Academic Year	Module	Assessment Number
2A	Concepts and Technologies of AI	3

BIKE STORE REVENUE PREDICTION USING REGRESSION MODELS

Student Id	: 2438466
Student Name	: Anup Maharjan
Section	: L5CG22
Module Leader	: Mr. Siman Giri
Tutor	: Mr. Ronit Shrestha
Submitted on	: 11-02-2025

Table of Contents

Table of Figures

ABSTRACT

In this study, two regression models (Linear Regression and Deep Tree) are created to predict the revenue of a bike store using the dataset of 113,036 sales record. With the two approaches that were implemented custom linear regression and gradient descent, R^2 scores of 1.0 and 0.975 respectively were achieved. The feature selection analysis determined the key features for predicting revenue to be cost, profit, and unit pricing. The gradient descent model that has been optimized for the hyperparameters has shown real world performance, while the linear regression had perfect accuracy. The findings can inform bicycle retail operation business decision and revenue forecasting.

1. INTRODUCTION

1.1 Problem Statement

The challenge to predict the bike store revenue is addressed by this research using regression models. In today's competitive retail environment, it is important to be able to predict accurately to undertake business planning, inventory management and strategic decision making. With comprehensive sales data available, it studies the development of predictive models that a bike retailer can use to forecast revenue and key revenue drivers.

1.2 Dataset

Comprehensive bike store sales database of 113,036 records with 18 features are used for analysis. The dataset contains customer statistical data, detailed product information and key financial metrics. This UN Sustainable Development Goal 8 dataset is consistent with this approach of growing the economy sustainably with data driven practices. The rich feature variety constitutes a solid base for developing good quality prediction models.

1.3 Objective

The main goal of this research is to develop and evaluate regression models for accurate revenue prediction based on the available features from sales data. The aim of the study includes identifying and analysis of key revenue drivers and producing actionable sense for making business conclusions. Our goal with this analysis is to develop a dependable model for forecasting bike revenue in the bike retail industry.

2. METHODOLOGY

2.1 Data Preprocessing

A critical first step occurred in the initial data preparation phase there, with many critical steps to prepare the data in such a way to ensure data quality and reliability. We removed 1,000 duplicate entries from the dataset during preprocessing and cleaned them. It was found that the data had high completeness in reality, and there were no missing values. Temporal analysis using date fields would be converted to datetime format. Feature scaling was used for the gradient descent model to normalize the data and improve convergence of the model.

2.2 Exploratory Data Analysis (EDA)

Our exploratory analysis revealed several significant patterns in the data.

