Electronic Device Rating Prediction

MS1 Report

Team: CS_27

Seat Number	Section	Name
2021170642	8	يوسف محمد سيد محمد
2021170480	6	محمد عوض طلعت محمد
2021170654	8	يوسف وجيه وديع ناشد
2021170649	8	يوسف مصطفى محمد يوسف
2021170483	6	محمد مبروك فكري عبدالفتاح
2021170641	8	يوسف عمرو احمد زهران

Preprocessing Techniques

Introduction

Data preprocessing is a crucial step in preparing raw data for analysis. It involves cleaning, transforming, and selecting relevant features to improve the quality and effectiveness of subsequent modeling. In this document, we explore various preprocessing techniques commonly used in data science.

1. Data cleansing:

Data cleansing aims to remove noise, inconsistencies, and irrelevant information from the dataset. Key steps include:

- **String Removal**: Eliminate extraneous strings (e.g., “th,” “GB”) from columns such as “processor_generation,” “ram_GB,” “SSD,” and “graphics_card_GB.
- **Handling Missing Values**: Address missing data points by imputing or removing them.
- **Outlier Treatment**: Detect and handle outliers using the interquartile range (IQR) method.

2. Data Transformation

Data transformation involves converting features into a suitable format for modeling. Notable techniques include:

- **Encoding**: Apply encoding to categorical columns like “ram_type” and “weight.”
- **One-Hot Encoding**: Convert nominal features (e.g., “OS_type,” “processor_brand”) into binary vectors.

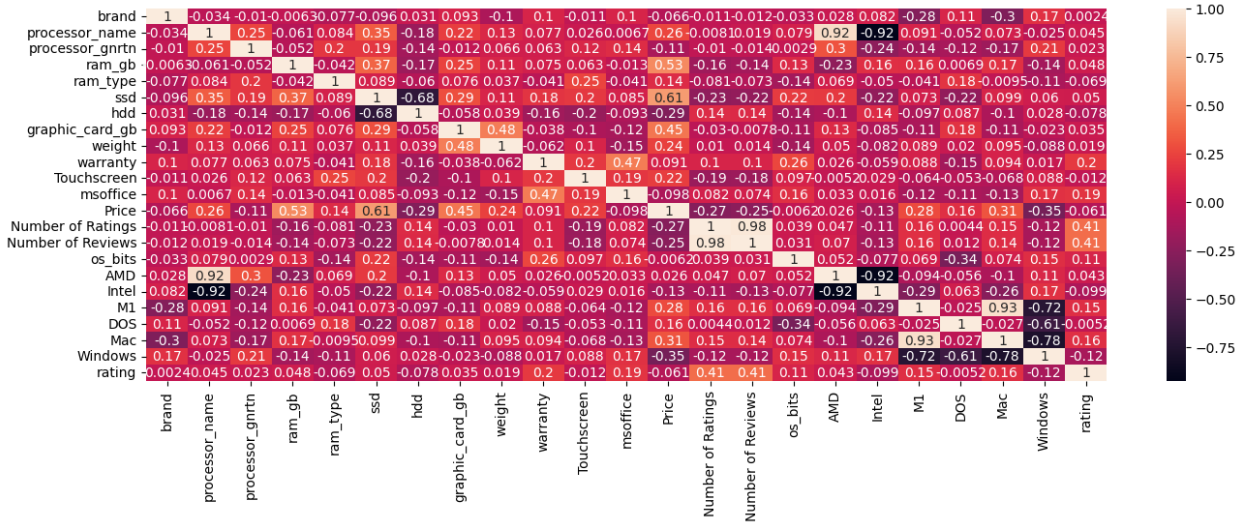
3. Feature Engineering

Feature engineering enhances model performance by creating new relevant features. Examples include:

