To study the attributes affecting the pricing of shoes

Kushal Motwani
Chennai Mathematical Institute

December 20, 2020

1 Introduction

Shoes are an essential daily wear. A basic pair helps us protect our feet and go around comfortably. And the advent of various brands, availability of shoes in various styles and designs as per our requirements and comfort level, has gives us a lot of choices. Online shopping adds a cherry on top exposing us to large collections, buyer reviews, etc.
Shoes have an incredible history. Anthropological research has revealed that humans may have started wearing shoes some 40,000 years ago. The oldest leather shoe found has been dated to be about 5500 years old, which gives you a fair idea of how old footwear is. However, shoes have greatly changed in the present modern times. Nike is now making auto-lacing shoes that can be laced from your smartphone and other fancy stuff!

As one of the famous saying goes,*"One can judge a person by the shoes he/she wears"*. Shoes are assumed to be representative of our personalities. A lot of thought, qualitatively and calculatedly go into designing and manufacturing as well as buying shoes. Here we try to assess how visible attributes constitute to shoe pricing as they act as an important parameter in a purchase.

It shall be interesting to understand how shoe styles, brand name, shoe colour/size, reviews and attractive sale offers on e-commerce sites affect the shoe pricing and sales. Comparison of prices over different shoe styles, brands, etc. shall supply us with obvious or intriguing inferences? Do people obsess over brands or colours? Are either of men or women shoes overpriced? Are classics or fashion trends more in demand?

2 Database Description

The study is based on shoe price datasets sourced from https://data.world/ provided by Datafiniti's Product Database. The 2 datasets consist of price data of 10000 unique men's and women's shoes at which they were sold respectively. The dataset includes shoe name, brand, price, colour, size, shoe style, review on e-commerce sites and more. Each shoe has an entry for each price found for it and some shoes may have multiple entries if sold in different prices.

Women's Shoes raw dataset consists of 33801 rows (sample points) and 52 columns. And Men's shoes dataset consists of 19315 rows and 48 columns. Following columns are present in the women's shoe dataset:

```
'id', 'asins', 'brand', 'categories', 'colors', 'count', 'dateAdded',
'dateUpdated', 'descriptions', 'dimension', 'ean', 'features',
'flavors', 'imageURLs', 'isbn', 'keys', 'manufacturer',
'manufacturerNumber', 'merchants', 'name', 'prices.amountMin',
'prices.amountMax', 'prices.availability', 'prices.color',
'prices.condition', 'prices.count', 'prices.currency',
'prices.dateAdded', 'prices.dateSeen', 'prices.flavor', 'prices.isSale'
'prices.merchant', 'prices.offer', 'prices.returnPolicy',
'prices.shipping', 'prices.size', 'prices.source', 'prices.sourceURLs',
'prices.warranty', 'quantities', 'reviews', 'sizes', 'skus',
'sourceURLs', 'upc', 'vin', 'websiteIDs', 'weight', 'Unnamed: 48',
'Unnamed: 49', 'Unnamed: 50', 'Unnamed: 51'],
```

The men's shoes dataset consists of same columns except the last four i.e. 'Unnamed' columns.

Since the data is directly scrapped from the e-commerce sites such as www.overstock.com there are no. of missing values spread over columns. As part of primary data cleaning, columns with more than 50% of missing values have been removed. Columns such as product description, manufacturer number, image urls, source urls, skus (id nos. of shoe specific to source urls) have also been dropped.

Some columns such as 'categories' has a list of comma separated values and 'features' column similarly consists of various key value pair of features embedded in a single column. Various information regarding shoe style, colour, etc. is extracted from these columns to form separate variables/columns. And average price column 'price.amountavg' has been added to get the average price of the shoes taking average from 'price.amountmin' and 'price.amountmax' column.

