April 10, 2024

IE7275 – Data Mining in Engineering

Bharath Raj Pragada
(Dept. of Industrial Engineering)

Northeastern University

Section 1

Project Findings

Section 2

Challenges Faced

Section 3

Key Takeaways

Northeastern University

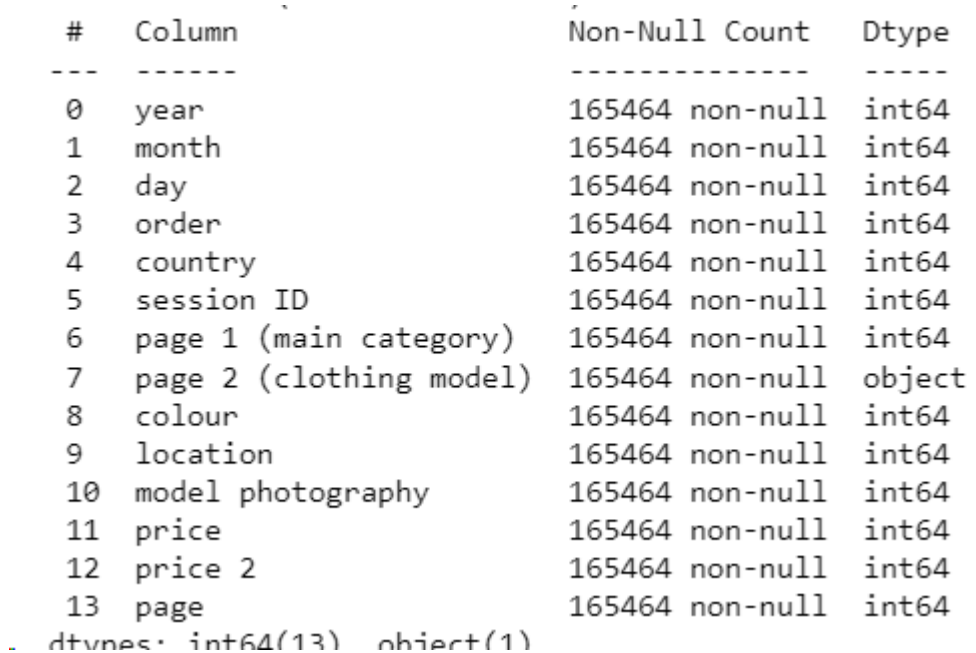
Data Collection

- https://fakestoreapi.com/docs
- JSON – CSV
- Query params – data of Interest

- shopping_data.rename(columns={

  'session ID': 'session-id',

  'price 2': 'price-higher-than-category',}
- lambda x: 1 if x == 1 else 0



#	Column	Non-Null Count	Dtype
0	year	165464 non-null	int64
1	month	165464 non-null	int64
2	day	165464 non-null	int64
3	order	165464 non-null	int64
4	country	165464 non-null	int64
5	session ID	165464 non-null	int64
6	page 1 (main category)	165464 non-null	int64
7	page 2 (clothing model)	165464 non-null	object
8	colour	165464 non-null	int64
9	location	165464 non-null	int64
10	model photography	165464 non-null	int64
11	price	165464 non-null	int64
12	price 2	165464 non-null	int64
13	page	165464 non-null	int64

dtypes: int64(13), object(1)

Clickstream data – suitable for Regression, Classification, and Clustering

Regression – Quantitative response – Price Prediction (In $)

Classification – Qualitative response – Price bucketing (4 categories – Budget, Value, Average, Premium)