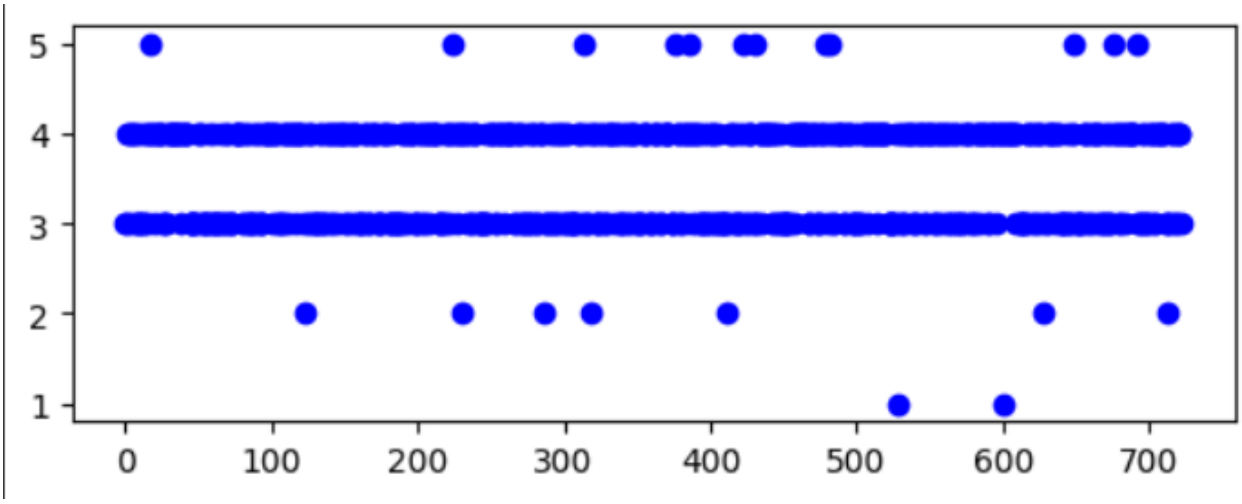
- **Splitting OS Information**: Divide the “OS” column into “OS_type” and “OS_bits.”

4. Feature Selection

- Feature selection reduces dimensionality and focuses on the most informative features:
- **Correlation Analysis**: Remove weakly correlated features (e.g., “brand,” “weight”).



