Team 8: Lifestyle, Merchandise & and Electronics Analytics and Price Predictor

December 9, 2022

Chauhan, Rahul
MS Data Science
University of Colorado, Boulder
Rahul.Chauhan@colorado.edu

Sawhney, Shivam
MS Data Science
University of Colorado, Boulder
Shivam.Sawhney@colorado.edu

1 Abstract

[1] Today, the lifestyle market is at the epicenter of Supply Chain transformation. Many fashion and clothing firms have been compelled to change their strategies and processes to meet the international demand of those rapidly growing economies. This opens up a lot of new options for the sector. There are a few projects about future production from prominent fashion businesses. Some companies are aiming to create their products entirely out of recycled materials.

[2] Is it possible to be trendy while remaining basic and high-quality? Yes, but clothing companies must join in and provide better and well-thought-out designs that will last for years and maintain financial margins.

[3] Companies in the electronics industry are constantly competing to implement novel ideas and offer cutting-edge technology to the market first. This puts a lot of pressure on design and engineering teams to develop and manufacture innovative goods and services as quickly and cheaply as possible. Sales and marketing teams are also under pressure to increase sales while maintaining profit margins well beyond production and operational costs. Many electrical companies are also experimenting with and implementing advanced analytics in order to generate insights that can assist improve processes and increase revenues.

Keywords— Sales Analytics, Customer Analytics, Consumer Behavior, Machine Learning, Regression Models, Random Forest, Decision Tree, EDA.

2 Introduction

Our product prioritizes both the customer and the organization. Customers can acquire an estimate of the cost of a laptop if they explain their needs in detail and vice-versa. For example, person A needs a laptop 16in long, is touch screen, and is within $1,500, then they can have an approximate idea which laptops fall in a similar range. Later, if person B needs a gist of similar products in respective price range, they can get an overview of the same as well.

Organizations, on the other hand, can make use of customer analytics for, let's say at what dates of a month, or during what months in a year, is a customer is inclined to purchase, which season during the year is best to launch a new one, which product is likely to skyrocket, and so on.

3 Related Work

The association between specifications and laptop price can be used to forecast future laptop costs and predict current laptop price

range. This is already available for instructional purposes on Kaggle and AnalyticsVidya. However, these projects are just intended to analyze the given statistics and anticipate laptop prices.

The other part of our project – customer analytics – on verticals such as merchandise, lifestyle, fashion, is something which we could not find much w.r.t. relevant and related work.

4 Research & Methodology

There are many models available across the internet for evaluation of our data, however, to find the correct model for our data was a real task, more like finding a needle in haystack. Below are some of the methods and concepts which we have incorporated and explored in our project:

1. Linear Regression: Simple linear regression is useful for finding the relationship between two continuous variables. One is the predictor or independent variable and the other is the response or dependent variable. It looks for a statistical relationship but a not deterministic relationship. A relationship between two variables is said to be deterministic if one variable can be accurately expressed by the other. For example, using temperature in degree Celsius it is possible to predict Fahrenheit accurately. The statistical relationship is not accurate in determining the relationship between two variables. For example, the relationship between height and weight.
2. KNN: The supervised machine learning technique known as the k-nearest neighbors (KNN) can be used to handle classification and regression issues. It is straightforward and simple to implement. Moreover, It is a supervised learning classifier that is non-parametric and employs proximity to classify or anticipate how a given data point will be grouped. It can be applied to classification or regression issues, although it is most frequently employed as a classification technique because it relies on the idea that comparable points can be discovered close to one another.
3. [11] Basic EDA: Before beginning any data-related work, such as machine learning, data analytics, etc., the first step is to view and analyze the data. To ensure that the solutions we came up with for our data- EDA is the best possible way to do the same, since it is extensively helpful in bringing out unknown data insights and trends. A machine learning model, for example, can be made simpler and the cost of computing the model's training can be reduced if you are clear about which input variables have the greatest impact on the response variable.
4. [13] Decision Tree: For classification and regression, decision trees (DTs) are a non-parametric supervised learning technique. By learning straightforward decision rules derived from the data attributes, the objective is to develop a model that predicts the value of a target variable. An approximate piece wise constant can be thought of as a tree.
5. [12] Random Forest: Many decision trees are built during the training phase of the random forests or random decision forests ensemble learning approach, which is used for classification, regression, and other tasks. The class that most of the trees chose is the output of the random forest for classification problems. The mean or average prediction of each individual tree is returned for regression tasks. The tendency of decision trees to overfit their training set is corrected by random decision forests. Although they frequently outperform decision trees, gradient boosted trees are more accurate than random forests. However, their effectiveness may be impacted by data peculiarities.