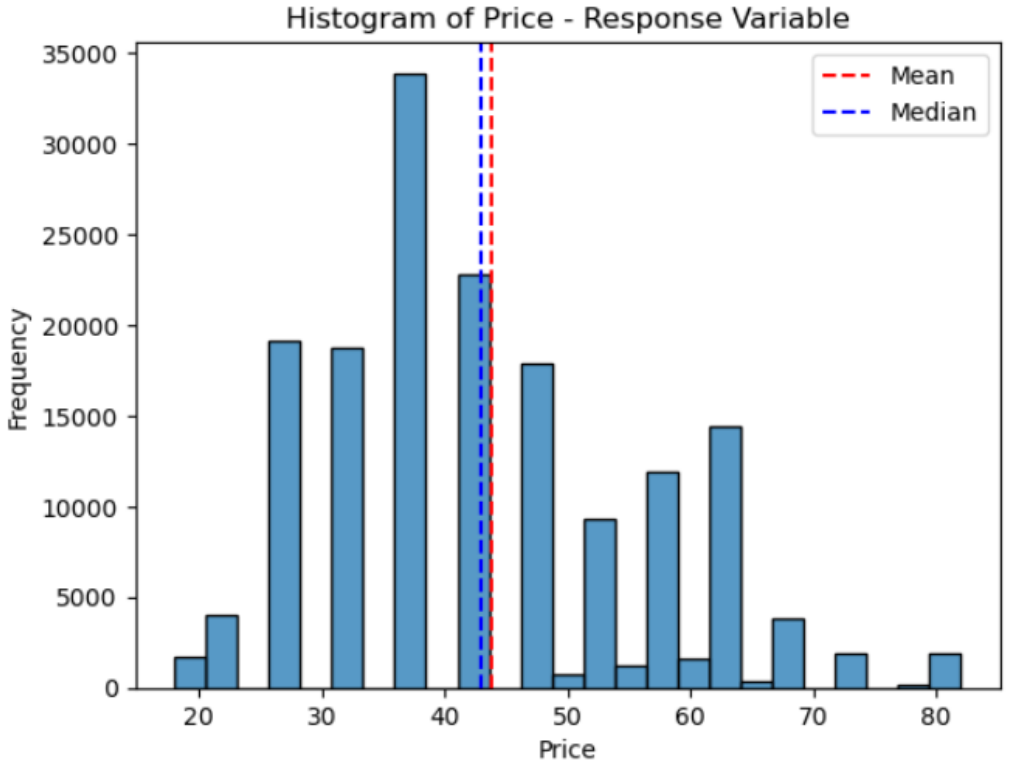
Rule of Thumb - 10 out of 165474 - aside

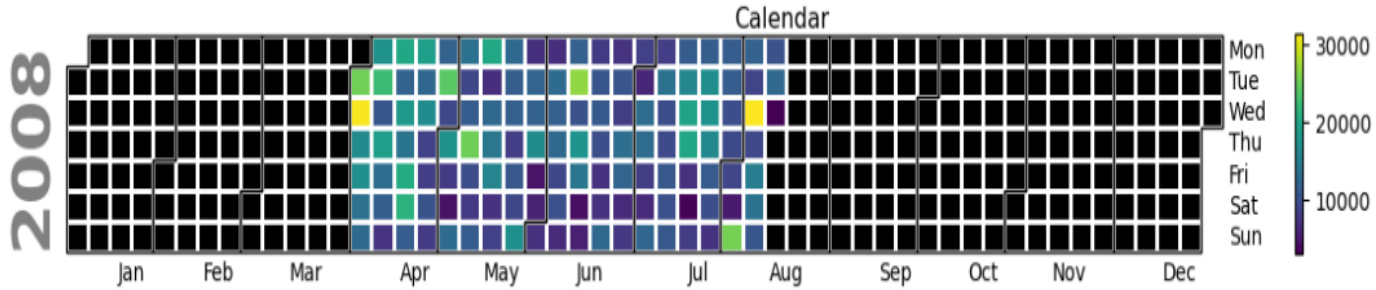
```
if price in range(0, 26):
    return 'budget'
elif price in range(26, 36):
    return 'value'
elif price in range(35, 66):
    return 'average'
elif price in range(66, 101):
    return 'premium'
else:
    return None
```

EDA

The year attribute only has one value - `2008` . Hence does not give any unique information about the response variable. Thereby can be dropped!

