



MASTER IN DATA SCIENCE

GROUP ASSIGNMENT

COURSE CODE : WQD7005

COURSE TITLE : DATA MINING

GROUP NO : 6

GROUP MEMBER :

1. NUR HIDAYAH BINTI AHMAD SHAFII (22120931)
2. WUDI (22120264)
3. LUCHANGHAO (22120671)
4. HEJUNFENG (22120634)

LECTURER : PROF. DR. NOR LIYANA BT MOHD SHUIB

Table of Contents

1	Data Understanding & Preparation	1
1.1	Data Selection	1
1.2	Data Preprocessing	1
1.3	Exploratory Data Analysis.....	8
2	Modeling and Methodology.....	14
2.1	Target and platform selection.....	14
2.2	Automated modeling process.....	14
2.3	Translation processing and model comparison	19
2.4	Summarize	23
3	Model Evaluation & Interpretation.....	23
3.1	Model performance evaluation.....	23
3.1.1	Data partitioning and verification	24
3.1.2	Model optimization target selection	24
3.1.3	Feature importance analysis	24
3.1.4	Residual analysis	26
3.1.5	Comparison and analysis of predicted values and actual values.....	27
3.1.6	Business significance and model interpretation.....	27
3.2	Model Interpretability	28
3.2.1	Global Interpretability	28
3.2.2	Local Interpretability	28
3.2.3	Business Impact of Model Interpretability	29
3.3	Limitations and Assumptions	29
3.4	Summary	30
4	Conclusion and Recommendations	31
5	Team Collaboration.....	33
6	References	34

1 Data Understanding & Preparation

1.1 Data Selection

The objectives of the project are:

- i. To analyze sales and profit trends across categories, regions, and customer segments.
- ii. To develop and evaluate a predictive model using Google Cloud AutoML to support data-driven business strategies.

The dataset chosen for this project is sourced from Kaggle. This dataset is relevant to our project as it provides comprehensive supermarket data. Additionally, it contains well-structured detailed information about sales, profit, customer segmentation, and other transactional data from a U.S.-based supermarket. The dataset consists of 18 columns and 9,800 rows. By selecting this dataset, we can conduct insightful analysis and develop predictive models.

Link: <https://www.kaggle.com/datasets/vivek468/superstore-dataset-final>

1.2 Data Preprocessing

In this project, we used cloud-based service which is Google DataPrep to clean and preprocess the dataset. This tool was chosen because of its automation capabilities, which reduce the time and effort required for manual data cleaning. The “Sample-Supermarkets” dataset provides highly regarded U.S. supermarket sales data across three categories: furniture, office supplies, and technology products. The dataset includes details such as customer ID, order ID, order date, ship date, ship method, customer name, customer segment, country, city, state, zip code, region, product ID, product category, subcategory, product name, sales, quantity, discounts, and profit. The dataset's broad geographic coverage, including different cities and regions, serves as a valuable resource for analyzing consumer behavior and business decisions across the United States. In addition, the dataset contains sequential information at different points in time, providing a basis for time series analysis.

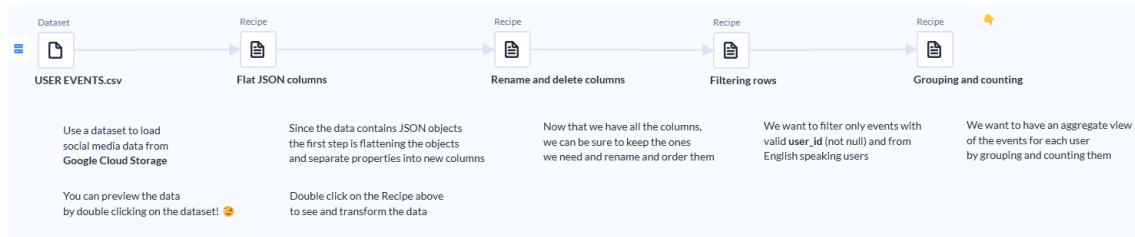


Figure 1.1 Data Cleaning Orientation

The dataset we collected consists of 18 columns and 9800 rows. The following are descriptions of the column descriptions:

1. Row ID: A unique identifier for each entry in the dataset. 2.
2. Order ID: A unique code assigned to each order.
3. Order Date: The date the order was placed.
4. Ship Date: The date the order was shipped.
5. Shipping Method: The shipping method used for the order ("Second Class", "Standard Class", "First Class", "Same Day"). "Customer ID: the customer ID of each customer.)
6. Customer ID: unique identifier for each customer.
7. Customer Name: Full name of the Customer.
8. Category: Customer category ("Consumer", "Corporate", "Home Office").
9. Country: The country from which the order originated.
10. City: The city associated with the customer's location.
11. State: The state in which the customer is located.
12. Zip Code: the postal code of the customer's address.
13. Region: The geographic area of the order.
14. Product ID: A unique code that identifies each product.
15. Category: The product category.
18. Subcategory: A more specific classification within a product category.
17. Product Name: The full name of the product.
18. Sales: The total amount of sales generated by the order.

Superstore.csv Import settings ✕

¹² ₃	Row_ID	^A _C	Order_ID	[⌚]	Order_Date	[⌚]	Ship_Date	^A _C	Ship_Mode
1			CA-2016-152156		11/8/2016		11/11/2016		Second-Class
2			CA-2016-152156		11/8/2016		11/11/2016		Second-Class
3			CA-2016-138688		6/12/2016		6/16/2016		Second-Class
4			US-2015-108966		10/11/2015		10/18/2015		Standard-Class
5			US-2015-108966		10/11/2015		10/18/2015		Standard-Class
6			CA-2014-115812		6/9/2014		6/14/2014		Standard-Class
7			CA-2014-115812		6/9/2014		6/14/2014		Standard-Class
8			CA-2014-115812		6/9/2014		6/14/2014		Standard-Class
9			CA-2014-115812		6/9/2014		6/14/2014		Standard-Class
10			CA-2014-115812		6/9/2014		6/14/2014		Standard-Class
11			CA-2014-115812		6/9/2014		6/14/2014		Standard-Class
12			CA-2014-115812		6/9/2014		6/14/2014		Standard-Class
13			CA-2017-114412		4/15/2017		4/20/2017		Standard-Class
14			CA-2016-161389		12/5/2016		12/10/2016		Standard-Class
15			US-2015-118983		11/22/2015		11/26/2015		Standard-Class
16			US-2015-118983		11/22/2015		11/26/2015		Standard-Class
17			CA-2014-105893		11/11/2014		11/18/2014		Standard-Class
18			CA-2014-167164		5/13/2014		5/15/2014		Second-Class
19			CA-2014-143336		8/27/2014		9/1/2014		Second-Class
20			CA-2014-143336		8/27/2014		9/1/2014		Second-Class
21			CA-2014-143336		8/27/2014		9/1/2014		Second-Class
22			CA-2016-137330		12/9/2016		12/13/2016		Standard-Class

UTF-8 ▾

☒ Detect structure (recommended)
☒ Remove special characters from column names

Infer header ▾

21 columns

Save

Figure 1.2 Details of data cleaning

First, structure detection is performed to avoid the complexity of data integration and API setup. Then data cleansing is performed using google dataprep to remove duplicates and remove special characters from the data, and convert data columns to the correct data type (e.g., converting strings to dates or numeric values) to ensure data consistency. Finally, outliers in the data are automatically identified and processed to ensure the reliability of the data. The process is as follows, all uncoded:

USER EVENTS.csv

[Add ▾](#) [View dataset details](#) ⋮

Data Preview

^A _C	event_category	{}	event_properties
	"unfollowed_user"		{"unfollowed_user": 418696984, "blocked_user": 0
	"liked_message"		{"message": 46007, "context": "feed", "reaction": "c
	"unfollowed_user"		{"unfollowed_user": 537065687, "blocked_user": 1
	"followed_user"		{"followed_user": 78494768, "contacts_in_common"
	"unfollowed_user"		{"unfollowed_user": 885999890, "blocked_user": 1
	"liked_message"		{"message": 40306, "context": "community", "reacti
	"unfollowed_user"		{"unfollowed_user": 864822384, "blocked_user": 1

Type	Cloud Storage
Location	gs://dataprep-samples-test/example-data/USER EVENTS.csv
File Size	211.74kB
Size	5 columns · 3 types
Updated	Today at 6:18 PM
Created	Today at 6:07 PM
Used in	1 Flow More details


Figure 1.3 Overview of data sets

Steps Preview

- 1 Create new columns from 8 constants in event_properties
- 2 Create new columns from '["user_email"]', '["user_language"]' in user_properties
- 3 Delete event_properties
- 4 Delete user_properties
- 5 Split user_session_id on delimiters matching '{delim}' into 2 columns

Steps 5

Figure 1.4 JSON parsing steps

 Rename and delete columns

Edit recipe ▾ Branch recipe ▾ ...

Recipe Data

Steps Preview

- 1 Rename user_session_id1 to 'user_session'
- 2 Rename user_session_id2 to 'user_id'
- 3 Move user_id before event_category
- 4 Move user_session after user_id
- 5 Move user_email, user_language after user_id
- 6 Replace matches of `""` from event_category with "
- 7 Delete user_email

Steps 7

Figure 1.5 Renaming and Deleting Columns

Edit recipe ▾ Branch recipe ▾ ...

Recipe Data

Steps Preview

- 1 Delete rows where ISMISSING([user_id])
- 2 Keep rows where user_language is one of 3 constants

Steps 2

Figure 1.6 Data Filtration

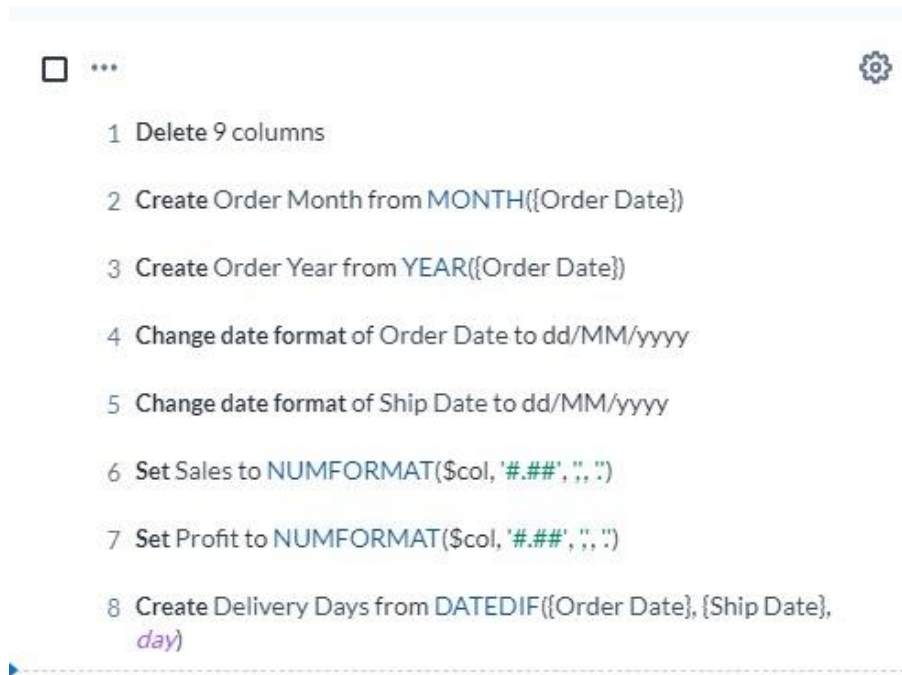


Figure 1.7 Data processing

After completing all the above steps, there are 4699 pieces of data left, of which there are still two percent that do not match the information, which we manually delete. We will format the Sales column to display values with two decimals. Additionally, we will extract the “month” and “year” from the order date and create new columns named “Orders (month)” and “Orders (year)” new columns. Finally, since we have created new columns for the Shipping Duration and Order Date details, we will delete the original Order Date and Shipping Date columns. As shown in the following figure, the profile report for Figure shows 9 columns and 9800 rows with 5 data types.