5 Evaluation & Results

5.1 Laptop Price Predictor

```
In [96]:  1 df.head()
```

Out[96]:

	Unnamed: 0	Company	TypeName	Inches	ScreenResolution	Cpu	Ram	Memory	Gpu	OpSys	Weight	Price
0	0	Apple	Ultrabook	13.3	IPS Panel Retina Display 2560x1600	Intel Core i5 2.3GHz	8GB	128GB SSD	Intel Iris Plus Graphics 640	macOS	1.37kg	71378.6832
1	1	Apple	Ultrabook	13.3	1440x900	Intel Core i5 1.8GHz	8GB	128GB Flash Storage	Intel HD Graphics 6000	macOS	1.34kg	47895.5232
2	2	HP	Notebook	15.6	Full HD 1920x1080	Intel Core i5 7200U 2.5GHz	8GB	256GB SSD	Intel HD Graphics 620	No OS	1.86kg	30636.0000
3	3	Apple	Ultrabook	15.4	IPS Panel Retina Display 2880x1800	Intel Core i7 2.7GHz	16GB	512GB SSD	AMD Radeon Pro 455	macOS	1.83kg	135195.3360
4	4	Apple	Ultrabook	13.3	IPS Panel Retina Display 2560x1600	Intel Core i5 3.1GHz	8GB	256GB SSD	Intel Iris Plus Graphics 650	macOS	1.37kg	96095.8080

Figure 1: Laptop Data set

Figure 1 above shows a brief snapshot of the data set we used for predicting laptop prices. We have further created more attributes from the data at hand - like PPI, Touch screen, and IPS. The most challenging part while working with this dataset was to pick and choose the correct attributes for analysis. It was quite difficult since everyone likes to have their favourite specifications for their "PC". It was hard to generalize.

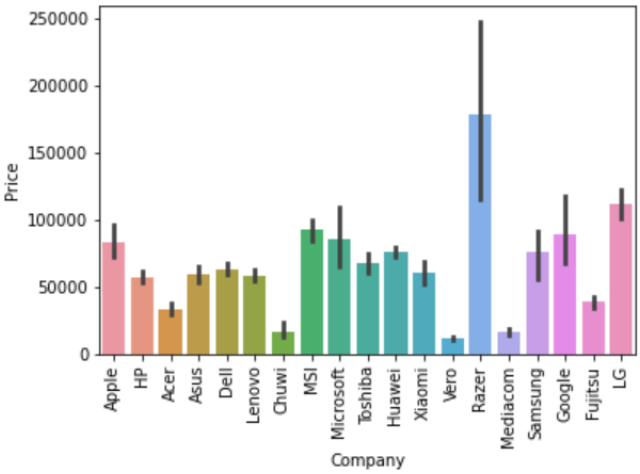


Figure 2: Company-wise Price distribution of Laptops

Once the data set was successfully pre-processed, we decided to perform some EDA on the data-set to uncover trends and insights. The figure 2 is one of the many plots which we generated. The figure above depicts the Price variation in a laptop w.r.t. the brand/company it belongs to. It is not at all surprising to see that the more established and recognized brands in the laptop industry, tend to price their laptops higher than their competition

