrlich, John C Haltiwanger, Ron S Jarmin, David Johnson, Ed Olivares, Luke W Pardue, Matthew D Shapiro, Laura Zhao, et al. Quality adjustment at scale: Hedonic vs. exact demand-based price indices. Technical report, National Bureau of Economic Research, 2023.
- Ivar Ekeland, James J Heckman, and Lars Nesheim. Identification and estimation of hedonic models. *Journal of political economy*, 112(S1):S60–S109, 2004.
- Oded Eltetö and Pal Köves. On a problem of index number computation relating to international comparison. *Statistikai Szemle*, 42(10):507–518, 1964.
- Corrado Gini. On the circular test of index numbers. *Metron*, 9(9):3–24, 1931.
- Austan D Goolsbee and Peter J Klenow. Internet rising, prices falling: Measuring inflation in a world of e-commerce. In *Aea papers and proceedings*, volume 108, pages 488–92, 2018.
- William M Gorman. A possible procedure for analysing quality differentials in the egg market. *The Review of Economic Studies*, 47(5):843–856, 1980.
- Yuriy Gorodnichenko and Oleksandr Talavera. Price Setting in Online Markets: Basic Facts, International Comparisons, and Cross-Border Integration. *American Economic Review*, 107(1):249–282, January 2017. ISSN 0002-8282. doi: 10.1257/aer.20141127. URL <https://www.aeaweb.org/articles?id=10.1257/aer.20141127>.
- Yuriy Gorodnichenko, Viacheslav Sheremirov, and Oleksandr Talavera. Price Setting in Online Markets: Does IT Click? *Journal of the European Economic Association*, 16(6): 1764–1811, December 2018. ISSN 1542-4766. doi: 10.1093/jeea/jvx050. URL <https://doi.org/10.1093/jeea/jvx050>.

- Michael Greenstone. The continuing impact of sherwin rosen's "hedonic prices and implicit markets: product differentiation in pure competition". *Journal of Political Economy*, 125(6):1891–1902, 2017.
- Z Griliches. Hedonic price indexes for automobiles: an econometric of quality change. Price Statistics Review Committee, 1961.
- Robert M. Groves and Brian A. Harris-Kojetin, editors. *Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy*. The National Academies Press, Washington, DC, 2017. ISBN 978-0-309-45428-5. doi: 10.17226/24652. URL <https://nap.nationalacademies.org/catalog/24652/innovations-in-federal-statistics-combining-data-sources-while-protecting-privacy>.
- Sukjin Han, Eric H Schulman, Kristen Grauman, and Santhosh Ramakrishnan. Shapes as product differentiation: Neural network embedding in the analysis of markets for fonts. *arXiv preprint arXiv:2107.02739*, 2021.
- Jessie Handbury, Tsutomu Watanabe, and David E Weinstein. How much do official price indexes tell us about inflation? Technical report, National Bureau of Economic Research, 2013.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- ILO, World Bank, et al. *Consumer price index manual: Theory and practice*. International Labour Organization, 2004.
- Lorraine Ivancic, W Erwin Diewert, and Kevin J Fox. Scanner data, time aggregation and the construction of price indexes. *Journal of Econometrics*, 161(1):24–35, 2011. Publisher: Elsevier.
- Ron S Jarmin. Evolving measurement for an evolving economy: thoughts on 21st century us economic statistics. *Journal of Economic Perspectives*, 33(1):165–184, 2019.
- Greg Kaplan and Sam Schulhofer-Wohl. Inflation at the household level. *Journal of Monetary Economics*, 91:19–38, 2017. Publisher: Elsevier.
- Patrick Kidger and Terry Lyons. Universal approximation with deep narrow networks. In *Conference on learning theory*, pages 2306–2327, 2020. tex.organization: PMLR.
- DP Kingma. Adam: a method for stochastic optimization. In *Int Conf Learn Represent*, 2014.
- Kelvin J Lancaster. A new approach to consumer theory. *Journal of political economy*, 74(2):132–157, 1966.