Session IDs are techincally unique IDs for each new user session that is generated. Hence this ID doesn't tell us anything about user's behaviour.

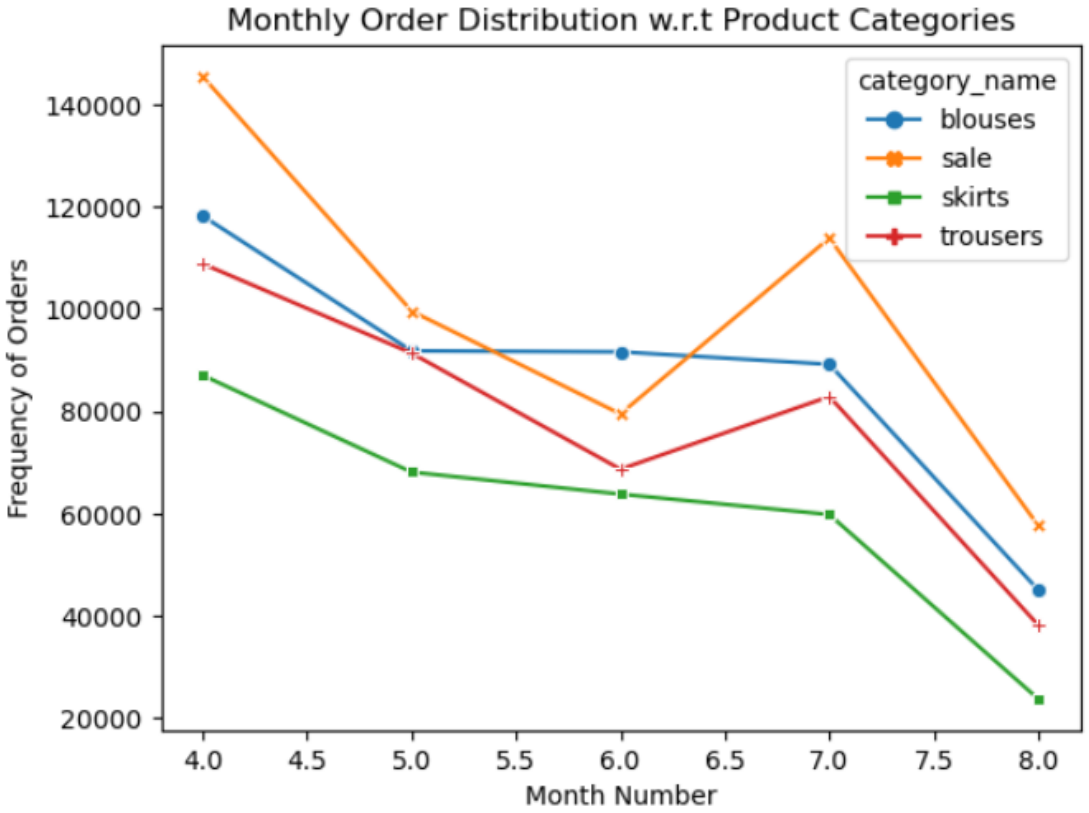




Clearly, the dates with lighter colors have highest number of orderes recieved. Rest seem to be recieve average number of orders. The dates with black or darker shades of green / blue have minimum number of orders. The following is observed:

- April has major number of orders (More lighter color cells)
- June has lowest amount of orders recieved
- May & July almost have equal amounts of orders recieved.
- Aug has one day with very high amounts of orderes recieved.

