Machine Learning Assessment

Pradip Basnet
Student ID: 23189643

Total Page Count: 23
Module : Machine Learning
Tutor Name: Rupak Koirala

Contents

List of Figures

1 Introduction

The dataset consists of various attributes related to backpacks, with the goal of predicting their price. It includes features such as the brand, material, size, and style of the bag, along with additional functional attributes like the number of compartments, whether there is a laptop compartment, and if the bag is waterproof. Other details such as color and weight capacity (in kilograms) are also provided. The target variable for prediction is the price of the backpack, which is influenced by these various attributes. By analyzing the relationships between these features and the price, the goal is to develop a model that can accurately predict the price based on the given characteristics of the backpacks.

2 Motivation behind the chosen domain

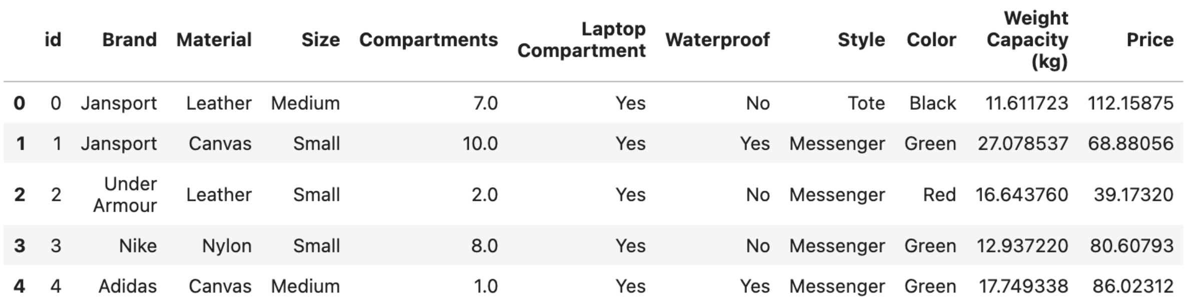
The motivation behind predicting backpack prices comes from the need to understand how various features, such as brand, material, size, style, and functionality, impact the pricing of consumer products. This ongoing competition on Kaggle provides a perfect opportunity to explore and sharpen my data science skills. By working on this challenge, I can apply machine learning techniques to a real-world-like dataset, gaining experience in feature engineering, regression models, and price prediction. This exercise will help enhance my ability to analyze complex datasets and make data-driven decisions, improving both my technical expertise and problem-solving abilities.

3 Scope of the visualizations

The scope for visualization based on this backpack dataset includes understanding customer behavior, product preferences, and pricing trends in the backpack market. Various visualization techniques can reveal insights into how different backpack attributes, such as brand, material, size, and functionality, are perceived by customers and how these attributes correlate with product pricing. Visualizations could also explore the relationship between features like compartments, laptop compatibility, and waterproofing with customer satisfaction and price recommendations. This would provide a clearer picture of what drives customer decisions, helping to optimize product designs, pricing strategies, and marketing efforts in the backpack industry.

4 Exploratory data Analysis and Data Visualizations

The dataset provides various attributes related to backpacks, including an identifier for each backpack (id), and key features such as the brand, material, size, and style. It also includes functional attributes like the number of compartments, the presence of a laptop compartment, and whether the backpack is waterproof. Additional details such as the weight capacity (measured in kilograms) and color are also recorded. The target variable in this dataset is the price of the backpack, which can be predicted using the other features. By analyzing these attributes, the goal is to build a model that can accurately predict the price of a backpack based on its design, functionality, and other relevant characteristics.

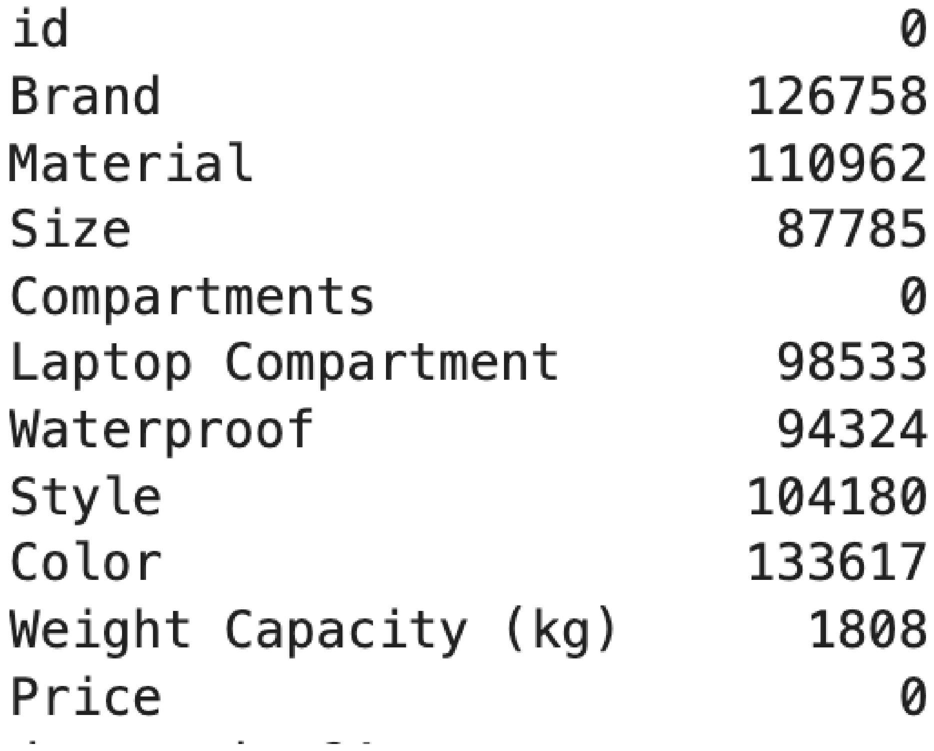


	id	Brand	Material	Size	Compartments	Laptop Compartment	Waterproof	Style	Color	Weight Capacity (kg)	Price
0	0	Jansport	Leather	Medium	7.0	Yes	No	Tote	Black	11.611723	112.15875
1	1	Jansport	Canvas	Small	10.0	Yes	Yes	Messenger	Green	27.078537	68.88056
2	2	Under Armour	Leather	Small	2.0	Yes	No	Messenger	Red	16.643760	39.17320
3	3	Nike	Nylon	Small	8.0	Yes	No	Messenger	Green	12.937220	80.60793
4	4	Adidas	Canvas	Medium	1.0	Yes	Yes	Messenger	Green	17.749338	86.02312

Figure 1: summary of a data

4.1 Data Summary and Replacing Null Value

The structure and summary of the data are inspected to understand the types of data, check for missing values, and gather basic statistical information. This step is going to be helpful in identifying potential problems that really need to be addressed. Replace all instances of ”null” or ”NULL” with either mean or median to standardize missing values. This allows easier handling of the missing data in subsequent steps. The dataset contains missing values across several attributes, which are addressed by using either the mean or median for imputation. This approach helps maintain the integrity of the dataset while minimizing the impact of missing data on the analysis and model predictions. For numerical features, the median or mean is applied depending on the distribution of the data, ensuring that the imputation method is appropriate for each variable.