- David Lebow. *Modernizing the Consumer Price Index for the 21st Century: National Academies of Sciences, Engineering, and Medicine: 2022*. Springer, 2023.
- Andrew Leicester. The potential use of in-home scanner technology for budget surveys. In *Improving the measurement of consumer expenditures*, pages 441–491. University of Chicago Press, 2014.
- Daniel McFadden. The measurement of urban travel demand. *Journal of public economics*, 3(4):303–328, 1974. Publisher: Elsevier.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.
- James Milton and Jeanine Treffers-Daller. Vocabulary size revisited: the link between vocabulary size and academic achievement. *Applied Linguistics Review*, 4(1):151–172, 2013.
- Brent R Moulton. The measurement of output, prices, and productivity. *Report, Productivity Measurement Initiative under The Hutchins Center on Fiscal and Monetary Policy, Brookings Institution, Washington, DC*, 2018.
- Lars Nesheim. Hedonic price functions. *CEMMAP, London*, 2006.
- Andrew Ng. Nuts and bolts of building ai applications using deep learning. *NIPS Keynote Talk*, 2016.
- U.K. Office of National Statistics. Using statistical distributions to estimate weights for web-scraped price quotes in consumer price statistics. Technical report, U.K. Office of National Statistics, September 2020.
- Ariel Pakes. A reconsideration of hedonic price indexes with an application to PC's. *American Economic Review*, 93(5):1578–1596, 2003.
- Sherwin Rosen. Hedonic prices and implicit markets: product differentiation in pure competition. *Journal of political economy*, 82(1):34–55, 1974.
- Statistics Bureau of Japan. Explanation of the consumer price index, Jan 2022. URL <https://www.stat.go.jp/english/data/cpi/1590.html>.
- James H Stock. Nonparametric policy analysis: an application to estimating hazardous waste cleanup benefits. *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, pages 77–98, 1991.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- David Wasshausen and Brent Moulton. The role of hedonic methods in measuring real gdp in the united states. *BEA Papers*, 67, 2006.

Dmitry Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94:103–114, 2017. Publisher: Elsevier.

Shipei Zeng. Hedonic imputation with tree-based decision approaches. *Working Paper, School of Economics, University of New South Wales, December, 2020.*

## APPENDIX A. OVERVIEW OF TEXT EMBEDDING MODELS

Here we provide a more technical overview of the text embedding models used in this paper.

**A.1. First generation: Word2Vec Embeddings.** We first recall some basic ideas underlying the Word2Vec algorithm (Mikolov et al., 2013). Here we rely on notation in Section 4. The goal is to find the  $r \times d$  matrix  $\omega = \{u_j\}_{j=1}^d$ , representing  $d$  words in  $r$  dimensional space. The columns are the embeddings for the words.

We can think of a word appearing in sentence as random variable  $T$ ; and we can let  $U$  denote the corresponding embedding. Word2Vec trains the word embeddings by predicting the middle word from the words that surround it in word sentences. Given a subsentence  $s$  of  $K + 1$  words, we have a central word  $T_{c,s}$  whose identity we have to predict and we have the words  $\{T_{o,s}\}$  that surround it. Collapse the embeddings for context words by a sum,

$$\bar{U}_o = \frac{1}{K} \sum_o U_{o,s},$$

where  $U_{o,s}$  is the element of  $\omega$  corresponding to the word  $T_{o,s}$ . This step imposes a drastic simplifying assumption that the context words are exchangeable.

The probability of middle word  $T_{c,s}$  being equal to  $t$  is modeled via multinomial logit function:

$$p_s(t; \pi, \omega) := P(T_{c,s} = t \mid \{T_{o,s}\}; \omega) = \frac{\exp(\pi'_t \bar{U}_s(\omega))}{\sum_{\bar{t}} \exp(\pi'_{\bar{t}} \bar{U}_s(\omega))},$$

where  $\pi = (\pi_1, \dots, \pi_d)$  is  $m \times d$  matrix conformable parameter vectors defining the choice probabilities. The model constraints  $\pi = \omega$ , and estimates  $\omega$  by using the maximum quasi-likelihood method:

$$\max_{\omega=\pi} \sum_{s \in \mathcal{S}} \log p_s(T_{i,s}; \pi, \omega),$$

where the summation is over many examples  $\mathcal{S}$  of subsequences  $s$ .

In summary, the Word2Vec algorithm transforms text into a vector of numbers that can be used to compactly represent words. The algorithm trains a neural network in a supervised manner such that the contextual information is used to predict another part of the text. For example, let's say that the title description of the item is: "Hiigoo Fashion Women's Multi-pocket Cotton Canvas Handbags Shoulder Bags Totes Purses". The

model will be trained using many  $n$ -word subsentence examples, such that the center word is predicted from the rest. If we just use  $K = 3$  subsentence examples, then we train the model using the following examples: (Hiigoo,Women's)  $\rightarrow$  Fashion, (Fashion,Multi-pocket) $\rightarrow$  Women's, (Women's,Cotton)  $\rightarrow$  Multi-pocket, and so on.

We can examine the quality of word embedding by assessing predictive performance for price prediction tasks. We can also qualitatively inspect whether embeddings capture word similarity. For example, we found that the embeddings for "necktie" and "bowtie" are most cosine-similar to the word "tie." The embeddings also seem to induce an interesting vector space on the set of words, which seems to encode analogies well. For example, the embedding for the word "briefcase" is most cosine-similar to the artificial latent word

$$\text{Word2Vec}(\text{men's}) + \text{Word2Vec}(\text{handbag}) - \text{Word2Vec}(\text{women's}).$$

Examples like this and others reported in Mikolov et al. (2013) supported the idea of summing the embeddings for words in a sentence to produce an embedding for sentences.

Word2vec embeddings were among the first generation of early successful algorithms. These algorithms have been improved by the next generation of NLP algorithms, such as ELMO and BERT, which are discussed next.

