Oxford DataPlan – Revenue Index Prediction Write-Up

Candidate: Vanshika Arora
Task: Predict revenue_index for ABC-Retail using historical order and transaction data

Overview

The task involved building a predictive model for the revenue_index of an online retailer ABC-Retail using three datasets:

- Order Data: Dates and sequential order numbers.
- Transaction Data: Daily total spend, gross orders, and active users
- Reported Revenue Data: Quarterly revenue index to be predicted.

I trained the model using data up to 2022 Q3 and evaluated on 2022 Q4.

Step 1: Data Cleaning

- Order Data: Removed 200 bad rows where order numbers decreased violating the monotonicity assumption.
- Calculated daily order_diff (number of orders between dates).
- Converted all dates to proper datetime formats and sorted them .

Step 2: Feature Engineering

- Aggregated daily transaction and order data into **quarterly features**, matching the format of revenue_index.
- Engineered features:
  - order_sum (total new orders in quarter)
  - spend_index_avg, gross_orders_avg, users_avg
  - normalized_spend and normalized_orders (adjusted by active users)
  - spend_per_order = spend_index / gross_orders_index
  - orders_per_user = gross_orders_index / weekly_active_users_index
  - days_in_period = number of days in each quarter
- All features were averaged or summed over the quarter to align with reported periods.

Step 3: Model Training

- Used Linear Regression to model revenue_index on 19 training observations.
- Also tried Ridge Regression, but it underperformed due to regularization on a small dataset.

Train-Test Split:

- Training: All periods up to 2022 Q3
- Test: Only 2022 Q4

Step 4: Evaluation

Metric	Value
In-sample R^2 (Linear)	0.8773
Predicted Q4 Revenue Index	348.78
Actual Q4 Revenue Index	512.08
Percentage Error	31.89%

Although the model captured overall patterns well (high R^2), Q4 2022 showed a significant jump.

Visualizations

- Actual vs. Predicted Revenue Index (Bar Chart)
- Forecast-style Line Plot (showing Q4 as future/test point)

How I Avoided Overfitting?

To prevent overfitting on a small dataset :

- Used a simple Linear Regression model
- Limited to a few interpretable features with business logic
- Preserved 2022 Q4 as a true hold-out test set
- Tried Ridge Regression as regularization but found no improvement
- Avoided excessive tuning or complexity