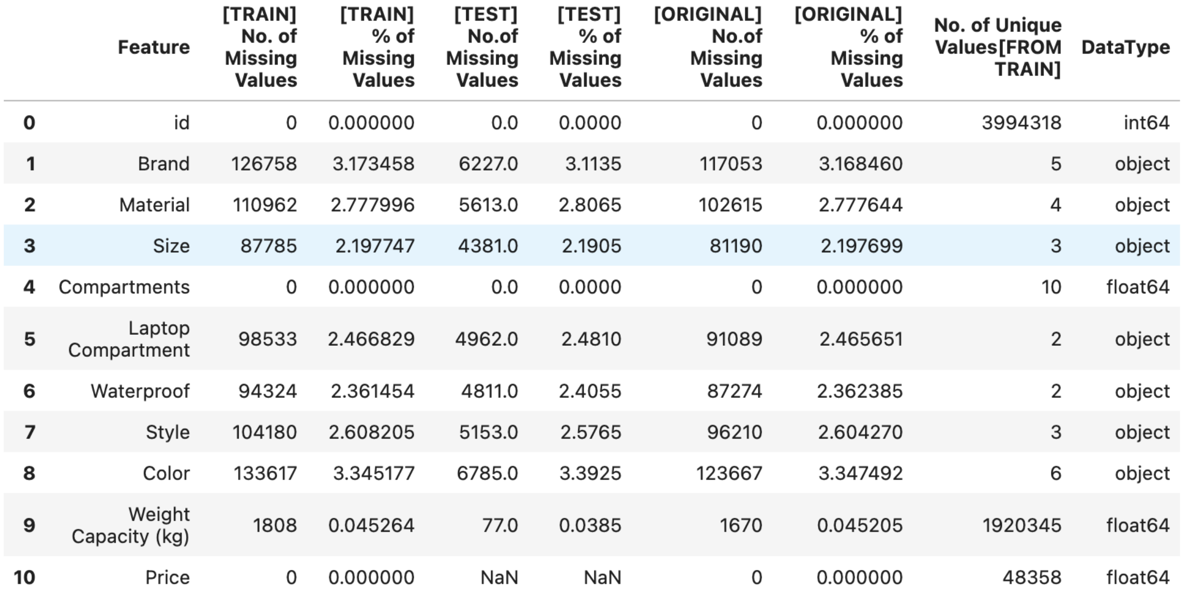


```
id                          0
Brand                  126758
Material               110962
Size                    87785
Compartments                0
Laptop Compartment      98533
Waterproof              94324
Style                  104180
Color                  133617
Weight Capacity (kg)     1808
Price                       0
```

Figure 2: summary of a data

4.2 Percentage of missing values

The dataset contains varying percentages of missing values across different features. For example, the “Brand” feature has 3.17% missing values in the training data, while the “Material” feature has 2.78%. Some features like “Compartments” and “Price” have no missing values at all. On the other hand, attributes such as “Style” and “Color” have 2.61% and 3.35% missing values respectively in the training set. Missing values are addressed by imputing with the mean or median depending on the nature of the feature, ensuring that the dataset remains consistent for model training.



	Feature	[TRAIN] No. of Missing Values	[TRAIN] % of Missing Values	[TEST] No.of Missing Values	[TEST] % of Missing Values	[ORIGINAL] No.of Missing Values	[ORIGINAL] % of Missing Values	No. of Unique Values[FROM TRAIN]	DataType
0	id	0	0.000000	0.0	0.0000	0	0.000000	3994318	int64
1	Brand	126758	3.173458	6227.0	3.1135	117053	3.168460	5	object
2	Material	110962	2.777996	5613.0	2.8065	102615	2.777644	4	object
3	Size	87785	2.197747	4381.0	2.1905	81190	2.197699	3	object
4	Compartments	0	0.000000	0.0	0.0000	0	0.000000	10	float64
5	Laptop Compartment	98533	2.466829	4962.0	2.4810	91089	2.465651	2	object
6	Waterproof	94324	2.361454	4811.0	2.4055	87274	2.362385	2	object
7	Style	104180	2.608205	5153.0	2.5765	96210	2.604270	3	object
8	Color	133617	3.345177	6785.0	3.3925	123667	3.347492	6	object
9	Weight Capacity (kg)	1808	0.045264	77.0	0.0385	1670	0.045205	1920345	float64
10	Price	0	0.000000	NaN	NaN	0	0.000000	48358	float64

Figure 3: Percentage of missing value

4.3 Missing Values

The extent of missing values across different features is clearly visible in the following plot. Some features, like ”Brand” and ”Color,” show significant gaps in data, while others, like ”Compartments” and ”Price,” have no missing values. This visualization highlights the areas of the dataset that require attention, with missing values addressed through imputation methods to maintain the consistency of the dataset for model training.