**Second Generation: ELMO.** The Embeddings from Language Models (ELMO) algorithm (Peters et al, 2018) uses the ideas of the Shannon game, where we guess the next word in the sentence  $m$  with  $n$  words, i.e.

$$p_{k,m}^f(t) = P[T_{k+1,m} = t | T_{1,m}, \dots, T_{k,m}; \theta]$$

and also uses reverse guessing as well:

$$p_{k,m}^b(t) = P[T_{k-1,m} = t | T_{k,m}, \dots, T_{n,m}; \theta],$$

where  $\theta$  is a parameter vector. Recursive neural networks with single or multiple hidden layers are used to model these probabilities. Parameters are estimated using quasi-maximum log-likelihood methods, where the forward and backward log quasi-likelihoods are added together.

To give a simple example, suppose we wanted incorporate more contextual information into our word embeddings. Instead of collapsing embeddings for the context word

by a sum, we could assign individual parameters to each preceding word. This would result in a model closely resembling the previous model, but where the order of context words would play a role. For example, we could model

$$P(T_{k,m} = t \mid \{T_{j,m}\}_{j=1}^{k-1}) = \frac{e^{\sum_{j=1}^{k-1} \pi'_{t,k} U_{k,m}(\omega)}}{\sum_{\bar{t}} e^{\sum_{j=1}^{k-1} \pi'_{\bar{t},k} U_{k,m}(\omega)}},$$

and similarly in reverse order. ELMO uses a more sophisticated (and more parsimonious) non-linear recursive nonlinear regression model to build these probabilities, shown in Figure 10. The resulting model is an example of a recurrent neural network (or RNN).

The basic structure of ELMO is as follows: Given a sentence  $m$  of  $n$  words, (1) words are mapped to context-free embeddings in  $\mathbb{R}^d$ . (2) A network is trained to predict each word  $T_{k,m}$  of a string given (a) words  $(T_{1,m}, \dots, T_{k-1,m})$  or (b) words  $(T_{k+1,m}, \dots, T_{n,m})$ . The objective is to minimize the average over the sum of the log-loss over the  $2n - 2$  prediction tasks, where the average is taken over all sentences. (3) The embedding of word  $T_{k,m}$  is given by a weighted average of outputs of certain hidden neurons corresponding to this word's entire context. Importantly, the same final logistic ("softmax") layer is used for prediction objectives (2a) and (2b). Thus the inputs to this layer, which represent the forward and backward context, are constrained to lie in "the same space."

**A.1.1. Training.** In Figure 10, the output probability distribution  $p_k^f$  is taken as a prediction of  $T_{k+1,m}$ ; similarly  $p_k^b$  is taken as a prediction of  $T_{k-1,m}$ . The parameters of the network  $\theta$  are obtained by maximizing the quasi-likelihood:

$$\max_{\theta} \sum_{m \in \mathcal{M}} \left( \sum_{k=1}^{n-1} \log p_{k,m}^f(T_{k+1,m}; \theta) + \sum_{k=2}^n \log p_{k,m}^b(T_{k-1,m}; \theta) \right),$$

where  $\mathcal{M}$  is a collection of sentences (titles and product descriptions) in our data set.

**A.1.2. Producing embeddings.** To produce embeddings from the trained network, each word  $t_k$  in a sentence  $m = (t_1, \dots, t_k)$  is mapped to a weighted average of the outputs of the hidden neurons indexed by  $k$ :

$$t_k \mapsto w_k := \sum_{i=1}^L \gamma_i w_{ki}^f + \bar{\gamma}_i w_{ki}^b.$$

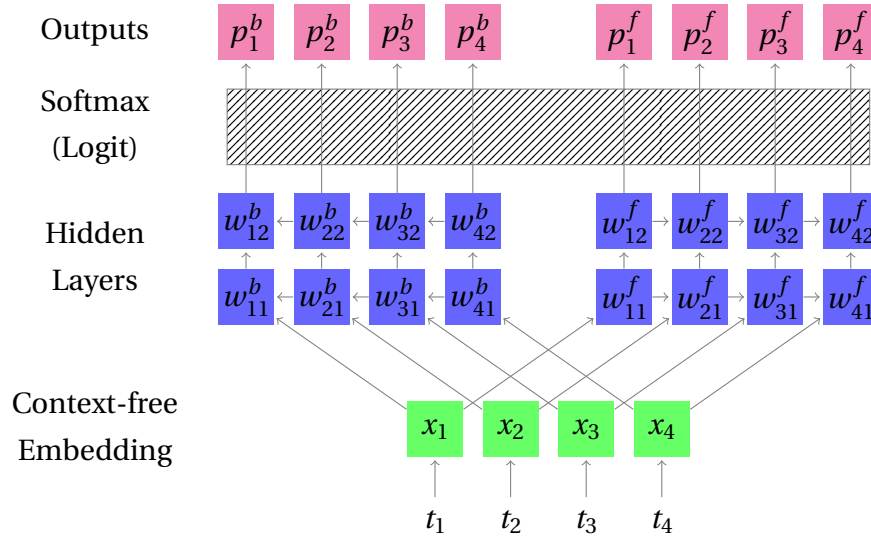


FIGURE 10. ELMO Architecture. This is an ELMO network for a string of 4 words, with  $L = 2$  hidden layers. Here, the softmax layer (multinomial logit) is a single function mapping each input in  $\mathbb{R}^d$  to a probability distribution over the dictionary  $\Sigma$ .

