Predicting Samsung Cell Phone Prices

Context

In this exercise, we focus on building regression models to predict the prices of Samsung cell phones based on various features. Accurate price prediction is crucial for companies in the mobile phone industry, as it helps in setting competitive prices, managing inventory, and planning marketing strategies. By analyzing features such as screen size, battery type, RAM, weight, storage capacity, screen resolution and time since release, we can understand how each attribute impacts the overall price. Understanding these relationships also provides insights into consumer preferences, which can guide future product development and feature enhancements.

You will find the dataset in the excel file 'data_Samsung_HW2.xlsx'. Here is a brief description of the variables.

- **inches**: Screen size in inches.
- **battery**: Numeric representation of battery capacity.
- **battery_type**: Type of battery, either Li-Ion or Li-Po.
- **ram(GB)**: RAM capacity in gigabytes.
- **weight(g)**: Weight of the phone in grams.
- **storage(GB)**: Internal storage capacity in gigabytes.
- **price(USD)**: Price of the phone in US dollars.
- **Resolution_product**: Natural logarithm of the product of the screen resolution dimensions.
- **days_from_release_date**: Number of days from the release date compared to January 1st, 2024.
- **up_to_8K**: Video recording capability, where:
  - 1 = up to 720p,
  - 2 = up to 1080p,
  - 3 = up to 4K,
  - 4 = up to 8K.

Questions:

1. **<u>Model 1</u>: Basic Linear Regression**

   Estimate a basic linear regression model using the provided variables to predict the price of Samsung cell phones. The model equation is:

$$\begin{aligned}\text{price(USD)}_i = \beta_0 &+ \beta_1 \times \text{Li-Ion}_i + \beta_2 \times \text{inches}_i + \beta_3 \times \text{battery}_i \\ &+ \beta_4 \times \text{ram(GB)}_i + \beta_5 \times \text{weight(g)}_i + \beta_6 \times \text{storage(GB)}_i \\ &+ \beta_7 \times \text{Resolution\_product}_i + \beta_8 \times \text{days\_from\_release\_date}_i \\ &+ \beta_9 \times \text{up\_to\_8K}_i + \epsilon_i,\end{aligned}$$

   in which Li-Ion is a dummy variable being equal to one when the battery type is Li-Ion and zero otherwise.

   (a) Interpret the significant parameters in the model at a 95% confidence level. Discuss how each significant variable affects the price. Note that a significant parameter is a parameter for which we reject the Null hypothesis of being equal to 0.

   (b) Use the Leave-One-Out Cross-Validation (LOO-CV) to compute the Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) for Model 1. Provide the values of the RMSE and the MAE.

2. **Model 2: Including Interaction and Quadratic Terms**

   Extend Model 1 by including interaction terms and quadratic terms for the numerical features (*excluding* the dummy variable). The model equation is:

$$\begin{aligned}\text{price(USD)}_i = \beta_0 + \beta_1 \times \text{Li-Ion}_i + \sum_k \beta_k \times X_i^{(k)} \\ + \sum_k \beta_{kk} \times (X_i^{(k)})^2 + \sum_{k<j} \beta_{kj} \times X_i^{(k)} \times X_i^{(j)} + \epsilon_i,\end{aligned}$$

   where $X_i^{(k)}$ represents each numerical explanatory variable: inches, battery, ram(GB), weight(g), storage(GB), Resolution_product, days_from_release_date, and up_to_8K.

   To compute the interaction and the quadratic terms, you can use the **PolynomialFeatures** function in the scikit-learn package. Here is an example on how to use the function.

   Python Code:

```
from sklearn.preprocessing import PolynomialFeatures
# Generate polynomial features (quadratic terms and interactions)
poly = PolynomialFeatures(degree=2, include_bias=False)
X_poly = poly.fit_transform(data_num)
```

   Explanation: The `PolynomialFeatures` class from the `sklearn.preprocessing` package is used to automatically generate a new feature matrix including all polynomial combinations of the features up to the specified degree (in this case, 2 for quadratic). By setting `include_bias=False`, we exclude the bias (intercept) term since you can add it separately.

   (a) Explain why you should not include the dummy variable (battery type) when adding quadratic terms.

   (b) Based on the in-sample estimation results (i.e., the summary output of the linear regression), do you expect this model to have better predictive performance compared to Model 1? Explain your reasoning.

   (c) Use the LOO-CV to compute the RMSE and MAE for Model 2. Discuss your results in comparison to Model 1.

3. <u>Model 3</u>: **Transforming the Dependent Variable**

Since the price variable is typically skewed, apply a natural logarithm transformation to the price to make it more symmetric. Estimate a linear regression model using the log-transformed price as the dependent variable with the same explanatory variables as Model 1. The model equation is:

$$\begin{aligned}\ln(\text{price(USD)})_i = \beta_0 &+ \beta_1 \times \text{Li-Ion}_i + \beta_2 \times \text{inches}_i + \beta_3 \times \text{battery}_i \\ &+ \beta_4 \times \text{ram(GB)}_i + \beta_5 \times \text{weight(g)}_i + \beta_6 \times \text{storage(GB)}_i \\ &+ \beta_7 \times \text{Resolution\_product}_i + \beta_8 \times \text{days\_from\_release\_date}_i \\ &+ \beta_9 \times \text{up\_to\_8K}_i + \epsilon_i.\end{aligned}$$

(a) Based on the in-sample estimation results (i.e., the summary output of the linear regression), do you expect this model to have better predictive performance than Model 1? Explain your reasoning.

(b) Use the LOO-CV to compute the RMSE and MAE for Model 3. Discuss your results in comparison to Model 1. In particular, does it improve over Model 1 ?

4. <u>Model 4</u>: **Improving Predictive Performance**

(a) Develop a linear regression model that achieves better predictive performance (in terms of RMSE) than the best-performing model from the previous questions. You may consider techniques like feature selection, adding or removing interaction terms, applying different transformations or any other method as long as you still use a linear regression model.

(b) Explain the changes you made to improve the model and why they might lead to better predictive performance.