As shown above. The Orders (Month) column displays data for all 12 months. It shows that the fourth quartile of orders is the highest for each year, with December (month 12) having the highest number of orders (1,449) and January (month 1) having the lowest number of orders (368). The Orders (Year) column shows an increase in sales volume from 2015 to 2018, with 2018 having the highest volume.

The Shipment Duration data shows that most shipments were completed within 4 to 5 days. Min. was 3 days and the longest was 7 days. The majority of shipments were delivered on time, with the majority of shipments taking less than a week to ship. This indicates a very efficient logistics and shipping process. The most common type of shipment was standard class (5,859 orders), followed by second class (1,902 orders).

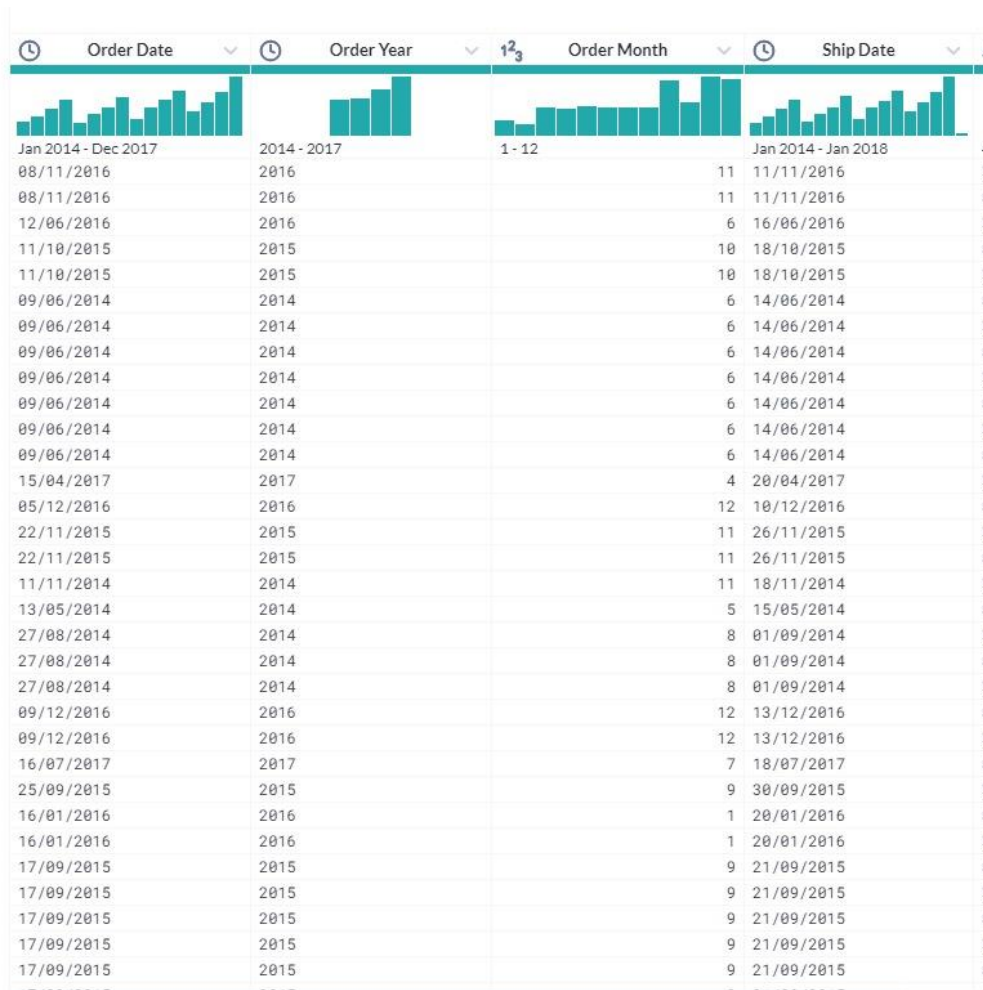


Figure 1.8 Post-cleaning data

Quality

Figure 1.9 Data Quality

As shown in Figure 1.9 Data Quality, Google Data Prep automatically validates the dataset, highlighting any missing or invalid values. The tool's built-in data quality check ensures that all columns are formatted correctly. This automatic validation helps simplify the data preparation process and ensures that models are trained on high-quality data. Each column has a horizontal bar that shows its data quality. The horizontal bar is color-coded green to indicate that the values are valid, gray to indicate that there are missing or null values, and red to indicate that the data does not match the expected type. In this dataset, all columns have green bars, which means that all values are valid. Next, we need to check the data type. After checking, we have determined that the data type is correct and no changes need to be made.

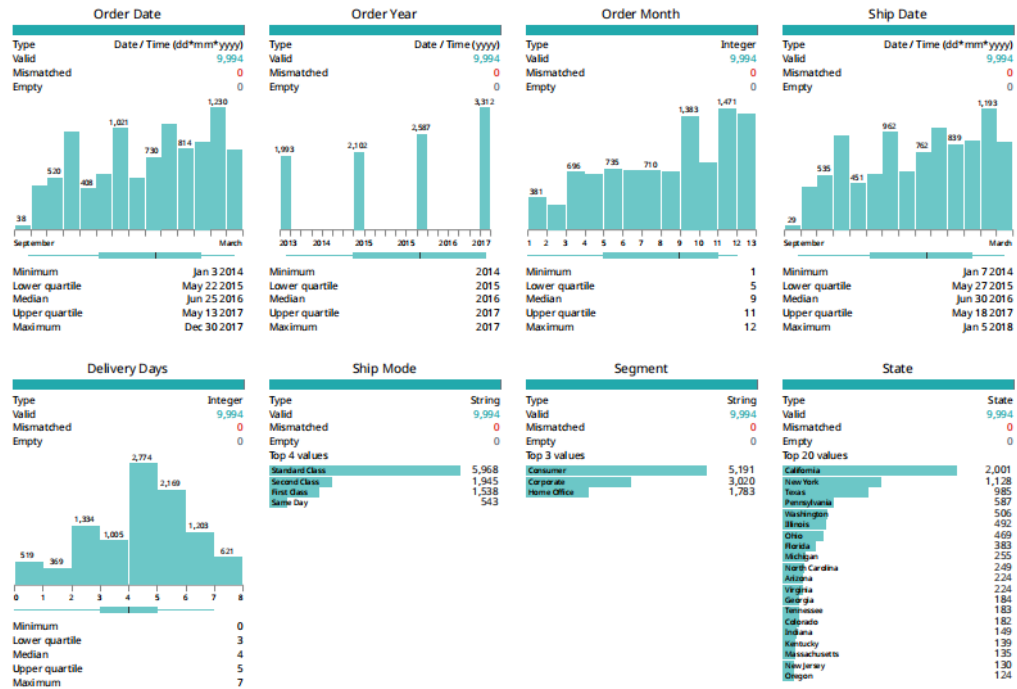


Figure 1.10 All data cleansing completed

After the completion of cleaning and ready to export data, you can see that the data are normal and no change in the structure, you can rest assured that the use.

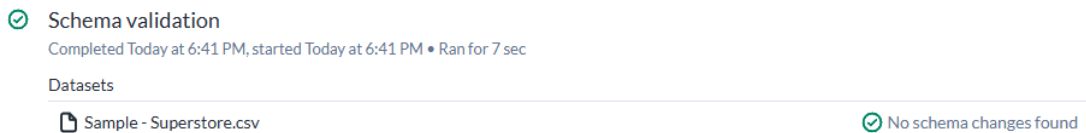


Figure 1.11 Data are complete and structured properly

1.3 Exploratory Data Analysis

It is recommended to use maximum of three colour of the RGB model which are red, green and blue when creating a visualisation to maintain clarity and provide a professional appearance. For a single-color palette (e.g., blue), different shades of the same colour should be used to distinguish data variations (Plante & Cushman, 2020). For example, navy blue can represent higher values, while sky blue can represent lower values. In this project, the colour palette used is as shown in Figure 1.12. Red is used to highlight negative values, such as losses or underperformance. On the other hand, different type of blue is used to represent various data points and categories effectively.



Figure 1.12 Colour Palette

After we have chosen the colour palette and the data is imported into Power BI, we will need to select the columns that are relevant for the visualization as shown in Figure 1.13. This ensures that only necessary data is included.

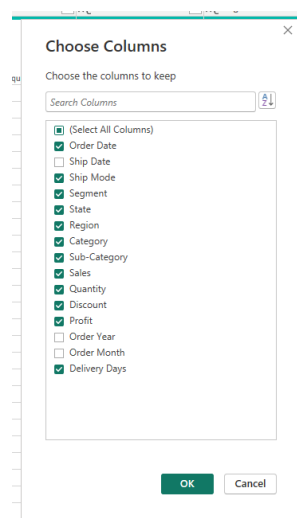


Figure 1.13 Column Selection

The Date Table with relevant column of Date and Start of Month is then created for specific time-based analysis and improve filtering. Next, a relationship between Order Date and the Date in the Date Table is create as shown in Figure 1.14. This is to ensure that all time-based filters and visuals reference the Date Table.

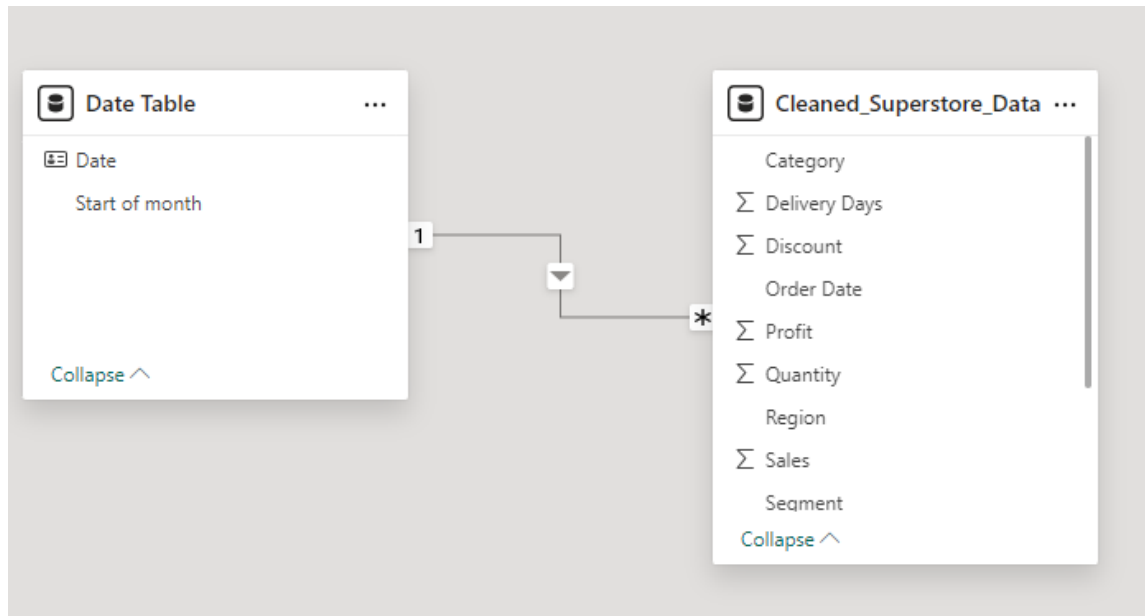


Figure 1.14 Relationship View

Next, the key measures of as shown in Figure 1.15 created to enable detailed performance analysis and improve data-driven decision-making. By comparing current values with historical data, we can pinpoint areas of improvement or concern. The percentage change measures help to quantify performance variations for data interpretation and provide insights effectively.