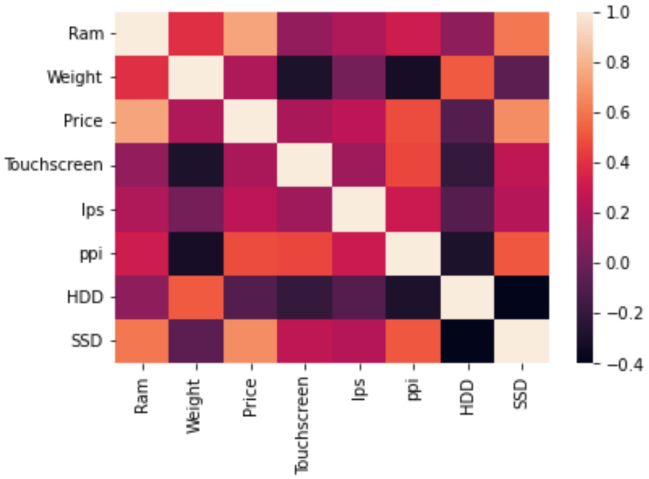


Figure 3: Correlation Matrix

Next, the figure 3 above is a heat map, depicting the correlation of all the attributes of our data set to the "Price" component. As can be seen, most of the attributes are having a positive co-relation except the last two. This correlation matrix is a good measure to detect the dependency amongst variables, and it also helps determine what attributes can be chosen for our model, so that it performs at an optimal level and gives the best results.

Finally, the figure below showcases the performance of our model. We used the R2 score as our evaluation metric. As can be seen, all the models which we built roughly lie around the 0.8-0.9 range. The best performing model was Random Forest, which was not surprising at all! Since the attributes being put into the model were both categorical and numerical in nature.

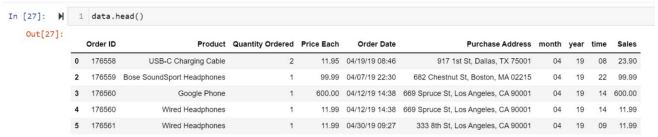
S.No.	Algorithm	R2 Score
1	Linear Regression	0.807328
2	KNN	0.802198
3	Decision Tree	0.830968
4	Random Forest	0.88734

Figure 4: Evaluation Matrix

NOTE - Our first model came back with an R2 score of 0.44. And we finally managed to get a score of 0.88 in the end, hence improving upon

our initial model score by quite a factor!

5.2 Sales Analytics



```
In [27]: 1 data.head()
Out[27]:
```

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address	month	year	time	Sales
0	176558	USB-C Charging Cable	2	11.95	04/19/19 08:46	917 1st St, Dallas, TX 75001	04	19	08	23.90
2	176559	Bose SoundSport Headphones	1	99.99	04/07/19 22:30	682 Chestnut St, Boston, MA 02215	04	19	22	99.99
3	176560	Google Phone	1	600.00	04/12/19 14:38	669 Spruce St, Los Angeles, CA 90001	04	19	14	600.00
4	176560	Wired Headphones	1	11.99	04/12/19 14:38	669 Spruce St, Los Angeles, CA 90001	04	19	14	11.99
5	176561	Wired Headphones	1	11.99	04/30/19 09:27	333 8th St, Los Angeles, CA 90001	04	19	09	11.99

Figure 5: Sales Analytics Data-Set (Electronics)

Talking about the other aspect of our project - Sales Analytics, figure 5 showcases the data-set belonging to the Electronics vertical of the shopping world. This data-set has around 175,000+ rows of data. This was the most challenging part of this aspect of our project. This data-set took huge amount of time to clean and pre-process due to its sheer volume and also due to the fact that the data was not tidy at all!