Dataset description post cleaning is as in Figure 1. Various feature variables such as Shoe type - Boots, Heels, Sandals, etc. were not present in scrapped men's shoe dataset emphasizing the fact these many styles aren't present in men's shoes or are not given much importance by men's or boys.

```
Data columns (total 14 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   id                18345 non-null  object
 1   brand             18345 non-null  object
 2   colors            18345 non-null  object
 3   merchants         13044 non-null  object
 4   prices.amountmin  18345 non-null  float64
 5   prices.amountmax  18345 non-null  float64
 6   prices.issale     18345 non-null  bool
 7   prices.merchant   13222 non-null  object
 8   brand_clean       18345 non-null  object
 9   prices.amountavg  18345 non-null  float64
 10  price.stat        18345 non-null  int64
 11  brand.count       18345 non-null  int64
 12  price_band        18345 non-null  int64
 13  Date              18345 non-null  object
```

(a) For Men's Shoes

```
Data columns (total 20 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   id                25976 non-null  object
 1   brand             25976 non-null  object
 2   colors            25976 non-null  object
 3   prices.amountmin  25976 non-null  float64
 4   prices.amountmax  25976 non-null  float64
 5   prices.issale     25976 non-null  bool
 6   prices.merchant   22859 non-null  object
 7   brand_clean       25976 non-null  object
 8   prices.amountavg  25976 non-null  float64
 9   price.stat        25976 non-null  bool
 10  isBoots           25976 non-null  int64
 11  isBooties         25976 non-null  int64
 12  isSandal          25976 non-null  int64
 13  isAthletic        25976 non-null  int64
 14  isCasual          25976 non-null  int64
 15  isDesigner        25976 non-null  int64
 16  isLeather         25976 non-null  int64
 17  isOther           25976 non-null  int64
 18  datetime          25976 non-null  datetime64[ns, UTC]
 19  Date              25976 non-null  datetime64[ns]
```

(b) For Women's Shoes

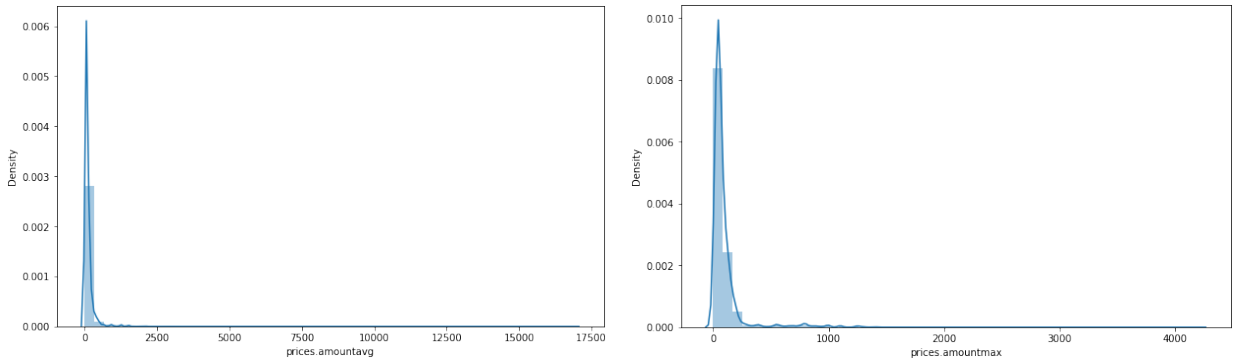
Figure 1: Dataset Column Description

3 Problem Statement

The study aims to explore the insights from men and women's shoe price data-set – the trend and differences among men and women shoe prices, determine attributes affecting the prices, and fitting an appropriate regression model to predict prices based on features significantly affecting the prices. We shall also explore the trend and seasonality affects over time for 2014 - 2017 data, validate our findings over 2018 data and finally test our analysis and fitted model on 2019 data. The plan is to check for regression assumptions over the significant features and applying required transformation and fit the appropriate regression model, compare various regression techniques for optimal fit and analyse the results.

4 Exploratory Data Analysis

Starting with the uni-variate analysis of the price data for both men and women shoes. Figure 2 depicts the density plots for men and women observed shoe prices respectively.