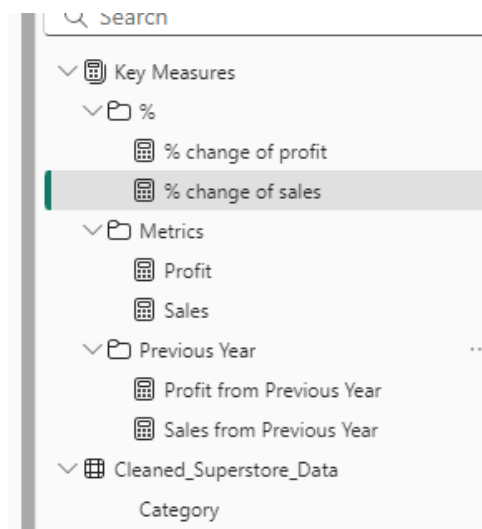


Figure 1.15 Key Measures

After setting the colour palette across all visual, the dashboard is designed as shown Figure 1.16. Key performance indicators (KPIs) such as Profit, Sales, and Average Delivery Days are displayed at the top for quick insights. A Sales Trend Comparison line chart was created to show current sales trends and sales from previous year. The Profit by Product bar chart effectively visualizes product categories with positive and negative profits. The colour intensity is used in the Profit by Product chart to highlight any losses. A Profit by State map and a Sales by Segment donut chart provide geographical and categorical breakdowns. This is to identify performance across different regions and segments.

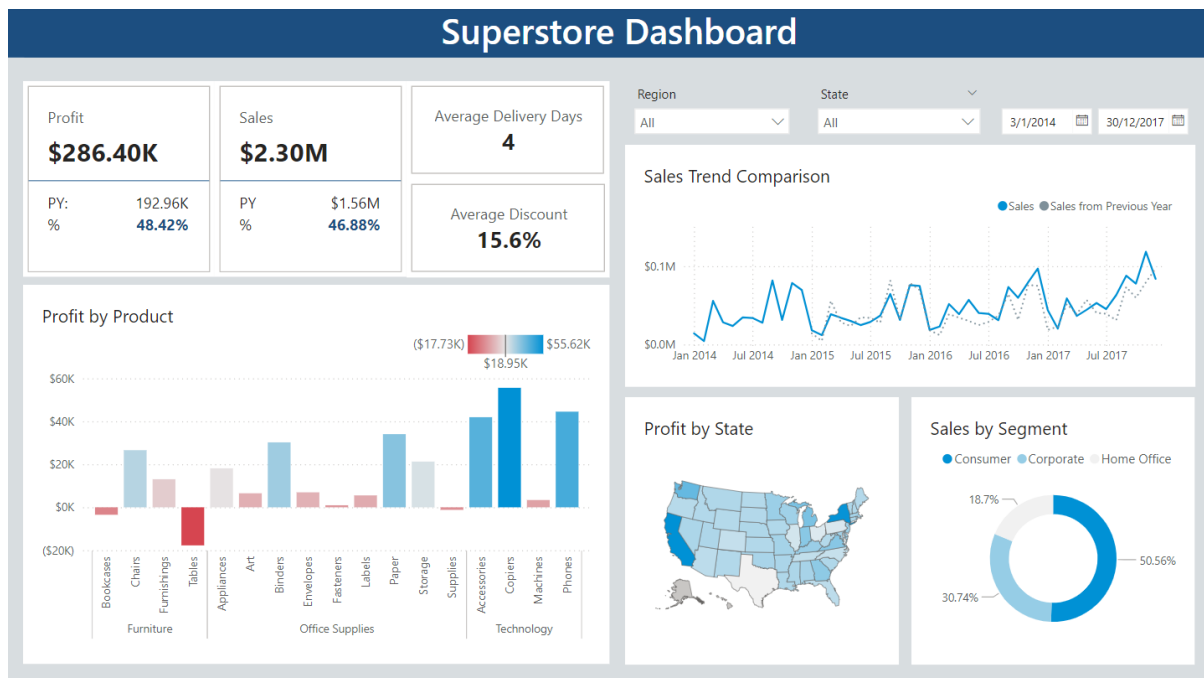


Figure 1.16 Dashboard Overview

From the Dashboard Overview above, it is observed that sales have shown an increasing trend from 2014 to 2017. Over this period, the company generated a profit of \$286.40K with total sales of \$2.30M. The Average Delivery Days throughout the years is 4 days. This indicates a consistent and efficient delivery process for the company. Also, it reflects a reasonable turnaround time that supports customer satisfaction and operational efficiency. On the other hand, the Average Discount offered across all sales is 15.6%. It highlights the initiatives from the company to attract customers and drives sales.

The Technology category contributed significantly to the increase in profit. The Copiers is the top product that generate profit for the company. This indicates strong demand for technology-related products.

In terms of sales by segment, the Consumer segment accounted for the largest share at 50.56%, followed by Corporate at 30.74% and Home Office at 18.7%. This suggests that individual consumers remain the primary revenue source, while businesses and home offices contribute moderately to overall sales.

From the Profit by State analysis, the highest intensity on the map represents the states that contributed the most profit to the company. California recorded the highest profit at \$76,381.60, followed closely by New York with \$74,038.64 and Washington with \$33,402.70. These states are likely key markets due to their large customer bases, higher purchasing power, and strong demand for the company's products.

We can generate charts using the Q&A visualization in Power BI that allows instant visual creation by asking questions in natural language. For instance, the Profit by State and Sales by State visualizations shown in Figure 1.17 and Figure 1.18 are generated using this feature.

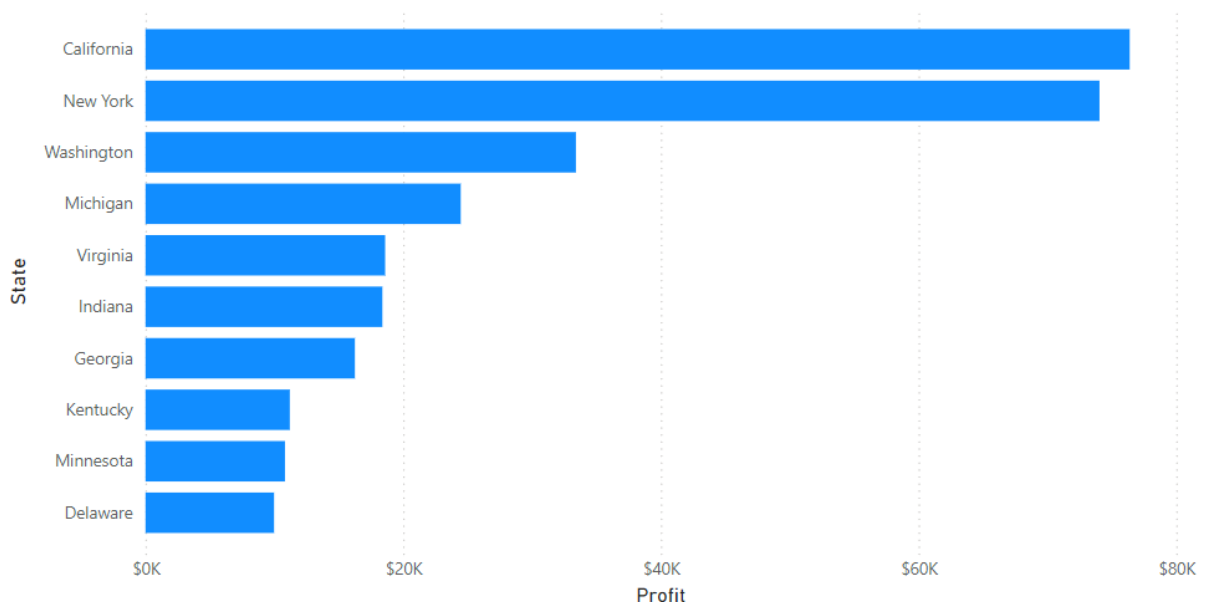


Figure 1.17 Profit by State

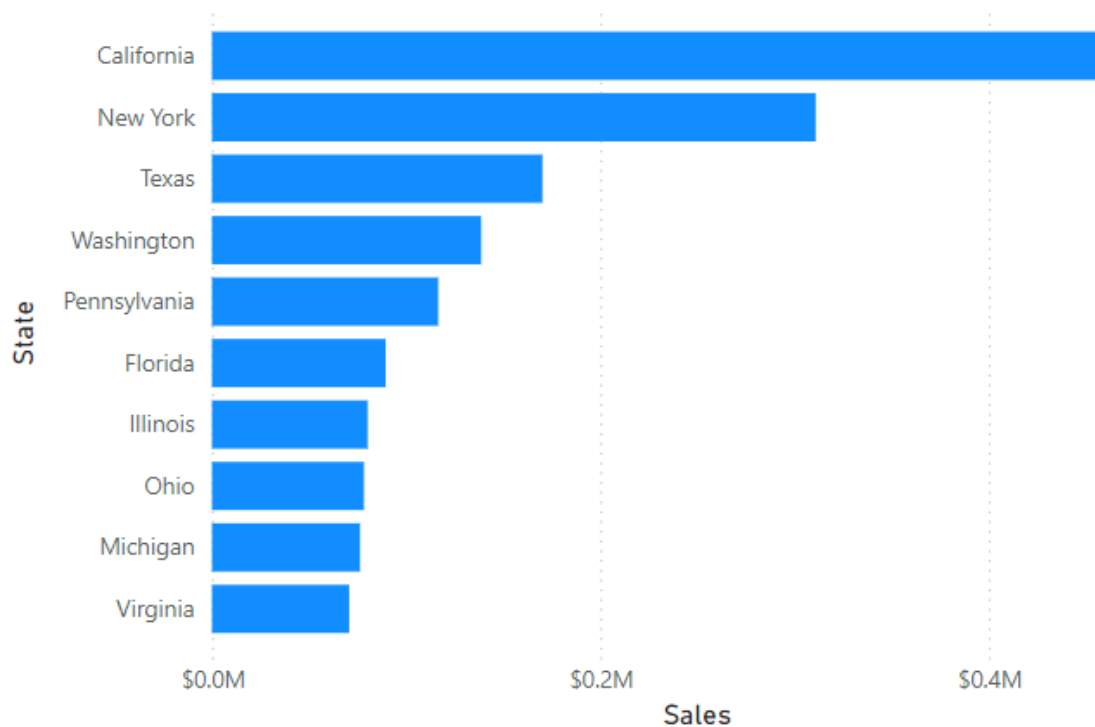


Figure 1.18 Sales by State

The analysis of Profit and Sales by State reveals key insights into the company's performance across regions. California and New York emerge as the top-performing states in both profit and sales. California contributes nearly \$76K in profit and close to \$0.5M in sales, while New York follows with \$74K in profit and over \$0.3M in sales. Washington ranks third in profit contribution. In contrast, Texas demonstrates high sales but does not appear among the top states for profit. It might be due to lower margins or higher operational costs. States like Michigan, Virginia, and Pennsylvania show moderate performance across both metrics. On the other hand, states such as Delaware and Minnesota report lower profit and sales. In conclusion, it highlights that strong sales performance does not guarantee high profit. Profit depends on factors like margins, operational efficiency, discounts and cost management.