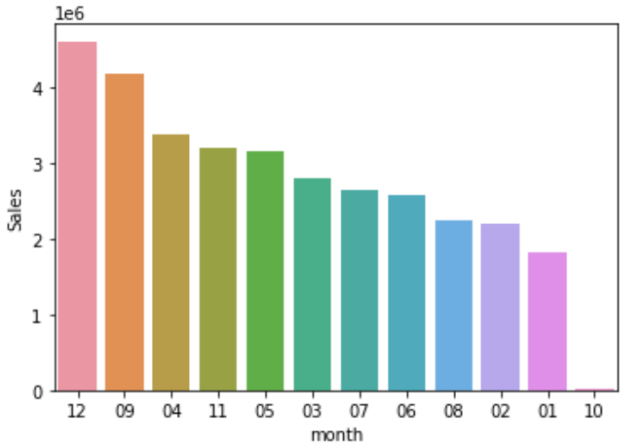


Figure 6: Month Wise Sales Count - Electronics

The figure above shows the count of monthly sales over a year, in the Electronics vertical. It can be seen that the month "12" or December has the most sales of electronic items across a year. This kind of information can be valuable to organizations as this enables to get customer insights and plan their sales and what not, so as to boost profit

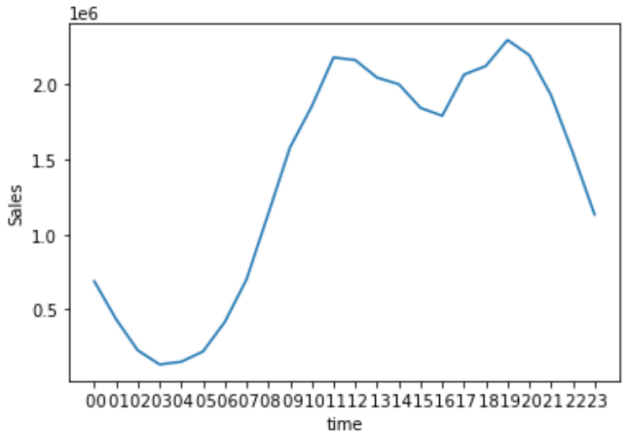
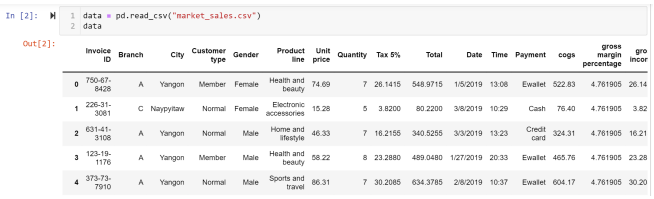


Figure 7: Hourly Analysis of Sales in a day - Electronics

The next figure; Figure 7 is one of the more interesting ones, that we were able to produce. This one highlights the daily sales trend of electronics spread across 24 hours. As can be inferred from the image, there is relative spike in the sales around 11 am in the morning and also around 7 pm in the evening. These kinds of insights can be made use of by companies to strategically plan when to show their ads, so as to maximize sales.



```
In [2]: 1 data = pd.read_csv("market_sales.csv")
        2 data
Out[2]:
```

	Invoice ID	Branch	City	Customer type	Gender	Product line	Unit price	Quantity	Tax 5%	Total	Date	Time	Payment	cogs	gross margin percentage	gro incor
0	750-67-8428	A	Yangon	Member	Female	Health and beauty	74.69	7	26.1415	548.9715	1/5/2019	13:08	Ewallet	522.83	4.761905	26.14
1	226-31-3081	C	Naypyitaw	Normal	Female	Electronic accessories	15.28	5	3.8200	80.2200	3/8/2019	10:29	Cash	76.40	4.761905	3.82
2	631-41-3108	A	Yangon	Normal	Male	Home and lifestyle	46.33	7	16.2155	340.5255	3/3/2019	13:23	Credit card	324.31	4.761905	16.21
3	123-19-1176	A	Yangon	Member	Male	Health and beauty	58.22	8	23.2880	489.0480	1/27/2019	20:33	Ewallet	465.76	4.761905	23.28
4	373-73-7910	A	Yangon	Normal	Male	Sports and travel	86.31	7	30.2085	634.3785	2/8/2019	10:37	Ewallet	604.17	4.761905	30.20

Figure 8: Sales Analytics Data-set (Lifestyle and Merchandise)