(a) For Men's Shoes

(b) For Women's Shoes

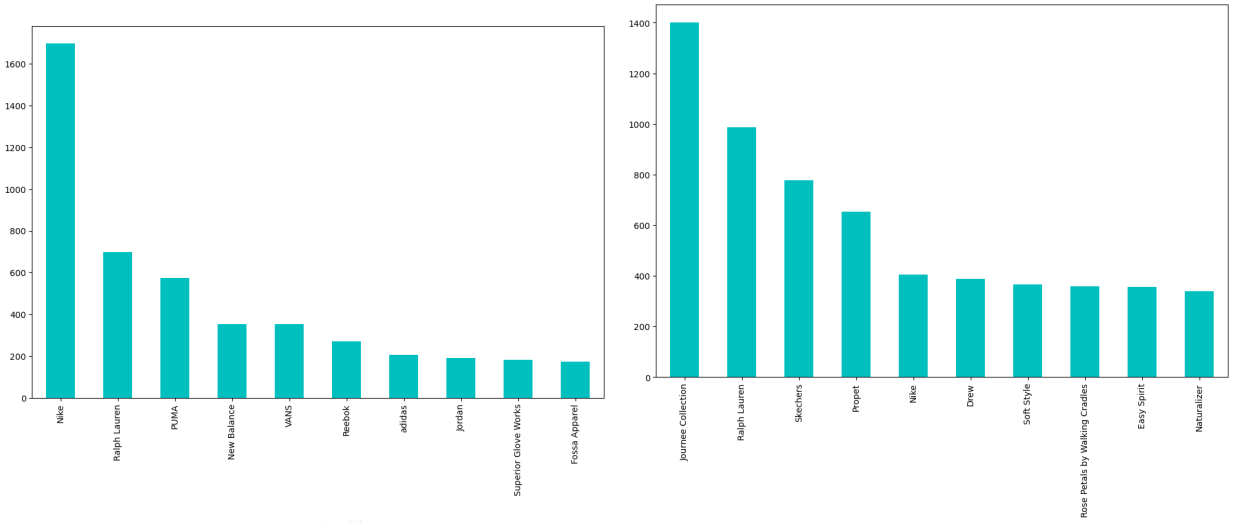
Figure 2: Mean Price Distribution

For both datasets we notice that the price distribution is highly skewed, an interesting point to note is the mean price for men's shoes = $ 113.65 which is more than mean value for women's shoe price = $ 88.30 and the maximum price range also exhibits a stark difference, $ 16949.00 for men (with just 5 samples with price greater than $ 5000) and $ 4270.00 for women despite the common trend that women tend to spend more on clothing and accessories.

Brands' name at times plays an important role in customer's choice and commodity value as well. The data includes shoes data across 1853 and 1116 brands for men and women shoes respectively. 23 and 67 brands respectively have more than 100 shoes in the datasets. Figure 3 shows the Top 10 dominant brands in Shoes category.

Checking on if these brands do affect the price significantly, and plotting them as per their mean price, we notice that there is approx. 0.7 correlation between the brand and the price value for shoes associated with them. *'jewelobsession'* and *'diamond wish'* brands sell the most expensive shoes. These brands are associated with stone studded handbags, shoes and accessories. **Gucci, Prada, Burberry, Ralph Lauren,** etc. are the most expensive ones whereas the ones with least price values are not much heard of and have very less product count in the scrapped sample.

