Backpack Prediction

Artificial Intelligence

Project 2 - Final Delivery

Group 131

Bruno Oliveira	202208700
Henrique Fernandes	202204988
Rodrigo Silva	202205188

FEUP - L.EIC029

Problem Description

Topic K: Student Bag Prediction Challenge

This project is related to the application of Supervised Learning techniques to real world problems. Specifically, we will develop and evaluate machine learning models that accurately predict the price of student backpacks (target variable), based on a variety of input attributes.

For this we will use a labeled dataset from Kaggle which contains different bag characteristics, such as brand, material and weight capacity, and we want to understand how they influence the price, therefore consisting of a **regression problem**. We will respect the following machine learning pipeline: data preprocessing, problem definition and target identification, model selection and parameter tuning, model training and testing, result evaluation and comparison.

Link to Backpack Prediction Challenge

Related Work

Backpack Prediction Code Submissions

Since this dataset was part of a Playground problem on Kaggle, there are multiple solutions and analysis of the dataset either from people who took part in this competition or who used this dataset to study and share their approaches with others.

Out of these submissions, we highlighted the following that were more aligned with our project:

https://www.kaggle.com/code/tarundirector/backpack-pred-baseline-ensemble-eda

https://www.kaggle.com/code/cdeotte/first-place-single-model-lb-38-81

Methodology

- Development Environment
  - Python -> Programming Language
  - NumPy/SciPy -> Numerical Calculations
  - Scikit-Learn -> Machine Learning Models and Evaluation Metrics
  - Matplotlib/Seaborn -> Data Visualization and Results
- Algorithms
  - Decision Tree
  - K-Nearest Neighbors
  - Support Vector Machines
  - Neural Networks
  - Ensemble Learning -> Random Forests
- Evaluation Metrics
  - Root Mean Squared Error, Mean Absolute Error, R-Squared Score
  - Training and Testing Time

NOTE: Measures such as accuracy, precision, F1-score, and such will not be used, as this is a regression problem, and these are characteristic of classification problems

Exploratory Data Analysis

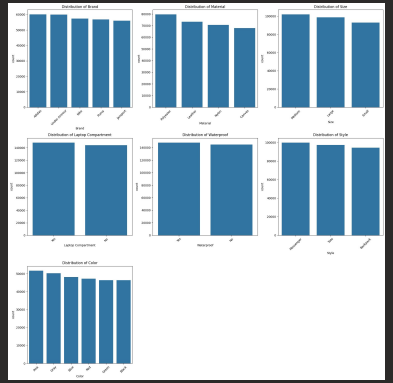
Steps

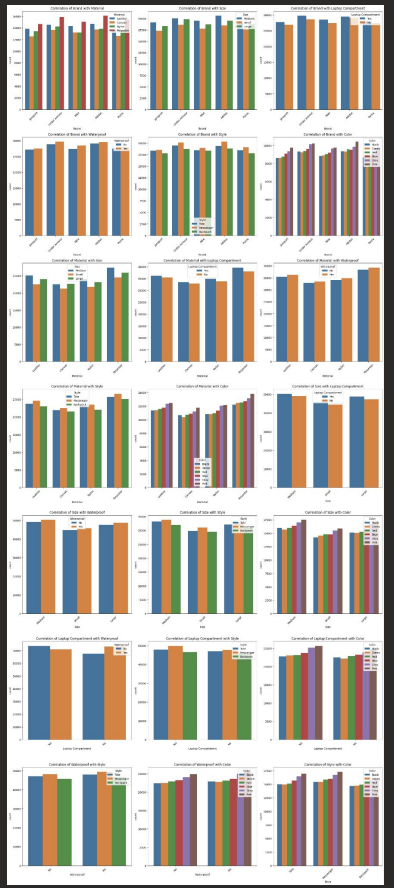
- Dataset Overview
- Missing & Duplicated Data
- Data Distribution & Correlation

```
Training Dataset Overview

Number of rows: 300000
Number of columns: 11

Data types of each column:
id                          int64
Brand                      object
Material                   object
Size                       object
Compartments              float64
Laptop Compartment         object
Waterproof                 object
Style                      object
Color                      object
Weight Capacity (kg)      float64
Price                     float64
dtype: object
```





