CIA 1

Application of various Regression models in Mobile Electronics sector

Submitted By

AELUKA VENKAT REDDY

Registration Number

2327504

Submitted to

Prof. Helen Josephine

MACHINE LEARNING ALGORITHMS-1

SCHOOL OF BUSINESS AND MANAGEMENT

CHRIST (DEEMED TO BE UNIVERSITY), BENGALURU

JULY 2024

Business Understanding

a. Problem Identification:

Predict the sales price of a smartphone using various features.

b. Variables

Dependent Variable

- **sales_price**: This is the target variable we aim to predict. It represents the sales price of the mobile phone.

Independent Variables

1. **brand**: The brand of the mobile phone (e.g., Samsung, Apple).
2. **model**: The specific model of the mobile phone.
3. **base_color**: The primary color of the mobile phone.
4. **processor**: The type of processor used in the mobile phone.
5. **screen_size**: The size of the mobile phone screen in inches.
6. **ram**: The RAM (Random Access Memory) size of the mobile phone in GB.
7. **rom**: The ROM (Read-Only Memory) size of the mobile phone in GB.
8. **display_size**: The display size of the mobile phone in inches.
9. **num_rear_camera**: The number of rear cameras on the mobile phone.
10. **num_front_camera**: The number of front cameras on the mobile phone.
11. **battery_capacity**: The battery capacity of the mobile phone in mAh (milliampere-hour).
12. **ratings**: The average user rating of the mobile phone (typically out of 5).
13. **num_of_ratings**: The number of user ratings received by the mobile phone.
14. **discount_percent**: The discount percentage applied to the sales price.
15. **sales**: The total number of sales of the mobile phone.
    -

c. Objectives

- To understand how individual features influence the sales price using simple

linear regression.

- To explore the relationship between multiple features and the sales price using multiple linear regression.
- To perform feature selection and predict the sales price using lasso regression.
- To handle multicollinearity and predict the sales price using ridge regression.

2. Business Understanding

1. **Problem Identification:** The primary challenge is to forecast Sales price of mobiles accurately based on its features.
2. **Variables:** Key predictors include color, RAM, RAM, battery capacity, ratings. Sales price serve as the dependent variable.
3. **Objectives:** Develop predictive models to assist customers in estimating charges and understanding the factors driving sales price, supporting informed decision-making.

3. Data Understanding

1. **Data Collection:** The dataset comprises data about Flipkart mobile sales, while ensuring data privacy and regulatory compliance.
2. **Data Exploration:** Exploratory data analysis techniques were employed to understand variable distributions, relationships, and potential correlations using statistical summaries and visualizations.
3. **Assessing Data Quality:** The dataset underwent rigorous checks for completeness, accuracy, and consistency to ensure reliability in subsequent analysis.

DATA DICTIONARY:

Column name	Description
Brand	The brand of the mobile phone (e.g., Samsung, Apple).
Model	The specific model of the mobile phone.
Base_color	The primary color of the mobile phone.
Processor	The type of processor used in the mobile phone.
Screen size	The size of the mobile phone screen in inches.
RAM	The RAM (Random Access Memory) size of the mobile phone in GB
ROM	The ROM (Read-Only Memory) size of the mobile phone in GB.
Display_size	The display size of the mobile phone in inches.
Num_Rear_Camera	The number of rear cameras on the mobile phone.
Num_Front_Camera	The number of front cameras on the mobile phone.
Battery Capacity	The battery capacity of the mobile phone in mAh (milliampere-hour).
Ratings	The average user rating of the mobile phone (typically out of 5)

No of Ratings	The number of user ratings received by the mobile phone.
Sales_price	The sales price of the mobile phone in the local currency.
Discount_percentage	The discount percentage applied to the sales price.
Sales	The total number of sales of the mobile phone.

4. Data Preparation

1. Data Integration: Integrated data from multiple sources to ensure consistency and compatibility for analysis.
2. Data Cleaning: Handled missing values using appropriate imputation techniques, standardized numerical variables, performed feature engineering, and addressed outliers to enhance data quality.

5. Modeling

1. Model Selection and Assumptions: Selected Multiple Linear Regression, Lasso Regression, and Ridge Regression based on their suitability and assumptions of linearity and error independence.
2. Model Output: Interpreted model equations, parameter coefficients, and fit indices

(e.g., R-squared) to assess predictive accuracy and variable impacts on insurance charges.

3. Model Interpretation from a Business Point of View: Provided insights into how demographic and lifestyle factors influence insurance costs, aiding insurers in pricing strategies and risk assessment.

6. Model Evaluation and Diagnostics

- **Performance Metrics:** Evaluated models using R-squared (variance explained) and Mean Squared Error (accuracy of predictions) to compare and recommend the best-performing model.
- **Comparison of Model Performance:**

MODEL	R-Squared	MSE	Recommended Model
Simple linear Regression	0.1685	327,537,719.1635	
Multiple linear regression	0.6503	146,311,377.5869	
Lasso Regression	0.6267	137,677,898.9946	Best fit Model due to balance between R square and MSE
Ridge Regression	0.6188	140,575,091.114	

Conclusion: Lasso Regression model is best fit because there is a balance between fit (R-squared) and predictive accuracy (MSE). Although its R-squared is slightly lower than that of Multiple Linear Regression, its MSE is the lowest, indicating superior predictive performance.

THANK YOU!