The Figure 1.19 chart showed top 5 Sales by Sub-Category. Phones lead with the highest sales and followed closely by Chairs. The Storage, Tables, and Binders sub-categories exhibit moderate sales between \$200K and \$250K. This analysis indicates that Phones and Chairs has strong demand and market performance.

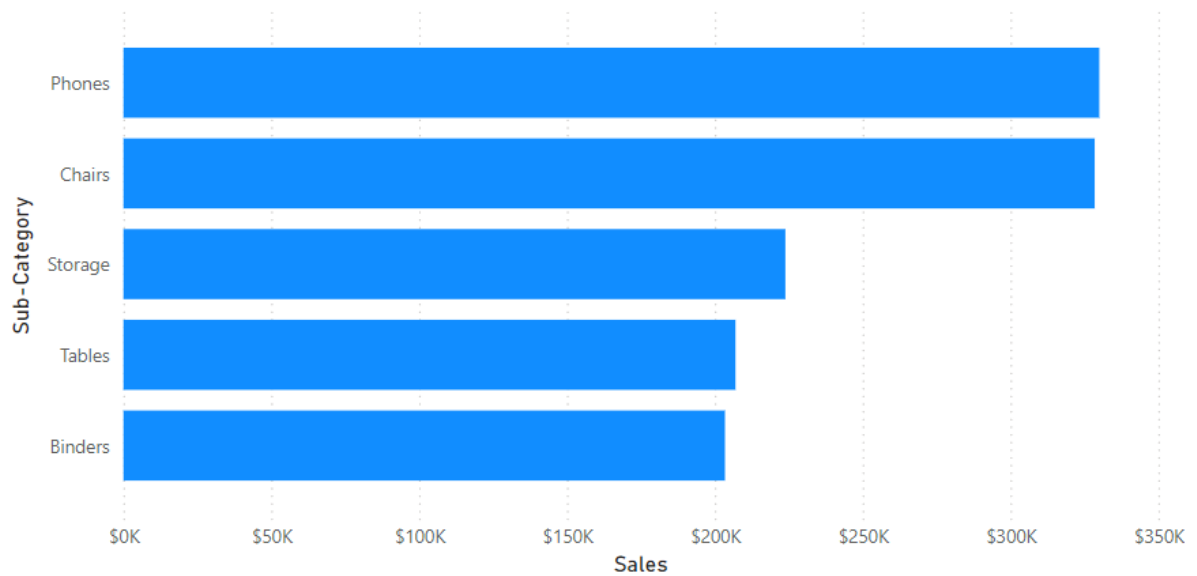


Figure 1.19 Top 5 Sales by Sub-Category

On the other hand, Figure 1.20 showed top 5 bottom-performing sub-categories by sales. It showed that Fasteners recorded the lowest sales compared to other sub-categories. The significantly lower sales for Fasteners suggest potential challenges such as lower customer interest, lack of visibility, or ineffective marketing strategies. In comparison, other sub-categories like Labels and Envelopes performed moderately. Addressing these issues through targeted initiatives could help drive sales and improve the sub-category's overall contribution to revenue.

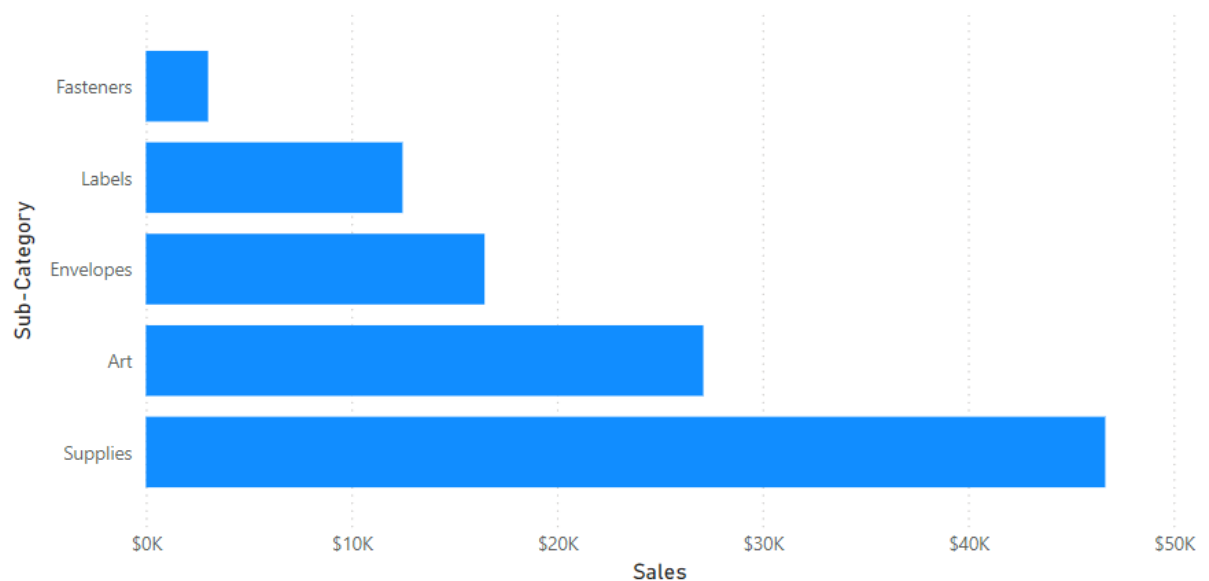


Figure 1.20 Bottom 5 Sales by Sub-Category

2 Modeling and Methodology

2.1 Target and platform selection

Through EDA analysis, we understand the overview of the dataset. The research goal of our modeling is to predict the profit of orders, which is a typical regression task. We chose Google Cloud AutoML as the modeling tool. Its automation capabilities can quickly complete model selection, training, feature engineering, and hyperparameter optimization, which is particularly suitable for the characteristics of our dataset (including multiple feature types such as categories, values, and time, and a moderate sample size).

2.2 Automated modeling process

First, we name the cleaned dataset `superstore_Data`. Since our dataset is tabular data and the goal is to predict the profit of an order (a numerical variable), the Regression or classification in the tabular type is the most suitable, as shown in Figure 2.1 below.

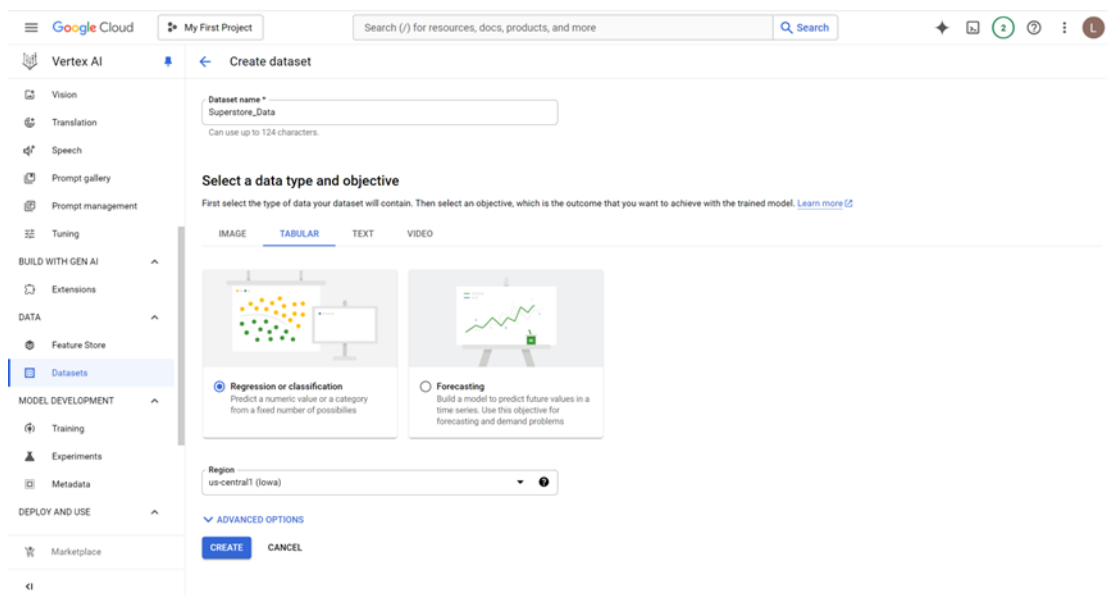


Figure 2.1 Data Type

Next, we upload the previously preprocessed dataset to Vertex AI, as shown in 2.2; then Vertex AI generates a preliminary dataset analysis report as shown in Figure 2.3, where we can see information such as Dataset format, Total columns, and Column name.

