DATA GENIUS: Machine Learning Engineer Test

Ngoc - Tram Nguyen

Developing Product Sales Prediction Model

Test Case Study for ML Engineer: Dr. Thêm Official Store on Shopee

+ ***Data Description*****:** Sales data over a three-week period (10/03/2024 – 31/03/2024)

+ **Issues**: Small dataset size, missing information, limited data fields.

+ Requirements: Develop a model that ***predicts the number of products sold*** in the upcoming periods.

PREPROCESSING

.

- Find and fill missing value (discount)
- Feature selection: Remove features not useful (shopid, shop_location, status, name)
- Feature engineering: Create new feature from exist features (revenue)
- Convert to datetime format for time feature (Date)

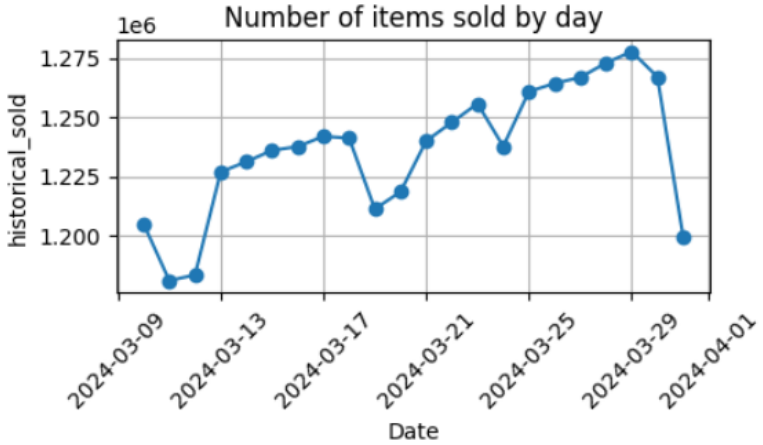
Goal: Clean Data

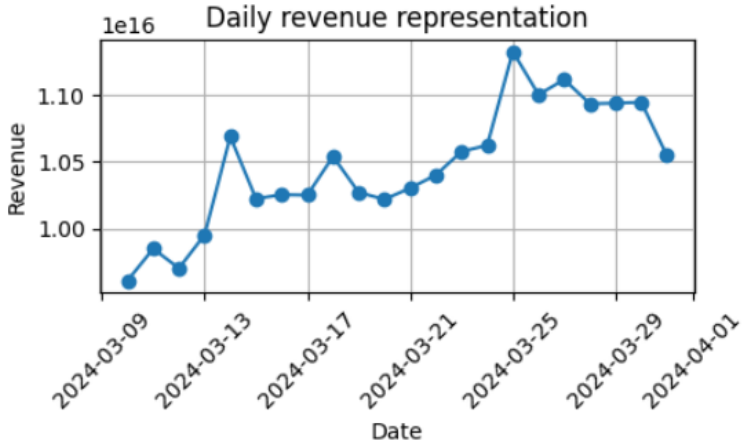
EDA (Exploratory Data Analysis)

