Simple Linear Regression Report

Introduction

This report explains how we used a statistical method called Simple Linear Regression to understand the relationship between two variables in a dataset.

- One variable is the target (Y) – what we want to predict.
- The other is the predictor (X) – what we use to make predictions.

We tested different types of linear regression models and compared their performance. This report summarizes our findings and gives recommendations on which model to use.

What is Simple Linear Regression?

Simple Linear Regression helps us find out if there's a straight-line relationship between two variables.

The formula for Simple Linear Regression is: **Y = b0 + b1 * X**

Where:

- Y is the target variable (what we are trying to predict).
- X is the predictor variable (what we are using to predict Y).
- b0 is the starting point of the line (intercept).
- b1 is how steep the line is (slope or coefficient).

Example:
If we are predicting house prices based on size, then:

- X = house size
- Y = house price
- b0 = base price even for a very small house
- b1 = how much the price increases with each extra square foot

Methodology (Steps We Took)

1. Data Preprocessing

Before building the model, we made sure the data was clean and ready to use.

- Checked for missing values: We made sure no data was missing.
- Exploratory Data Analysis (EDA): We looked at graphs and statistics to understand how the data was spread out.
- Split data into training and testing sets:
    - Training set (70%) : Used to build the model.
    - Testing set (30%) : Used to check how well the model works on new data.

2. Model Training

We used a Python library called sklearn.linear_model.LinearRegression to train the model.
This means we let the computer find the best possible line that fits the data.

3. Performance Evaluation

We used three main ways to measure how good the model is:

- R^2 Score (R-squared): Tells us how well the model explains the changes in the target variable.
    - A score of 1.0 means perfect prediction.
    - A score of 0 means the model doesn't help at all.
- Mean Absolute Error (MAE): Average error in predictions.
- Mean Squared Error (MSE): Similar to MAE but punishes bigger errors more.

Model Comparison Results

We tested several versions of linear regression models. Here's what they do and how they performed:

NO	MODEL TYPE	BEST USE CASE	R^2 SCORE
01	Standard Linear Regression	Basic model for straight-line relationships	0.9358
02	Feature Scaled Regression	When features have different sizes	0.9358
03	Ridge Regression	To avoid overfitting (when model memorizes training data)	0.9358
04	Lasso Regression	To pick only the most important features	0.9358
05	Elastic Net Regression	Combines Ridge and Lasso	0.9358
06	Polynomial Regression	For curved (non-linear) relationships	0.9358
07	Robust Regression	When there are outliers (weird data points)	0.9358
08	Bayesian Ridge Regression	When using probability-based methods	0.9358
09	Quantile Regression	For predicting ranges (like percentiles)	0.9358

All models gave the same R^2 score of 0.9358 , which is very high. That means the model explains about 93.58% of the variation in the target variable.

Conclusion

Since all models had the same performance, we can say:

1. The data has a linear pattern.
   All models worked the same, so the relationship between X and Y is likely a straight line.
2. There's no overfitting.
   Regularization models like Ridge and Lasso didn't improve results, so the model isn't memorizing the training data.
3. No major outliers or non-linear patterns.
   Models like Robust Regression and Polynomial Regression also didn't help.
4. Feature scaling wasn't needed.
   Even without scaling, the model worked fine, so all the features were already on similar scales.

Recommendations

Based on our analysis, here's what we suggest:

1. Use Standard Linear Regression

- It's simple and easy to understand.
- Since the data is linear and clean, there's no need for more complex models.

2. Don't Overcomplicate Things

- Only use Ridge, Lasso, or Polynomial models if you see problems like overfitting or non-linear patterns.

3. Keep Practicing Good Habits

- Even though feature scaling didn't change results, it's still a good idea when working with regularization or different-sized features.

4. Check Other Metrics

- Look at Mean Absolute Error (MAE) or Root Mean Squared Error (RMSE) to better understand prediction errors.

5. Make Sure Data is Clean

- Always double-check for missing values, strange data points, or incorrect entries.

6. Try New Features

- If the model isn't accurate enough, try adding or changing features to get better predictions.

7. Upgrade Models When Needed

- If the data becomes more complex (e.g., many features or curves), consider more advanced models like Random Forest or Neural Networks.

8. Use Cross-Validation

- Test the model on different parts of the data to be sure it works well across all cases.

Final Recommendation

For this dataset, Standard Linear Regression is the best choice because it's simple, fast, and performs well. However, always test the model on new data and keep checking its performance over time. If the data changes or gets more complex, you may need to switch to a different model.