Conclusions

- The training dataset has 300000 rows (+ 3700000 in the extra)
- Less than 18% of the rows have missing data (below 3.5% per column)
- There are no duplicates in the entire dataset
- All features appear to be evenly distributed without correlation
- The target variable has a spike at the maximum value (150.0) which we might indicate a truncated maximum

Data Preprocessing

Steps

- Data Dropping & Imputation
- Data Encoding and Normalization

```
[5 rows x 23 columns]
Testing data
0    112.15875
1     68.88056
2     39.17320
3     80.60793
4     86.02312
Name: Price, dtype: float64
Training data shape
(246686, 23)
```

Conclusions

- To account for the missing values, we dropped the rows with missing data for training the models, and we did **mode imputation** for the categorical features and **median imputation** for the numerical features for the final model
- We also encoded the features using **binary mapping** (for most categorical features) and **ordinal mapping** (since “size” has a natural ordering)
- The encoded columns were also normalized to prevent unwanted influences on some of the models

Training data

	Brand_Adidas	Brand_Jansport	Brand_Nike	Brand_Puma	Brand_Under Armour
0	0	1	0	0	0
1	0	1	0	0	0
2	0	0	0	0	1
3	0	0	1	0	0
4	1	0	0	0	0

	Material_Canvas	Material_Leather	Material_Nylon	Material_Polyester	\
0	0	1	0	0	
1	1	0	0	0	
2	0	1	0	0	
3	0	0	1	0	
4	1	0	0	0	

	Style_Backpack	...	Color_Blue	Color_Gray	Color_Green	Color_Pink	\
0	0	...	0	0	0	0	
1	0	...	0	0	1	0	
2	0	...	0	0	0	0	
3	0	...	0	0	1	0	
4	0	...	0	0	1	0	

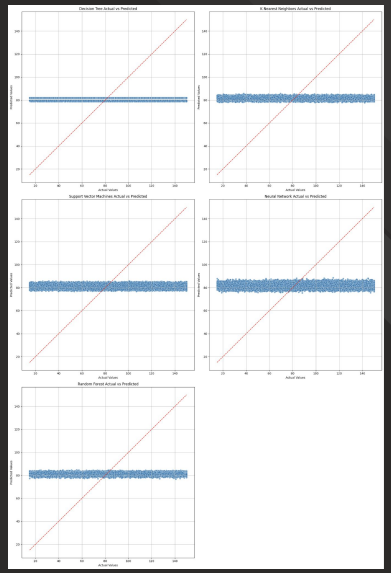
	Color_Red	Size_Ratio	Has_Laptop_Compartment	Is_Waterproof	\
0	0	0.5	1	0	
1	0	0.0	1	1	
2	1	0.0	1	0	
3	0	0.0	1	0	
4	0	0.5	1	1	

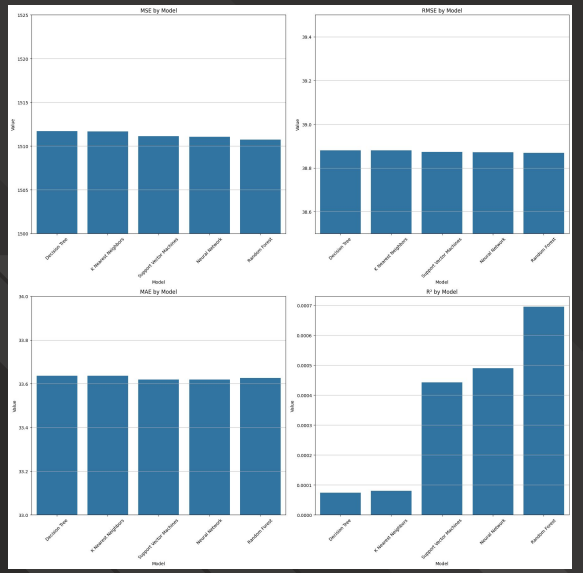
	Compartments_Ratio	Weight_Capacity_Ratio
0	0.666667	0.264471
1	1.000000	0.883149
2	0.111111	0.465754
3	0.777778	0.317491
4	0.000000	0.509978