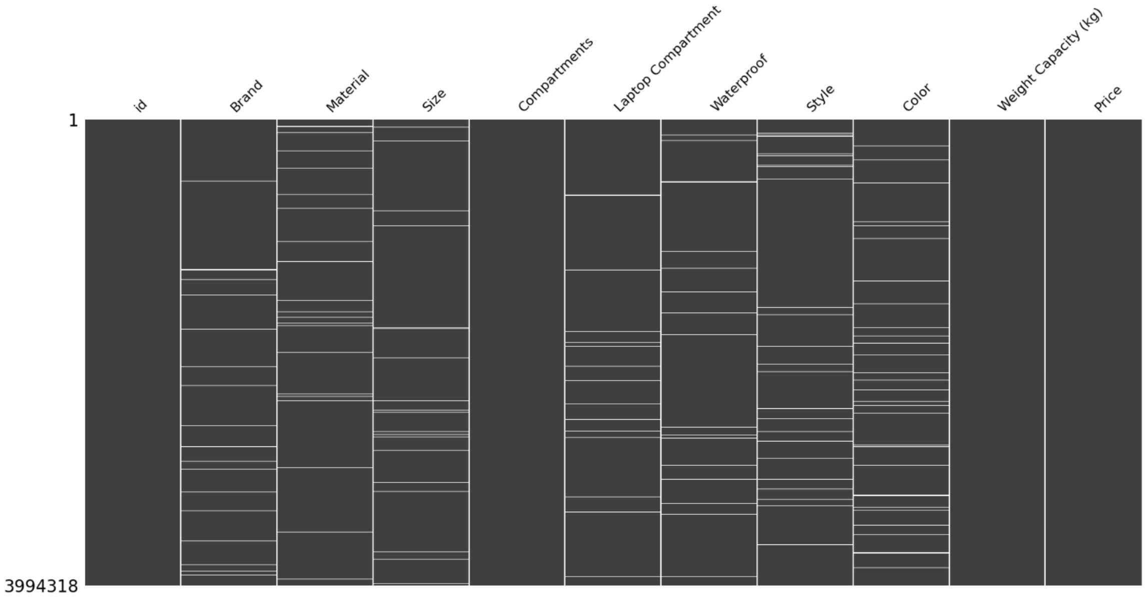


Figure 4: Visual representation of missing values across features.

4.4 Categorical features distribution

The figure presents the distribution of bags based on four categorical attributes: brand, material, size, and style. The brand distribution shows that the number of bags across different brands, including Jansport, Under Armour, Nike, Adidas, and Puma, is fairly even, with some brands having slightly higher counts. In terms of material, polyester appears to be the most common choice, while canvas is the least used. When considering size, medium-sized bags have the highest count, followed by large, while small-sized bags are the least common. Lastly, in terms of style, messenger bags are slightly more prevalent compared to tote and backpack styles. Overall, the dataset appears well-balanced across these attributes, with no significant skewness in distribution.

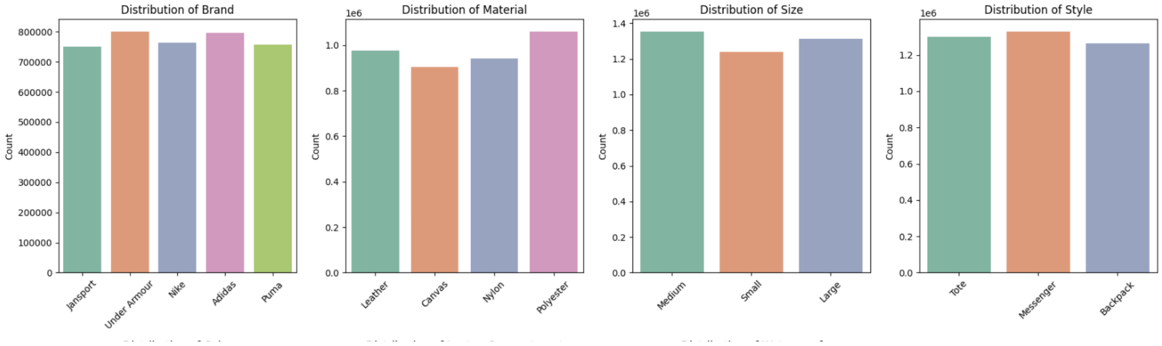


Figure 5: a glance at categorical features

Additionally, this figure visualizes the distribution of color, laptop compartments, and waterproofing in the dataset. The first chart shows a balanced distribution of colors, with Pink being the most common. The second chart reveals an almost equal split between bags with and without a laptop compartment. Similarly, the third chart indicates that waterproof and non-waterproof bags are nearly equal in number. These insights help in understanding product availability and customer preferences.

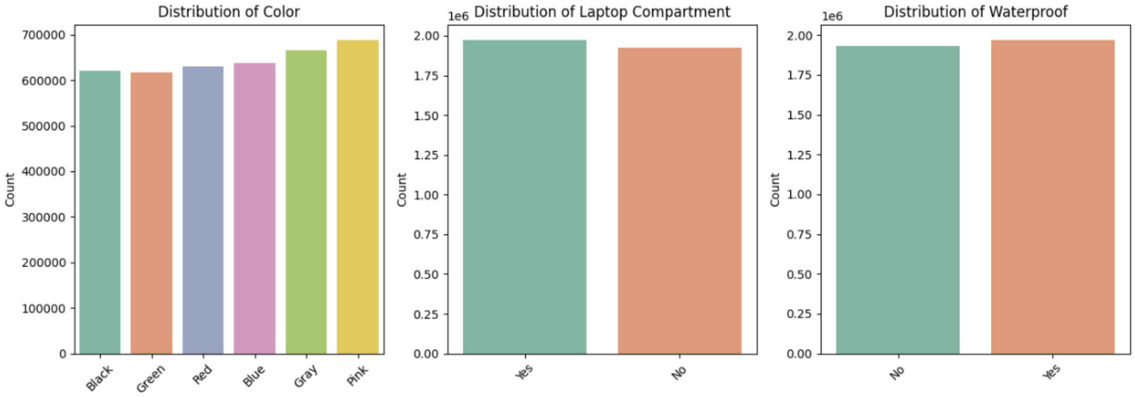


Figure 6: categorical values distribution

4.5 Distribution of weight capacities in kilograms

The figure illustrates the distribution of weight capacities in kilograms, highlighting key trends and frequencies. The tallest bar represents the most common weight capacity, indicating which capacity appears most frequently in the dataset. Conversely, the shortest bar shows the least common weight capacity, revealing which capacity is rare. The overall pattern of the bars provides insight into how the weight capacities are distributed, whether they are evenly spread, concentrated in certain ranges, or follow a specific trend. For instance, if the bar at 10 kg is the tallest, it suggests that 10 kg is the most prevalent capacity, while a shorter bar at 30 kg would indicate that 30 kg is less common. .