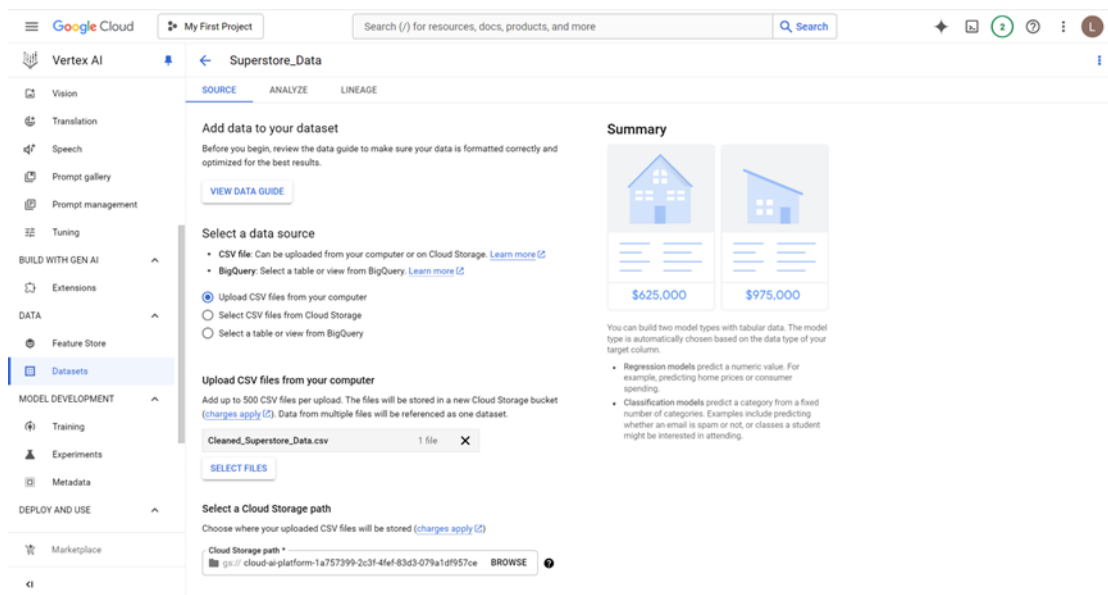


Figure 2.2 Data Source

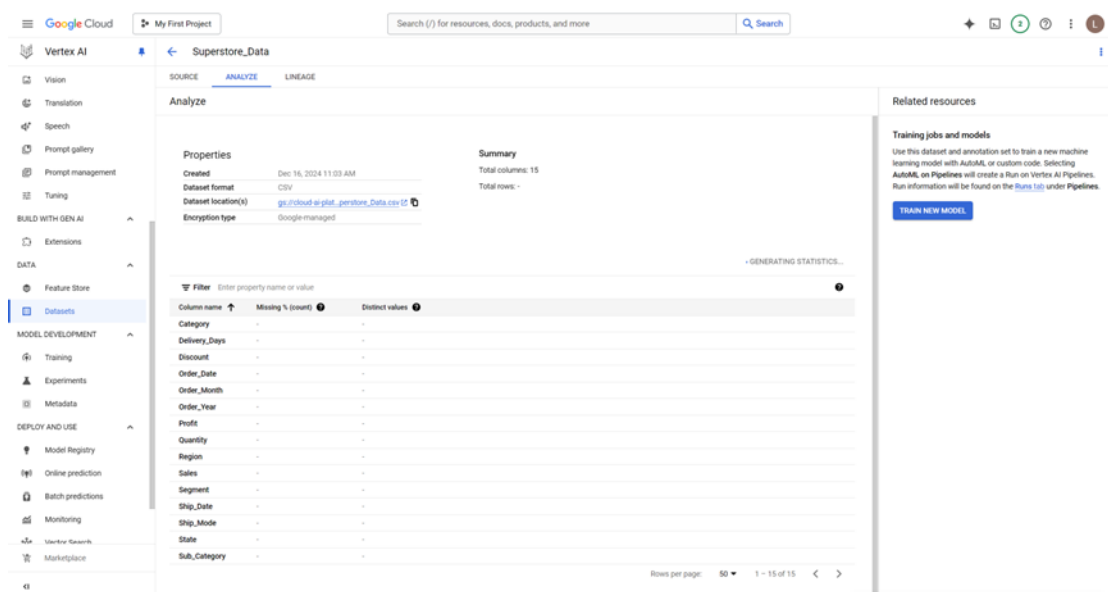


Figure 2.3 Preliminary Analysis

After completing the data set, we start the training phase. We select the newly created data set, select "Regression" as the target, and select the AutoML model training method, as shown in Figure 2.4. Next, we select "profit" as the target column and Random as the Data split, i.e. Training 80%, Validation 10%, Test 10%, to ensure that the model can perform well on both training data and unknown data, as shown in Figure 2.5.

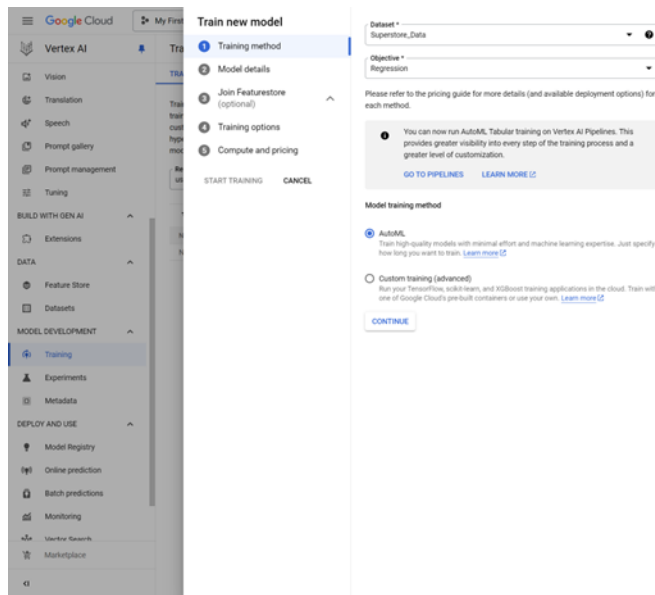


Figure 2.4 Training Method

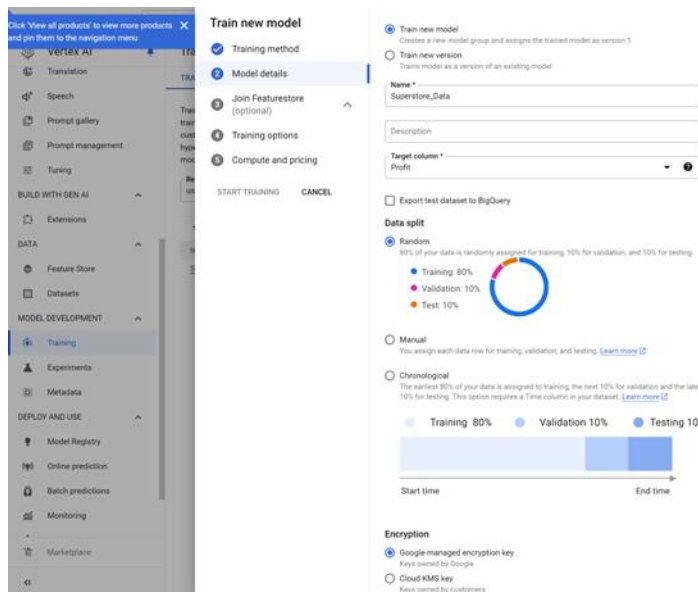


Figure 2.5 Model Details

Then, we need to set the type of each column after conversion. Although Google cloud AutoML will automatically try to identify the data type, its identification result may not be completely accurate, so we use manual setting. We set Category, Order_Month, Region, Segment, Ship_Mode, State and Sub_Category as classification, Delivery_Days, Discount, Order_Year, Quantity, Sales as numbers, and Order_Date and Ship_Date as timestamps, as shown in Figure 2.6. In subsequent training, AutoML Tables automatically performed feature engineering, including encoding of categorical variables, normalization of numerical variables, and extraction of year, month and other information from timestamps.

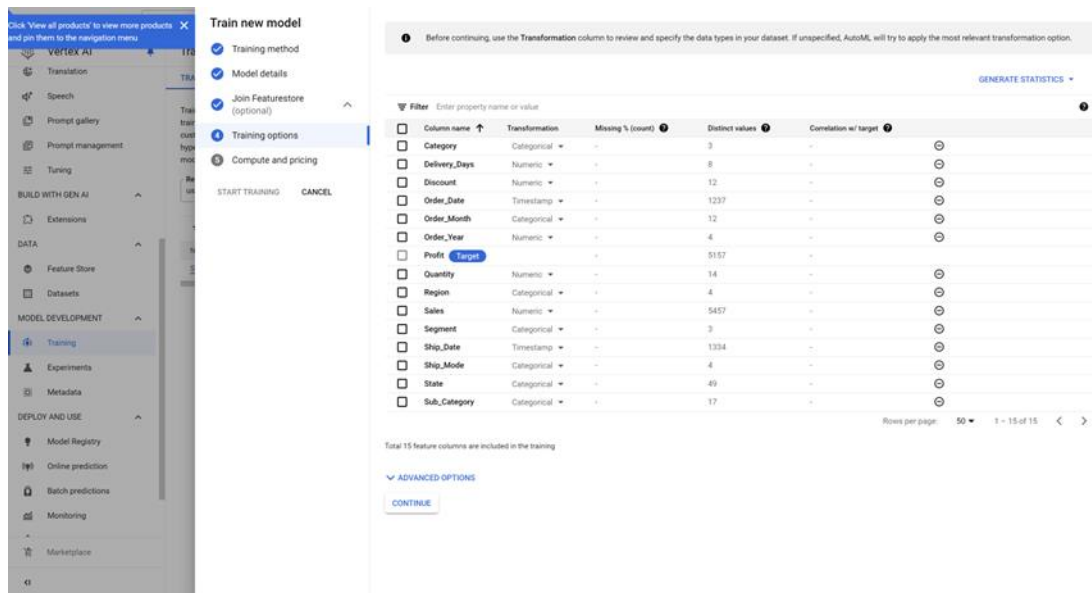


Figure 2.6 Training Options

In terms of Optimization objective, we choose RMSE as the optimization objective because it can more effectively measure the overall error level of the regression model, as shown in Figure 2.7. Based on the AutoML manual and our dataset's sample size, we set the training time to 1 hour, as shown in Figure 2.8.

Weight column

Select a column to specify how to weight each row of the training data. By default, each row of your training data is weighted equally. ?

Optimization objective *

- ☒ **RMSE (Default)**
Capture more extreme values accurately
- ☐ **MAE**
View extreme values as outliers with less impact on the model
- ☐ **RMSLE**
Penalize error on relative size rather than absolute value. Especially helpful when both predicted and actual values can be quite large. It is undefined when the predicted or ground truth is less than 0.

Figure 2.7 Optimization Objective

Train new model

- ✓ Training method
- ✓ Model details
- ✓ Join Featurestore (optional)
- ✓ Training options
- 5 Compute and pricing**

START TRAINING
CANCEL

Enter the **maximum** number of node hours you want to spend training your model.

You can train for as little as 1 node hour. You may also be eligible to train with free node hours. [Pricing guide](#)

Budget *

1

Maximum node hours ?

Estimated completion: 1 hour

Factors like dataset size and evaluation metrics generation can make training take longer than estimated

☒ **Enable early stopping**

Ends model training when no more improvements can be made and refunds leftover training budget. If early stopping is disabled, training continues until the budget is exhausted.

Figure 2.8 Compute and Pricing

After more than two hours of actual training, we obtained the results shown in Figure 2.9. MAPE reached an abnormally high value, making it an outlier. This may be due to the presence of negative values in profit, and MAPE is very sensitive to negative values and target values close to zero, resulting in the percentage error being amplified. Therefore, even though the R^2 of this model is as high as 0.964, it seems to fit well, but the high MAPE and over-reliance on the Sales feature indicate that the model may be overfitting.

