Predicting The Price Of Backpacks, A Rigurous System Analysis

Daniel Esteban Camacho Ospina
20231020046
Universidad Distrital FJDC

Edgar Julian Roldan Rojas
20241020041
Universidad Distrital FJDC

Juan Esteban Rodriguez Camacho
20241020029
Universidad Distrital FJDC

***Abstract*—Kaggle presents us with new challenges, offering us accessible and interesting datasets that allow the community to practice and improve their machine learning skills. In this case, the objective is to predict the price of backpacks considering several attributes specified in the dataset.The proposed solution used data preprocessing with the pandas library to extract and analyze the dataset. Subsequently, linear regression was used to identify the most relevant variables and, finally, scikit-learn was applied to perform multiple linear regression to relate the predictor variables to the target variable (backpack price).As a result, important findings were obtained on the relevance of the detailed analysis of each aspect of competition. The sensitivity of the model and the accuracy of its predictions were evaluated, excluding external factors that could influence the system. In addition, the effects of external variables on the proposed solution were analyzed, exploring the interconnection between the different elements through a general mapping and the study of the system properties.**

I. COMPETITION OVERVIEW

The objective of the selected competition will be the analysis and design of a practical and viable solution that will allow us to predict the prices of backpacks, based on the evaluation of the dataset provided, which offers a wide range of records under which we will formulate our solution. The dataset presented will be composed of 11 columns which will reflect the main characteristics of the backpacks (ID, Brand, Material, Size, Compartments, Waterpoof, etc), in addition to this, we will have an amount of 300,000 records, this by identifying the final ID that we can find in the dataset, among the restrictions or limitations that we can find will focus on the sense of the information provided by the dataset, On the one hand we can exemplify from the weight, which does not present a specific restriction against the maximum weight or minimum weight that can have the backpack, in turn, the number of compartments does not present a specific limit, which can generate problems with the interpretation of the number of compartments.

II. SYSTEMIC ANALYSIS

At the time of our systemic analysis, we will start from the collection of features that the system will have as inputs, among them we will have: brand, material, size, number of compartments, computer compartment, waterproof, style, color and weight capacity. The following is a brief description of each of the features in general and the relationship of this with the chosen competition will be presented:

- Brand (String): The main insignia that denotes the distinction between products with similar purposes, this represents an intangible asset, due to the relevance of renown when choosing the most appropriate product, given this, we can indicate the importance that this aspect can have in the final valuation of the product in question. In this case, 5 brands are presented along the various records found (Jansport, Under Armour, Nike, Adidas, Puma).

- Material (String): Raw material of which the primary fabric of the product in question will be composed, the presentation of variations due to the type of material can result in fluctuations in the market value of these products, highlighting key aspects such as the difficulty of obtaining raw materials, import costs, as well as the relevance of labor in the manufacture of the goods created, highlighting the risk and discomfort that may involve the handling of these raw materials, in this particular case, we can highlight 4 types of raw material (Nylon, Leather, Canvas and Polyester).

- Size (String): Presents the quantification of the available space that can be used in the product, in this case, use is made of a qualitative type quantification, because it presents the size of the product in an estimate valued by the perception of buyers/sellers, thus excluding the need to present an estimate of the size of the product in numerical terms, In this case, it is common to recognize the variation of the price in relation to the increase of the size of the product, in spite of this, it can be affected by other characteristics such as the raw material or the load capacity, in this case 3 types of size were taken into account: Small, Medium and Large.

- Compartments (Integer): Expresses the amount of useful spaces (subdivisions of the product) that the product has, it highlights the directly proportional relationship between the number of compartments and the price, where the increase in this amount reflects a higher price in the product, this due to the utility and the benefits it can offer in terms of organization to have a greater amount of usable spaces, the allocation of this data can be seen in a range of 1 to 10 compartments.

- Laptop Compartment (Boolean): Indicates the presence or not of an exclusive compartment for portable devices (given the dimensions and the place where the compartment is located), given the relevance that this space may have for customers (comfortable transport of necessary device in the performance of daily activities, etc.) we can present a relationship in the price change given the existence or not of this space, to indicate whether the chosen product has this

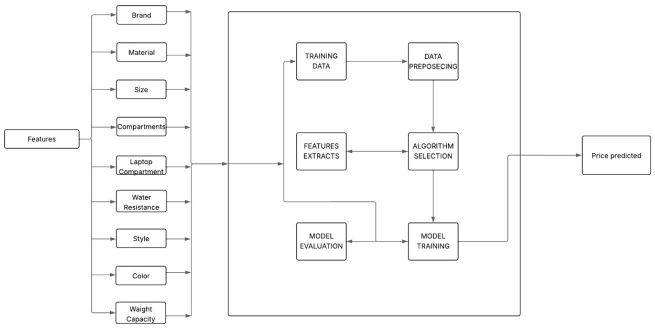

Fig. 1. Problem mapping

feature, we handle a Boolean data that indicates whether this compartment is available or not.

