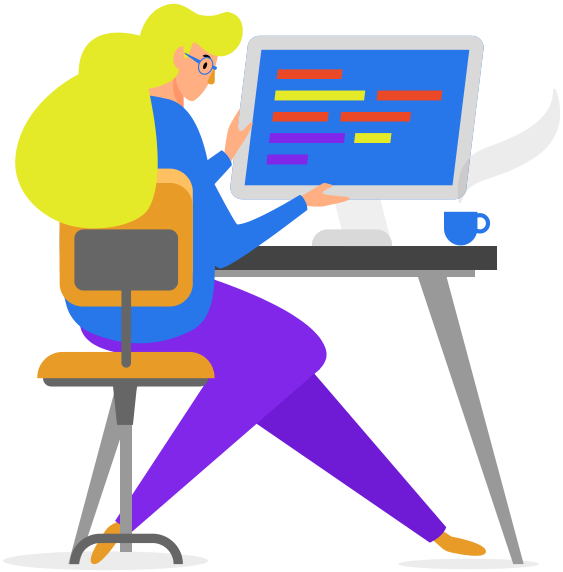
Zara Fashion Trends Prediction Model

Elsa Figueroa
Gauri Gupta
Miga Budaasuren
Termeh Mohebbie

Table of Content



Introduction and Goal

- The main goal of this analysis is to **predict** the **best-performing products** for the last week in terms of revenue using the previous week's data.
- This dataset is from a real-world Machine Learning Contest from a major retailer, **ZARA**, in 2019 with a cash prize.
- The data includes 3 months worth of sales, stock, and positioning of items for ZARA online store in a specific country.
- Due to business confidentiality, all information is coded.

Exploratory Data Analysis

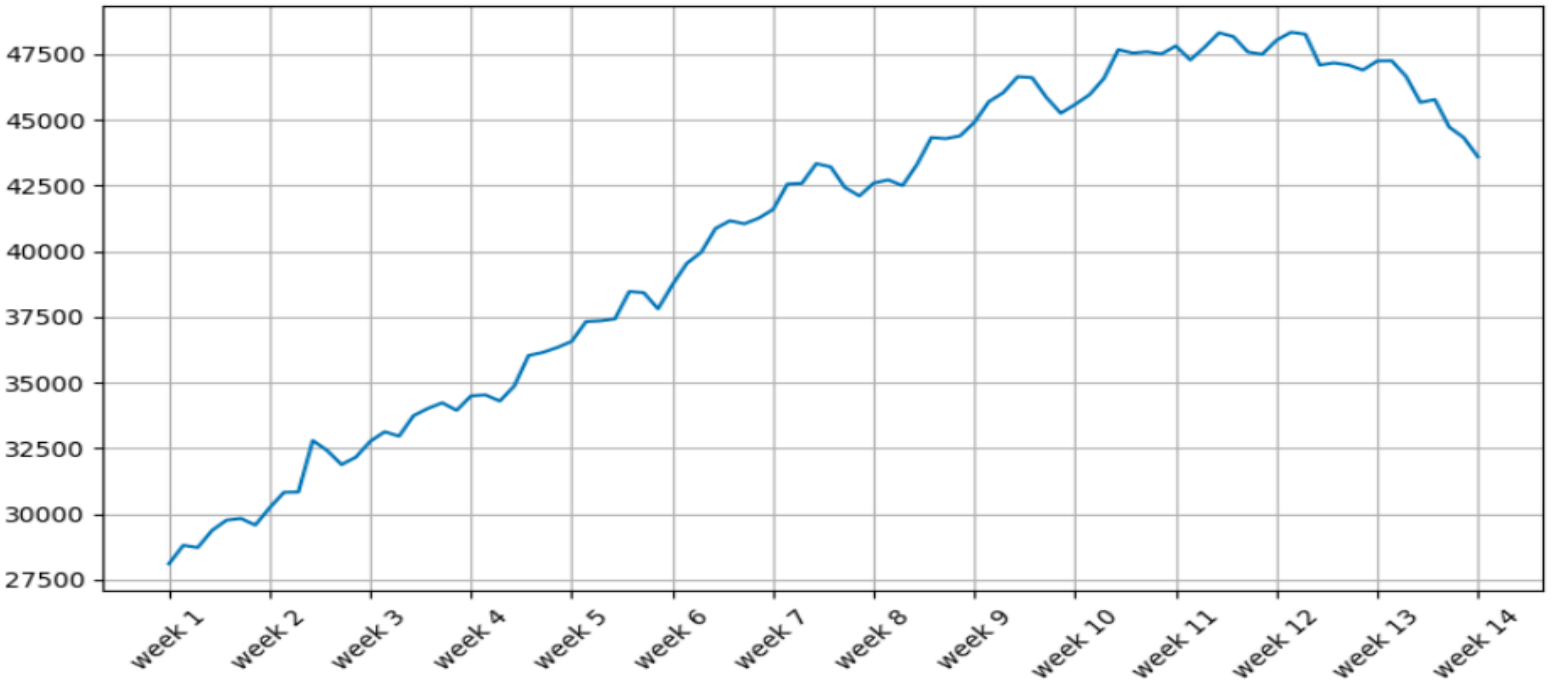
	product_id	block_id
0	612967398	0
1	296892108	0
2	139541214	0
3	963923934	0
4	938230141	0
5	172045154	0
6	663552768	0