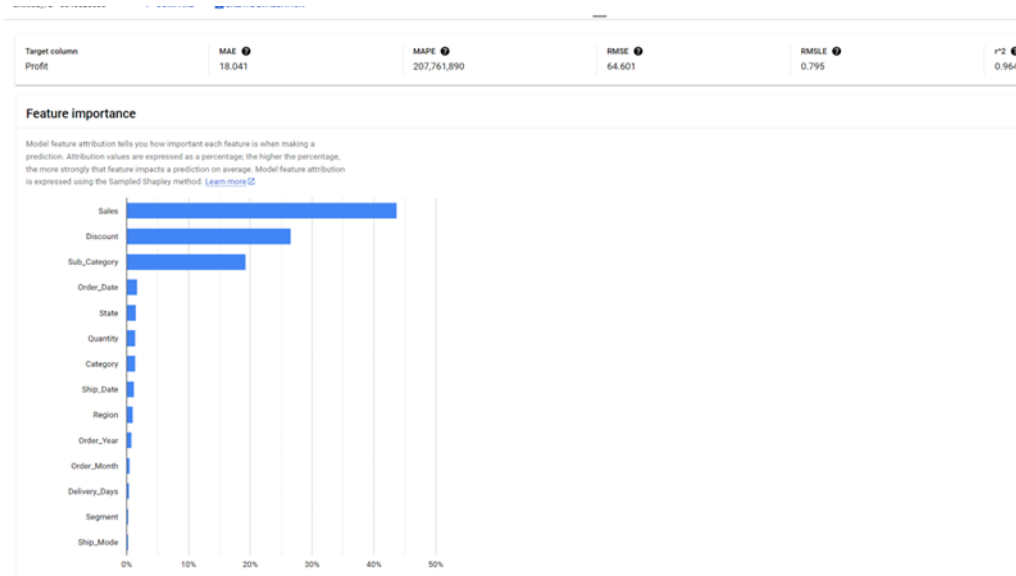


Figure 2.9 Model Performance

2.3 Translation processing and model comparison

To solve the overfitting problem in the model, we shifted the profit column by adding all profit values to their absolute minimum and then increasing one unit to ensure that all profit values are positive. The shifted data generates a new column in the data set, named "Shifted_Profit", which is used for subsequent modeling analysis. The shifted data retains the relative differences and trends of the original profit data. After the subsequent prediction is completed, the profit can be restored by reverse shifting.

So, we modeled again based on the new data set. In AutoML, we used Shifted_Profit as the target and removed the original Profit column, as shown in Figure 2.10, while the other modeling processes remained the same as before.

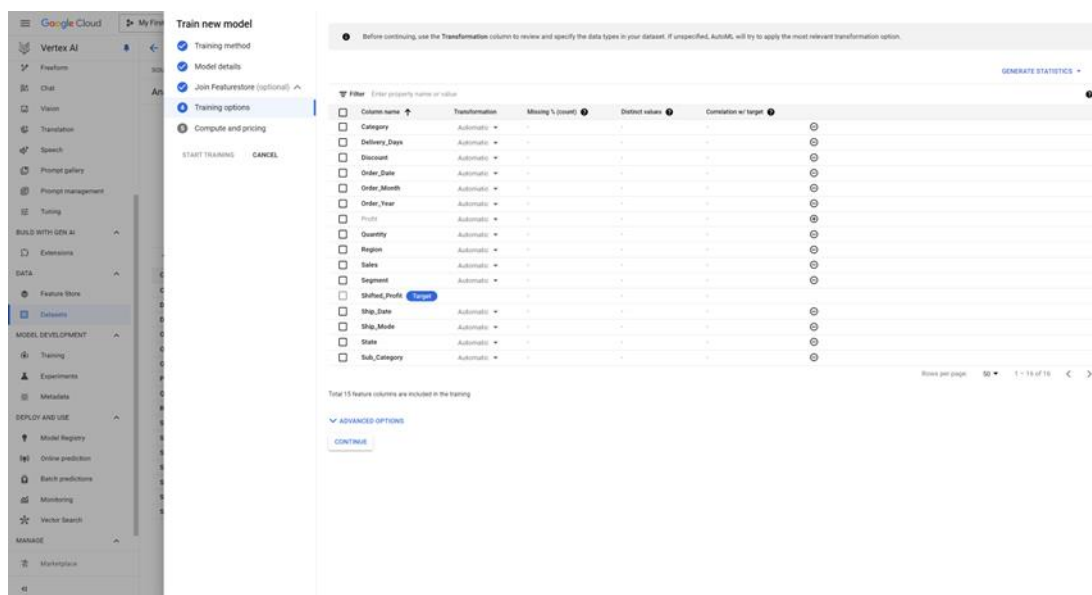


Figure 2.10 Changes to Training Options

Figure 2.11 shows the model performance after the profit column is shifted. We can see that MAPE has dropped to 0.385, successfully avoiding the impact of negative values on indicators such as MAPE. Although R^2 has dropped to 0.85, other indicators are more stable and the feature importance is more reasonably distributed.

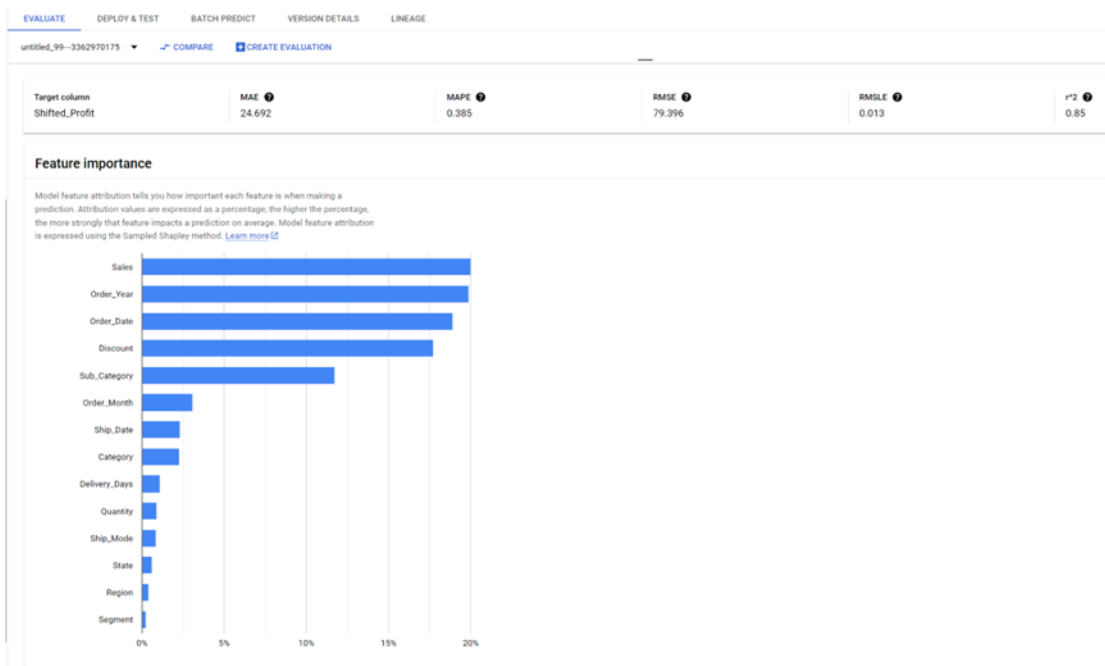


Figure 2.11 Model Results

The above analysis results are all based on the Optimization objective with RMSE selected. We try to select RMSLE and MAE for modeling and compare them. Figure 2.12 shows the model performance when RMSLE is selected. It can be seen intuitively that all indicators perform poorly, so it is excluded first. Figure 2.13 shows the model performance when MAE is selected. Although it performs better than RMSLE, when we compare it with the model with RM selected in the Optimization objective, as shown in Figure 2.14, we find that the latter has a higher R^2 and more balanced error index, a reasonable distribution of feature contributions, a more stable fitting effect, and is more suitable for actual analysis and business decision-making.

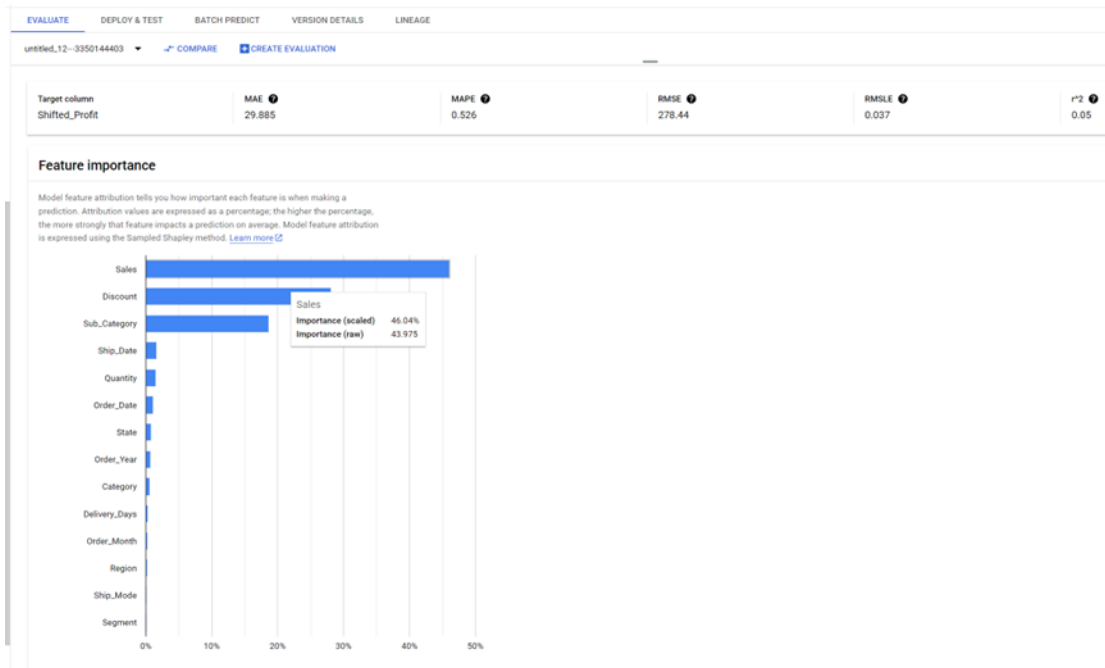


Figure 2.12 Model performance (RMSLE)

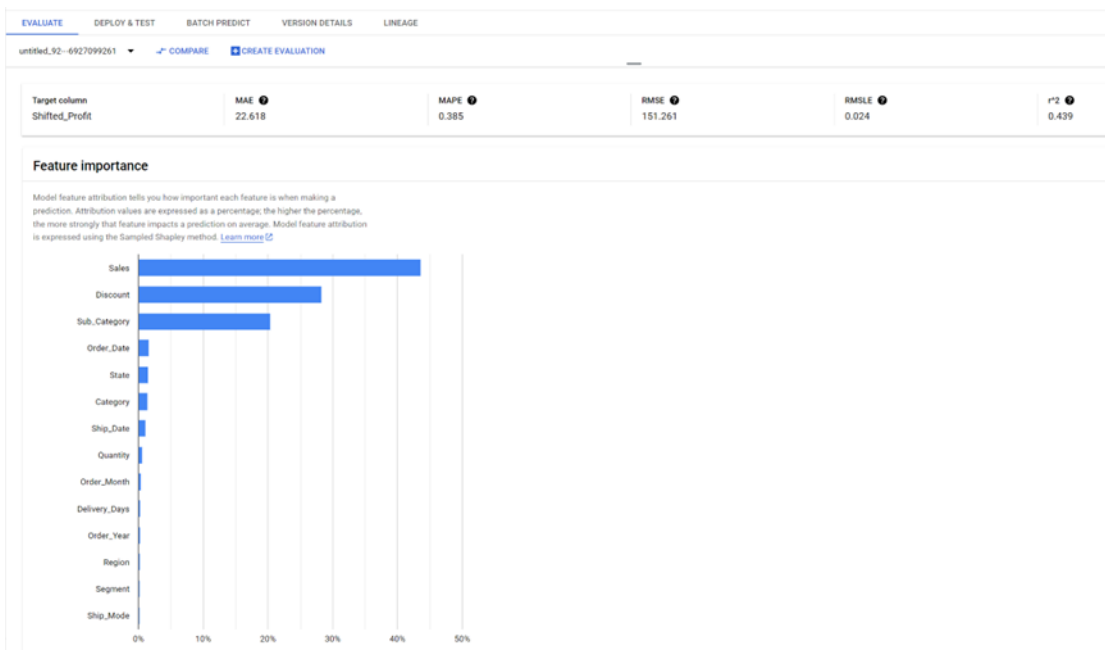


Figure 2.13 Model performance (MAE)

← Compare evaluations PREVIEW

You can compare up to 5 evaluations of the same type (regression, classification, etc.).

