110th VitrA Data Science Bootcamp
Pharma Sales Prediction Project

Project Group Members:

- Erkan Çetinyamaç
- Yağmur Uzun
- Arda Can Baysar
- Eren Kaya
- Melike Demirdağ

Business Understanding

Data Understanding

Feature Engineering

Data Visualization

Data Pre-Processing

Modelling

Conclusion

Business Understanding

Since predicting to future sales holds crucial importance for a company to decide how to take action about production and purchasing operations, it is vital to build a model to predict next sale amounts in order to company can benefit from this significant intel.

Our project group aims to build a model that can predict next sale amounts.

Data Understanding

	Year	Period	Product	Province	Quantity
46891	2018	201811	PRODUCT_X	KAYSERİ	479
11994	2018	201811	PRODUCT_A	ANTALYA	63
17366	2019	201905	PRODUCT_A	RİZE	16
40389	2018	201807	PRODUCT_C	ZONGULDAK	-4
18364	2017	201703	PRODUCT_B	KAHRAMANMARAŞ	160
15687	2019	201911	PRODUCT_A	ISTANBUL	15
22169	2017	201709	PRODUCT_B	ANKARA	30
36818	2017	201705	PRODUCT_C	ADANA	72
16574	2019	201912	PRODUCT_A	ANKARA	17
21618	2017	201711	PRODUCT_B	ISTANBUL	435

Fig. 1 – Raw Data

➔ Product:
- Product A : Chronic Gastroenterology, Sales volume proportional to the number of patients.
- Product B : Acute: Painkiller.
- Product C : Acute : Digestive System, for children 0-4 years old.
- Product V : Vitamin.
- Product X : Chronic: Urology, Patient group with a high average age.

➔ Period and Year Features: Dates for the sales.

➔ Province: The city information that product was sold.

➔ Quantity: The number of drug that was sold.

Feature Engineering

Region	Season	Metropol	USD-TL	TUFE_Annual_Change	TUFE_Monthly_Change	Total_Sale_Volume	Quantity_M3	Quantity_M6	Quantity_M9	Quantity_M12
Karadeniz	spring	0	3.672548	11.29	1.02	3	7.0	20.0	97.531593	99.074121
Karadeniz	autumn	1	5.741548	10.56	0.38	3	14.0	-5.0	22.000000	46.000000
Marmara	winter	1	5.848150	11.84	0.74	3	39.0	136.0	57.000000	32.000000
Karadeniz	summer	1	5.634906	15.01	0.86	3	77.0	24.0	22.000000	41.000000
İc_Anadolu	winter	1	5.848150	11.84	0.74	5	13.0	18.0	15.000000	16.000000

Fig. 2 – New Features

Web Page of the Data Source:

1) https://www.tcmb.goconnect/TR/TCMB+TR/Main+Menu/Istatistikler/Enflasyon+Verileri/Tuketici+Fiyatlariv.tr/wps/wcm/
2) https://www.tcmb.gov.tr/wps/wcm/connect/tr/tcmb+tr/main+menu/istatistikler/doviz+kurlari/reel+efektif+doviz+kuruu