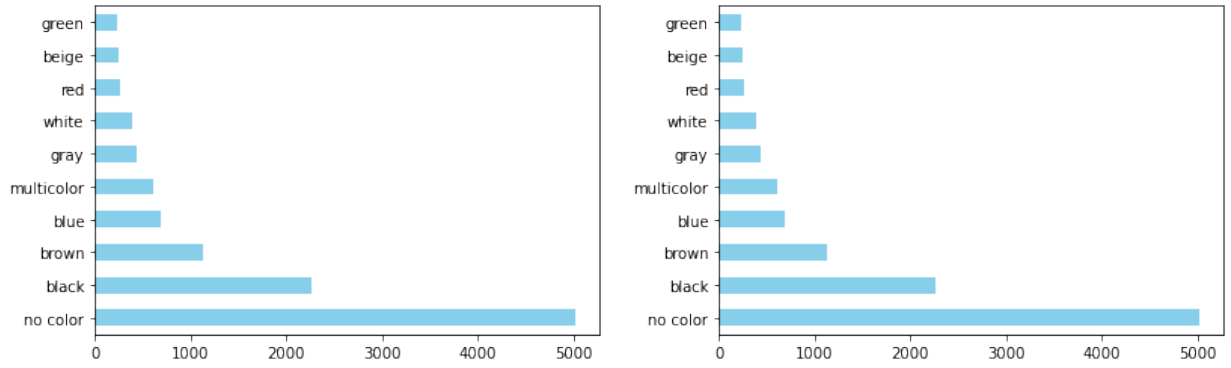
Diving more into the colour distribution for both datasets, we notice that *'black'* is the most popular among both of them. Classic colours such as *'brown', 'blue', 'grey',* etc. is preferred more than others. High price range shoes had *'Blue', 'Silver'* colours and *'Black', 'crystal'* for Men and Women footwear respectively.



(a) For Men's Shoes

(b) For Women's Shoes

Figure 3: Top 10 Brands by count



(a) For Men's Shoes

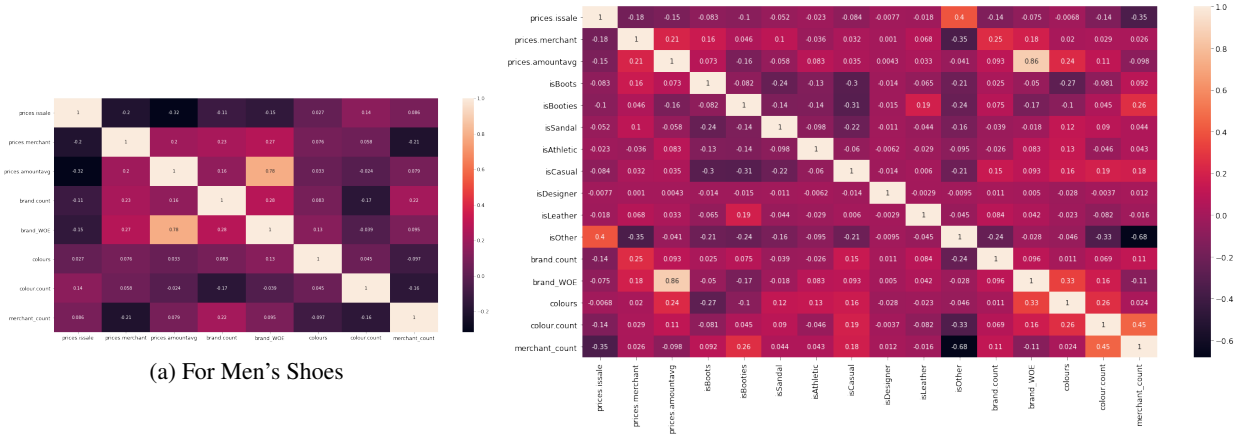
(b) For Women's Shoes

Figure 4: 10 Most Frequent Colours

Prior to modelling, we also look at the correlation plots for both datasets in order to eliminate columns with high correlation among themselves to avoid multicollinearity issues as well as to check if we have any spurious relation among independent columns. Figure 5 represents the correlation plots for respective datasets. We observe that there is no inter-correlation among the independent variables and brands have maximum correlation with average shoe prices.

Various other points to consider are the following:

- 30% and 14.65% of listed shoes were on sale when sold in both datasets respectively. For obvious reasons, mean price for shoes on sale is fairly lesser than their counterparts.
- There is no significant trend or inflation observed in shoe prices over time. This implies shoe prices aren't affected much over a short time span of 2-3 years, there might be seasonal affect over sales but prices of a particular pair doesn't change noticeably over short span of time.
- Styled shoes such as Boots, Leather, Designer ones are expensive than the Casual ones in case of Women Shoes.Work and Athletic shoes are fairly priced.