	date_number	product_id	category_id	position
0	0	4450020	4461548	17
1	0	42147334	4461548	4
2	0	81131830	4461548	35
3	0	84035833	4461548	38
4	0	125252584	4461548	39

```
Number of positions :954
Number of products :15238
Number of blocks :2776
Number of family :84
Number of subfamily :288
Number of days :92
```

Exploratory Data Analysis

Sales Trend



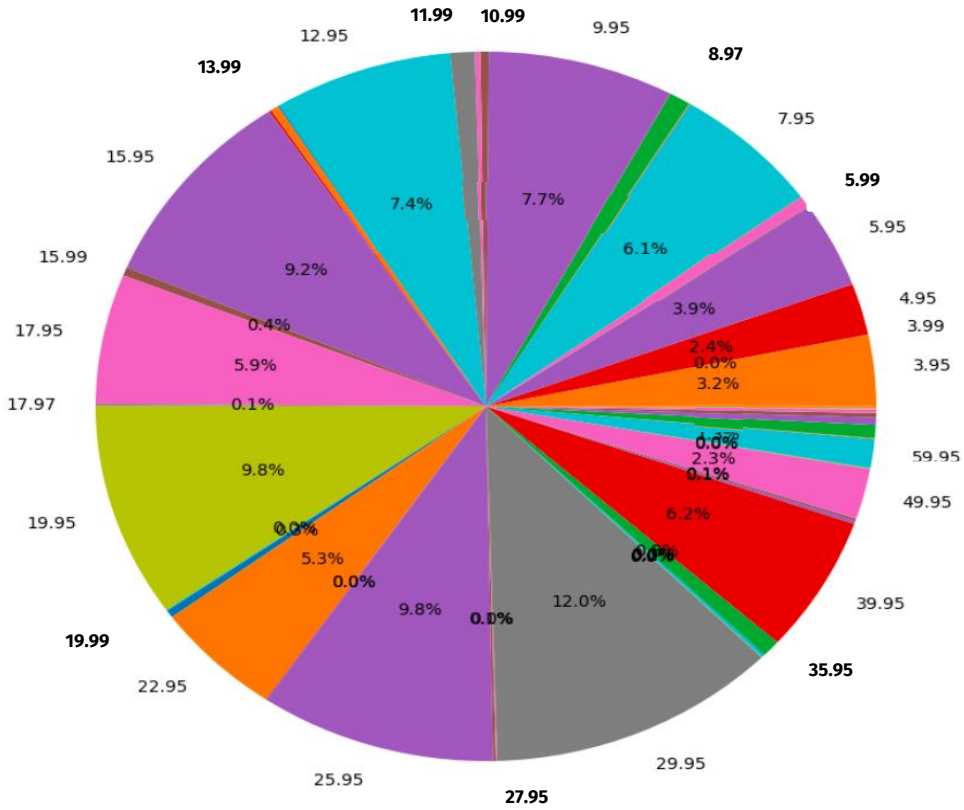
Exploratory Data Analysis

Correlation Map

- Strong correlation between **sales** and **stock**
- No other significant correlation