Graph - Total Quantity and TUFE Annual Change

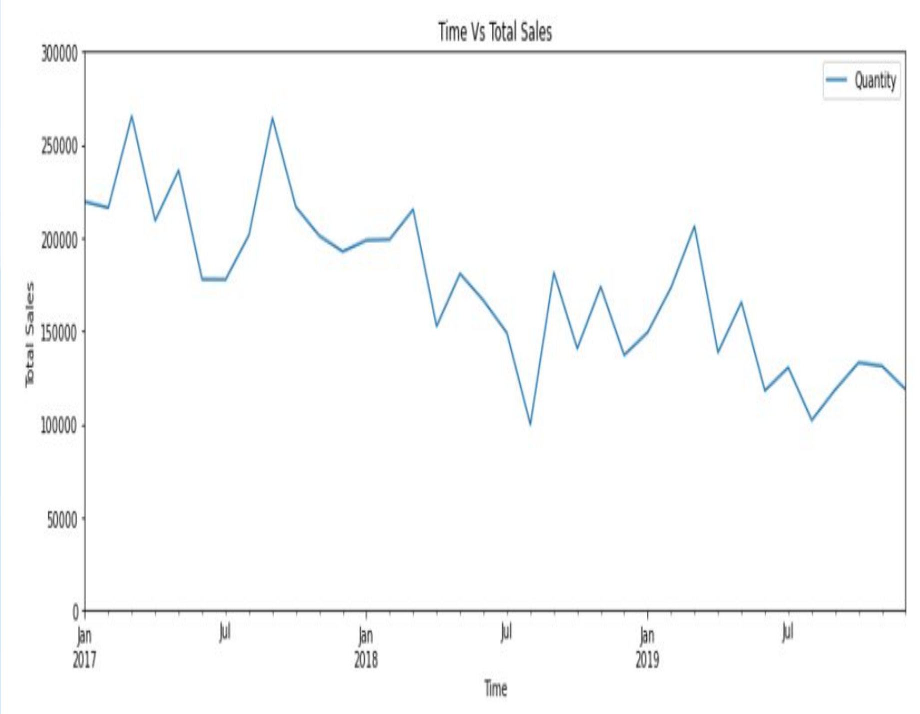


Fig. 3 – Total Quantity vs Time Line Plot