(a) For Men's Shoes

(b) For Women's Shoes

Figure 5: Correlation Matrix

- Shoes with same unique id are listed for approximately same prices over different merchant sites implying that merchant sites don't have a significant affect on price determination, though they control the 'Sale' factor, whether to put the shoes on sale or not.

Keeping these exploratory analysis in mind, we will move further in building price prediction model for a pair of shoes.

5 Data Preparation for Model Fitting

Since we have certain categorical features in the data, first step towards modelling for pricing is to encode these via appropriate label encoding approach.

5.1 Label Encoding for Brands

Since we have 1116 different brand names in the column, one-hot encoding would lead to large sparse matrix not much feasible for efficiency and time consumption in model compilation, general label encoding is not a preferable as it is in general used for ordinal data and our column 'brand' data is nominal. Thus, we have opted for **Weight Of Evidence (WOE) Encoding/Transformation.**

The goal of such transformation is to get the maximum difference among the categories relating to the target variable, and then assigns a numeric value to each of the categories. In this transformation the information of the target variable has been utilized. Here, the WOE for each of the unique brand names is calculated based on the weight calculated by the following formula

$$WOE = ln(\frac{\%price}{\%brand}) * 100$$

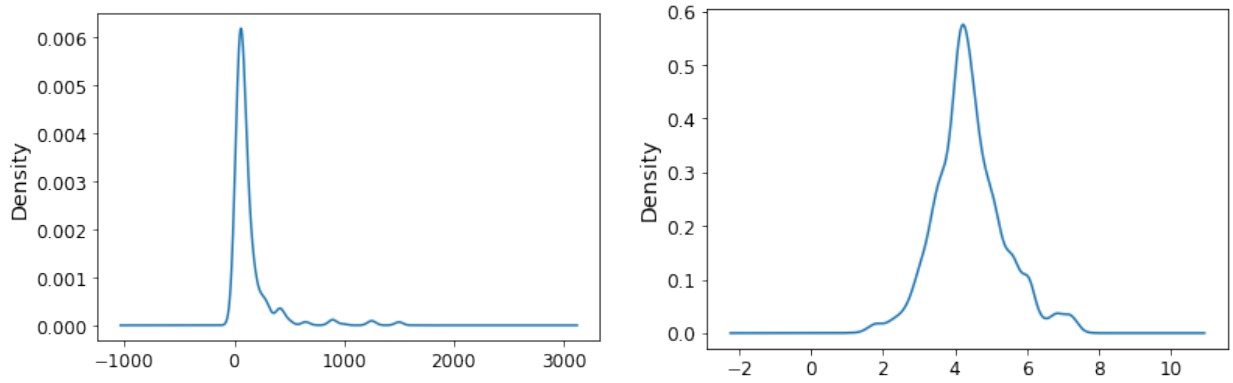
5.2 Label Encoding for Colours and Merchants

Several samples have multiple/shades of colours listed in the column, for modelling purposes, we have opted only for the first colour listed, i.e. the most dominant colour on the shoe. Both of the variables were transformed via standard label encoding.

5.3 Target Variable - 'prices.amountavg'

Since the distribution of shoe prices is highly skewed. We use log transform of the target variable for regression modelling. This is done keeping in mind the linear regression assumptions. Figure 6 illustrates the density

plot of transformed target variable for men shoe data. The kurtosis and skewness of the actual distribution are 20.732729747821107 and 4.193393655107598 respectively while for the transformed distribution, we have kurtosis = 0.7796208735702432 and skewness = 0.3977371036155497.Similar plots are obtained for women shoe data as well. Removing outliers could be an option but right now we don't want to exclude the modelling flexibility that shall include predicting high prices for shoes with exceptional features.