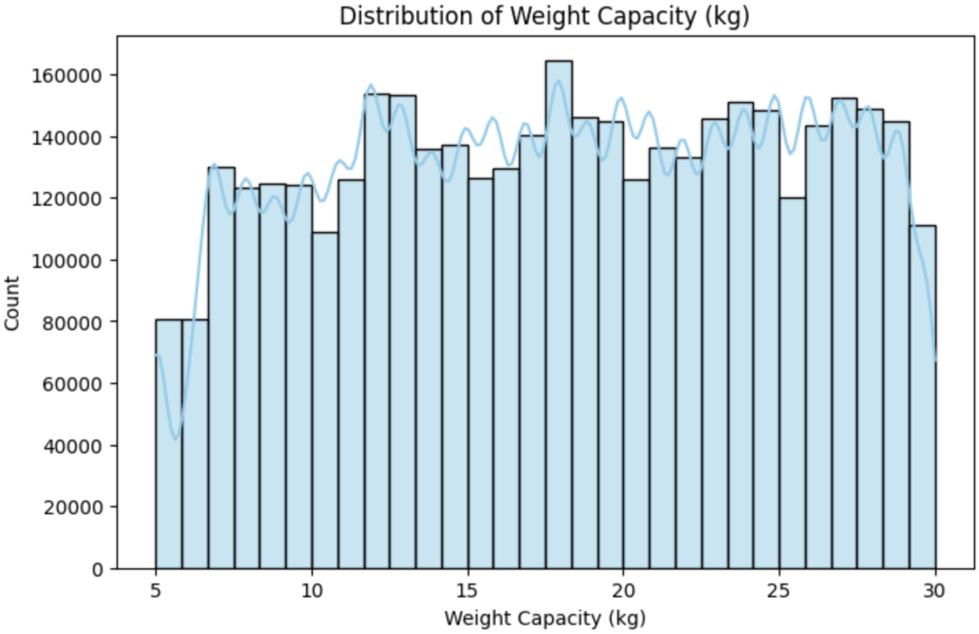


Figure 7: weight distribution

4.6 Price distributions across brands

The boxplot reveals distinct price distributions across brands like Jansport, Under Armour, Nike, Adidas, and Puma. Jansport has the lowest median price, indicating it's the most affordable, while Nike and Adidas have higher median prices, positioning them as more premium options. Under Armour and Puma share similar median prices, slightly above Jansport. Price variation within each brand also differs, with taller boxes suggesting greater price variability. Overall, the brand influences price, with Jansport being more budget-friendly and Nike and Adidas reflecting a higher-end market segment, helping consumers make informed purchasing decisions based on brand and budget.

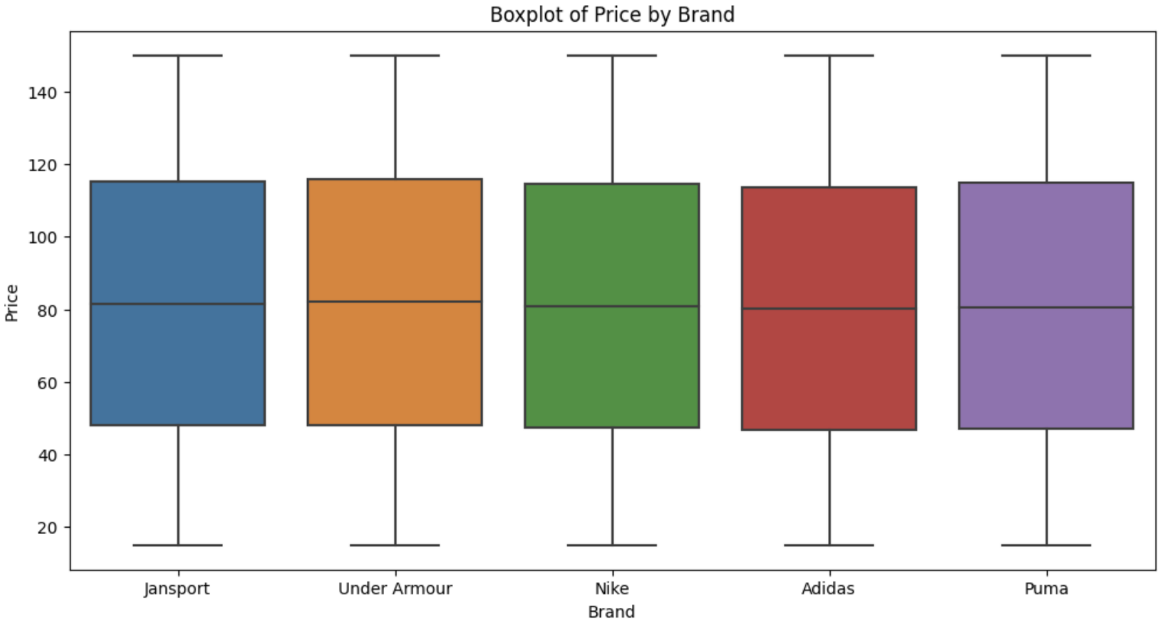


Figure 8: price distributions across brands

4.7 Comparing prices between bags with and without laptop compartments

The box plot comparing prices between bags with and without laptop compartments reveals some interesting insights. The median prices for both groups are very similar, around $80. Additionally, the spread of prices and the variability are almost identical, indicating comparable price ranges for both types of bags. The whiskers show that the minimum and maximum prices are also similar, ranging from approximately $15 to $150 for both groups. The middle 50% of prices, represented by the boxes, span from about $45 to $115 for both. The key takeaway is that the presence of a laptop compartment does not seem to significantly influence the price, which might be surprising as one would expect a laptop compartment to increase the cost of production and, consequently, the price.