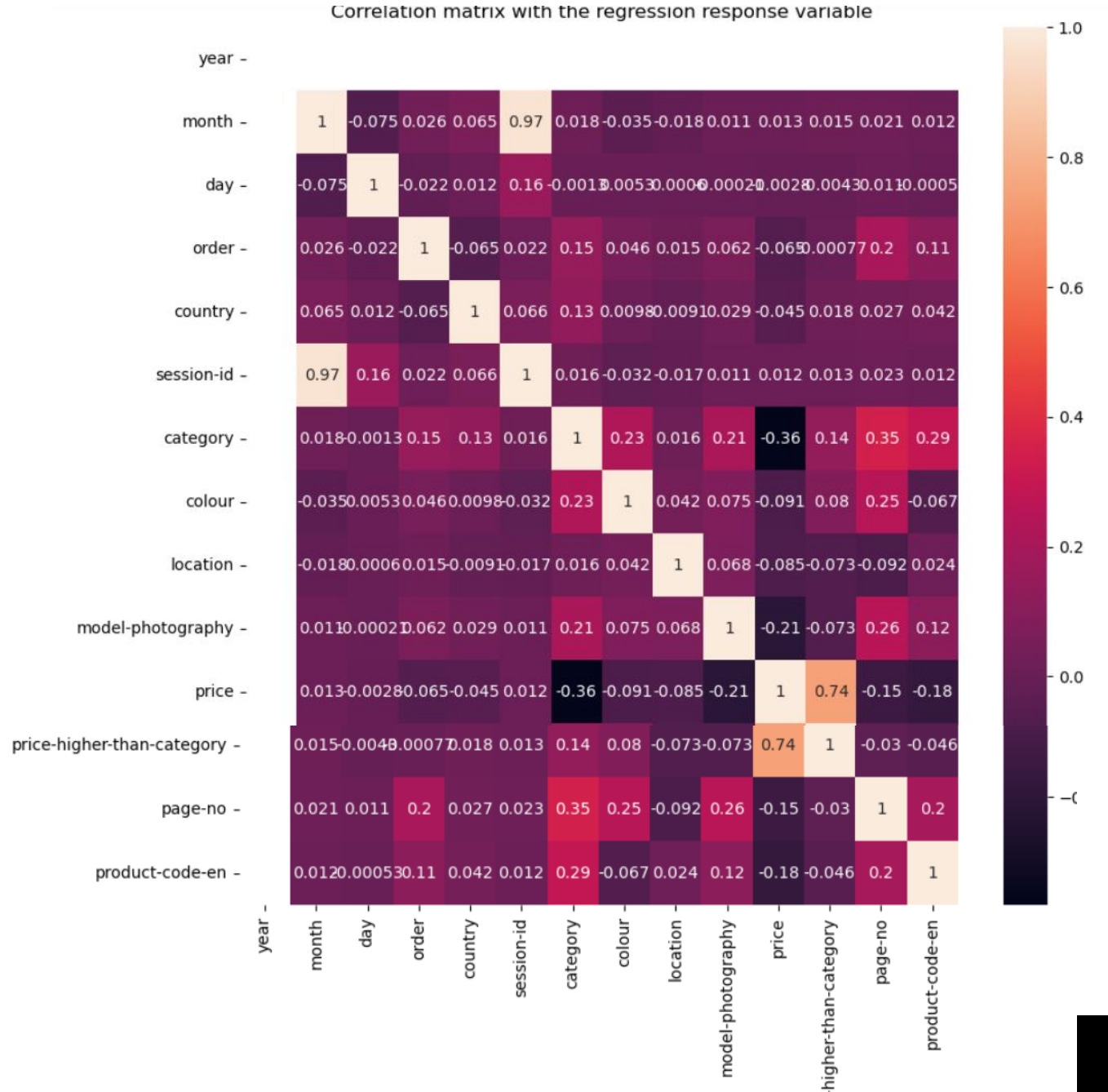
Clearly, There seems to be a sale in the months of April & August. Though trousers have higher number of orders collectively, we find blouses have higher number of orders each month. Skirts have relatively less frequency of orders per month.