The embedding for the sentence (or product description) is produced by summing the embeddings for each individual word. The weights  $\gamma$  and  $\bar{\gamma}$  can be tuned by the neural network performing the final task. In principle, however, the whole network can be merged with the network performing the final task and jointly optimized.

**A.2. Second generation: BERT.** Bidirectional Encoder Representations from Transformers (BERT) is another contextualized word embedding learned from deep language model (Devlin et al, 2018). It is a successor of ELMO and achieved state-of-art results on multiple NLP tasks, improving somewhat on ELMO. Instead of using Recurrent Neural Network as in ELMO, BERT uses the Transformer structure with attention mechanism (Vaswani et al., 2017) that pays attention to whole sentence or context.

Unlike the language model in ELMO which predicts the next word from previous words, the BERT model is trained on two self-supervised tasks simultaneously:

- Mask Language Model: randomly mask certain percentage of the words in the sequence and predict the masked words

- Next Sentence Prediction: given a pair of sentences, predict whether one sentence proceeds another.

The structure of the BERT model is as follows: (1) Each word in the input sentence is broken to standardized word fragments called “tokens.” (2) For each token, its input representation consists of i) a binary encoding of the token from (1), and ii) a position embedding indicating the position of the token in the sentence (3) The input representation of tokens in the sequence is fed into the main model architecture:  $L$  layers of Transformer-Encoder blocks. Each block consists of a multi-head attention layer, followed by a feed forward layer.

To initially train the network, a special token [cls] is added to the beginning of the sequence, and a fraction of the tokens are replaced by [mask] *at random*. The output representation of the mask token [mask] is used to predict the masked word via a softmax layer, and the output representation of the special [cls] token is used for next sentence prediction. The loss function is a combination of the two losses. Once trained, the final hidden layer (excluding the softmax layer used in prediction) is used as the sequence of word embeddings.

We next focus in detail on the main structure in step (3), especially the “multi-head attention” layer.

*A.2.1. Computing The Attention.* We begin with  $n$  word embeddings  $(x_1, x_2, \dots, x_n)$ , with each  $x_k \in \mathbb{R}^d$ . Let  $X$  denote the matrix whose  $k$ th row is  $x_k$ . The Multi-Head Attention mapping is applied on  $X$  directly:

$$X \mapsto \text{MultiHead}(X, X, X),$$

where

$$\text{MultiHead}(Q, K, V) = \text{Concatenate}(\text{Head}_1, \dots, \text{Head}_h) \omega^O,$$

$$\text{Head}_i = \text{Attention}(Q \omega_i^Q, K \omega_i^K, V \omega_i^V),$$

$$\text{Attention}(\tilde{Q}, \tilde{K}, \tilde{V}) = \text{softmax}\left(\tilde{Q} \tilde{K}^T / \sqrt{d_k}\right) \tilde{V},$$

where  $\omega^O$  and  $(\omega_i^Q, \omega_i^K, \omega_i^V)$  are matrix parameters, which are optimized to maximize the model performance. In other words, each word embedding is replaced by a weighted average of embeddings for all other words, and the weights are learned from the scaled

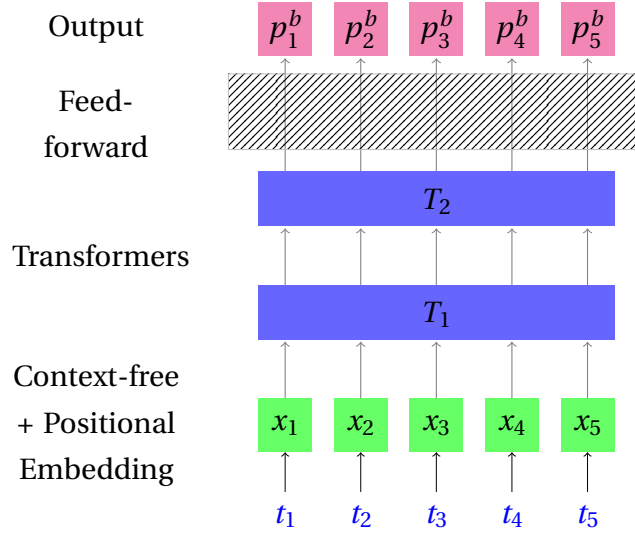


FIGURE 11. BERT Architecture

dot-product of different projections of the word embeddings themselves. The projection matrices are parameters learned during training.

