Sales Prediction Hacking Process - Approach Notes

Objective

To minimize RMSE in predicting `Item_Outlet_Sales` using advanced feature engineering and multiple regression models (Linear, Polynomial, Random Forest, XGBoost).

Baseline Models

Without feature engineering, models quickly saturated:

- **XGBoost RMSE:** 1089
- **Linear Regression RMSE:** 1212
- **Polynomial Regression RMSE:** 1090
- **Random Forest RMSE:** 1132

Observation: Baseline models could not capture complex item-outlet interactions.

Feature Engineering Process

Feature engineering was done iteratively, using domain knowledge and experimental tuning. Each feature was added, tested, and refined based on RMSE reduction.

1. Item_Sales_Frequency

Goal: Capture item popularity relative to outlet age, price, visibility, and weight.

Formulas:

Trial 1:

```
Item_Sales_Frequency = (Outlet_Age × (Item_MRP - Item_Visibility)) / (Item_Weight + 1)
```

Trial 2:

```
Item_Sales_Frequency = log(1 + Outlet_Age) × (Item_MRP / (Item_Weight + 1)) × Item_Popularity
```

Trial 3 (Best - Polynomial Regression RMSE: 1038.26):

```
Item_Sales_Frequency = log(1 + Outlet_Age) × ((Item_MRP - mean(Item_MRP)) / (std(Item_MRP) + 1)) ×
(Item_Popularity + 0.01)
```

Explanation: Combines outlet age effect, standardized price, and item popularity with smoothing to prevent zero-frequency bias.

2. Customer_Outlet_Preference

Goal: Model customer preference for outlet types based on item price, visibility, and outlet popularity.

Final Formula (Polynomial Regression RMSE: 1038.26):

```
Customer_Outlet_Preference = √(Item_MRP / median_MRP) × (1 / (1 + log(1 + Item_Visibility))) ×
(Outlet_Type_Percentage / Outlet_Location_Type)
```

Explanation: This feature captures the relationship between item pricing relative to market median, adjusted for visibility effects, and weighted by outlet type preferences in different location categories.

Key Insights

1. **Feature Engineering Impact:** Advanced feature engineering reduced RMSE from ~1089-1212 to 1038.26, representing a significant improvement in prediction accuracy.
2. **Iterative Refinement:** Multiple trials of feature formulation were necessary to achieve optimal performance, with Trial 3 of Item_Sales_Frequency providing the best results.
3. **Domain Knowledge Integration:** Successful features incorporated business logic around outlet age, item popularity, price positioning, and customer preferences.
4. **Model Performance:** Polynomial Regression emerged as the best-performing model with the engineered features, achieving RMSE of 1038.26.

Technical Notes

- All features included appropriate smoothing terms (e.g., +0.01, +1) to prevent division by zero and extreme values
- Standardization was applied where necessary to ensure feature stability
- Logarithmic transformations were used to handle skewed distributions and reduce the impact of outliers

Appendix: Explanation of Feature Engineering Equations

1. Item_Sales_Frequency

Trial 1:

Item_Sales_Frequency = (Outlet_Age × (Item_MRP - Item_Visibility)) / (Item_Weight + 1)
• Outlet age scales demand over time.
• Item MRP adjusted for visibility (visibility reduces effective pricing power).
• Item weight in denominator acts as a smoothing factor.

Trial 2:

Item_Sales_Frequency = log(1 + Outlet_Age) × (Item_MRP / (Item_Weight + 1)) × Item_Popularity
• Log transformation dampens the effect of very old outlets.
• Price-to-weight ratio represents cost efficiency.
• Popularity is added to capture demand dynamics.

Trial 3 (Best):

Item_Sales_Frequency = log(1 + Outlet_Age) × ((Item_MRP - mean(Item_MRP)) / (std(Item_MRP) + 1)) × (Item_Popularity + 0.01)
• Outlet age still captured via log scaling.
• Standardized price (mean-centering and scaling by std) ensures comparability across items.
• Small constant prevents zero-frequency bias.

2. Customer_Outlet_Preference

Customer_Outlet_Preference = √(Item_MRP / median_MRP) × (1 / (1 + log(1 + Item_Visibility))) × (Outlet_Type_Percentage / Outlet_Location_Type)
• Price effect normalized against market median (square root moderates extremes).
• Visibility penalized logarithmically (items too visible may be discounted in preference).
• Outlet preferences weighted by type distribution across locations (captures customer heterogeneity).