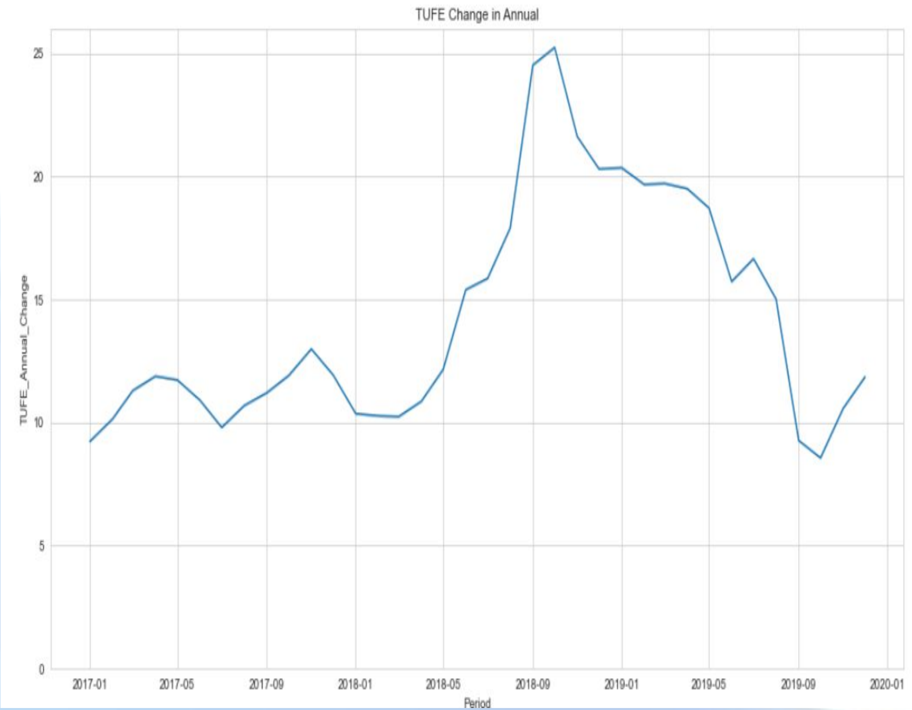
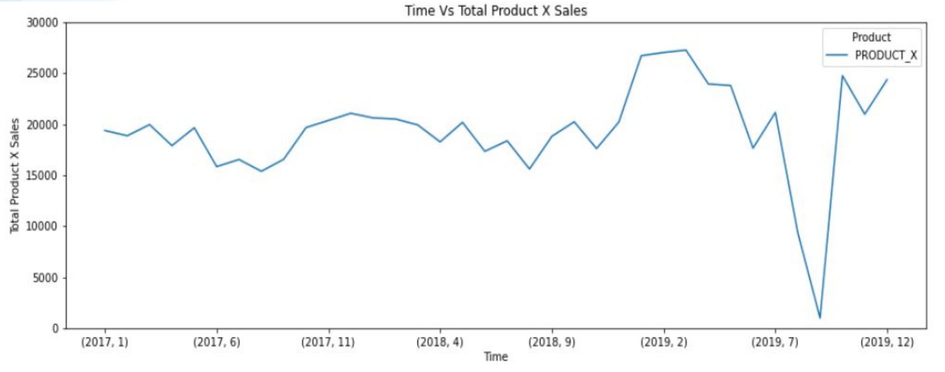
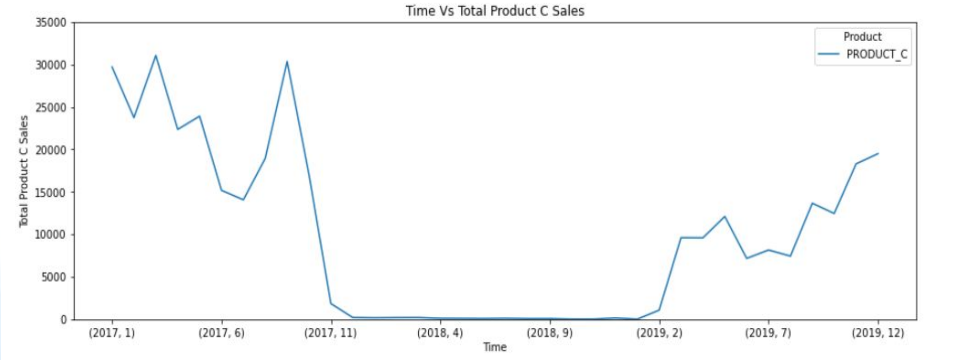
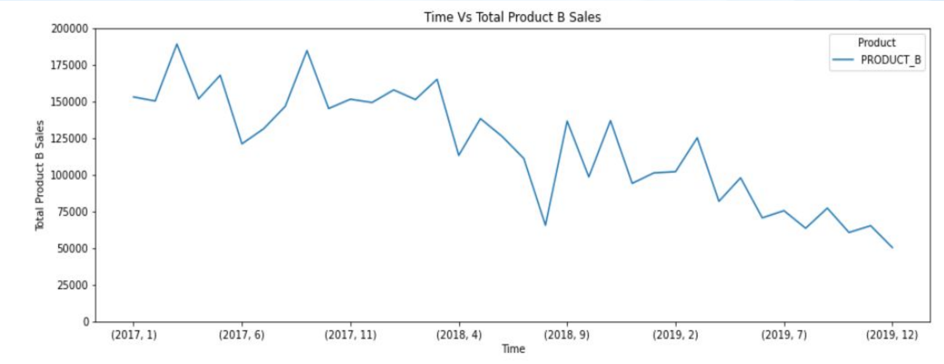