- Waterproof (Boolean): Characteristic that indicates if the product in question has the ability to repel liquids which can damage the objects that are loaded in it, we can reflect this utility as an added value to the products due to the utility that this can provide to the customers in question, this can be interpreted from a Boolean data that indicates whether it has this feature or not.

- Style (String): Indicates the type of style under which the product was produced, this mainly exposes if the product in question was produced under certain conditions to create a product focused on certain customer segmentation, taking into account external factors that may draw more attention regarding its utility, practicality and / or design, for this case, we have 3 types of main styles: handbag, shoulder bag and backpack, in addition to providing the opportunity to not have a defined style.

- Color (String): Type of dye that the product can take, sometimes it can change the price significantly due to its resemblance to other brands or directly because the dye is more difficult to obtain, this attribute has 6 types of colors (Black, Green, Red, Blue, Gray, Pink) .

- Weight capacity (Float): The resistance that comes to have the set of materials that make up the product, this attribute contains 181. 596 types of capacities which its interval is between (5kg and 30kg).

Already with the inputs of our system, we will give way to the mapping of the relationships, showing how the elements interact with each other:

Concluding, the formulation of the perspective from systems engineering, applying its principles allows us to frame the competition problem, thus, we obtained that, from the following properties, we will perform a more detailed analysis that will lead us to pose a better solution to the problem:

1. Homeostasis: The backpack price prediction model faces a critical challenge when prices in the real market experience abrupt increases not reflected in the training data. This disconnect between historical data and current reality can generate systematically low predictions, eroding the reliability of the system. To counteract this effect, strategies such as moving training windows that prioritize recent data, or adaptive weighting systems that assign greater importance to more current observations can be implemented. The key is to design automatic recalibration mechanisms that maintain model accuracy without constant human intervention.

2. Adaptability: Markets for products such as backpacks exhibit seasonal variations, fashion trends, and fluctuations in material costs that require a model capable of continually evolving. Advanced techniques such as meta-learning allow the system to identify patterns of change and adjust its internal parameters proactively. For example, we could implement a conceptual drift detection system that triggers retraining processes when it identifies significant changes in data distribution. This dynamic adaptation capability is especially valuable in commercial environments where delays in updates can translate into economic losses.

3. Resilience: In practice, price data sets often contain registration errors, outliers or outdated information that can significantly distort predictions. An effective approach combines anomaly detection techniques with inherently robust algorithmic models, such as quantile-based methods or neural networks with adaptive loss functions. In addition, the implementation of data sanitization layers that filter out obvious inconsistencies before the main processing adds an additional protective barrier. This redundancy of mechanisms ensures that the system maintains its performance even when faced with imperfect or corrupted data.

4. Emergence: Complex relationships between seemingly simple characteristics can generate non-intuitive pricing patterns that defy traditional models. For example, the interaction between seasonality, special materials and collaborations with famous designers can create multiplier effects on the final price. To capture these dynamics, sophisticated modeling architectures that include higher-order interaction learning capabilities are required. Techniques such as attention mechanisms or characteristic interaction networks allow these emerging relationships to be discovered and leveraged without requiring extensive manual specification.

Finally, we will have the objective variable of the problem, which is the price of the backpacks, for which the price is defined as follows:

Price (Float): It is the value by which the product is distributed depending on the other established variables, this is the target variable which is sought after considering all the different variables mentioned above, this final variable has 48212 types of prices which have a range between (15 dollars and 150 dollars).

III. COMPLEXITY AND SENSITIVIY

As mentioned above in the description of each of the Features, the minimum change in value of at least one of them will cause a change that may be minimal or too large in the price or target variable, since each of these characteristics of the backpack define its value in the market along with other external variables that are not taken into account for the realization of the prediction.

For example, with the amount of weight of the backpack, assuming it is Under Armour, Medium, 8 compartments, has a computer compartment, is not waterproof, is made of polyester and has a capacity of 10.20 kg has an approximate price excluding external variables of 25.98 dollars and now we simply increase its capacity to 14.74kg has an approximate price excluding external variables of 44.68 dollars, with this we can understand that the minimum change in one of its characteristics can significantly or drastically affect its price.

IV. CHAOS AND RANDOMNESS

In competition, the behavior of the chaotic attractor is reflected in external factors that have no relationship with the attributes set to arrive at the predicted price. For example, a social impact generated by a famous figure promoting the brand may increase the price of its backpacks in an unpredictable way. These types of phenomena are not possible to predict through the program, since it depends on external variables that alter the market in a non-linear way.

Chaos theory can also be visualized when you look at the prices and compare similar variables except for one and you realize that they do not follow a specific pattern, for example, a backpack with all similar variables but with a weight capacity of 27kg has an approximate price in the data of 25.82 dollars and another with all the same variables but with less weight capacity has a value of 142.73 dollars, so we can intuit that its price was affected by external variables.