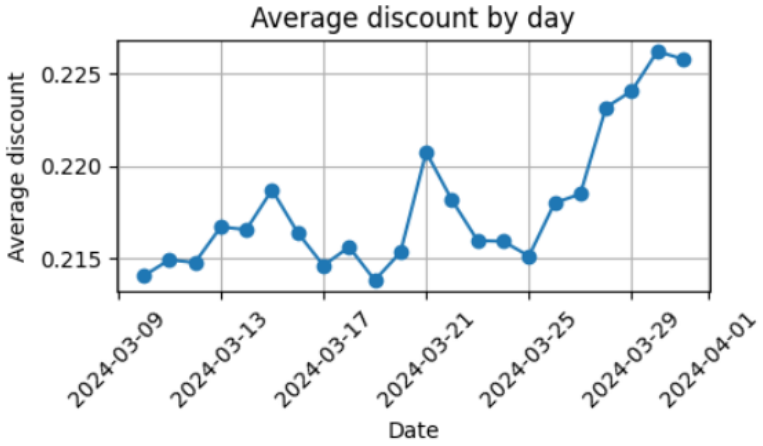
Overview of the distribution of data

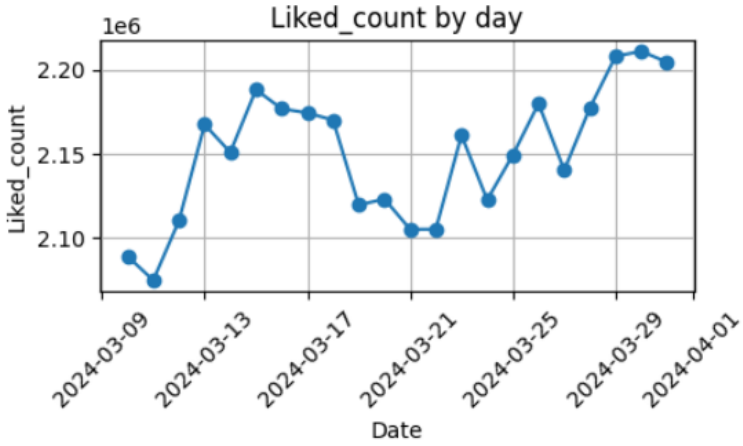
	itemid	liked_count	cmt_count	shop_rating	historical_sold	price	rating_count	rcount_with_context	Date
count	3.867000e+03	3867.000000	3867.000000	3867.000000	3867.000000	3.867000e+03	3867.000000	3867.000000	3867
mean	1.466942e+10	12234.599948	2032.896302	4.931021	7035.889061	2.451592e+10	2032.798293	1210.578485	2024-03-20 16:10:47.944142592
min	1.985670e+09	0.000000	0.000000	4.930905	0.000000	9.000000e+08	0.000000	0.000000	2024-03-10 00:00:00
25%	5.481377e+09	130.000000	55.000000	4.930969	225.000000	1.125000e+10	55.000000	25.000000	2024-03-15 00:00:00
50%	1.823627e+10	16565.000000	198.000000	4.931027	673.000000	2.050000e+10	198.000000	107.000000	2024-03-21 00:00:00
75%	2.215632e+10	18615.500000	909.500000	4.931054	3643.500000	3.490000e+10	909.500000	494.000000	2024-03-26 00:00:00
max	2.591996e+10	71117.000000	106400.000000	4.931238	322223.000000	1.199000e+11	106413.000000	64046.000000	2024-03-31 00:00:00
std	8.170483e+09	10035.877379	8781.247212	0.000066	27425.013404	1.760156e+10	8782.279788	5334.233921	NaN

EDA (Exploratory Data Analysis)









EDA (Exploratory Data Analysis)

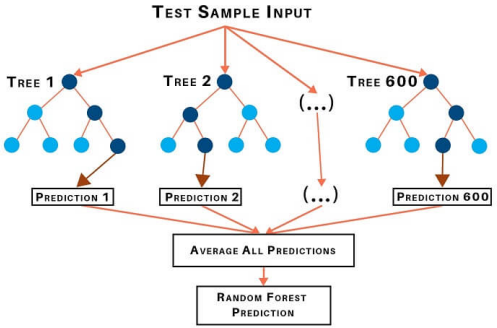
Top 5 Most Sold Items

	items	quantity	price
0	2421653980	7043524	9.500000e+09
1	5451541710	2724989	1.270000e+10
2	10001549800	1576811	1.090000e+10
3	23232932577	1318263	6.790000e+09
4	16870222597	1140525	4.200000e+09

Top 5 Least Sold Items

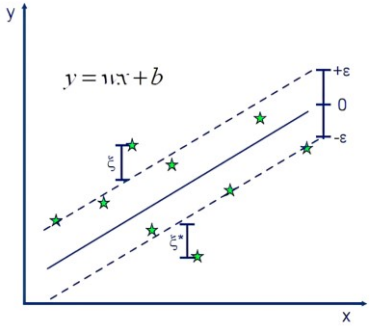
	items	quantity	price
0	25919955014	0	6.500000e+10
1	15689885742	0	7.700000e+10
2	24070912680	0	1.990000e+10
3	25617038549	0	5.000000e+10
4	24320863474	0	1.990000e+10

PROPOSE MODEL AND EVALUATE



Random Forest Regression

```
Train RMSE: 28.882367353420427
Test RMSE: 58.13478817586936
Mean Squared Error: 3379.6535962532
```



- Minimize:

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*)$$

- Constraints:

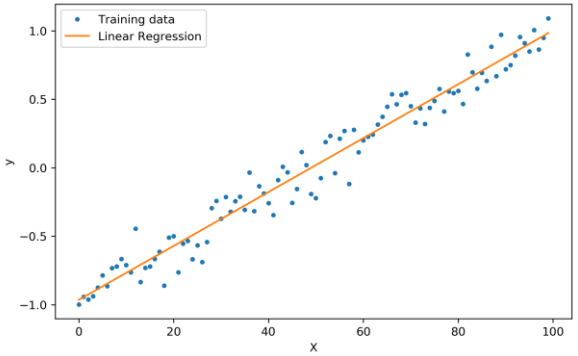
$$y_i - wx_i - b \le \varepsilon + \xi_i$$

$$wx_i + b - y_i \le \varepsilon + \xi_i^*$$

$$\xi_i, \xi_i^* \ge 0$$

Support Vector Regression

```
Mean Squared Error: 885733717.3245064
Train RMSE: 27729.12107954742
Test RMSE: 29761.27882542191
```



Linear Regression

```
Train RMSE: 9.717994879822407e-11
Test RMSE: 1.0465745881020919e-10
```

CHECK OVERFITING

```
Cross-Validation RMSE Scores: [6.91524811e+02 2.12237437e-10 3.35303403e-11 5.51623611e-11
 4.48795140e+02]
Mean RMSE: 228.06399033514867
```

=> REGULARIZATION TECHNIQUE

Ridge Regression (L2 regularization)	Lasso Regression (L1 regularization)
RMSE = 0.033	RMSE = 0.030

CONCLUSION

- Small data should use machine learning models for forecasting is reasonable.
- Perform preprocessing to clean data and EDA to visualize data effectively.
- Among three recommended methods (Random forest Regression, SVR and Linear Regression), Linear Regression has the best forecast results.
- Implement overfitting prevention techniques with regularization technique (Ridge and Lasso Regression).
- Besides, some technique such as feature selection, feature important or add data from review data (26/03/2019-16/04/2024) also can improve model learning performance.

THANK YOU FOR LISTENING