(a) Actual Distribution

(b) Transformed Distribution

Figure 6: For Men's Shoe Data

6 Regression Modelling

We have modelled both men and women shoe data separately due to difference in independent parameters obtained. After finishing the necessary encoding of categorical variables, standard scaling of independent feature columns, initial checking for multicollinearity via Variance Inflation Factor (VIF) and required transformation of dependent variable. We proceed to fit a Ordinary Linear Regression model to the train data. Additionally we have also tried to check other regression models such as decision trees and support vector regression which does not make any assumptions regarding the target parameter distribution.

6.1 Ordinary Linear Regression - OLS Model

Fitting the training data to linear regression, and predicting on test data we have the results as depicted in Figure 7. Several variables were removed over the course of fit if they had significant p-value. Adjusted R-squared value for women shoe price regression fit = 0.82. This could be due to more explainable features present in the data. The linear regression fit does a fair job fitting the data.

Checking for predicted value plots, we observe from Figure 8, that most outlier price values are not captured by the OLS model. This shall be due to violation of the target value distribution assumption. Even after log/box-cos transformations, the resultant distribution fails to follow normal distribution. Double log transformation result in negatively skewed distribution. Figure 9 shows up the residual plots for Men and Women Shoe price prediction model respectively.

Though the residual plots seem random, they do not follow normal distribution in either of the datasets, the same property was tested via Kolmogorov-Smirnov Test done on the residual data. The resultant p-value in both cases is less than 0.01 [KS test Result(statistic=0.2802011539648952, pvalue=0.0) for Men's Shoe Price data]. We have not checked fitting the polynomial features in the linear regression setup as that will will result in same assumption violation while modelling. Using Breush-Pagan test for checking hetroscedasity, we obtained a significant p-value (=

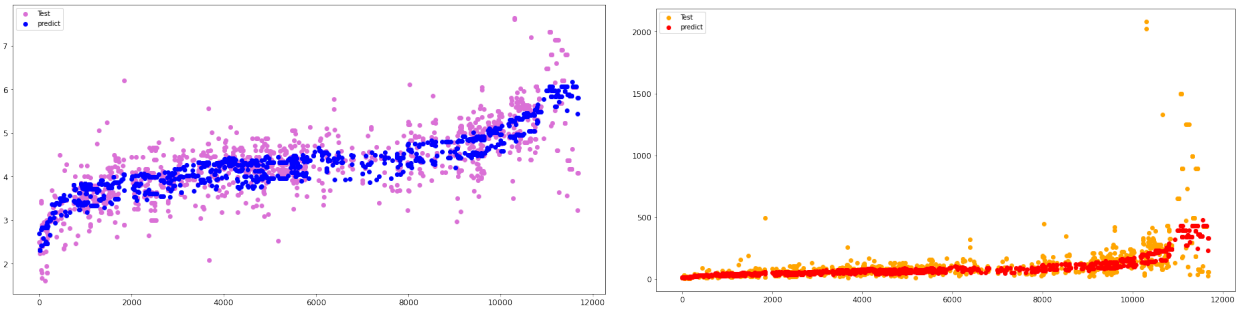
OLS Regression Results

Dep. Variable:	prices.amountavg	R-squared:	0.683
Model:	OLS	Adj. R-squared:	0.683
Method:	Least Squares	F-statistic:	1703.
Date:	Thu, 24 Dec 2020	Prob (F-statistic):	0.00
Time:	00:39:26	Log-Likelihood:	-3843.6
No. Observations:	4750	AIC:	7701.
Df Residuals:	4743	BIC:	7746.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.8612	0.022	221.645	0.000	4.818	4.904
prices.issale	-0.3753	0.018	-21.353	0.000	-0.410	-0.341
prices.merchant	-0.0018	0.001	-3.505	0.000	-0.003	-0.001
brand.count	-0.0001	2.26e-05	-4.655	0.000	-0.000	-6.1e-05
brand_WOE	0.9020	0.010	87.761	0.000	0.882	0.922
colours	-0.0059	0.001	-8.579	0.000	-0.007	-0.005
colour.count	2.779e-05	6.85e-06	4.055	0.000	1.44e-05	4.12e-05