Exploratory Data Analysis

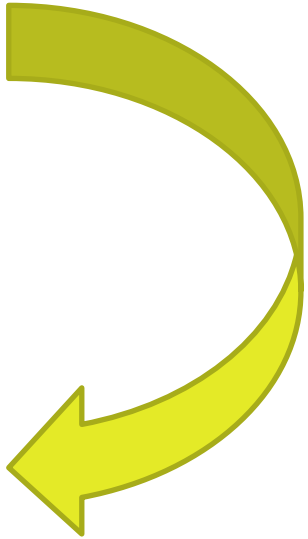
Stock Level by Price



We can see the price points for products that have the **largest stock** are **$29.95** and **$25.95**

Data Transformation & Feature Engineering

	date_number	product_id	category_id	position
0	0	4450020	4461548	17
1	0	42147334	4461548	4
2	0	81131830	4461548	35
3	0	84035833	4461548	38
4	0	125252584	4461548	39



	date_number	product_id	pos_count	pos_min	pos_max	pos_mean
0	0	310130	3	3	80	54.333333
1	0	1178388	1	19	19	19.000000
2	0	1561460	3	3	38	20.000000
3	0	1874414	6	12	190	64.666667
4	0	2094841	4	48	204	130.250000

Data Transformation & Feature Engineering continues

	product_id	block_id
0	612967398	0
1	296892108	0
2	139541214	0
3	963923934	0
4	938230141	0
5	172045154	0
6	663552768	0

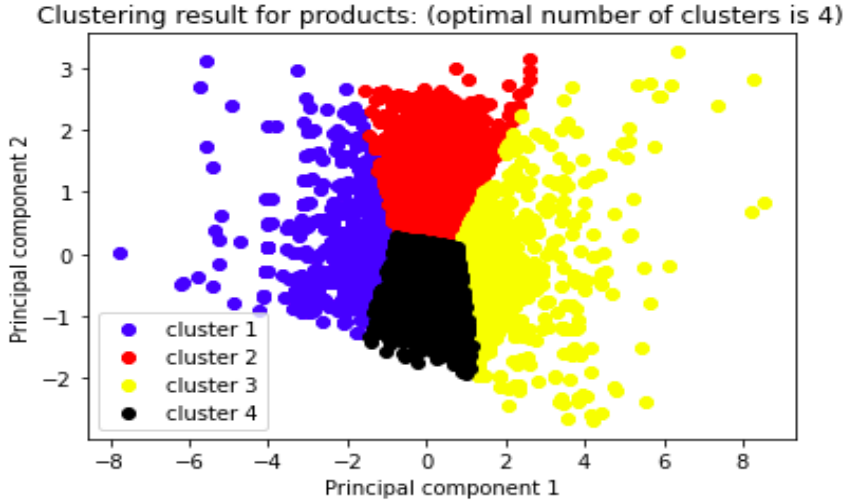
	product_id	block_id	num_items
0	612967398	0	7
1	296892108	0	7
2	139541214	0	7
3	963923934	0	7
4	938230141	0	7

Data Transformation & Feature Engineering continues

	date_number	product_id	sales	stock
0	0	310130	11	461
1	0	1178388	0	60
2	0	1561460	7	791
3	0	1874414	4	281
4	0	2436420	0	245

- As we are interested in product level, we removed the **color_id** and **size_id** columns. Then aggregated the sales and stock columns by product only.

Data Transformation & Feature Engineering continues



	product_id	num_items	stock	price	pos_count	cluster_labels	principal_component_1	principal_component_2
0	310130	8	341	12.95	1	3	0.025018	-0.161967
1	1561460	9	448	29.95	3	3	0.092808	-0.757716
2	1874414	3	246	25.95	6	1	1.280072	1.035497
3	2644529	4	201	7.95	2	1	0.564468	1.178173
4	3176725	9	331	29.95	3	3	-0.057877	-0.797549