Next up, the figure 8 is the data-set for the sales analytics in the lifestyle and merchandise vertical. Re-iterating this again, the same issue we encountered - the sheer volume of the data-set was enough for us to spend hours and hours cleaning the data and getting it ready.

```
In [ ]: 1 sns.relplot(data=the_hours.reset_index(),x='Hours',y='Quantity',col='Product line',kind='line')
Out[ ]: <seaborn.axisgrid.FacetGrid at 0x1bbd7c70190>
```

Figure 9: Hourly Analysis of sales for different Product lines over 24 hours

Figure 9 is similar to Figure 7 - on it being interesting! It represents the hourly sales in day, of different product lines - Fashion, Food, Lifestyle etc. Just like the previous instance of a similar figure, organizations can make use of such insights to optimize advertisement times to maximise sales

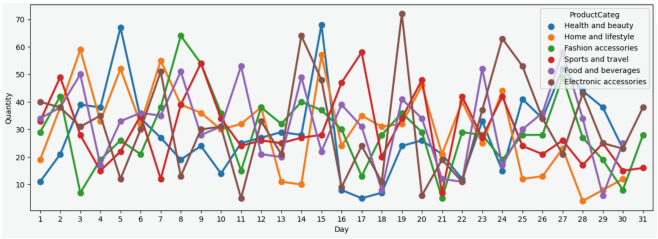


Figure 10: Daily Sales Analysis of Product Categories

The figure above displays the daily sales trends of different product categories over a month. There are sparks for every category but Electronics is the one which the highest rise. Over the period of a month there are specific days over which specific products peak and low.

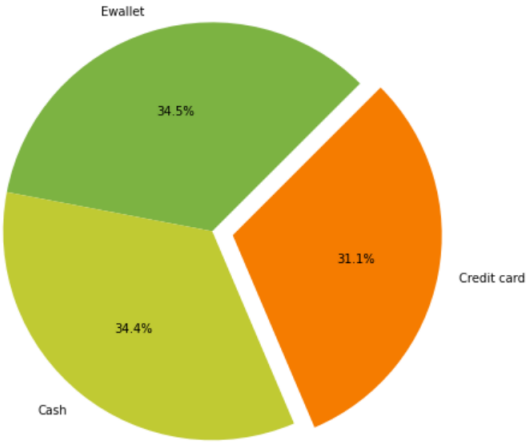


Figure 11: Sales distribution across different product lines using various payment methods

In the modern world today, there are n number of payment methods one can use to buy anyhing. We found such segregation in our dataset and went on to analyse sales made from each of these payment modes. The pie chart shows the sales distribution across different product lines using varous payment methods

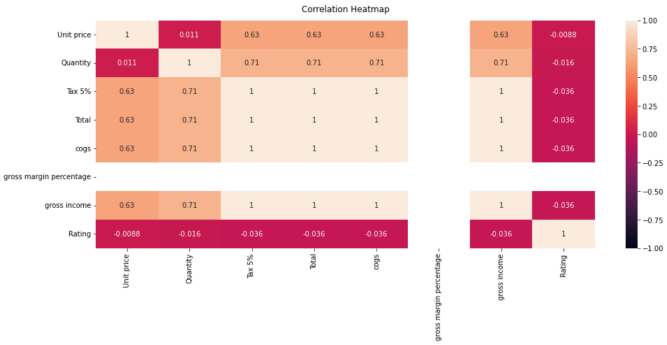


Figure 12: Heatmap for various attributes

Next, we wanted to get a gist of all the factors and attributes which may or may not affect sales of any of the product lines under consideration. Hence, we plotted a correlation heatmap to know more about which attributes to study and investigate

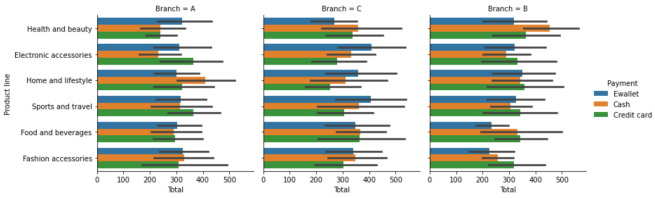


Figure 13: Sales for different product line with varying payment methods

The last figure - Figure 13 showcases the sales made by different product line with varying payment methods across all the branches. It can be inferred from the graph that all the three branches have different preferences w.r.t. different product lines. There is no single clear cut winner here!