V. USEFUL CODE

For the analysis of the information, we made use of lines of code that gave us a better perspective of the situation, here is the code for parts:

```
import pandas as pd
import numpy as np
from sklearn.preprocessing import LabelEncoder
```

```
df=pd.read_csv("train.csv")
df
```

	id int64	Brand object	Material object	Size object	Compartments fl...	Laptop Compart...	Waterproof object
0	0	Jansport	Leather	Medium	7	Yes	No
1	1	Jansport	Canvas	Small	10	Yes	Yes
2	2	Under Armour	Leather	Small	2	Yes	No
3	3	Nike	Nylon	Small	8	Yes	No
4	4	Adidas	Canvas	Medium	1	Yes	Yes
5	5	Nike	Canvas	Medium	10	No	Yes
6	6	Nike	nan	Large	3	No	No
7	7	Puma	Canvas	Small	1	Yes	Yes
8	8	Under Armour	Polyester	Medium	8	Yes	No
9	9	Under Armour	Nylon	Medium	2	Yes	Yes

300000 rows, 11 cols / page Page 1 of 30000 Format

Fig. 2. Import of useful libraries (Panda and numpy)

```
variables_por_columna = df.nunique()
variables_por_columna
```

```
id                    300000
Brand                      5
Material                   4
Size                       3
Compartments              10
Laptop Compartment         2
Waterproof                 2
Style                      3
Color                      6
Weight Capacity (kg)  181596
Price                   48212
dtype: int64
```

Fig. 3. Number of unique values for each column

```
for col in df.columns:
    print(col)
    print(df[col].unique())
```

```
id
[     0      1      2 ... 299997 299998 299999]
Brand
['Jansport' 'Under Armour' 'Nike' 'Adidas' 'Puma' nan]
Material
['Leather' 'Canvas' 'Nylon' nan 'Polyester']
Size
['Medium' 'Small' 'Large' nan]
Compartments
[ 7. 10.  2.  8.  1.  3.  5.  9.  6.  4.]
Laptop Compartment
['Yes' 'No' nan]
Waterproof
['No' 'Yes' nan]
Style
['Tote' 'Messenger' nan 'Backpack']
Color
['Black' 'Green' 'Red' 'Blue' 'Gray' 'Pink' nan]
Weight Capacity (kg)
[11.61172281 27.07853658 16.64375995 ...  9.55993494 26.63318223
  6.1757379 ]
Price
[112.15875  68.88056  39.1732  ...  78.7574  131.37288  41.96325]
```

Fig. 4. Output with the unique values of each column

```
max_price = df['Price'].max()
min_price = df['Price'].min()
max_weight_capacity = df['Weight Capacity (kg)'].max()
min_weight_capacity = df['Weight Capacity (kg)'].min()
print("capacidades:"," min", min_weight_capacity,"y max", max_weight_capacity,)
print("precios:","y min", min_price,"y max", max_price,)
```

```
capacidades:  min 5.0 y max 30.0
precios: y min 15.0 y max 150.0
```

Fig. 5. Limits of price and weight capacity

```
import numpy as np
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score

# Filtramos las columnas numéricas para analizar la linealidad
data_numeric = df[['Compartments', 'Weight Capacity (kg)', 'Price']].dropna()

# Dividir X (predictores) y y (variable objetivo)
X = data_numeric[['Compartments', 'Weight Capacity (kg)']]
y = data_numeric['Price']

# Crear modelo lineal
model = LinearRegression()
model.fit(X, y)

# Predicción
y_pred = model.predict(X)

# R2 score para medir linealidad
r2 = r2_score(y, y_pred)
r2
```

```
0.0003248369439911844
```

Fig. 6. Predictions using linear progression

VI. CONCLUSION

In conclusion, we can indicate the functionality of the proposed solution thanks to the fact that it was able to provide us with relevant information, which is useful to fulfill the main objective of predicting the value of the backpacks according to the variation of the inputs proposed by the user, therefore, we can affirm the functionality and success of the program in question for the requested need, on the other hand, we highlight the strengthening of activities focused on the analysis and design of solutions from key concepts, as well as, on the other hand, the work of programming skills and approach of computer solutions using tools that facilitated the obtaining of a practical and viable solution.

In another instance, it is essential to recognize the functionality of the solution, this can be seen in obtaining key data from the dataset delivered, with this, we were able to raise a more focused analysis to the need that is posed to us.

VII. BIBLIOGRAPHY

REFERENCES

[1] Carlos Andres Sierra, Systems Thinking, Systems analysis, Universidad Distrital FJDC, Bogota, Slides, 2025.

[2] Kaggle, Playground Series - Season 5, Episode 2, Kaggle, [En línea]. Disponible en: https://www.kaggle.com/competitions/playground-series-s5e2 [Fecha de acceso: abril 4, 2025].