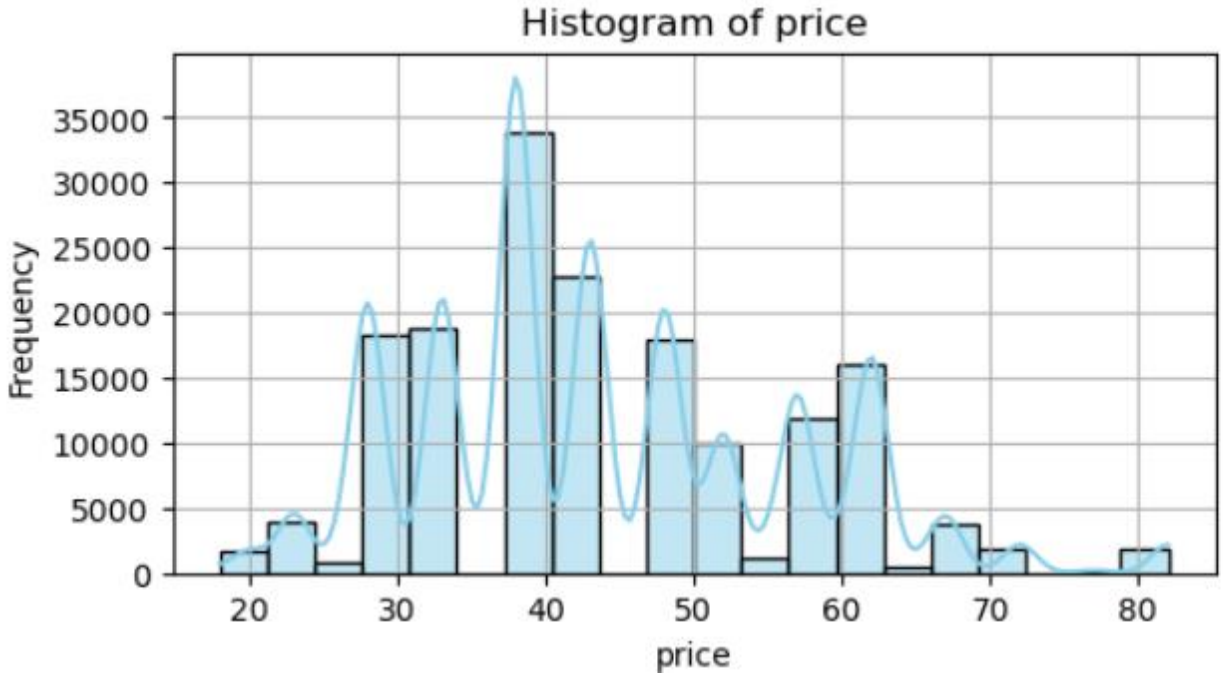
Feature Selection

- Correlation Matrix
- Feature importance – Random
- Forest
- OLS

p-value higher than
0.05 (significance level).
It means that we fail to reject the null hypothesis, that the coefficients of these attributes are zero –
thus implying these coefficients are not effective in predicting the price.



Distribution



- Standardization – mean around ‘0’ and standard deviation ‘1’

Base Model

Regression

- Linear Regression
- Ridge Regression
- LASSO Regression
- KNN Regressor
- Decision Tree Regressor
- Random Forest Regressor
- Bagging Regressor
- MLP Regressor

Classification

- Logistic Regression
- Gaussian Naive Bayes
- K-Nearest Neighbors(KNN)
- Decision Tree Classifier
- Random Forest Classifier
- Bagging Classifier
- MLP – NN Classifier

Section 1

Project Findings

Section 2

Challenges Faced

Section 3

Key Takeaways

Northeastern University

- Polars – Errors related to reading csv,
- MLP Regressor -
- Grid Search CV – Time Complexity is more - lack of Computational Power
- Customer behavior is not static and may evolve over time due to various factors such as market trends, promotions, or external events

Section 1

Project Findings

Section 2

Challenges Faced

Section 3

Key Takeaways

Northeastern University

No Free Lunch theorem in machine learning states that no single machine learning algorithm is universally superior for all tasks

Future Work

- Feature Expansion – X^2, log(X)
- Enhance the User Experience
- Interactive Visualizations
- A/B Testing and Experimentation

Thank you for listening

Questions?