Data Transformation & Feature Engineering continues

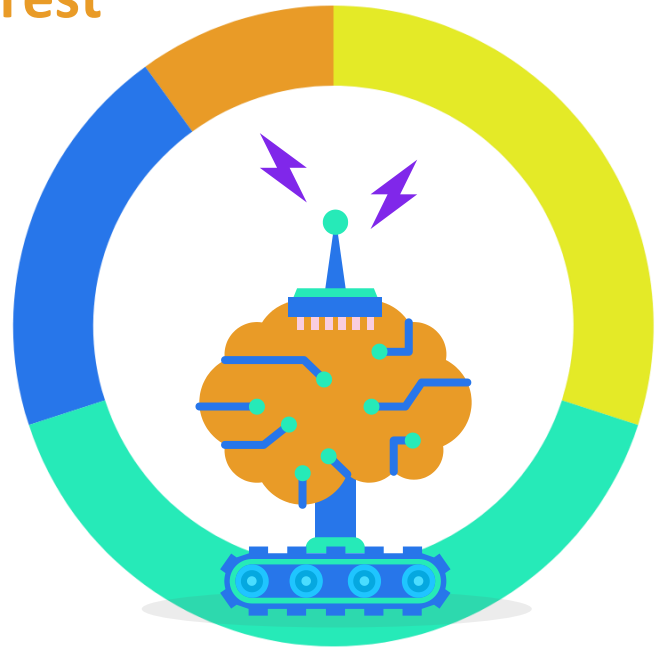
	date_number	product_id	sales	stock	pos_count	pos_min	pos_max	pos_mean	block_id	num_items	family_id	subfamily_id	price	cluster_labels
0	0	310130	11	461	3	3	80	54.333333	1726	8	679611953	533441312	12.95	3
1	1	310130	13	437	3	2	3	2.666667	1726	8	679611953	533441312	12.95	3
2	2	310130	14	435	3	2	8	6.000000	1726	8	679611953	533441312	12.95	3
3	3	310130	15	410	3	4	8	6.666667	1726	8	679611953	533441312	12.95	3
4	4	310130	17	538	3	4	8	6.666667	1726	8	679611953	533441312	12.95	3