Omnibus:	693.780	Durbin-Watson:	2.008
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1560.870
Skew:	-0.855	Prob(JB):	0.00
Kurtosis:	5.228	Cond. No.	5.62e+03

Figure 7: OLS for Men Shoe price data



(a) Predictions for log(price)

(b) Transformed Predictions over actual distribution

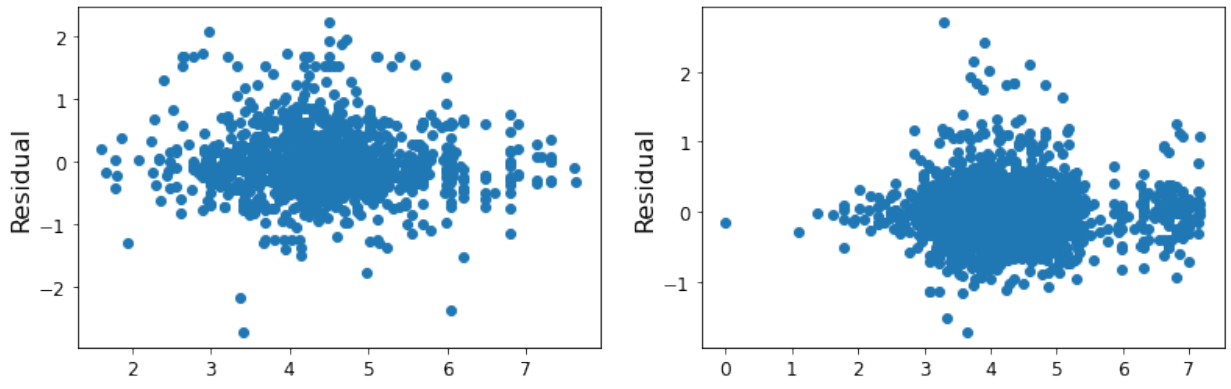
Figure 8: For Men's Shoe Data

0.9) leading to conclusion, that our residual data is homoscedastic.

7 Conclusion

We conclude that the estimated regression coefficients obtained though are BLUE estimators (Best Linear, Unbiased Estimator) for the given dataset regardless of the distribution, they are not the Maximum Likelihood estimates. This is due to the following:

- Least Squares Regression guarantees BLUE Estimation



(a) For Men's Shoe Prices

(b) For Women's Shoe Prices

Figure 9: Residual Plots

- Gauss-Markov Theorem states, 'A normal distribution is only used to show that the estimator is also the maximum likelihood estimator.'

Thus Linear Regression Modelling for the given data shall not be the best regression technique. A better approach could be to use Decision Tree / Support Vector Regression which do not have assumptions regarding the residuals and thus target distribution. (Discussed in Section 8.2). Generalized Linear Regression can also be an option if we develop more idea about the target distribution.

Also comparing the overall trend for Men and Women shoe prices, we notice that although there may be some pricing difference, the overall distribution is similar, and features mentioned in most of listed shoes is limited and on average less than features listed on women shoes.

8 Appendix

8.1 Average Price Time Series for Shoe Dataset

Figure 10 depicts the trend of mean shoe prices for women over time. Since the collected data is not uniformly distributed over time and we do not derive any significant trend from the given sample set, we have not taken date-time of the shoes sold while price prediction modelling in the current approach.