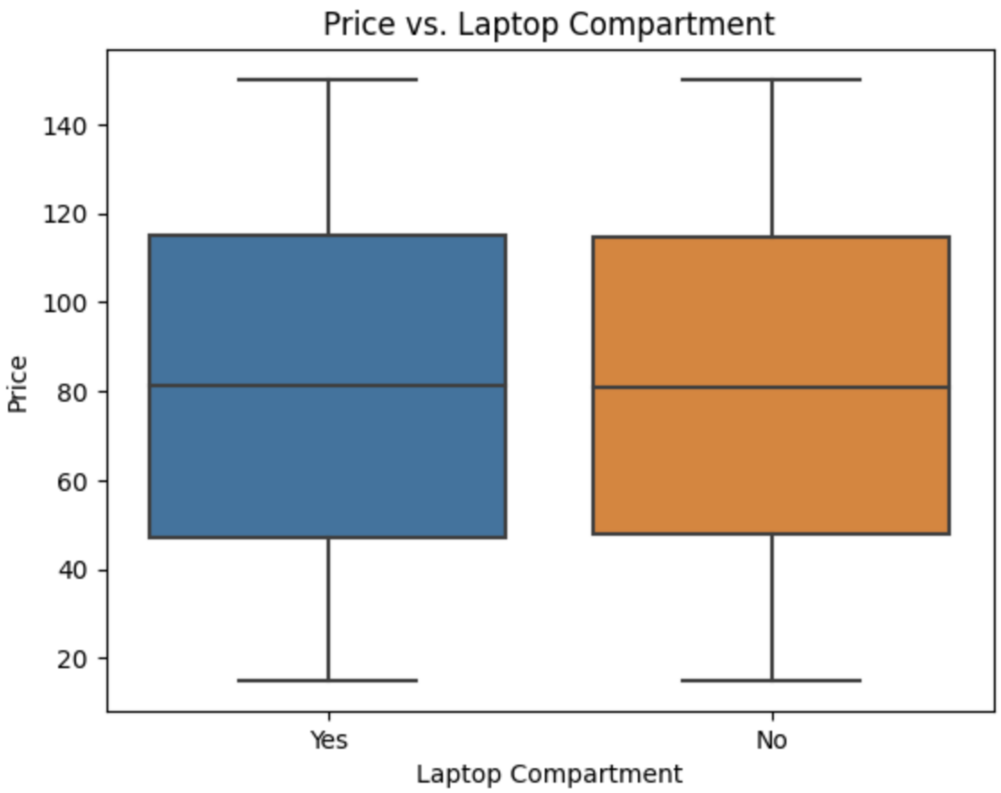


Figure 9: Comparing prices between bags with and without laptop compartments

4.8 Count of different bag styles

This bar chart shows the count of different bag styles (Tote, Messenger, and Backpack) across five major brands. Under Armour appears to have the highest overall production numbers across all three styles, with their messenger bags being particularly numerous (around 330,000 units). Messenger bags (shown in orange) consistently have the highest count across all brands, suggesting this style is the most popular or most produced bag type. Totes (shown in blue) and backpacks (shown in green) tend to have similar production numbers within each brand, though slightly lower than messenger bags. Puma seems to have the lowest overall production numbers among the five brands, but they maintain the same pattern of messenger bags being their highest-volume product. The difference between counts is relatively consistent across brands, with messenger bags ranging from 250,000 to 330,000 units, totes from 240,000 to 300,000 units, and backpacks from 230,000 to 290,000 units. This data suggests that while there are differences between brands in total production volume, they all follow a similar pattern in terms of style distribution, with messenger bags being the predominant style across all brands.

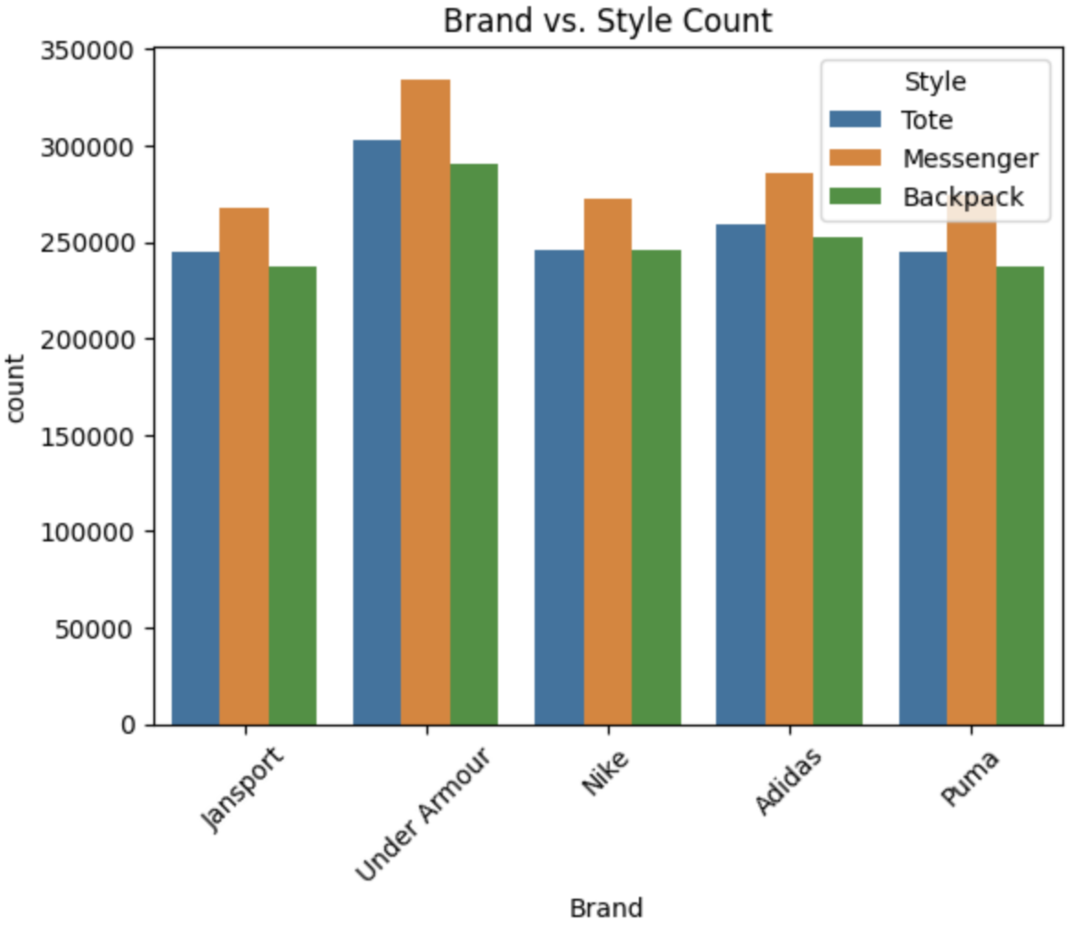


Figure 10: Count of different bag styles

4.9 Price distributions across three different bag sizes

The box plot comparing price distributions across three different bag sizes (Small, Medium, and Large) reveals some interesting insights. The median prices for all sizes are quite similar, around $80-85. The price ranges are comparable, spanning from about $20 to $140, and the interquartile ranges (the boxes) are consistent, ranging roughly from $50 to $110. Interestingly, larger bags don't seem to command higher prices, which goes against the common assumption that they would due to increased material costs. The price variability is also similar across all sizes, suggesting that other factors like brand, materials, or design play a more significant role in price determination than size. Additionally, the relatively regular whisker lengths indicate there are few outliers, showing consistent pricing practices across the market. This challenges the idea that larger bags would naturally be more expensive, highlighting that size may not be a major factor in determining bag prices in this market.