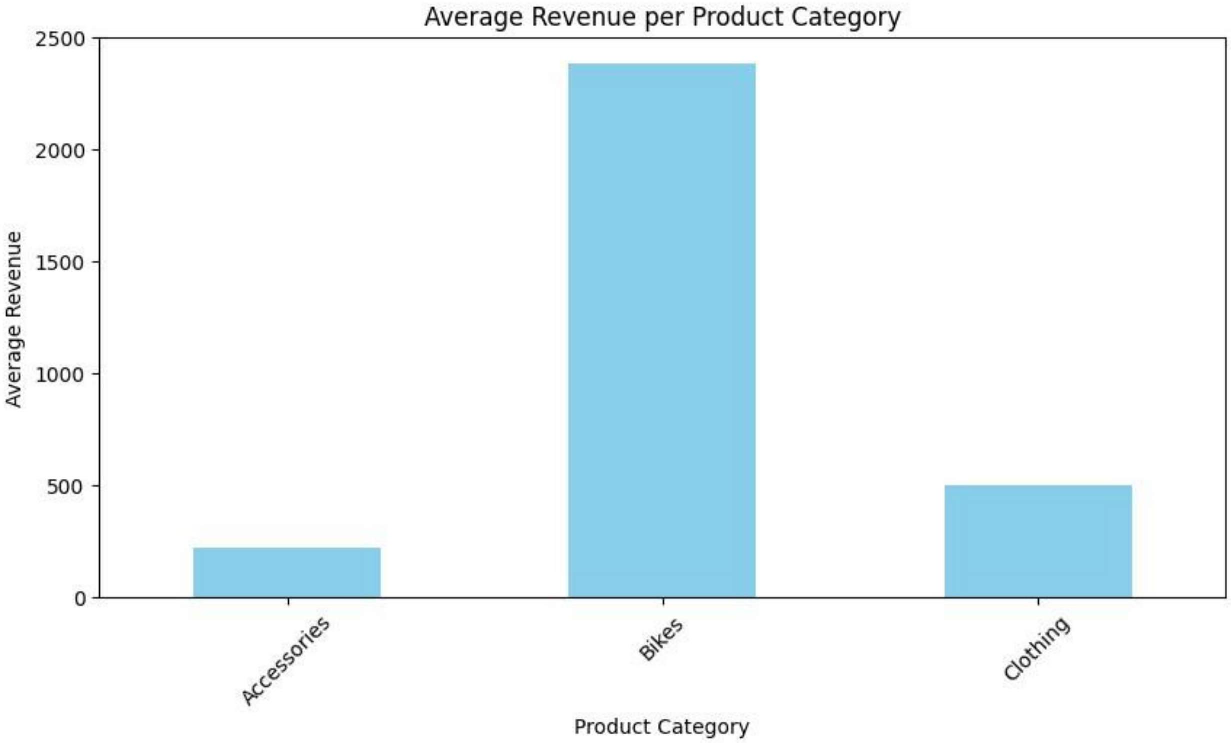

Figure 1 Bar Chart

The bar chart shows the average revenue for different product category. We can observe that highest revenue is gained from the bikes, secondly clothing and at last accessories.

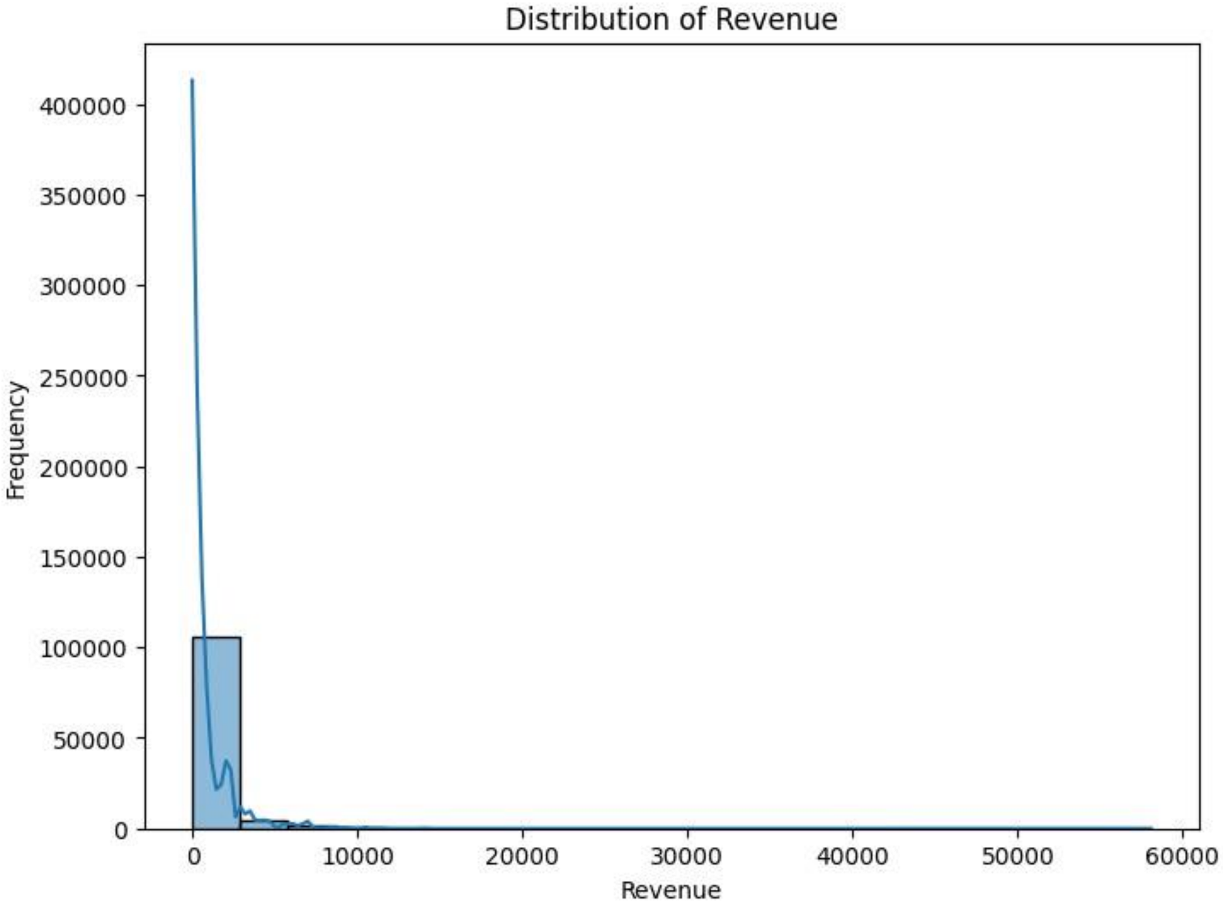

Figure 2: Histogram

The distribution of revenue showed a Right Skew / Positively Skew thus most of the revenue is concentrated at a lower end with the outliers. This means that most of the sales are of low revenue; however, some sales earn very high revenue.