**A.2.2. Generating product embeddings.** Depending on specific tasks and resources, Devlin et al. (2018) suggested using the BERT embeddings in various ways: 1) use the last layer, second-to-last layer or concatenate last 4 layers of the encoder outputs from the pre-trained BERT model, 2) fine tune the whole BERT model on the downstream task, or 3) train the BERT language model from scratch on the new data. For this study, we choose the feature-based approach and extract the second-to-last layer as embeddings from the pre-trained BERT model. Each product's text embedding is the average of the embeddings of each word/token from the input text field.

**A.3. Comparing ELMO and BERT.** While ELMO and BERT both represent recent breakthroughs in natural language processing, the former was an initial, highly effective, contextual word embedding model trained using deep learning, while the latter was the first contextual word embedding to fully exploit the Transformer architecture. Since the BERT paper was published second and could respond directly to the ELMO paper (but not vice-versa), its comparisons may be somewhat biased towards BERT.

There are several technical differences between the two proposals: (1) They use different initial context-free embeddings. ELMO applies an initial convolutional layer to



a *character* embedding, while BERT augments a binary encoding of sub-words (i.e., tokens) with positional data. (2) ELMO is based on the recurrent neural network architecture depicted in Figure 10 while BERT is based on Transformer architecture in Figure 11. (3) The ELMO implementation only allows the averaging weights to be fine-tuned, whereas BERT proposes fine-tuning the whole network. Of these, the biggest difference lies in the choice of the model architecture. The transformer architecture appears to be more effective than the RNN at capturing long-range dependencies in text. Furthermore, ELMO creates context by using the left-to-right and right-to-left language model representations, while in BERT models the entire context simultaneously.

## APPENDIX B. OVERVIEW OF IMAGE EMBEDDINGS VIA RESNET-50

The central idea of the ResNet is to exploit “partial linearity”: traditional nonlinearly generated neurons are combined (or added together) with the previous layer of neurons. More specifically, a building block is to take a standard feed-forward convolutional neural network and add skip connections that bypass two (or one or several) convolution layers at a time. Each skipping step generates a residual block in which the convolution layers predict a residual. Formally each  $k$ -th residual block is a neural network mapping

$$v \longmapsto (v, \sigma_k^0(\omega_k^0 v)) \longmapsto (v, \sigma_k^1 \circ \omega_k^1 \sigma_k^0(\omega_k^0 v)) \longmapsto v + \sigma_k^1 \circ \omega_k^1 \sigma_k^0(\omega_k^0 v),$$

where  $\omega$ ’s are matrix-valued parameters or “weights”. This can be seen as a special case of general design pattern. Putting together many blocks like these sequentially, we obtain the overall architecture depicted in Figure 12.

Convolutional Neural Networks (CNNs) are a type of deep learning model especially effective for analyzing visual imagery. They work by using convolutional layers that apply filters to input data, capturing spatial hierarchies and patterns like edges, textures, and shapes. CNNs efficiently learn features from images, making them widely used in tasks such as image classification, object detection, and facial recognition.

Initial prototypes of deep CNNs (those with many layers of neurons) were challenging to optimize: at a certain depth, additional layers often resulted in much higher validation and training errors. The residual network architecture addressed this by using the residual block architecture outlined above. This allowed for effective optimization even for very deep networks.

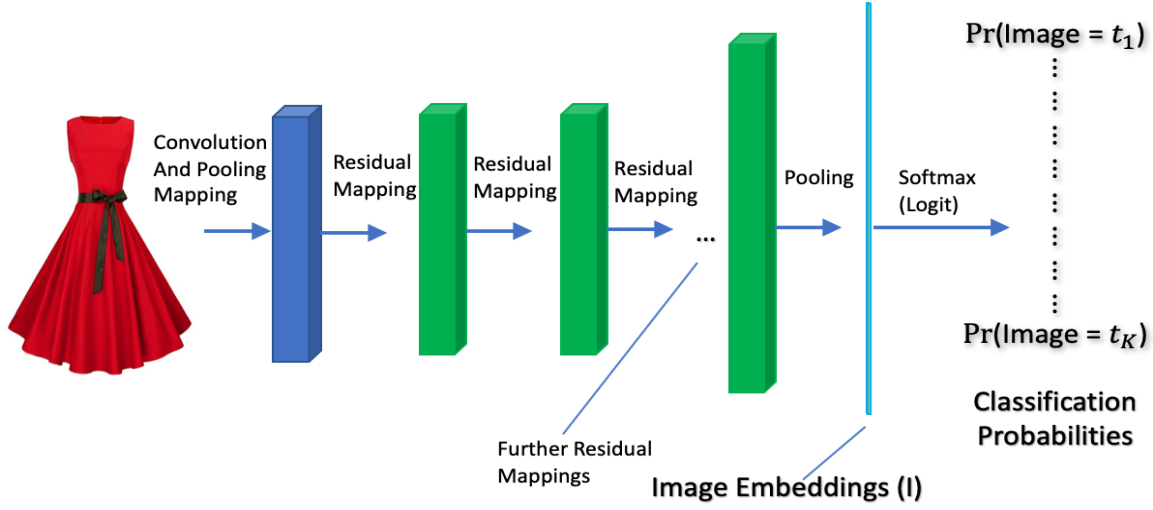


FIGURE 12. The ResNet-50 operates on numerical 3-dimensional arrays representing images. It first does some early processing by applying convolutional and pooling filters, then it applies many residual block mappings, producing arrays shown in green. The penultimate layer produces a high-dimensional vector  $I$ , the image embedding, which is then used to predict the image type.