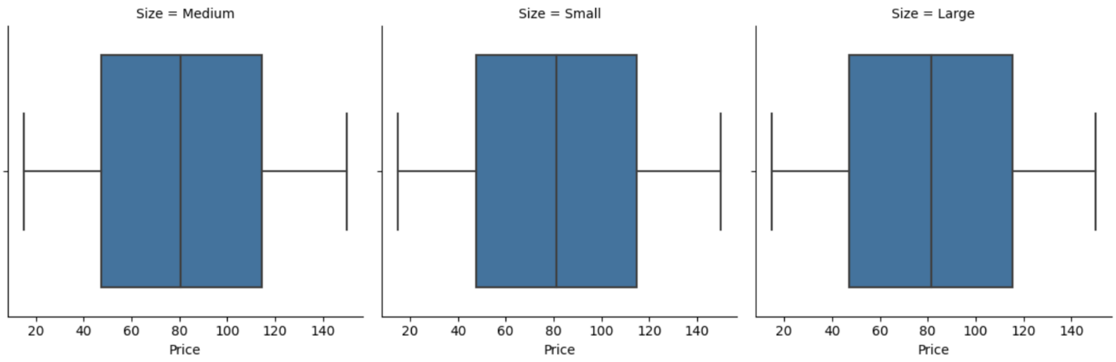


Figure 11: price distributions across three different bag sizes

4.10 Feature importance through Random Forest

Feature importance through a Random Forest model is done to assess how much each feature contributes to predicting bag prices. The results reveal that the most significant feature is weight capacity, with an importance score of about 55%, indicating that the weight a bag can carry is the primary factor influencing its price. Other important features include compartments, material, color, and brand, each contributing between 7% and 10% to the price prediction. On the other hand, features such as size, style, waterproofness, and laptop compartment have relatively low importance, with scores ranging from 1% to 3%. Interestingly, some features that might be expected to have a larger impact, like size and laptop compartments, have surprisingly low scores. Size, in particular, aligns with the earlier box plot analysis, where price distributions were similar across different bag sizes. The minimal impact of the laptop compartment, despite being perceived as a premium feature, suggests that functional attributes such as weight capacity are more significant than design features. Additionally, brand influence is less than anticipated, further reinforcing the idea that practical features matter more in determining bag prices than brand perception. This analysis provides insight into pricing strategies in the bag market, highlighting the importance of functionality over aesthetics and brand in price determination.

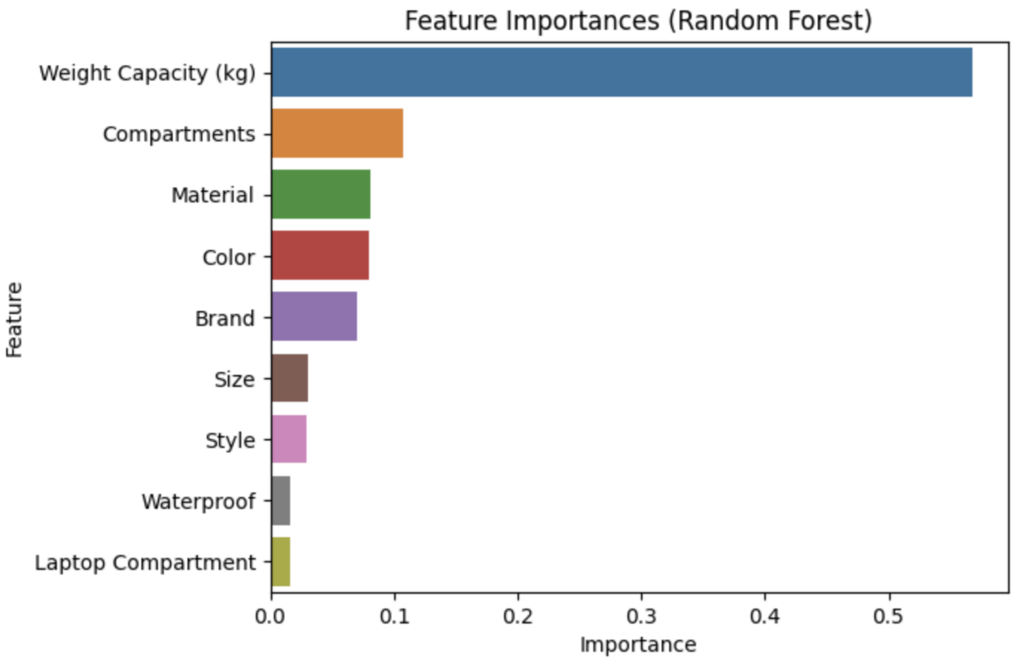
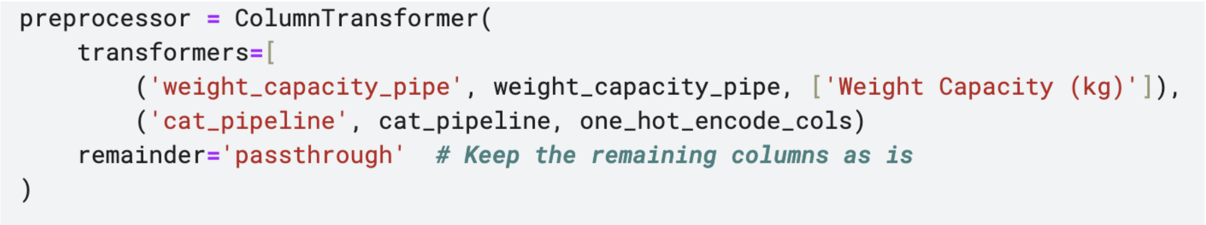


Figure 12: Feature importance through a Random Forest model

5 Feature Engineering

5.1 One-Hot Encoding Scaling

The `ColumnTransformer` is used to apply different preprocessing steps to different types of features. In this case, `one_hot_encode_cols` (which includes categorical features like `Brand`, `Material`, etc.) are processed using `cat_pipeline`, which applies one-hot encoding to convert categorical variables into numerical form. Meanwhile, `weight_capacity_pipe` applies scaling (using `StandardScaler`) to the `'Weight Capacity (kg)'` column to normalize numerical values, ensuring they are on a similar scale. The `remainder='passthrough'` argument ensures that any other columns not specified in the transformer list remain unchanged.

