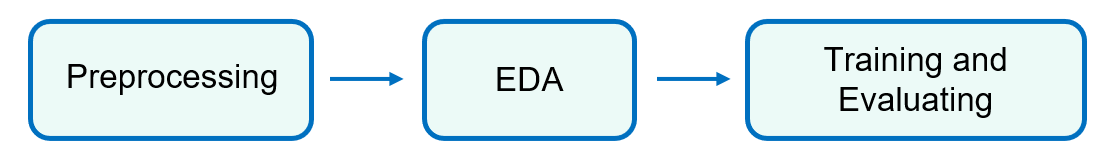
**DATA GENIUS: Machine Learning Engineer Test**

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.

**SOLUTION**



1. **Preprocessing to get clean data**

* Identify and handle missing values: Calculate % missing value in dataframe to fill missing values or remove rows or columns with missing data.
* Feature selection: Remove feature that are not useful for modeling process = > Reduce dimensionality and improve model performance.
* Feature engineering: Create new features from existing ones that may provide additional information for the model.

Example: revenue = history\_sold\*price\*(1-discount)

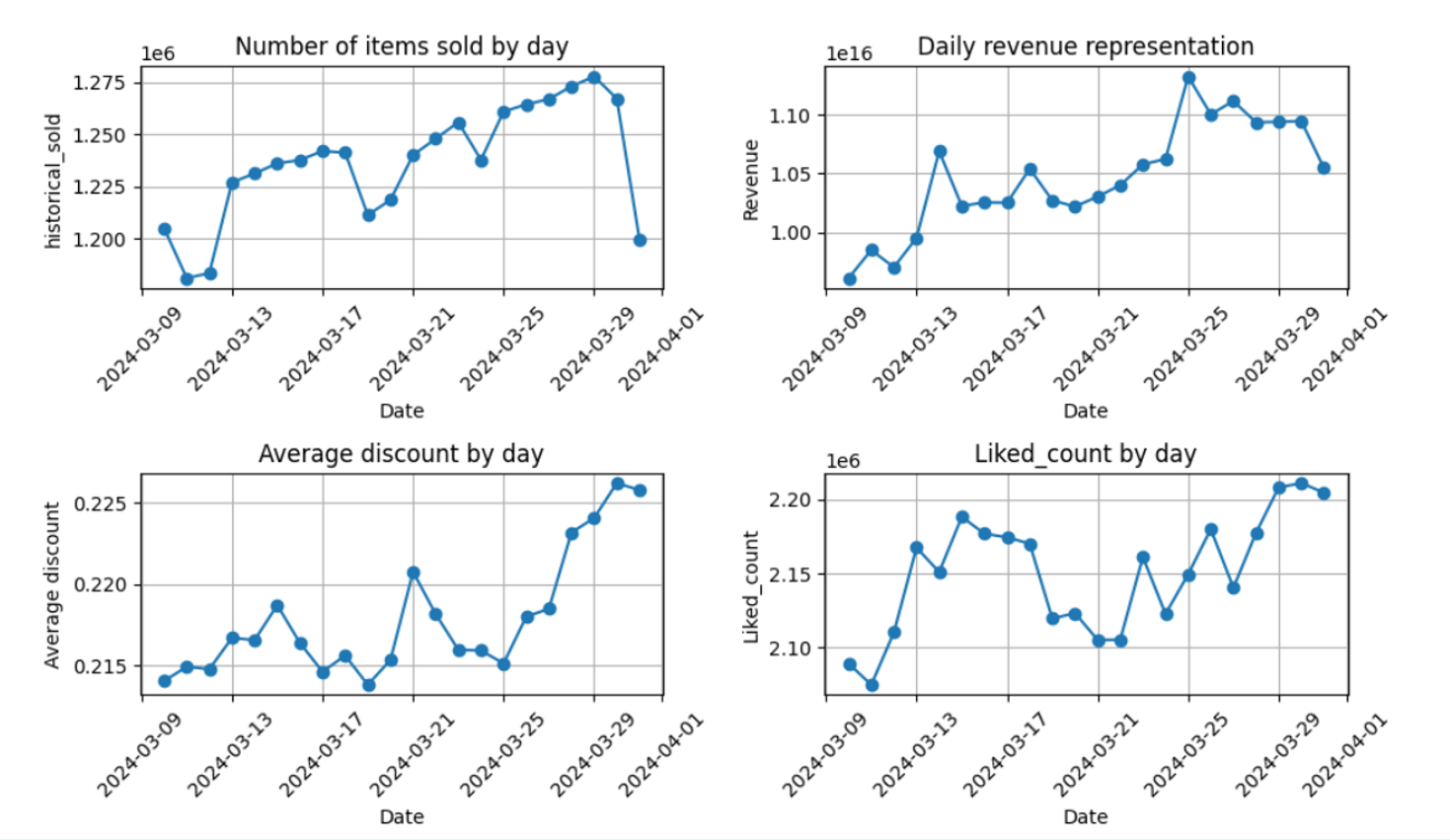
* Convert relevant feature to datetime format for easier analysis.