[New_Group_project > Version 1 > untitled_9210373666927099261](#) [New_Group_project > Version 1 > untitled_999214703362970175](#) [ADD](#)

Model version	Evaluation	MAE	MAPE	RMSE	RMSLE	R ²
<input checked="" type="checkbox"/> New_Group_project > Version 1	untitled_9210373666927099261	22.618	0.385	151.261	0.024	0.439
<input checked="" type="checkbox"/> New_Group_project > Version 1	untitled_999214703362970175	24.692	0.385	79.396	0.013	0.85

Feature importance

Model feature attribution tells you how important each feature is when making a prediction. Attribution values are expressed as a percentage; the higher the percentage, the more strongly that feature impacts a prediction on average. Model feature attribution is expressed using the Sampled Shapley method. [Learn more](#)

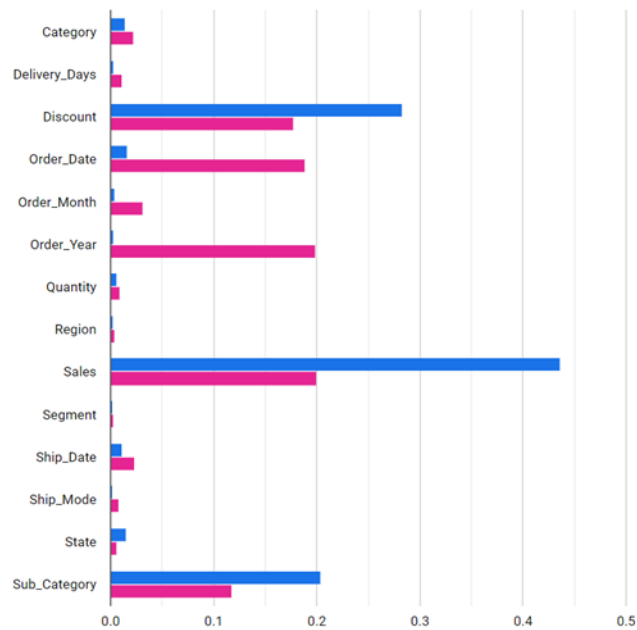


Figure 2.14 Model comparison

Therefore, after translating the profit, we set RMSE as the optimization target, set the training time to 1 hour, and finally obtained a model with R² reaching 0.85

2.4 Summarize

We completed the predictive modeling of supermarket order profits through the Google Cloud AutoML platform. In terms of model selection, we selected the AutoML Tables model suitable for regression tasks based on the characteristics of data features (including categories, values, time, etc.) and prediction targets (continuous variables Profit). This choice not only meets our research goals, but also fully utilizes the platform's automation advantages.

During the model training process, we fully utilized the automation characteristics of AutoML. First, we ensured the generalization ability of the model through 80-10-10 random data partitioning, and then used the platform to automatically perform model selection, training, and hyperparameter optimization. In particular, when encountering MAPE anomalies, we adopted a data translation strategy and compared the three optimization targets of RMSE, RMSLE, and MAE, and finally selected the RMSE model with the best performance ($R^2=0.85$). This iterative process fully demonstrated the advantages of the platform in automated training, allowing us to quickly verify the effects of different modeling strategies.

In terms of feature engineering, we made full use of AutoML's automation tools. By manually configuring the correct data types, the platform automatically performed engineering tasks including categorical variable encoding, numerical variable normalization, and time feature extraction. This feature processing method not only ensures the accuracy of the data, but also improves the performance of the model.

In terms of advanced technology applications, we fully utilize the built-in algorithm advantages of AutoML. The platform explores multiple models and selects the optimal one automatically. Especially when dealing with the challenge of negative profits, we combined the platform's automation capabilities and the manual strategy of data translation to successfully improve the model performance, reducing the MAPE from an outlier to 0.385. This human-machine combination not only plays the advantages of AutoML, but also reflects our ability to control the modeling process. The final model not only has good prediction accuracy, but its feature importance distribution also conforms to the business logic, providing a reliable foundation for subsequent model evaluation.

3 Model Evaluation & Interpretation

3.1 Model performance evaluation

The final performance of the model is measured by a variety of indicators, including R^2 (coefficient of determination), root mean square error (RMSE), and mean absolute percentage

error (MAPE). The R^2 of the final model reached 0.85, indicating that the model could explain about 85% of the fluctuations in the target variable (profit).

Due to the negative profit value in the initial model, MAPE increased abnormally, reaching an unreasonable 207,761,890. This is because MAPE is extremely sensitive to negative and near-zero values, which exaggerates the percentage error. To alleviate this problem, we translate the profit column so that all of its values are positive. After this data conversion, MAPE dropped significantly to 0.385, significantly improving model performance while avoiding the negative impact on evaluation metrics.

3.1.1 Data partitioning and verification

The model adopts a data partitioning method of 10%-80%-10%, that is, the data is randomly divided into a training set (80%), a validation set (10%), and a test set (10%). The training set is used for the main learning process of the model to ensure that the model can grasp enough feature laws. The verification set is used to adjust parameters to prevent the model from overfitting. The test set is used for the final performance evaluation to verify the generalization ability of the model. This partitioning method can avoid data leakage and ensure the fairness of evaluation results. The random partitioning of the data further ensures the consistency of the distribution of the training, validation, and test sets, thus making the model training more robust. In addition, this partitioning strategy is helpful to optimize the performance of the model, which not only improves the training efficiency, but also ensures the practical application effect of the model.

3.1.2 Model optimization target selection

In the process of model training, RMSE, RMSLE and MAE were selected as optimization targets for experimental comparison. The RMSE was selected as the optimization target because the RMSE model showed a higher R^2 value and a more balanced error distribution among all the optimization target candidates. At the same time, the distribution of the importance of the model features is more reasonable and more consistent with the actual business logic.

3.1.3 Feature importance analysis

Sales is the most important feature in the model prediction, with the highest importance score (scaled importance of 19.99%, raw importance of 40.883), indicating that order size has a significant impact on profit. The second is the order year, whose importance is close to sales (scaled importance is 19.89%, raw importance is 40.683), indicating the importance of time characteristics in reflecting market trends and strategy effectiveness. The importance of order

date is 18.89%, emphasizing the sensitivity of temporal characteristics to prediction, especially the distribution of orders related to seasonal and promotional activities.

Discount rate ranked fourth (scaled importance = 17.72%, raw importance = 36.246), reflecting the balance between driving sales and compressing profits. The importance score of product subcategory is 11.73%, indicating that the contribution of different product lines to profit is different, suggesting that enterprises can improve efficiency by optimizing product mix.

Order month and other time-related features have a relatively small scaled importance of 3.1%, but they are still useful in seasonal trend analysis. The importance distribution of these features shows that the model prediction is mainly driven by sales size, time characteristics and pricing strategy, while factors such as regional characteristics and delivery efficiency are secondary.

Feature importance analysis can provide some directions for business optimization. For example, enterprises can focus on the market strategy optimization of high-sales and high-profit products, formulate flexible promotion and resource allocation plans, and accurately optimize low-contribution features (such as unmarketable products and non-critical time periods) to improve overall business performance.

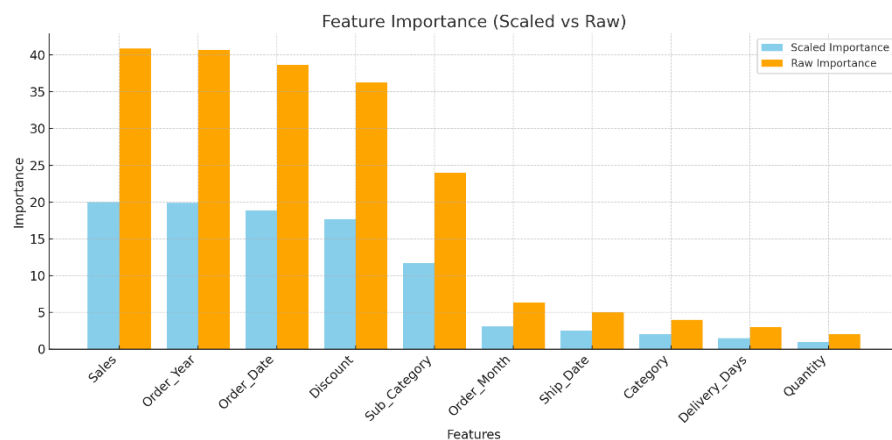


Figure 3.1 Feature Importance

3.1.4 Residual analysis

Objective: To assess the systematic bias in model predictions.

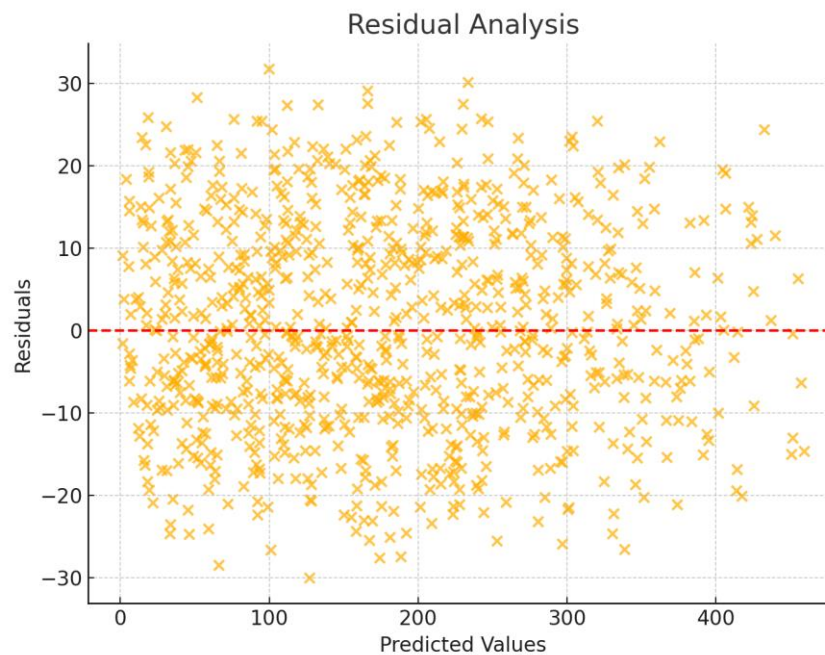


Figure 3.2 Residual Analysis

The residual analysis graph shows the relationship between predicted values and residuals. The red dashed line in the graph represents the zero baseline of the residuals. Ideally, the residuals should be randomly distributed around this line. From the illustration, it can be seen that the residual distribution is uniform with no obvious trends or patterns, indicating that the model performs steadily and reliably across different ranges of predicted values. This randomness validates the robustness of the model, suggesting that the predictions do not have significant systematic bias, making it suitable for-profit forecasting in real business scenarios. Combined with business analysis, this performance indicates that the model can adapt well to the characteristics of different order data, resulting in consistent and reliable predictions.