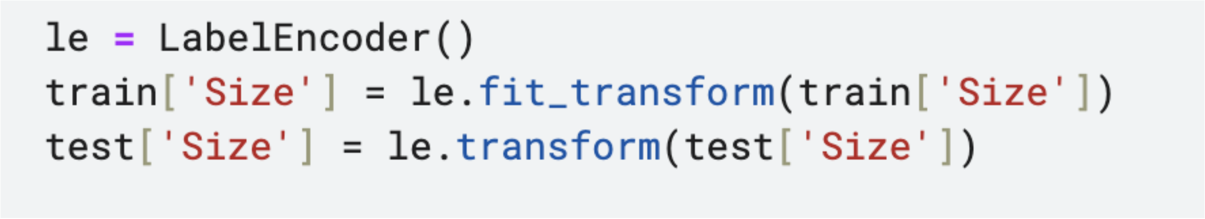


```
preprocessor = ColumnTransformer(
    transformers=[
        ('weight_capacity_pipe', weight_capacity_pipe, ['Weight Capacity (kg)']),
        ('cat_pipeline', cat_pipeline, one_hot_encode_cols)
    remainder='passthrough'  # Keep the remaining columns as is
)
```

Figure 13: encoding and scaling the features

5.2 Label Encoding for 'Size'

The `'Size'` column is being label-encoded using `LabelEncoder()`, which assigns a unique integer to each category. This is useful when the feature represents ordinal data (e.g., Small, Medium, Large), where the categories have a meaningful order. Since `LabelEncoder` assigns numbers based on the fitted dataset, the train dataset is first fitted and transformed, while the test dataset is only transformed to maintain consistency. Unlike one-hot encoding, which creates multiple binary columns, label encoding keeps the feature as a single numerical column, making it more efficient for models that can handle ordinal relationships.



```
le = LabelEncoder()
train['Size'] = le.fit_transform(train['Size'])
test['Size'] = le.transform(test['Size'])
```

Figure 14: lebel encodiing for ordinal data

6 Models Used

Various models, including XGBoost, CatBoost, LightGBM, AdaBoost, Random Forest, Linear Regression, Ridge Regression, Partial Least Squares Regression, and ElasticNet, are used to identify the most efficient model for predicting price values. These models represent different approaches to regression and ensemble learning, each with its strengths in capturing complex patterns and relationships in the data. The performance of these models is evaluated using the Root Mean Squared Error (RMSE), a common metric for regression tasks. RMSE is calculated using the following formula:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

where y_i represents the actual values, $\hat{y}_i$ represents the predicted values, and n is the number of data points. By comparing the RMSE of each model, the most accurate and efficient model for price prediction can be selected.

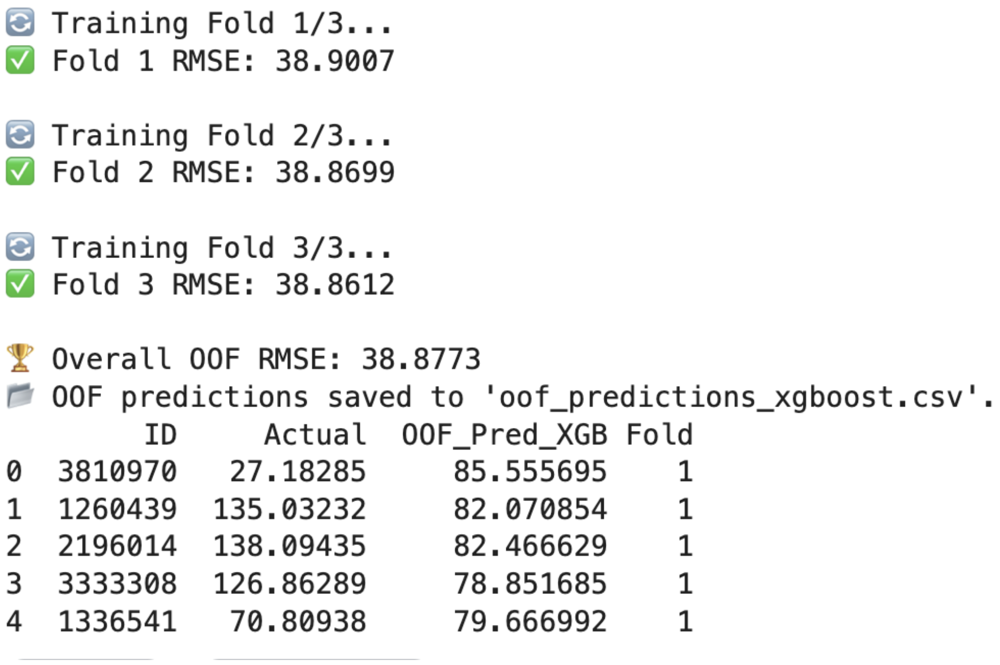
List of Models used

- XGBoost
- CatBoost
- LightGBM
- AdaBoost
- Random Forest
- Linear Regression
- Ridge Regression
- Partial Least Squares Regression
- ElasticNet

6.1 Best Model (XGBoost

For the backpack price prediction task, the best model selected was **XGBoost**, which achieved the optimal result through a combination of *K-Fold Cross Validation* and *hyper-parameter tuning*. Prior to training the model, several preprocessing steps were applied to the data to ensure it was in the best form for machine learning. Categorical variables were {one-hot encoded using a column transformer, while the *'Weight Capacity (kg)'* feature underwent scaling to normalize numerical values and bring them to a similar scale. Additionally, the *'Size'* column was label-encoded to convert ordinal categories (such as *Small, Medium, and Large*) into numerical values. This step was critical for ensuring that the model could handle both categorical and continuous features correctly. The model was trained using K-Fold Cross Validation, which helps assess the model's generalization ability by splitting the data into multiple subsets (or folds). The final performance of the model was evaluated using Root Mean Squared Error (RMSE), a standard metric for regression tasks that quantifies the difference between predicted and actual values. The final Out-of-Fold (OOF) RMSE for the XGBoost model was calculated to be 38.8773, which reflects the model's ability to generalize well to unseen data.