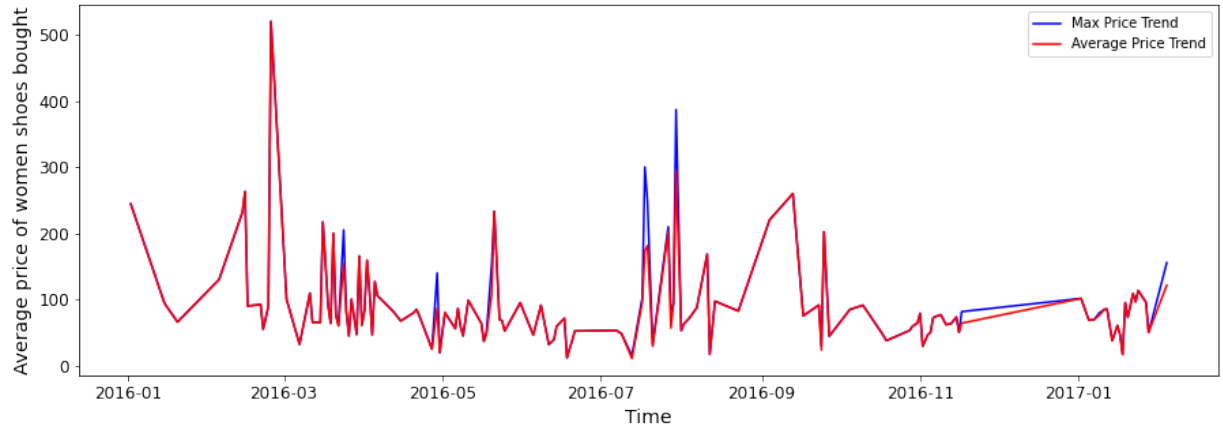
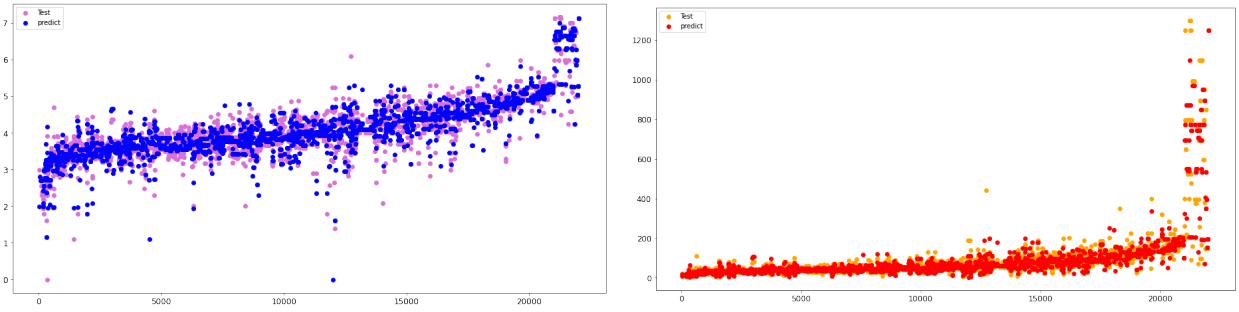


Figure 10: Women Shoe price over time

8.1.1 Decision Tree Regression

Decision Tree Regression works similar to classification analogous. In case of continuous targets variable it divides the whole range into parts while adding a leaf node to the tree. It doesn't result in a straight regression line. Instead, the resultant line shall be more like consisting of steps up and down. Those points of discontinuity indicate division at that place. Decision Tree Regressor fit for the given dataset, resulted in a relatively high R-square value(= 0.875) and a fairly accurate fit for test data as in Figure 11 (for Women's Shoe price).



(a) Predictions for log(price)

(b) Transformed Predictions over actual distribution

Figure 11: For Women's Shoe Data

9 References

1. Men's Shoe Data:
https://data.world/datafiniti/mens-shoe-prices

2. Women's Shoe Data:
https://data.world/datafiniti/womens-shoe-prices

3. Shoes perceived as a persona representative:
https://news.yahoo.com/blogs/sideshow/judge-90-percent-people-personalities-shoes-researchers-192903995.html

4. Label Encoding:
https://towardsdatascience.com/a-data-scientists-toolkit-to-encode-categorical-variables-to-numeric-d17ad9fae03f

5. Least Squares Regression for Distributions which are not Normal
https://stats.stackexchange.com/questions/75054/how-do-i-perform-a-regression-on-non-normal-data-which-remain-non-normal-when-tr