Fig. 4 – Annual TUFE Change vs Time Line Plot

Graph - Product vs Time







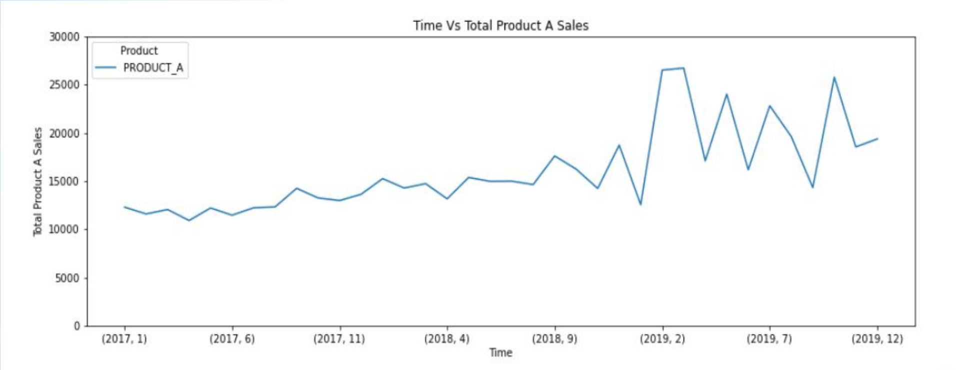


Fig. 5, 6, 7, 8 – Product Types vs Time

Graph - Metropol vs Anadolu Product Sales Comparison

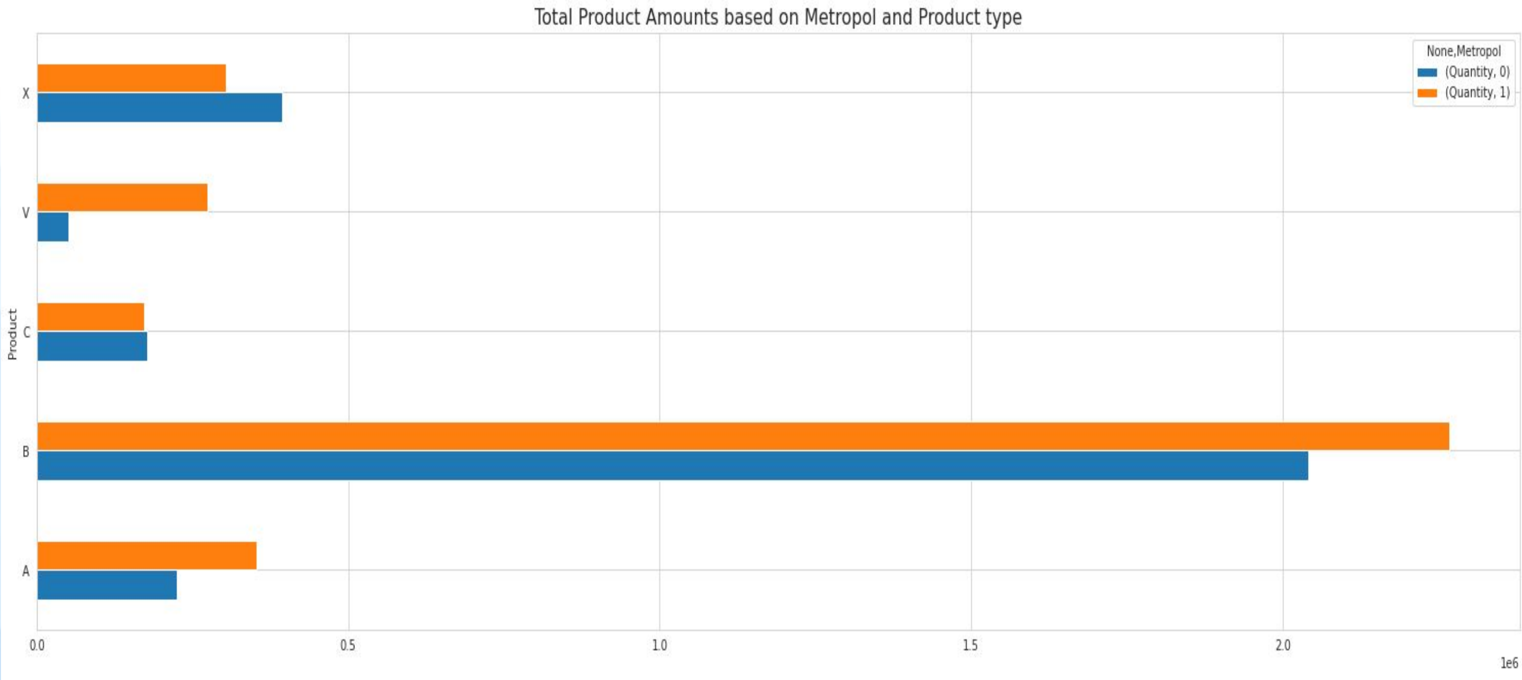


Fig. 9 – Product Sales of Metropol and Anadolu Comparison Bar Plot

Graph - Sales With Respect To Regions

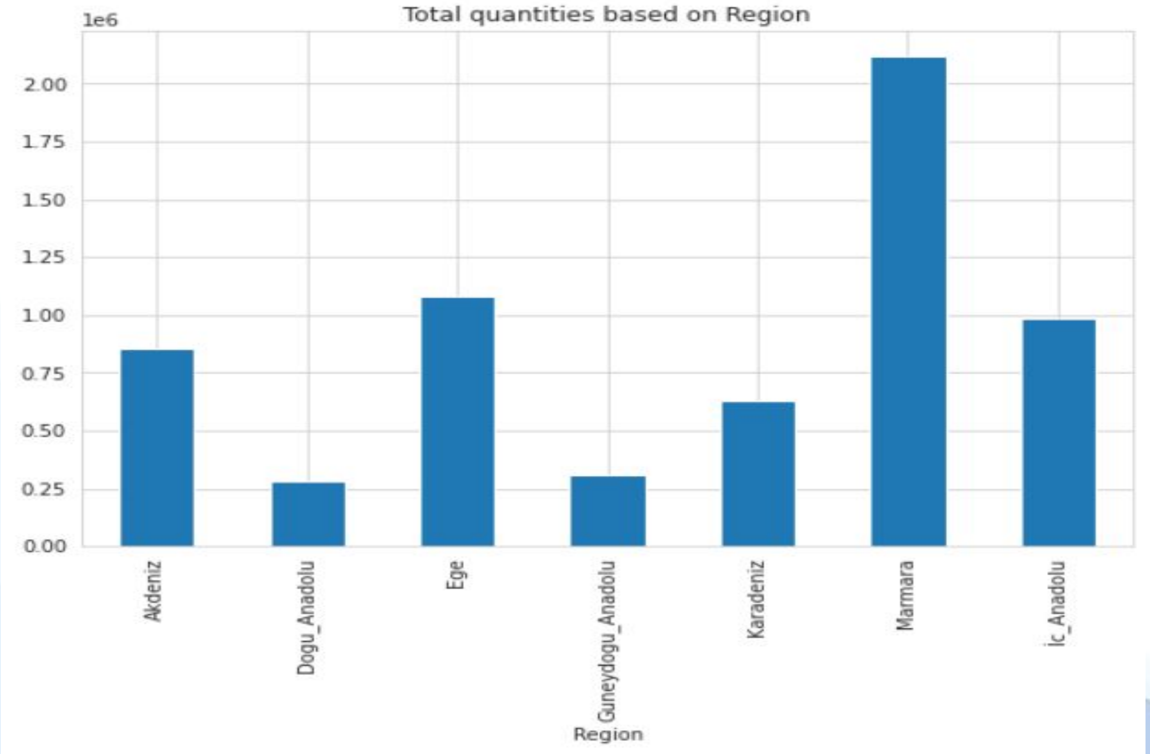


Fig. 10 – Total Sales vs Regions Bar Plot

Graph - Product Sale Quantities vs Season

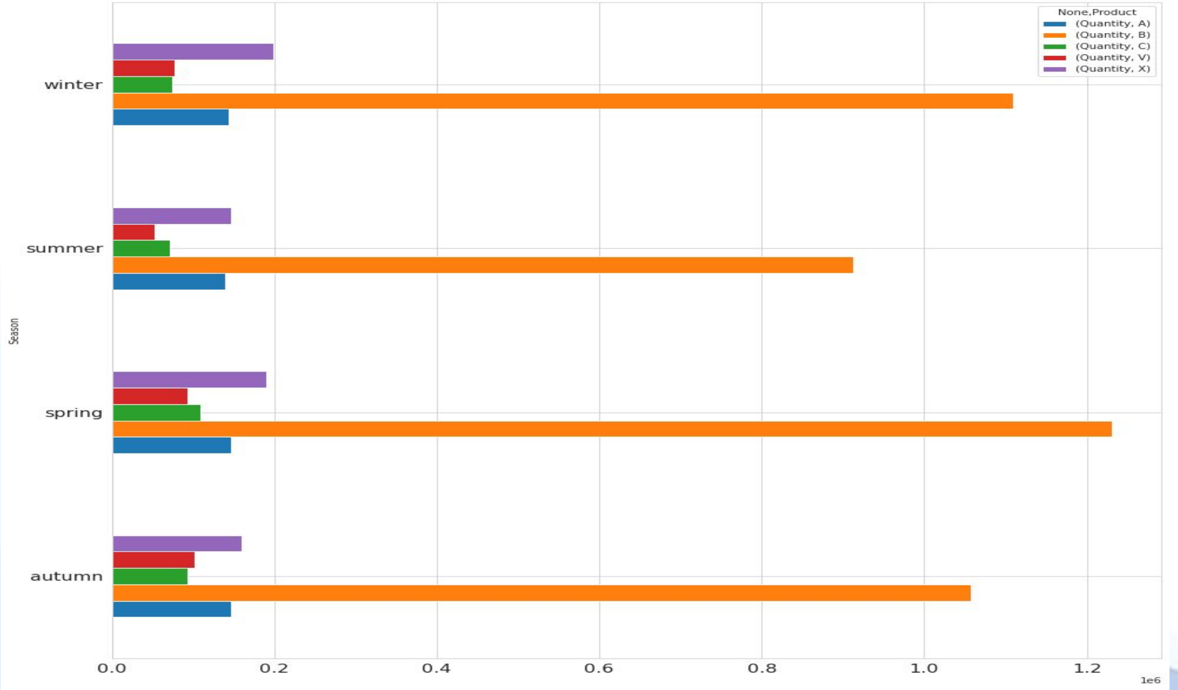


