PAPER NAME

2407749_SubashBikramTamang_Regression_Report.pdf

AUTHOR

-

WORD COUNT

842 Words

CHARACTER COUNT

4660 Characters

PAGE COUNT

6 Pages

FILE SIZE

107.9KB

SUBMISSION DATE

Feb 11, 2025 5:14 PM GMT+5:45

REPORT DATE

Feb 11, 2025 5:14 PM GMT+5:45

20% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 1% Internet database
- 1% Publications database
- Crossref database
- Crossref Posted Content database
- 20% Submitted Works database

UNIVERSITY PARTNER

UNIVERSITY OF WOLVERHAMPTON

HERALD COLLEGE KATHMANDU

Concepts and Technologies of AI

AI: Regression Analysis Report

Student ID: 2407749

Student Name: Subash Bikram Tamang

Course Code: 5CS037

Abstract

The goal of the following project is to predict a continuous variable using regression techniques.

Methods: Diamonds Dataset is selected, containing a range of diamond attributes for price predictions. This study will require processes such as hyperparameter tuning, feature selection, model implementation using Linear Regression and Decision Trees, and EDA.

Important Outcomes: The model performance was observed through the performances by Mean Squared Error 'MSE' and R-squared. Their performances varied: at some instances, Decision Trees outperformed Linear Regression in performance.

Two key lessons from the performance of regression models, though good, revolve around how important feature selection plays a role in model performance, as well as the necessity for hyperparameter tuning.

1. Introduction

1.1 Problem Statement

The regression models will use the characteristics of the diamond to predict the price of each in this project. Price prediction is a big determinant in the pricing strategy and value within the diamond industry

1.2 Dataset

The data for the study came from the Diamonds Dataset that was obtained from Kaggle. It consists of 53,940 rows and 11 columns. Each column represents features in carat, cut, color, clarity, etc. This dataset contributes to supporting the UN SDGs by enabling ethical production and consumption in the jewelry sector.

1.3 Aim

This paper aims to design a predictive regression model that would be able to calculate the price of various kinds of diamonds given their characteristics.

2. Methodology

2.1 Data Cleaning and Preprocessing

Data cleaning with respect to handling outliers and missing values was followed by model construction. Other preparations included scaling and encoding categorical variables in preparation for modeling.

2.2 Exploratory Data Analysis

EDA was done by visualizing the data using scatter plots, histograms, and summary statistics to understand the data more intuitively. Following are some key findings derived from EDA:

- Price and carat weight are highly correlated.
- Cut, color, and clarity are examples of categorical variables that greatly affect price.
- High carat diamonds have some outliers that need to be handled carefully.

2.3 Building the Models

For this problem, the following two regression models were considered:

- Regression Linearity
- Trees of Decisions

To create the model, the data had to be divided into both training and testing datasets, an 80-20 split, respectively. Later, the training data was used to train these models.

2.4 Model Evaluation

As measures to find how good and generalized the model performed, it considers:

- R-squared: It calculates what % of the variance in a dependent variable could be explained by the independent variable(s).
- The Mean Squared Error-MSE, is a method of measuring the average of the squared differences between actual and expected values.

2.5 Hyperparameter Optimization

GridSearchCV for hyperparameter tuning to further optimize the performance of the model. The best parameters for the Decision Tree model proved to be:

- Maximum Depth: 6
- Split of Min Samples: 8

2.6 Feature Selection

For feature selection, RFE has been employed in order to check which attributes had been more important to forecast the target variable. The selected features are:

- Carats
- Slice
- Colour
- Clarity

3. Conclusion

3.1 Key Findings

MSE and R-squared were calculated to test the model's performance on the test dataset. It was observed that Decision Trees were better than Linear Regression in terms of handling complex relationships.

3.2 The Final Model

Based on the analysis, the best model to predict the prices of the diamond was a decision tree-based model. The model returned an R-squared value of 0.85.

3.3 Challenges

The following are some of the challenges that arose during the project:

- Outliers affect the model's predictions.
- how to maintain interpretability after encoding categorical varia

3.4 Future Work

Future work that can be done to enhance this model is:

- applying high-level regression, such as Random Forests.
- checking deep learning models for more accuracy.

4. Discussion

4.1 Evaluation of Model Performance

The performance of the model has been evaluated on MSE and R-squared. The results support that Decision Trees performed better while finding non-linear relationships in data compared to Linear Regression.

4.2 Hyperparameter Tuning and Feature Selection

This involved feature selection and tuning of hyperparameters that stood to better the performance of models. After having both strategies in place, model performance went up significantly.

4.3 Result Interpretation

The model performance and performance of the selected features met expectations. The primary inferences to be made from the data indicate that carat weight is the highest determinant of price.

4.4 Limitations

Despite the efficient modelling, the approach has a number of weaknesses, including:

- Generalisability is reduced by the limited dataset size.
- There is a risk of biases in encoding category variables.

4.5 Limitations of Future Research

The following aspects may be considered in further research:

- more models can be applied to the ensemble regression techniques.
- more diamond qualities can be added into the data.
- feature engineering methods can be used to enhance the prediction accuracy.

20% Overall Similarity

Top sources found in the following databases:

- 1% Internet database
- Crossref database
- 20% Submitted Works database
- 1% Publications database
- Crossref Posted Content database

TOP SOURCES

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	University of Wolverhampton on 2025-02-11 Submitted works	9%
2	University of Wolverhampton on 2025-02-11 Submitted works	6%
3	University of Wolverhampton on 2025-02-11 Submitted works	1%
4	fastercapital.com Internet	1%
5	Zhihao Lei. "A Comprehensive Statistical Analysis of COVID-19 Trends:... Crossref posted content	1%
6	Saint Andrews Lutheran College on 2024-03-25 Submitted works	1%