Just like with text embeddings, we are not interested in the final predictions of these networks but rather in the last hidden layer, which is taken to be the image embedding.

#### APPENDIX C. DETAILS OF DIFFERENT ARCHITECTURES FOR PREDICTION USING EMBEDDINGS

Here we summarize different configurations of neural networks that we considered.

**C.1. Model 1: ELMO + Single Task Models.** The first neural network model we tried is the Single Task model, i.e. predicting product prices one period at a time. For text, we used pre-trained ELMO embedding. For images, we used pre-trained ResNet 50 embedding. The structure is shown in Figure 13. This neural network takes the pre-computed ELMO text embedding (of dimension 256) and the pre-computed ResNet-50 image embedding (of dimension 2048) as input, transforms through 1 to 3 fully connected hidden layers with dropout, and a final linear layer maps the last hidden layer of neurons to the

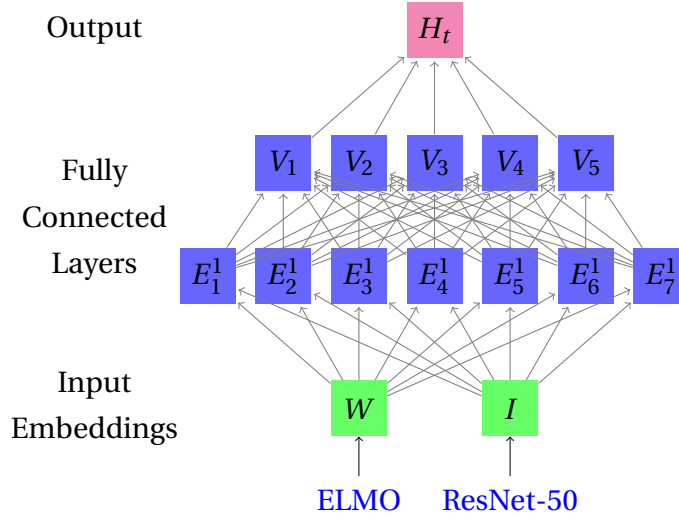


FIGURE 13. SingleTask + ELMO model. Product text is mapped to  $W$  and the image is mapped to  $I$  of dimensions 256 and 2048. For illustration purposes, we only show two hidden layers with dimensions 7 and 5 respectively. In practice, we use three layers with dimensions 2048, 1024, and 256. The output is price for one time period  $t$ .

one-dimensional output, which is the predicted hedonic price. We trained one model for each time period, so in total there are  $T$  neural networks for  $T$  time period.

**C.2. Model 2: BERT + Multitask model.** The second neural network model that we tested is the Multitask NN with pre-trained BERT embeddings. The BERT embeddings are precomputed from a multi-lingual BERT model trained by Google. The input is ResNet-50 image embeddings (of dimension 2048) and concatenated BERT sentence embeddings for the title, brand, description, and bullet points (of dimension  $768 \cdot 4 = 3072$ ). The multitask NN has 1 to 3 dense layers and the output is a  $T$  dimensional vector, which represents the hedonic price for each of the  $T$  time periods.

**C.3. Fine-Tuned BERT + Multitask model.** In the last experiment, we used the end-to-end training framework to fine-tune a BERT model for hedonic price prediction. The model takes raw product text as input, tokenized using the WordPiece tokenizer and truncated / padded to a maximum sequence length  $d_S$ , and run through a BERT base model which consists of 12 transformer blocks. Then the sequence output (of dimension  $d_S \times d_T$ ) from the transformer blocks is aggregated through a Global Average Pooling

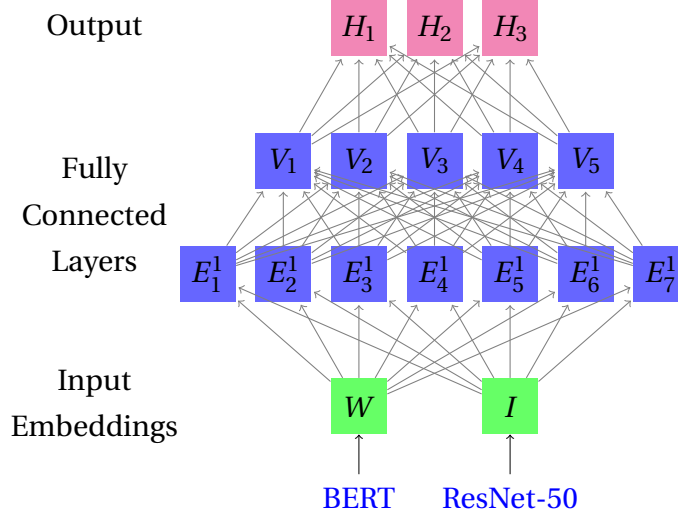


FIGURE 14. MultiTask + BERT. Product text is mapped to  $W$  and image is mapped to  $I$  of dimensions 3072 and 2048 respectively. For illustration purpose, we only show two hidden layers with dimension 7 and 5 respectively, and output is of dimension 3. In practice, we use three layers with dimension 2048, 1024, and 256, and the output is a price vector over  $T = 72$  time periods.