Fig. 11 – Sale Quantities vs Season Bar Plot

Libraries

pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$

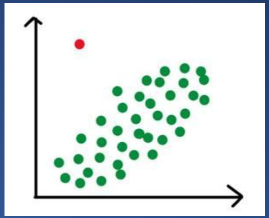
matplotlib

scikit learn

seaborn

NumPy

Pre-Processing



Target feature 'Quantity' is highly skewed and has many outliers in it. Also since minus values of quantity feature that represents refund information therefore we should deal with them before building ML models by log10 transforming the target and IQR Method for outlier elimination.

Before the Process

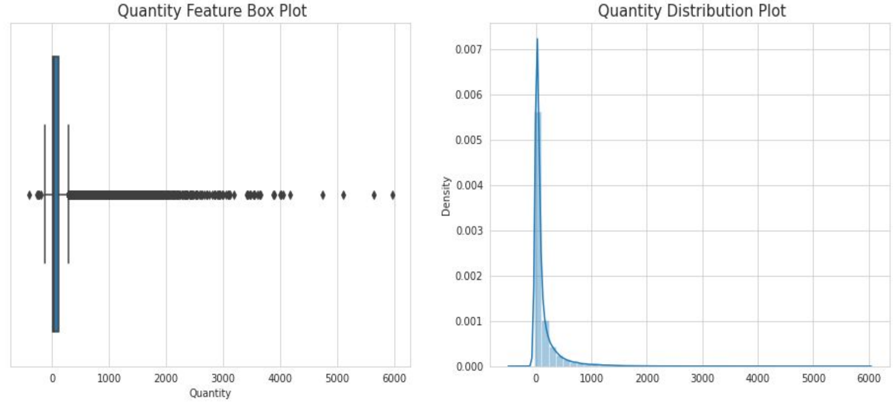


Fig. 12 – Outliers and Skewness Before the Process

After the Process

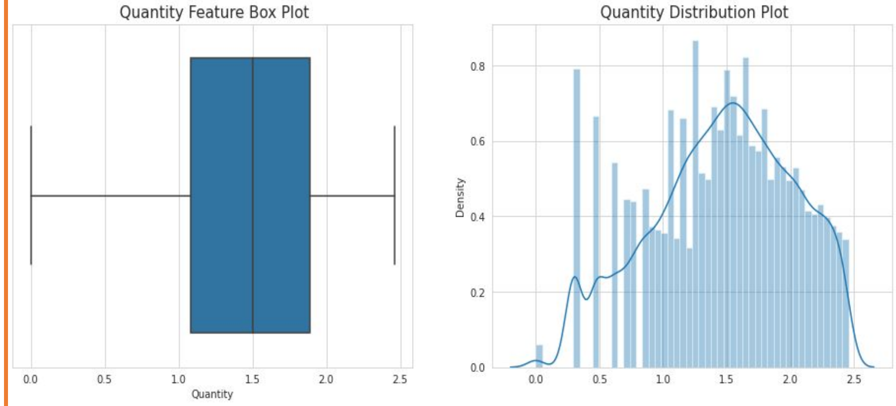


Fig. 13 – Outliers and Skewness After the Process

Pre-Processing

Due to creating new features using by ‘Quantity’ column we got NA values in the ‘Quantity’ feature which was not originally in the data. We thought that since the target feature is right skewed, imputing NA values with mean is more logical choice than imputing median here.