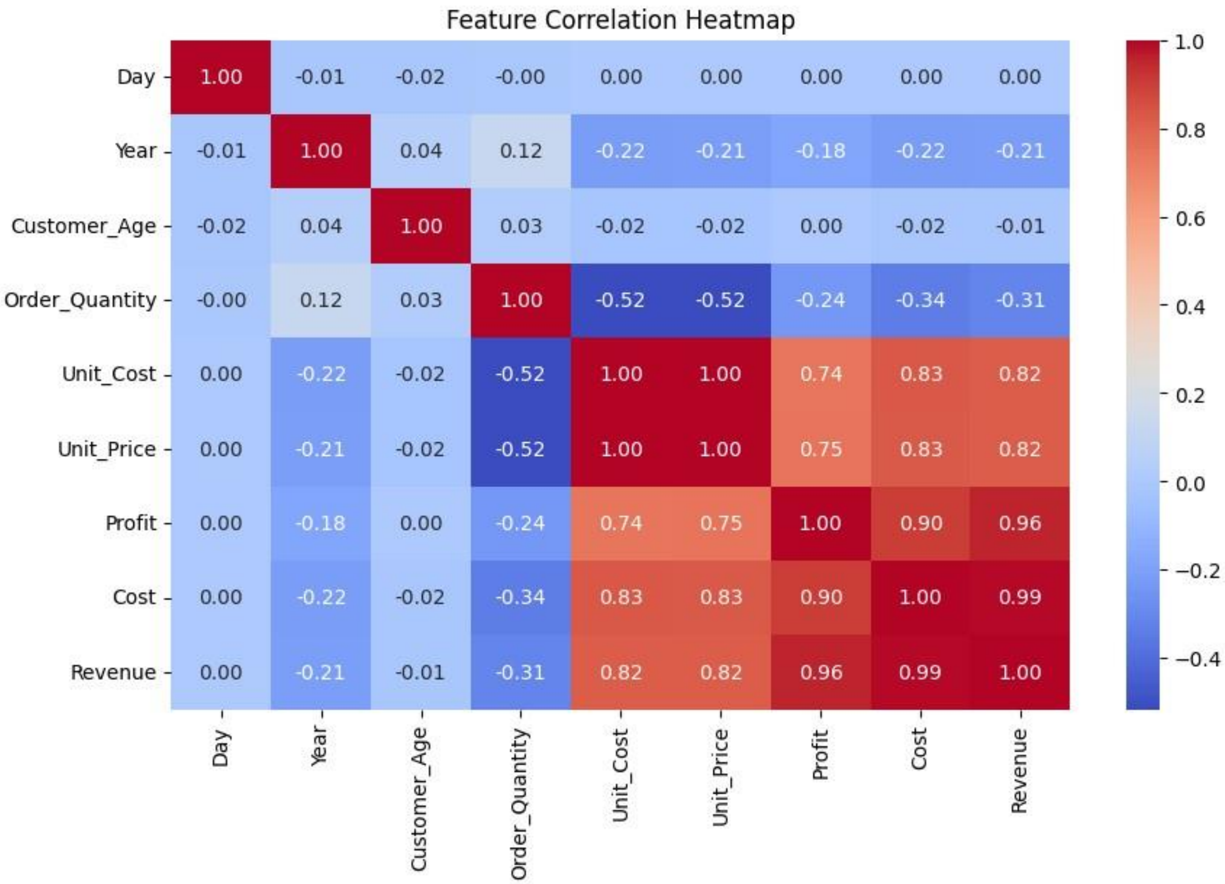

Figure 3: Heat Map

The Heatmap shows the correlation analysis uncovering the relationships strength between the different features in the dataset. It shows perfect correlation between Unit _Cost and Unit_Price (1.0), and very strong correlations among profit, cost, and revenue (0.90-0.99). Order_Quantity have moderate positive correlations with Unit_Cost, Unit_Price, Profit, Cost, and Revenue (0.24 to 0.52).