layer to product embedding (of dimension  $d_T$ ). Then the product embedding is linearly mapped to the output which is  $T$ -dimensional hedonic price vector for  $T$  time periods. In our experiment, we use  $d_S = 512$  and  $d_T = 768$ .

The loss function is the same as in Model 2, which combines the weighted squared error term and a regularization term that controls volatility. The weights from all or some layers of the transformer blocks are fine-tuned for the pricing task. Figure 15 is a simple illustration of the end-to-end BERT + Multitask model structure.

We have experimented with different numbers of fine-tuned layers. Results show that fine-tuning more layers improve model performance by a large margin, but it also takes much longer to train. This part of the work is not included in this paper, and we are continuing to explore this research direction.

## APPENDIX D. ADDITIONAL DETAILS

**D.1. Identification of average marginal willingness to pay via the hedonic price function.** We briefly sketch identification of the average derivative of a structural function

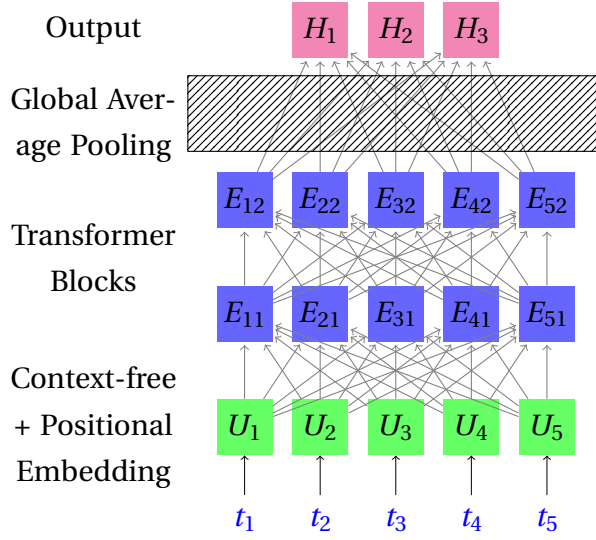


FIGURE 15. MultiTask + Fine-tuned BERT. Product text (sentence) is tokenized and padded to  $X$ , where components represent the context-free input embedding plus a positional encoding for a token (word). Then the input is fed into a BERT model which consists of 12 layers of transformer blocks and outputs  $T$  dimensional price vector. For illustration purposes, we only show 5 tokens and two transformer blocks.

under suitable assumptions. For illustration, we assume with some loss of generality that unobservable product characteristics  $U$  are independent of observable characteristics  $X$  at equilibrium.

This argument supports our characterization (made in Section 2) that the average partial derivative of the hedonic price function with respect to the equilibrium distribution of observed and unobserved characteristics is equal to the average marginal willingness to pay for a particular characteristic.

In the hedonic equilibrium (see e.g. Ekeland et al., 2004), quantities  $q_t(x, u)$  and prices  $h_t(x, u)$  are well-defined as functions of observed characteristics  $x \in \mathcal{X}$  and unobserved characteristics  $u \in \mathcal{U}$ , for  $\mathcal{X}, \mathcal{U} \subset \mathbb{R}^d$ . Moreover, at equilibrium, the distribution of characteristics  $(X, U) \sim F_t$  is fixed. We denote by  $F_t^X$  and  $F_t^U$  the associated marginal distributions.

In order to make the sketch, we assume independence of  $U$  and  $X$  under the distribution  $F_t$ . We further assume that the function  $h_t$  is bounded and that the functions  $\{\nabla_x h_t(x, u)\}_{u \in \mathcal{U}}$  are equicontinuous.

By independence of  $X$  and  $U$  under  $F_t$ , the average structural function

$$h_t(x) = \int h_t(x, u) dF_t^U(u)$$

coincides with the hedonic regression function, in the sense that  $h_t(X) = \mathbb{E}_{F_t}[h_t(X, U)|X]$  almost surely. To prove this claim, let  $\psi$  denote any bounded,  $\sigma(X)$ -measurable test function, and note that by independence the joint distribution induced by  $F_t$  coincides with that of the product measure  $F_t^X \otimes F_t^U$ . Thus,

$$\begin{aligned} \mathbb{E}_{F_t}[\psi(X)h_t(X, U)] &= \int h_t(x, u)\psi(x) dF_t(x, u) \\ &= \int h_t(x, u)\psi(x) d\{F_t^X \otimes F_t^U\}(x, u). \end{aligned}$$

