Capstone Project Final Report: Predictive Analysis of Diamond Prices

Introduction

Problem Identification

The diamond industry represents an interesting segment of the luxury goods market, characterized by its complex relationship between supply and demand and the high value of its products. The primary challenge addressed in this project is the prediction of diamond prices based on various features such as carat weight, cut quality, color, and clarity. Accurate price prediction can aid sellers and buyers in making informed decisions.

Data Description

The dataset utilized in this project is sourced from Kaggle and titled "The Largest Diamond Dataset." It comprises records of over 50,000 diamonds, with features including carat weight, cut, color, clarity, dimensions (x, y, z), depth percentage, table percentage, and price. The data were cleaned and prepared for analysis, ensuring the models are built upon a robust foundation.

Methodology

Data Wrangling

The initial step involved cleaning the data by handling missing values, removing duplicates, and correcting any anomalies in the features. For example, diamonds with a zero value in any of the dimension attributes (x, y, z) were excluded from the analysis.

Exploratory Data Analysis (EDA)

EDA figures and visualizations were created to understand the relationships between features. This process included analyzing the distribution of key features like carat weight and price, and exploring how categorical features such as cut, color, and clarity affect diamond prices.

Feature Engineering

Dummy features were created for categorical variables (cut, color, clarity) to facilitate their inclusion in the regression models. Features were magnitude standardized to ensure that variables on different scales did not unduly influence the model.

Data Splitting

The dataset was split into training and test subsets, with 80% of the data allocated for training and 20% for testing. This split was performed to evaluate the models' performance on unseen data accurately.

Model Building and Evaluation

Models Built

Three different models were built and evaluated:

1. **Linear Regression** - Served as the baseline model.

2. **Random Forest Regressor** - A non-parametric model chosen for its ability to handle complex, non-linear relationships between features and target.

3. **Gradient Boosting Regressor** - Selected for its precision and performance in handling diverse datasets.

Model Performance Comparison

A comparison table was created to evaluate each model based on metrics such as Mean Squared Error (MSE) and R-squared (R^2). The Random Forest Regressor was selected as the final.

Final Model Application

The Random Forest model, fine-tuned through GridSearchCV, was applied to the test data. The results demonstrated the model's effectiveness in predicting diamond prices, with a notable R^2 score indicating a high level of variance explained by the model.

Results and Recommendations

Key Findings

Carat weight is the most significant predictor of diamond price, followed by clarity and color.

The Random Forest model, with optimized parameters, provides accurate and reliable price predictions.

Recommendations

1. **Pricing Strategy**: Utilize the model's predictions to refine pricing strategies, ensuring competitive and fair pricing for both buyers and sellers.

2. **Inventory Management**: Apply insights from feature importance to prioritize the acquisition of diamonds with attributes most positively associated with higher prices.

3. **Customer Engagement**: Develop personalized recommendations for customers based on the model's insights, enhancing customer satisfaction and loyalty.

Further Research

Investigating the impact of market trends and economic factors on diamond prices could enhance the model's predictive accuracy.

Exploring if advanced machine learning techniques might offer improvements in modeling complex relationships in the data.

Conclusion

This project demonstrates the potential of machine learning models to predict diamond prices effectively, providing valuable insights for stakeholders in the diamond industry. The methodology and findings offer a foundation for informed decision-making, with the potential for further enhancements through continued research and model refinement.

Model Metrics: RandomForestRegressor

Model Overview

Model Type: RandomForestRegressor

Features

Numerical Features: Carat Weight, Depth Percent, Table Percent, X (Length), Y (Width), Z (Depth)

Categorical Features: Cut, Color, Clarity

Note: Categorical features were one-hot encoded.

Model Parameters

n_estimators:100

max_depth20:

min_samples_split: 2

min_samples_leaf: 2

max_features: 'auto'

Hyperparameters Tuning

Method Used: GridSearchCV

Parameter Grid:

n_estimators: [100, 200]

max_depth: [10, 20]

min_samples_split: [2, 5]

min_samples_leaf: [1, 2]

Cross-Validation Folds: 5

Scoring Metric: Negative Mean Squared Error (MSE)