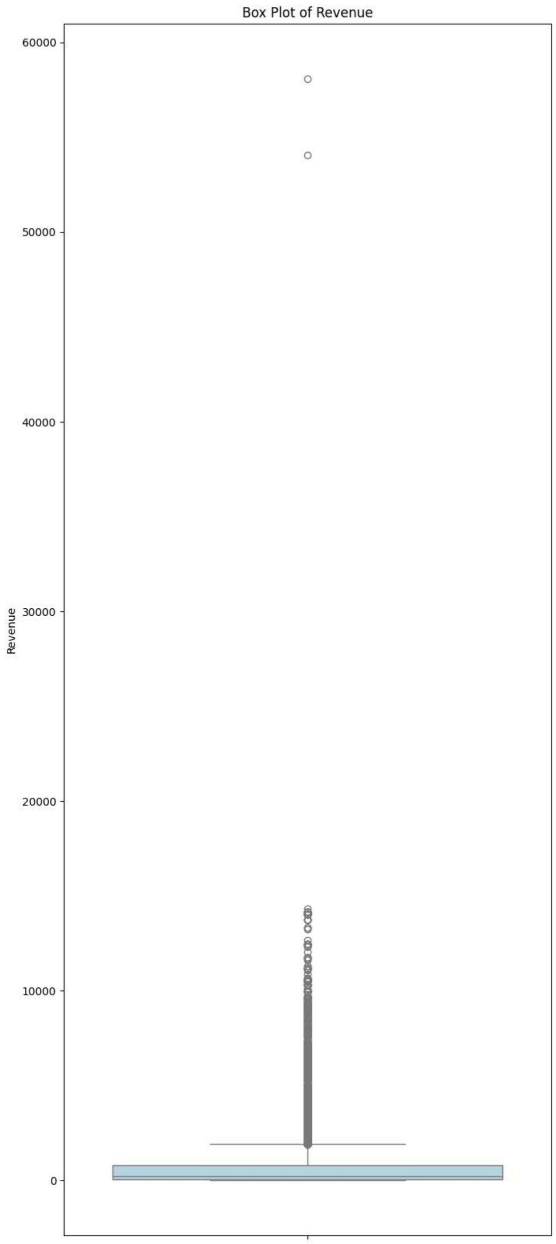

Figure 4: Box Plot

The box plot of revenue shows that most of the data are concentrated near the end and have a large number of outliers. The IQR is small compared to the spread of the data, clustering most of the values at the lower end.