By Fubini's theorem, using boundedness of  $\psi$  and  $h_t$ , this is the same as

$$\begin{aligned} &= \int \left[ \int h_t(x, u) dF_t^U(u) \right] \psi(x) dF_t^X(x) \\ &= \int h_t(x)\psi(x) dF_t^X(x) = \mathbb{E}_{F_t}[h_t(X)\psi(X)]. \end{aligned}$$

By the definition of the conditional expectation, we conclude that

$$h_t(X) = \mathbb{E}_{F_t}[h_t(X, U)|X].$$

Under standard conditions (say, if the functions  $\{\nabla_x h_t(x, u)\}_{u \in \mathcal{U}}$  are equicontinuous), we may then differentiate under the integral to obtain

$$\frac{\partial}{\partial x_k} h_t(x) = \frac{\partial}{\partial x_k} \int h_t(x, u) dF_t^U(u) = \int \frac{\partial}{\partial x_k} h_t(x, u) dF_t^U(u).$$

Thus, after integrating both sides with respect to  $dF_t^X$ , we find that the average derivative of the hedonic price function (left-hand side) coincides with the average marginal willingness to pay for characteristics  $x_k$  (right-hand side).

**D.2. Impact of product variations.** In Amazon's data, products are distinguished by their so-called "Amazon standard identification number," or ASIN. These numbers are assigned at the finest-possible granularity; for example, multiple editions of the same book are assigned different ASINs, as are different colors or sizes of the same jacket.

Such “child ASINs” are then grouped together into “parent ASINs” which include every size and color of a given jacket, or every edition of a particular book.

As we discussed in Section 5.1.1 in the main text, we define a distinct product for each “child ASIN.” This seems sensible for two reasons. For one, in the apparel sector, different sizes and colors can often sell for different prices. For another, this choice gives the neural network freedom to capture important price differences between similar products when data are sufficiently rich, while preserving the possibility of lumping together similar products when fewer transactions are available. Such trade-offs will naturally occur due to regularization of the network’s weights.

In principle, the low out-of-sample error we report could be driven by good performance for such “sibling” products that happen to be split across training and test datasets. We first remark that good performance in these cases is still helpful, since product entry and exit often occurs for some colors/sizes and not others. Fortunately, Amazon tabulates “sibling” products (e.g. variations in size and color, or newer versions of existing products). We found that of the roughly 2.6 million products in the test dataset, only 2.93% (roughly 60 thousand) of those products had a relative in the training dataset, limiting the potential impact of these examples on the reported accuracy.

As a final check, we plotted the difference in  $R^2$  within each period when all products in the test dataset which have a relative in the training dataset are removed from consideration. We found that the  $R^2$  was reduced by a negligible amount, on the order of 0.001. This is detailed in Figure 16.

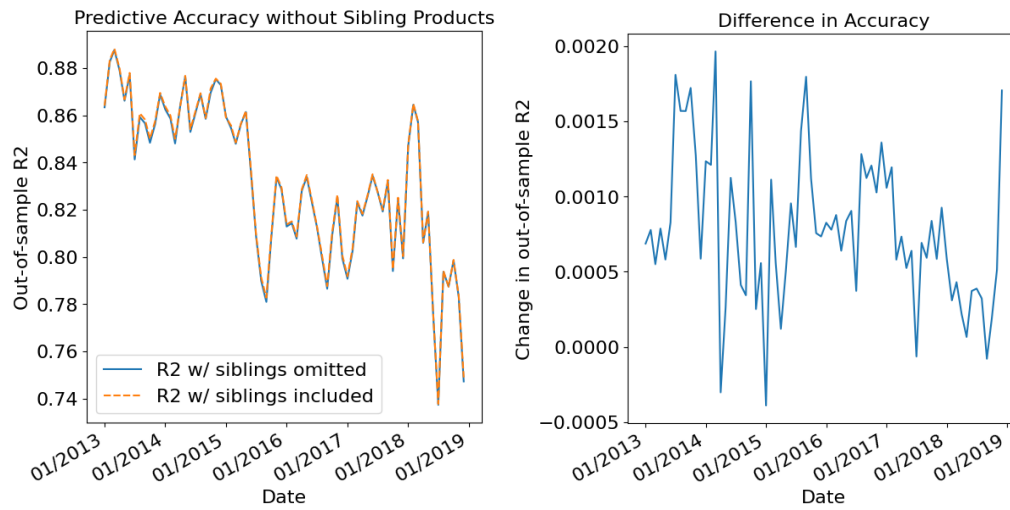


FIGURE 16. Predictive accuracy with and without sibling products. On the left, we have plotted the  $R^2$  obtained both with (orange, dashed) and without (blue, solid) including products that have a sibling in the training dataset (e.g., a variation in size or color, or a newer version). The lines differ imperceptibly—the difference between the two is plotted on the right hand side; it hovers around 0.001.