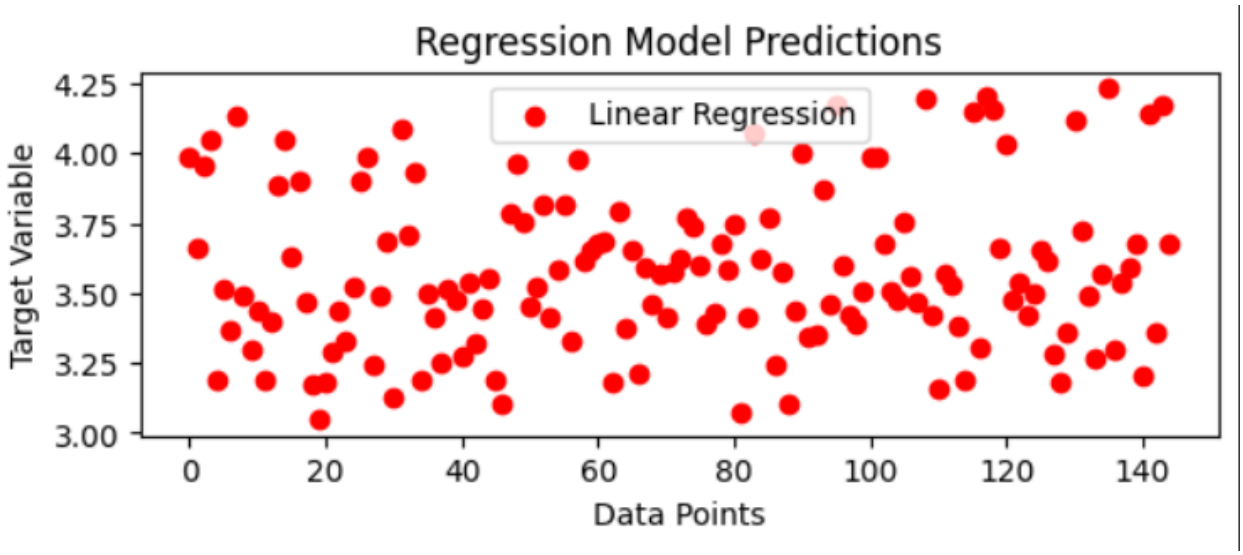
Data



Model Evaluation

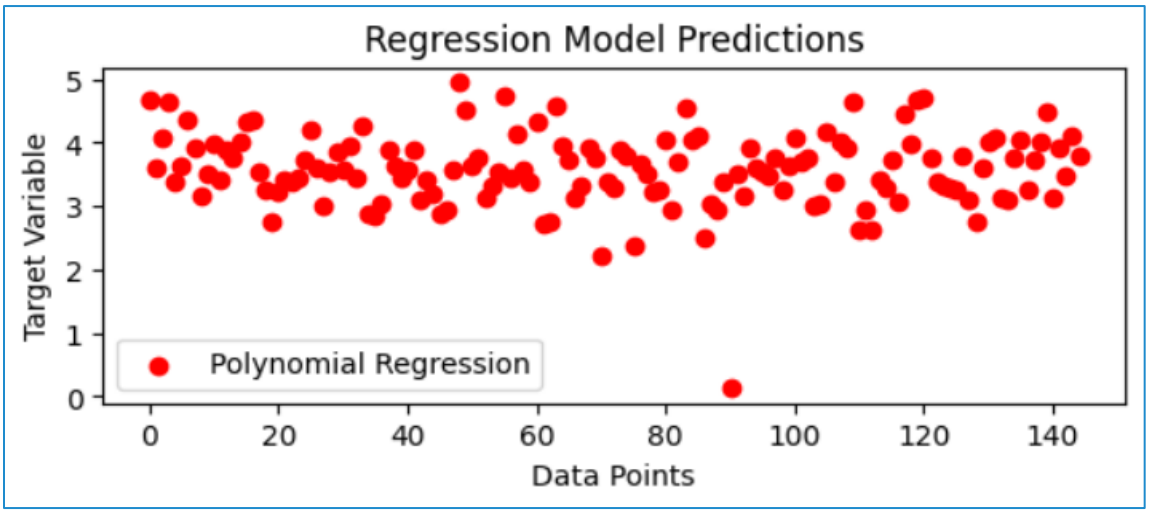
1. Linear Regression:

- Training MSE: 0.24
- Testing MSE: 0.20
- Provides a linear equation describing variable relationships.



2. Polynomial Regression:

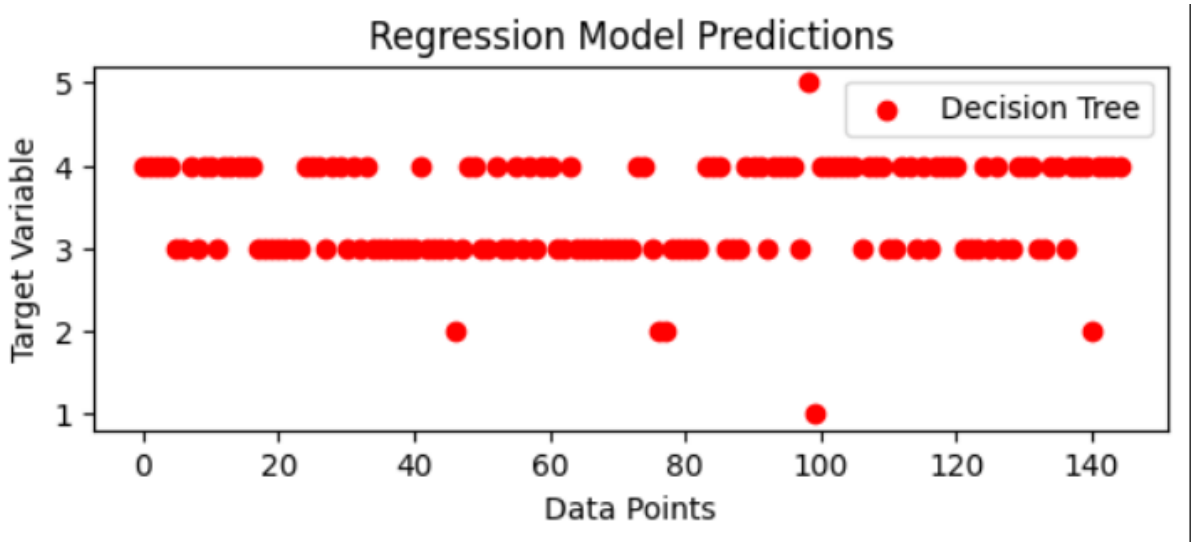
- Training MSE: 0.16
- Testing MSE: 0.38
- Polynomial regression is an extension of linear regression that models the relationship between the independent and dependent variables as a nth-degree polynomial. It can capture non-linear relationships between.



3. Decision Tree Regression:

- Training MSE: 0

- Testing MSE: 0.24
- Non-parametric approach based on tree-like splits.



Features used:

Used features according to the heights correlation:

- warranty
- msoffice
- Number of Ratings
- Number of Reviews
- os_bits
- M1
- Mac
- Windows

Sets size:

1. Training set: 80%
2. Testing set: 20%

Conclusion:

Effective preprocessing significantly impacts model performance. By applying these techniques, data scientists can enhance the quality of their analyses and build more accurate predictive models.