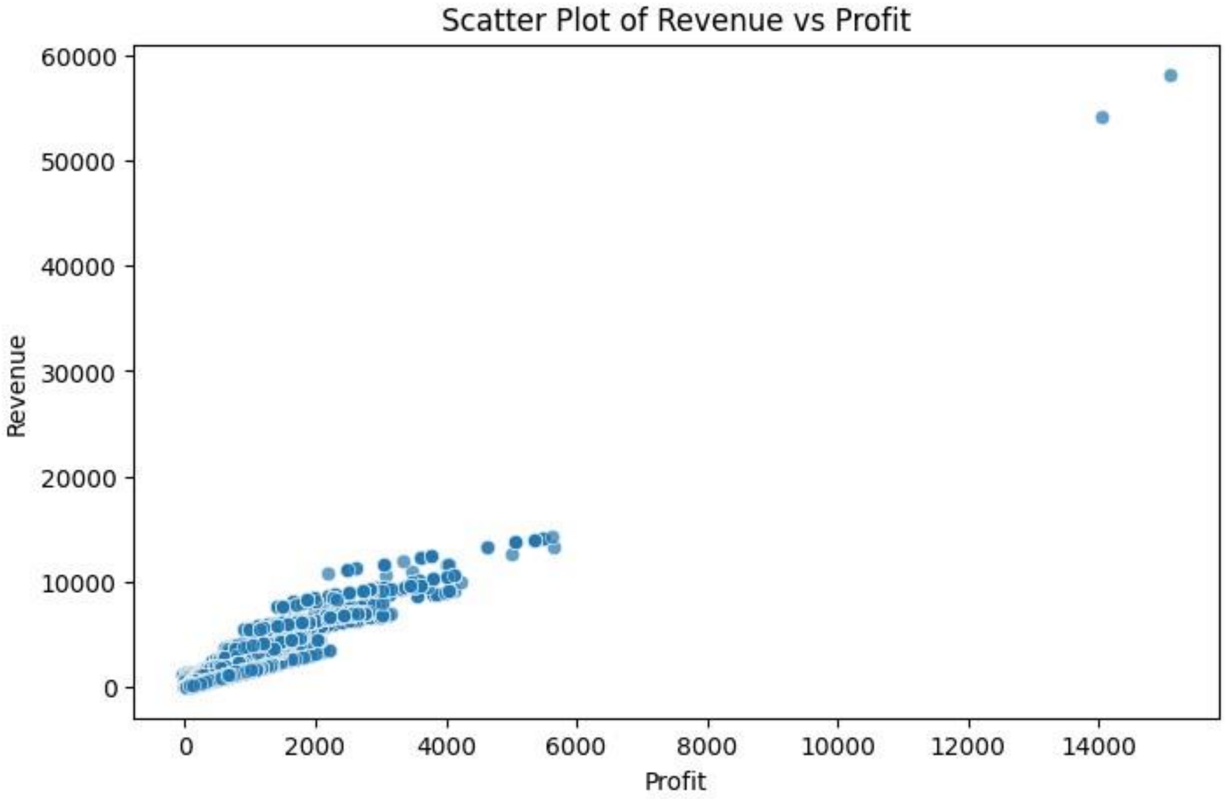

Figure 5: Scatter Plot

The scattered plot of revenue by profit shows positive correlated, i.e., when profit increases, revenue also increases in general. A few outliers may have greater revenue and profit than most of the data with more. The lower range contains most of the data.

2.3 Model Building

Two different regression schemes to estimate revenue were implemented. The custom linear regression with direct matrix solution with an intercept term was used in the first approach. The second approach was gradient descent linear regression combined with learning rate optimization and scaling the features for improved convergence. A model performance can be evaluated efficiently by splitting the dataset into a training (80%) and a testing (20%) set.

2.4 Model Evaluation

A comprehensive evaluation was carried out with a set of metrics including Mean Absolute Error (MAE), Mean Squared Error (MSE), R-squared (R^2) score for both models. The choice of these metrics was made to give a balanced view of model performance in different ways of prediction accuracy.

2.5 Hyperparameter Optimization

We got the optimized parameter for both models through GridSearchCV best performed when doing the linear regression, whereas the gradient descent model has the best performance with a learning rate of 0.1, 1,500 iterations. Based on the cost function criteria and model stability, these parameters were chosen.

2.6 Feature Selection

The feature importance analysis suggested the cost (score 3,999,636) followed by profit (977,262), unit price (180,407) and unit cost (179,561). The financial metrics were found to have greater power with regard to predicting revenue, as order quantity was much less important (9,709).

3. RESULTS AND DISCUSSION

3.1 Key Findings

The best accuracy by the custom linear regression model was an R² score of 1.0, MAE of 2.55e10, and MSE of 7.09 e-20. The gradient descent model also showed a good accuracy with an R² score of 0.975, MAE of 149.389, and MSE of 37976.867. It was found that the predictive capability of both models is high, and the perfect score of the linear regression definitely should warrant some further investigation for overfitting.

3.2 Final Model

The performance metrics gave us some insight as to how capable we were in predicting. Finally, a Linear Regression with chosen features-based model was the most effective to predict the revenue of sales. The test set accuracy corresponded MAE: 2.561-12, MSE: 8.903-24 and R2 score: 1.0.

3.3 Challenges

The main challenges were in dealing with perfect accuracy of the linear regression model as well as fine tuning features for the gradient descent implementation.

3.4 Future Work

Future research direction is to analyze the perfect accuracy of the linear regression, implement additional algorithms such as Random Forest and XGBoost, and other techniques.

4. Discussion

4.1 Model Performance

This research had successfully developed the two regression models for the prediction of revenue in the bike retail sector. In particular, the gradient descent implementation was also very real world. Perfect accuracy was achieved by custom linear regression.

4.2 Impact of Hyperparameter Tuning and Feature Selection

Feature selection and hyperparameter optimization significantly improved model stability, and the key revenue drivers for business insights were found from it. On implementing the tuned hyperparameter and feature selection the results were MAE: 2.55e-10, MSE: 7.095e-20 and R2 score: 1.0

4.3 Interpretation of Results

The results indicate that financial metrics like cost and profit are the best predictors of revenue in bike retail operations. The performance metrics gave us some insight as to how capable we were in predicting. Though perfectly impressive mathematically, perfect score of linear regression further brings the issue of overfitting and generalizability. Most likely the performance of the gradient descent model is strong but imperfect, which is a more realistic prediction scenario. The product category was influential regarding revenue generation whereas cost and profit were found to be the strongest revenue predictors.

4.4 Limitations

There are several limitations that should be considered such as, the perfect score achieved by the linear regression model has to be further investigation to ensure generalizability. Since the model is linear and assumes that there are linear relationships between the variables, it is not ready to adjust to other available feature sets.

4.5 Suggestions for Future Research

The use of more advanced feature engineering techniques, collecting other relevant sales factors and further investigation of application of sets methods in future research.