6 Final Findings

If the buyer knows what they want to buy, what factors influence the pricing, and what similar products would be there in a said price range, they will be able to make an informed decision. Thus, Sales Analytics is of utmost importance to detect buyer trends and buyer sentiments. Since, we are working with three different industries: Lifestyle & Beauty, Merchandise & Clothing and Electronics, the

target audience for our project is massive, giving us enormous growth potential. That gives us a lot of opportunity to do more, maybe use neural networks to improve R2-score beyond the current value and further, improve the efficiency and prediction.

7 Future Work

Sales Analytics is a crucial aspect to be considered when talking about buying trends. If a person knows which features, do they need, what factors influence the pricing, and what are similar products that provide the same features at an affordable rate, God help the sellers, they can get the cognizance to buy the product with maximum outcome. Thus, it helps prospective buyers by giving them proper knowledge, which helps them in buying better products.

Moreover, the buyer also gets a gist of similar products and their respective price range. If we integrate this concept with an application, customers can also get notifications across various seasons (Thanksgiving, Black Friday, Christmas etc.) regarding sales, cheap products, newly launched products and so on.

Let's talk about organizational perspective now. [7] Collecting customer data has been one of the foremost priorities for all the businesses to analyze consumer behavior to further monetize, draw insights and understand it. .

So how do companies collect the data? Well, most of the companies ask for the consent in the start of the relationship with the customers. Usually, a form is supposed to be filled out at the start, "Please accept these terms and conditions in order to proceed." Well, as we know, no one reads the fine print as we are in hurry to spends some bills.

Due to their technological capabilities, large corporations are also skilled at collecting client data from a wide range of internet sources. Naturally, the most obvious source is their own websites, the majority of which are now fitted with cookies and web beacons. These technologies allow businesses to monitor the surfing habits of customers, even after they leave the company's website and move to other websites. With the help of cookies, businesses may track their customers' movements, what they've looked at, and where they go after leaving their website. As a result, companies are able to re-target clients with adverts. For example, let's say you went to Amazon, you searched for laptop sleeve. You added it to cart but didn't buy it. Two days later, you will get a notification saying "Have you still bought your laptop sleeve?", Well, it's $2 cheap now, would you consider it? Of course you would!

Companies of nearly any size can use the majority of the methods mentioned above to get client data. But collection is only the beginning—in many ways, it's the simple part. The real difficulty lies in gathering all of this data, analyzing it, and translating that analysis into useful insights.

Since, there is no relative work in this particular area, except for "Laptop price predictor" on various sources, we can definitely auction our product to a company, so they may use it for further purposes.

8 Acknowledgement

We would like to thank our professor, Dr. Di Wu (University of Colorado Boulder, teaching assistants Ajay Sadananda (University of Colorado Boulder), and Bhawneet Singh (University of Colorado Boulder) for providing us with the necessary knowledge, tools and comments on the project.

9 References

You may click on the any of the references provided to open it:

1. Lifestyle Industry and it's sophisticated and long supply chain structure

2. Taking a good look at the beauty industry
3. Industries at Glace: General Merchandise Stores
4. Kaggle: Supermarket Sales
5. Laptop Price Prediction – Practical Understanding of Machine learning project lifecycle
6. Medium: Laptop Price Prediction using machine learning
7. How do big companies collect consumer data
8. IBM: Linear Regression
9. Deepnote: Linear Regression tutorial
10. Basics of Machine Learning with k-nearest neighbors algorithm
11. Basic EDA for dummies
12. Random Forest - Wikipedia
13. Decision Tree for Beginners