The categorical features that are ‘Product’, ’Region’, ’Season’ are encoded with dummy encoding method.


```
X = pd.concat([pd.get_dummies(df[["Product","Region","Season"]],drop_first=True)
```

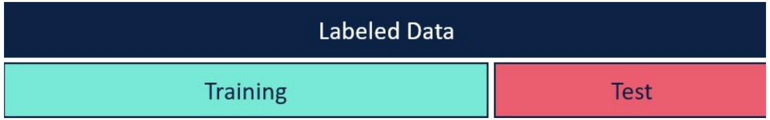
	Product_B	Product_C	Product_V	Product_X	Region_Dogu_Anadolu	Region_Ege	Region_Guneydogu_Anadolu	Region_Karadeniz	Region_Marmara	Region_İc_Anadolu	Season_spring	Season_summer	Season_winter
875	0	0	0	0	0	0	0	0	1	0	0	0	1
18157	1	0	0	0	0	1	0	0	0	0	0	0	1
25122	1	0	0	0	1	0	0	0	0	0	0	0	0

Fig. 14 – Dummy Encoding to Categorical Features

VitrA KODLUYORUZ Patika.dev

Model Building and Comparison

Train/Test Split



- Split with respect to the date.
- Split with respect to sample.

	CastBoost	XGBoost	KNN	MLP	RandomForest	GBM	LinearRegression	Ridge	Lasso	ElasticNet
R^2:	0.34	0.34	0.21	0.37	0.33	0.31	0.06	0.06	-0.69	-27.24
Adjusted R^2:	0.34	0.34	0.21	0.37	0.32	0.30	0.06	0.06	-0.69	-27.29
MAE:	31.53	31.58	35.77	35.15	31.91	32.45	36.65	36.65	40.00	49.66
MSE:	2711.55	2723.78	3250.28	2606.33	2771.98	2856.95	3855.82	3855.84	6931.01	116113.40
RMSE:	52.07	52.19	57.01	51.05	52.65	53.45	62.10	62.10	83.25	340.75
MAPE:	98.52	98.01	131.12	191.85	101.10	101.74	117.23	117.23	133.36	166.36
CoV:	0.91	0.92	1.00	0.90	0.92	0.94	1.09	1.09	1.46	5.97
Explained Variance:	0.39	0.39	0.26	0.37	0.38	0.37	0.14	0.14	-0.59	-27.19

Fig. 15 – Regression Models Comparison Table

Model Building For Each Product

After the GridSearchCV and Model tuning process, the models that below returns slightly better scores compared to other regression models for predicting only the product X quantity amount.

Testing The Previous Model With Only Product X Test Data

	CastBoost	XGBoost	GBM
R^2:	0.70	0.67	0.69
Adjusted R^2:	0.69	0.66	0.68
MAE:	25.98	27.31	26.33
MSE:	1567.67	1732.50	1640.00
RMSE:	39.59	41.62	40.50
MAPE:	54.63	55.43	54.97
CoV:	0.45	0.48	0.46
Explained Variance:	0.71	0.68	0.70

Fig. 16 –Previous Model's Evaluation Metrics For Product X Test Data

Testing New Trained Model For Only Product X Data

	CastBoost	XGBoost	GBM
R^2:	0.80	0.80	0.79
Adjusted R^2:	0.80	0.79	0.79
MAE:	41.09	42.14	42.22
MSE:	4785.73	4905.61	5045.73
RMSE:	69.18	70.04	71.03
MAPE:	57.79	59.24	59.21
CoV:	0.43	0.43	0.44
Explained Variance:	0.81	0.80	0.80

Fig. 17 – Evaluation Metrics For the Trained Model With Only Product X

Grouping by the Month and Predicting Monthly Total Sales of All Products

	RandomForrestRegressor	CatBoostRegressor	XGBoostRegressor	GBM
R^2:	0.90	0.89	0.87	0.90
Adjusted R^2:	0.86	0.84	0.81	0.86
MAE:	8836.32	8525.69	9093.57	8454.38
MSE:	200764305.41	228530762.56	271664174.62	193616039.10
RMSE:	14169.13	15117.23	16482.24	13914.60
MAPE:	63.41	54.96	51.21	58.45
CoV:	0.40	0.42	0.46	0.39
Explained Variance:	0.90	0.89	0.87	0.91

Fig. 18 – Hyperparameter Tuned Model Comparison Table

Time Series Analysis with ARIMA

We used Univariate Time Series Forecasting to predict sales by using arima package in python. Sales quantities were grouped by month then processed.