Through careful data preprocessing, cross-validation, and hyperparameter tuning, the XGBoost model delivered the best performance for predicting backpack prices, showing its effectiveness in handling both categorical and numerical features. The final RMSE of 38.8773 indicates a reliable and accurate model for this prediction task.



```
🔄 Training Fold 1/3...
✅ Fold 1 RMSE: 38.9007

🔄 Training Fold 2/3...
✅ Fold 2 RMSE: 38.8699

🔄 Training Fold 3/3...
✅ Fold 3 RMSE: 38.8612

🏆 Overall OOF RMSE: 38.8773
📂 OOF predictions saved to 'oof_predictions_xgboost.csv'.
        ID     Actual  OOF_Pred_XGB Fold
0  3810970   27.18285     85.555695    1
1  1260439  135.03232     82.070854    1
2  2196014  138.09435     82.466629    1
3  3333308  126.86289     78.851685    1
4  1336541   70.80938     79.666992    1
```

Figure 15: XGB output

7 Unsupervised Learning (K-MEANS Clustering)

The K-Means Clustering Visualization reveals several key insights about the data. The algorithm successfully divided the data into three distinct clusters, each represented by a different color (yellow, dark blue, and teal-ish blue). The red ”X” symbols indicate the centroids of these clusters, which represent the average location of data points within each group. The clusters themselves exhibit elongated, curved shapes, suggesting that the data may have non-linear relationships that were captured through PCA for dimensionality reduction. Additionally, there is a noticeable replication of the cluster pattern along the horizontal axis (PCA Component 1), indicating that the first principal component is likely capturing a repeating feature or pattern. Overall, the clusters are well-separated, highlighting that K-Means was effective in grouping similar data points while distinguishing them from dissimilar ones.

Particularly Cluster 2, shows how feature combinations influence cluster placement. Cluster 2’s location on the PCA plot is determined by the bags’ characteristics in the table, such as the ”Messenger” style, laptop compartments, and varying weights. These features correspond to specific values in PCA Component 1 and PCA Component 2, which place the data points within Cluster 2 on the plot.The centroid, marked with a red ”X,” represents the average bag in Cluster 2, reflecting the typical combination of features. PCA simplifies the data by reducing the original features to two components, making it easier to visualize the clustering. Thus, the table provides insights into why the bags are grouped in Cluster 2, while the plot shows their position in the reduced PCA space.

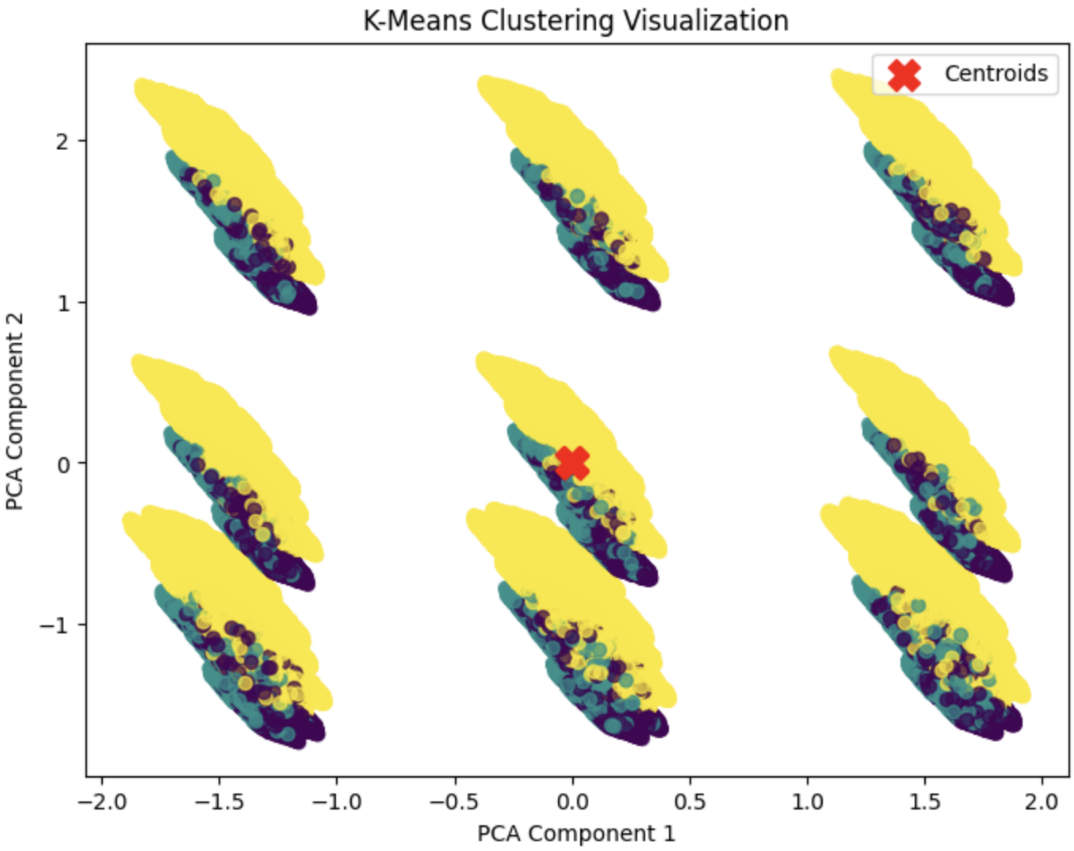


Figure 16: Feature importance through a Random Forest model

8 Conclusion

This study analyzed factors influencing backpack prices, revealing that weight capacity is the most significant driver, while size and laptop compartments have minimal impact. Among regression models, XGBoost performed best, achieving an RMSE of 38.8773 after tuning.K-Means clustering identified distinct product groups based on features like style and weight. Overall, functional attributes outweigh brand perception in price determination, offering valuable insights for pricing strategies and market positioning. Future research can explore additional factors like consumer reviews and seasonal trends.