1. **Exploratory Data Analysis (EDA) to visualize the data**

* Visualize distribution of data with descriptive statistics:



* Understanding distribution and identify potential outliers or anomalies.
* Visualize Daily sale metrics:



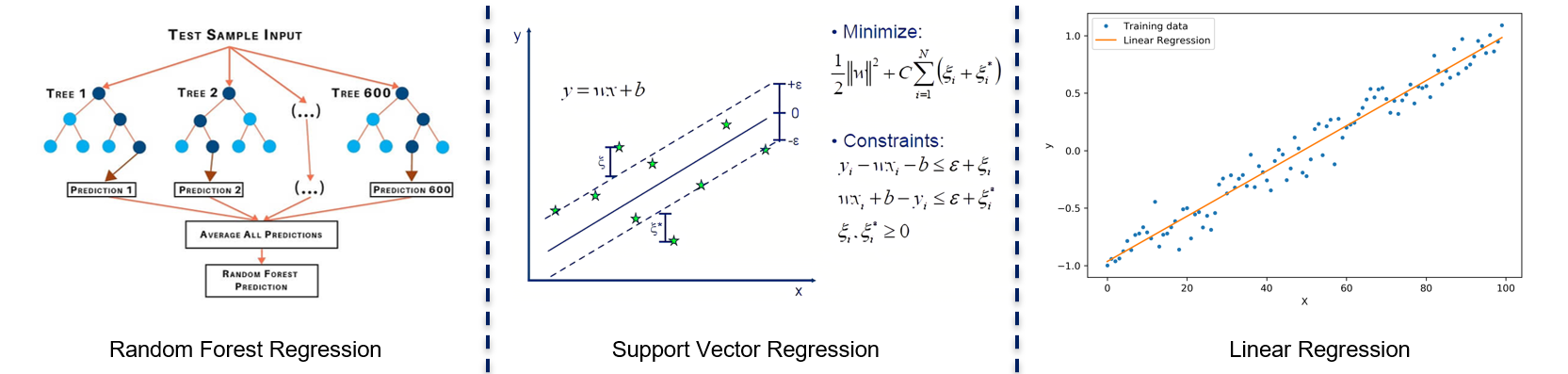
= > Promotional discounts on Shopee often occur during the last days of the month, particularly from the 25th to the 27th. As a result, there is typically a noticeable increase in the number of items sold, revenue, discount rates, and liked counts on these days. Conversely, the beginning of the month tends to experience lower activity.

* Visulize top 5 most sold items and top 5 least sold items:



= > By implementing targeted advertising campaigns and strategies tailored to both the top-selling and least-selling items, we can optimize sales volume and maximize overall revenue for the business.

1. **Proposing and training three models, follow model evaluation and addressing overfitting issues**

****

**3.1 Random Forest Regression**

|  |  |
| --- | --- |
| Training Result | Feature important |
|  |  |

* 1. **Support Vector Regression**

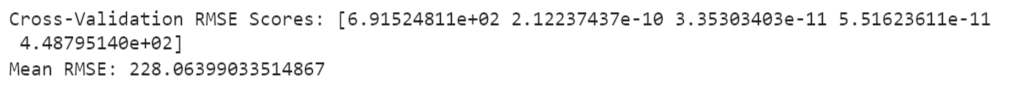
|  |
| --- |
| Training Result |
|  |

* 1. **Linear Regression**

|  |  |
| --- | --- |
| 1. Training Result | Feature important |
|  |  |

=> Linear Regression is better than Random forest regression and SVR.

Using Cross-Validation technique to check overfitting:

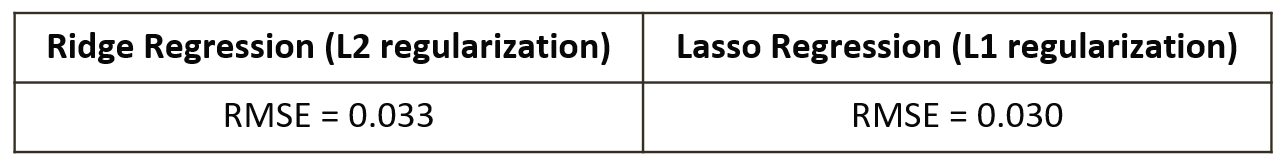


=> Based on the results of RMSE cross-validation, it can be seen that some folds have a very small RMSE (close to 0) while others have RMSE many times larger. This may be a sign of overfitting.

* 1. **Solution to prevent overfitting**

**Regularization techniques:**

Ridge and Lasso regression help prevent overfitting in machine learning models by adding penalty terms to the model's cost function. Ridge regression adds a penalty term proportional to the square of the weights (L2 regularization), while Lasso regression adds a penalty term proportional to the absolute value of the weights (L1 regularization). These penalties help to constrain the magnitude of the weights, reducing the model's complexity and improving its generalization performance on unseen data.

**  
  
Summary:**

* 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.