Best model parameters for product V: ARIMA(4,1,0)(1,1,0)[12]

	Arima_X	Arima_B	Arima_C	Arima_A	Arima_V
mape	3.21	0.32	0.40	0.19	0.57
me	2,468.05	-13,839.50	-6,261.17	676.03	-2,783.64
mae	5,833.34	20,477.83	6,458.97	3,448.00	2,783.64
mpe	3.06	-0.22	-0.37	0.07	-0.57
rmse	8,123.85	25,788.08	8,700.83	4,350.32	3,176.24
corr	0.67	0.51	-0.61	-0.26	0.92
evs	0.22	-4.74	-0.75	-0.45	0.82
r2	0.14	-7.06	-2.63	-0.49	0.23

Fig. 19 – Error Metrics for Arima Predictions

Time Series Analysis with ARIMA

Best prediction performance is achieved for Product V.

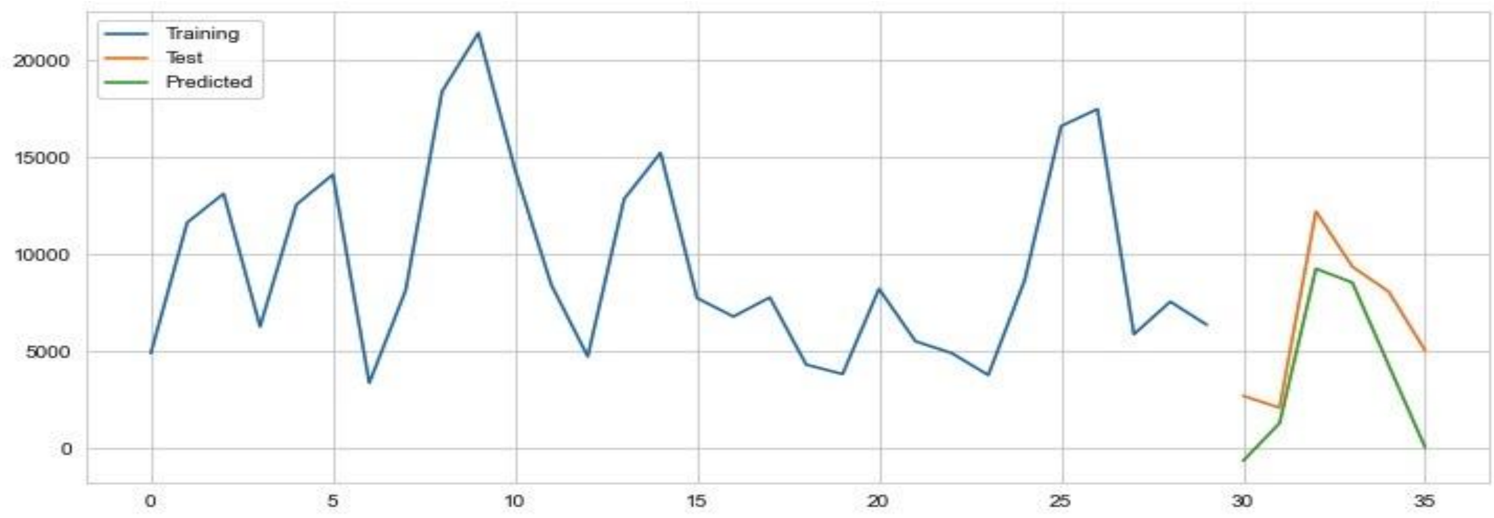


Fig. 20 – Train, Test and Prediction Plot for Arima

Conclusion

- Model for all product at once
- Models for each product
- Model for all product after group by period and aggregate sum of quantity
- Time series analysis(ARIMA) for each product

We achieved the best prediction performance with grouping by month for all products.

To-Do List

The next steps would be;

➔ Grouping the data by Province and make prediction.

➔ Building a separate model for high sale volume products.

➔ And building another model for low sale volume products

➔ Local and Global explainability with SHAP library.

THANK YOU FOR YOUR ATTENTION!