Result Analysis

Steps

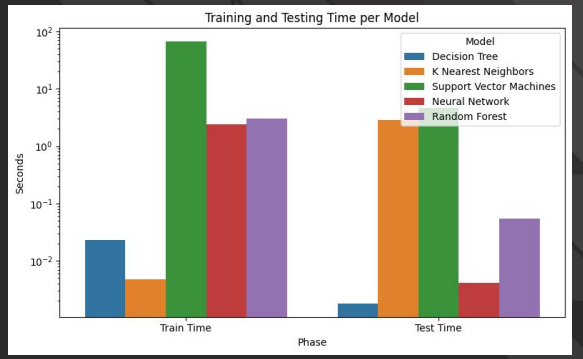
- Model Metric Comparisons
- Learning Curves
- Expected vs. Actual Plot
- Residual Plots





Conclusions

- ❖ In terms of evaluation metrics and execution time, models ranked as follows: Decision Tree → KNN → SVM → Neural Networks → Random Forest
- ❖ The artificial dataset had uniform distributions and weak feature-target correlations, limiting model learning.
- ❖ Expected vs real and residual plots showed that all models had poor predictive performance, often predicting around the average price independently of the features.



Predictions and Conclusions

In this project, we explored the application of supervised techniques applied to regression problems to predict the prices of student backpacks based on the Kaggle Playground Competition Season 5 Episode 2.

We followed the common machine learning pipeline starting with the Exploratory Data Analysis, which allowed us to get an overview of the dataset we were working with. Based on some observations from the previous step, we performed a Data Preprocessing step to improve the quality of the dataset, specified the target variable and defined methods to tune the parameters of each model.

As for the models implemented, we went with Decision Tree, K-Nearest Neighbors, Support Vector Machines, Neural Networks and Random Forests. All of the models were tuned, trained and compared based on chosen metrics (RMSE, MAE and R^2 Score) as well as Training and Testing times. Based on this, we ranked the models based on suitability for this problem (considering both results and execution times) in following manner: Decision Tree → KNN → SVM → Neural Networks → Random Forest.

Despite this, we observed that all of our models had very bad results, usually making predictions centered around the average value despite the actual features of the data they were using. We conclude that this has to do with the dataset itself, since it was generated using artificial intelligence and all the columns have uniform distributions without significant correlations that allow the models to accurate predictions.

For the final part of our project, we used the best model from the ones that were trained and predicted the target variable (price) of the data in the test.csv dataset. This was the data that was used to test the models of the Kaggle competition, so we exported the results and uploaded our solution to Kaggle, having achieved a score of 38,94 placing us just over 2000th place. Nonetheless, the top results were not far from the ones obtained, further indicating that the main issue was the training dataset in question.

References

Kaggle. (2024). Playground Series - Season 5, Episode 2. Retrieved from https://www.kaggle.com/competitions/playground-series-s5e2/data

Tarun Director. (2024). Backpack Pred - Baseline + Ensemble + EDA. Kaggle Notebooks. Retrieved from https://www.kaggle.com/code/tarundirector/backpack-pred-baseline-ensemble-eda

Scikit-learn Developers. (n.d.). Scikit-learn: Machine learning in Python. Retrieved from https://scikit-learn.org/stable/

IBM. (2024). What is supervised learning? Retrieved from https://www.ibm.com/think/topics/supervised-learning

Stanford University CS 229. (n.d.). Supervised Learning Cheatsheet. Retrieved from https://stanford.edu/~shervine/teaching/cs-229/cheatsheet-supervised-learning