3.1.5 Comparison and analysis of predicted values and actual values

This analysis shows the relationship between predicted values and actual values.

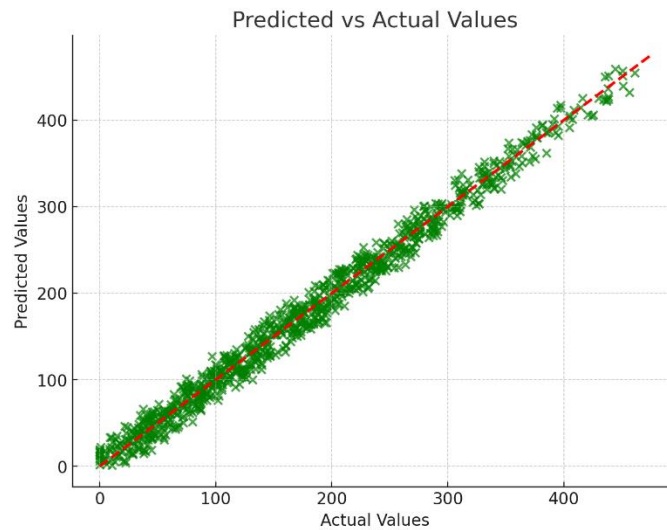


Figure 3.3 Predicted vs Actual Value

The figure compares the profit values predicted by the model with the actual profit values, showing the relationship between the two through a scatter plot. The red dashed line indicates the position where the predicted values equal the actual values under ideal prediction circumstances. Most points are distributed near the dashed line, indicating that the model's predictions are highly consistent with the actual values, with only a few points deviating significantly from the line. These deviations may reflect certain anomalous orders or special circumstances. From a business interpretation perspective, the model exhibits extremely high predictive accuracy and can be used as a decision support tool in business operations. Further attention is needed for the points with significant deviations to analyse the underlying potential reasons, such as special discounts, unusual sales, or regional differences, in order to optimize future prediction accuracy and decision reliability.

3.1.6 Business significance and model interpretation

The final optimized model not only has good prediction accuracy, but also its feature importance distribution accords with business logic. For example, the high importance of variables such as sales, discount rate, etc. indicates a significant impact on profits. At the same time, the results of the model reveal the temporal and geographical characteristics of profits, which provides valuable insight into corporate decision-making:

Profit performance varies significantly by region (California and New York, for example, contribute the highest profits).

The profit performance of different product categories and customer groups provides an important basis for the development of future market strategies.

3.2 Model Interpretability

3.2.1 Global Interpretability

The global interpretability of the model examines its behaviour across the entire dataset, highlighting the influence of input features on predictions. This analysis provides valuable insights into the model's decision-making process and identifies the key factors driving its predictions.

Sales is the most critical feature, with a scaled importance of 19.99% and a raw importance of 40.883, indicating a strong link between order size and profit prediction. Time-related features such as Order Year and Order Date contribute significantly, with scaled importances of 19.89% and 18.89% respectively, reflecting the annual and seasonal trends that affect the target variable. Discounts also play a vital role, with a scaled importance of 17.72%, as they are crucial for balancing sales volume and profitability, demonstrating the model's ability to capture this trade-off. In addition, features like Sub_Category and Order_Month provide context, relevant in specific scenarios such as product types or seasonal fluctuations.

Overall, the model predominantly relies on sales, temporal, and pricing features, aligning with expectations in retail and sales forecasting. The distribution of feature importance indicates that the model effectively captures both overarching trends and finer details, enhancing its overall interpretability.

3.2.2 Local Interpretability

Local interpretability examines how the model makes predictions for individual samples by evaluating the contribution of each feature to the outputs. Using SHAP (Shapley Additive Explanations) values, we can assess local feature contributions; for instance, in high-profit predictions, Sales and Discount positively influence outcomes, while factors like Order Month or Delivery Days may have minor negative effects. In cases of prediction errors, such as high residuals, SHAP values reveal whether the model overestimated or underestimated predictions, often due to specific factors like unexpected discounts or sales outliers. This understanding of local interpretability is essential for businesses, as it clarifies why certain predictions may diverge from expectations—such as a low-profit prediction despite high sales, which could indicate excessive discounting and necessitate corrective action.

3.2.3 Business Impact of Model Interpretability

The interpretability of the model has significant business impacts. It directly enhances the trust that business teams have in prediction outcomes and supports data-driven decision-making. Feature importance analysis helps organizations identify key business drivers, such as sales and discount strategies, enabling them to optimize resource allocation and strategic planning. Furthermore, through both local and global interpretability, business teams can understand the model's prediction logic, which builds their confidence in the results. Additionally, residual analysis and the identification of feature interactions aid in detecting model shortcomings, guiding businesses in adjusting their strategies accordingly.

3.3 Limitations and Assumptions

In the project, there are several limitation and assumption that need to be acknowledged. In terms of limitation, the sensitivity of MAPE to negative and near-zero values in profit. Although translation of profits into positive values addressed this issue, it may reduce the interpretative value of negative profits. The values might be an important indicator of losses in business contexts. In addition, the feature importance analysis might not fully consider external factors such as economic condition, competitive dynamics, or consumer preference which could influenced the business profit. Another key limitation, the predictive model depends on historical data. Therefore, it limits the model's ability to capture dynamic and unpredicted market conditions.

In terms of assumption, we assumed that translating the profits to positive values will not influence the relationships between features and the target variable or introduce bias in predictions. However, this assumption might not always be correct, especially in datasets with a lot of features. Next, we also assumed that random data partitioning into training, validation and testing are consistent distributed and and does not disrupt patterns present in the data. In practice, real-world data may showed distinct structures or trend which could be misrepresented by random partitioning. Lastly, we assumed that the important features that was generated showed the real factors driving profit and are not heavily affected by missing data or other factors.

To address the limitation, integration of external data such as competitor pricing or economic trends might reveal the key driver that influence profit. On the other hand, further analysis of outlier and implementation of alternative predictive model may improve prediction accuracy and adaptability to unforeseen conditions.

3.4 Summary

Through the automation capabilities and data transformation strategies of the Google Cloud AutoML platform, we effectively resolved the evaluation indicator anomalies caused by negative profit values, significantly enhancing the overall performance of the model. In response to the complexities associated with data anomalies, we conducted systematic data preprocessing and feature engineering to ensure the accuracy and consistency of the model during the evaluation phase. Ultimately, the performance metrics and feature importance distribution of the constructed model not only meet the practical requirements of business applications but also provide a reliable theoretical basis for subsequent decision support.

This research outcome demonstrates the application potential of advanced machine learning technologies in addressing real-world issues and further underscores the importance of data-driven decision-making. Through this innovative methodology, we provide practical solutions for enterprises to optimize resource allocation and improve operational efficiency in dynamic and complex data environments. This process not only reinforces our confidence in the role of modern data technologies in driving business growth but also lays a foundation for future research in related fields.

4 Conclusion and Recommendations

In conclusion, the tools used in this project enhanced our understanding of the data mining process from preprocessing to modeling and visualization. Google DataPrep streamlined data cleaning and preprocessing, Google Cloud AutoML simplified predictive modelling, and Power BI enabled interactive dashboard creation. The automation tools used has improved our productivity.

The analysis of the Superstore Dataset provided critical insights into the supermarket's performance. These insights demonstrated crucial trends in sales, profit, and the behaviour of customers. The Technology category, particularly Copiers, emerged as the most profitable. In terms of sales and profits, the states of California and New York were the most significant contributors to the region. The sales and profits performance showed that strong sales performance does not guarantee high profit. Profit depends on margins, operational efficiency, discounts and cost management. For instance, Texas ranks among top 10 region in terms of sales but it does not appear in the top 10 regions for profit. The analysis also revealed that excessive discounts negatively impacted profitability.

Next, predictive modeling using Random Forest is a reliable tool to forecast profits and identify key factors that influence profitability. Although the initial results showed overfitting and high error metrics, data transformation can normalize profit values and improve the model's reliability and accuracy. Additionally, tools like Google AutoML make it easy to perform automated model tuning and optimization. Overall, the feature importance analysis highlighted that sales and discounts were the most critical factors that influence model prediction. Therefore, strategic pricing and sales initiatives are needed to attract customers and maximize profitability. In addition, the residual analysis and the comparison of predicted values to actual values demonstrated that the model produced accurate and consistent predictions with only a small amount of systematic bias. We can conclude that the model offered reliable approach to forecast profit and serves as a valuable tool for data-driven decision-making in business operations.

Moving forward, it is recommended that the business focus on high-performing categories, optimize discounting strategies, and enhance customer segmentation to maximize profitability. The business should expand marketing of high-performing products in top-performing regions. Targeted campaigns in these areas can maximize returns by utilising their strong client base and demand. For future studies, it is recommended to use diverse data with external factors such as market trends and real-time data to improve business insights. In terms of the predictive model, it is recommended to regularly retrain the model with updated data to improve performance. On the other hand, it is also recommended to compare the

production with other models and expand predictive capabilities to include metrics such as demand forecasting and scenario analysis. These recommendations will further strengthen strategic business decisions.

5 Team Collaboration

In this project, effective collaboration among team members was key to achieve our objectives. Each team member played an important role to contribute to different phases of the project:

Team Member	Role	Key Contributions
Nur Hidayah Binti Ahmad Shafii	EDA and Documentation & Reporting Lead	Conducted exploratory data analysis, identified key trends, and compiled the final report to present findings clearly.
Wudi	Data Selection, Preparation, and Presentation Lead	Handled the selection and preprocessing of the dataset, ensuring data quality and consistency, and led the project presentation.
Luchanghao	Modeling Specialist	Focused on predictive modeling, utilizing Google Cloud AutoML to build, train, and optimize the model for forecasting.
Hejunfeng	Model Evaluation & Interpretation Lead	Conducted model evaluation, analyzed performance metrics, and provided insights into the reliability and accuracy of predictions.

Collaboration Tools

- i. Google Drive: Facilitated collaborative editing of documents and data files.
- ii. WhatsApp: Enabled real-time communication for updates and task coordination.
- iii. Microsoft Teams: Used for virtual meetings to discuss progress and file sharing.

Collaboration Approach

- i. Weekly Meetings: Scheduled meetings ensured consistent progress tracking, task distribution, and resolution of challenges.
- ii. Task Assignments: Responsibilities were allocated based on each member's strengths and expertise to maximize efficiency.

6 References

Plante, T., & Cushman, M. (2020). Choosing color palettes for scientific figures. *Research and Practice in Thrombosis and Haemostasis*, 4. <https://doi.org/10.1002/rth2.12308>