Machine Learning Models

Random forest

Deep Learning

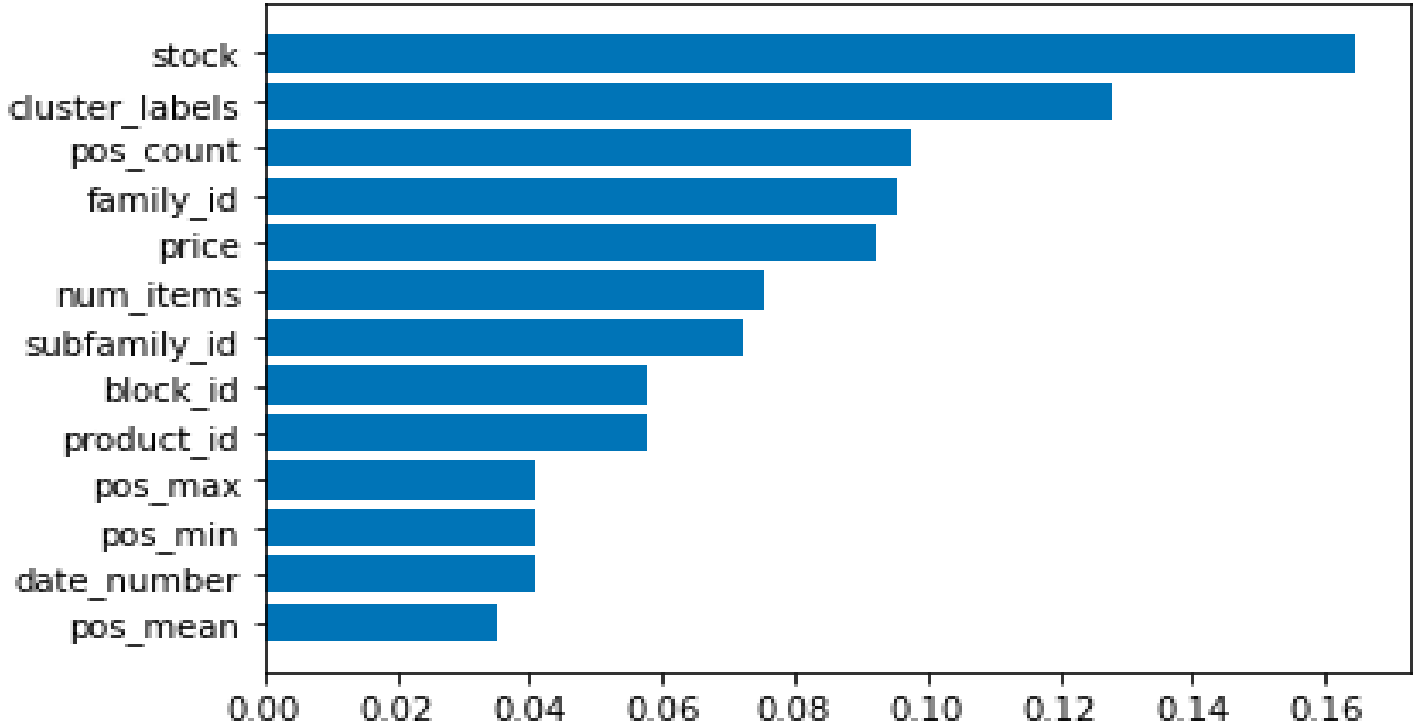
XG Boost



Experiment Results

	R score
Random Forest n_estimators=100, random_state=42	0.40
XGBoost experiment 1: n_estimators=100, max_depth=7	0.50
XGBoost experiment 2: n_estimators=200, max_depth=10	0.59
XGBoost experiment 3: n_estimators=300, max_depth=10	0.60
XGBoost with fewer features experiment 4: n_estimators=300, max_depth=10	0.61
Deep learning -2 hidden layers and a total of 50 nodes, 1 output layer -Activation='relu' , loss='mse', optimizer='rmsprop', epoch =10	-3.340532347850811e-06

The Feature Importance of the XG Boost 3rd Experiment



Accuracy for top 100 products (by revenue) predicted vs actual

Day	Accurately Predicted Percentage
85	66%
86	61%
87	50%
88	49%
89	45%
90	45%
91	51%

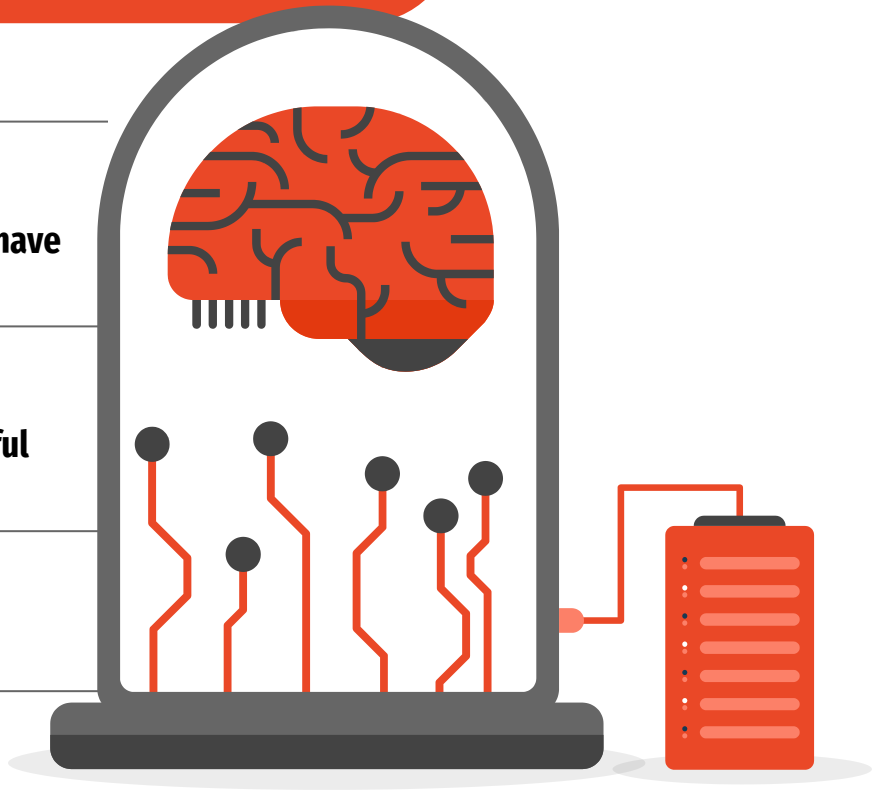
Limitations

01 R score is 0.61 and there is a room to improve

02 Did not include color and size features which might have an impact on the model performance

03 Data confidentiality limitation (did not provide useful insights to use in a real business world)

04 Only 3 months data



Conclusion

01 **EDA :** EDA showed us the complexity of the data. More specifically, the sudden change in the trend and scale of features, etc. This makes the model prediction more difficult.

02 **Feature Engineering :** Added new features like cluster labels (2nd), position count (3rd), number of items (5th) were shown as very important features. (Total 13 features)

03 **Model Training :** XGBoost provided the best performing model and Deep Learning model was the worst performing model. Random Forest was the medium performing one. Logistic Regression and linear regression will not work for our case.

04 **Prediction :** About Top 100 best performing products, we predicted 50-60% accurately. The further forecasting horizon, the lower the accuracy.