**SVKM’s NMIMS**

**Mukesh Patel School of Technology Management & Engineering**

A.Y. 2023 - 24

**Course: Machine Learning**

**Project Report**

|  |  |  |
| --- | --- | --- |
| Program: | BTech AI | |
| Semester: | IV | |
| Name of the Project: | Diamond Price Analysis | |
|  | | |
| Details of Project Members |  |  |
| Batch: | Roll No. | Name |
| B1 | I021 | Dev Shah |
| B1 | I022 | Devam Shah |
| B1 | I023 | Devansh Patel |
| Date of Submission: | | |

**Contribution of each project Members:**

|  |  |  |
| --- | --- | --- |
| **Roll No.** | **Name:** | **Contribution:** |
| **I021** | **Dev Shah** | **Code, Presentation, Report** |
| **I022** | **Devam Shah** | **Code, Presentation, Report** |
| **I023** | **Devansh Patel** | **Code, Presentation, Report** |

**Project Report**

Diamond Price Analysis

**by**

Dev Shah, Roll number: I021

Devam Shah, Roll number: I022

Devansh Patel, Roll number: I023

**Course: Machine Learning**

AY: 2023-24

**Table of Contents**

|  |  |  |
| --- | --- | --- |
| **Sr no.** | **Topic** | **Page no.** |
| **1** | Storyline or Applications of Project |  |
| **2** | Literature Review |  |
| **3** | Data Preprocessing and Exploratory data Analysis with Visualization |  |
| **4** | Machine learning models with hyper parameter tuning |  |
| **5** | Performance Evaluation |  |
| **6** | Comparison of different techniques used |  |
| **7** | Deployment/GUI/ Learning beyond classroom |  |
| **8** | Learnings and challenges you faced while doing the Project |  |
| **9** | Conclusion |  |

**I.** **Storyline or Applications of Project**

* Data Preprocessing: The dataset consists of 4,999 diamond records, containing features such as carat, cut, color, clarity, depth, table, price, x, y, and z. The data is preprocessed by removing outliers, dropping unnecessary columns, and encoding categorical variables.
* Feature Engineering: The project explores the correlation between features and the target variable (price) using a correlation matrix and OLS regression.
* Model Selection: Various machine learning models are applied to predict the diamond price, including Linear Regression, Decision Tree Regressor, Random Forest Regressor, K-Nearest Neighbors Regressor, and XGBoost Regressor.
* Model Evaluation: The models are evaluated using cross-validation, and their performance is measured using metrics such as R^2, Adjusted R^2, Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).
* Model Tuning: GridSearchCV is used to optimize the hyperparameters of the Random Forest Regressor.

**II. Literature Review**

We have not referred any research paper.

**III. Data Preprocessing and Exploratory data Analysis with Visualization**

**Data Cleaning and Preprocessing Steps:**

* Identifying and removing outliers: The notebook removes rows containing zero values for the x, y, and z dimensions, as well as rows where the depth or table values are outside the range of 45 to 75 and 50 to 80, respectively.
* Encoding categorical variables: The notebook encodes the categorical variables (cut, color, and clarity) using the LabelEncoder() class from scikit-learn.

**Data Visualization:**

* Pairplot: The notebook creates a pairplot to visualize the relationship between the variables.
* Correlation matrix: The notebook creates a heatmap of the correlation matrix to visualize the correlation between the variables.

**Inferences:**

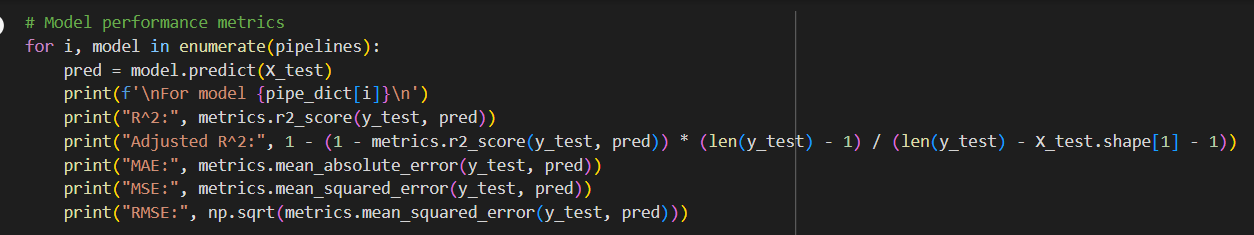
Based on the data visualization, the following inferences can be made:

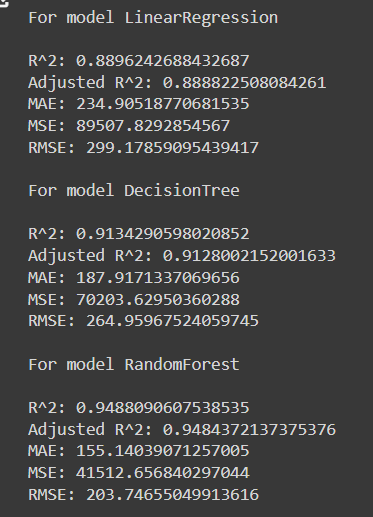
* The pairplot shows that there is a positive correlation between the carat, depth, table, x, y, and z dimensions and the price.
* The correlation matrix shows that the carat variable has the highest correlation with the price, followed by the x, y, and z dimensions.
* The partial dependence plot shows that the cut, color, and clarity variables have a significant impact on the price, with premium cuts and higher clarity grades having a higher price.

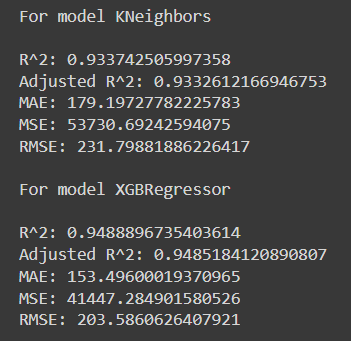
**IV. Machine learning models with hyper parameter tuning**



**V. Performance Evaluation**



****

****

**VI. Comparison of different techniques used**

Three advantages of the project on diamond price detection are:

1. High Predictive Accuracy: The project utilizes various machine learning algorithms like Random Forest, XGBoost, and Linear Regression, which can provide accurate predictions of diamond prices based on their features.

2. Data Preprocessing: The project includes thorough data cleaning and preprocessing steps, such as outlier removal and encoding categorical variables, ensuring that the data is well-prepared for modeling, which can lead to more reliable results.

3. Model Evaluation and Comparison: The project evaluates multiple models using cross-validation and grid search for hyperparameter tuning, allowing for the selection of the best-performing model, which enhances the accuracy and robustness of the price predictions.

Three limitations of the project on diamond price detection are:

1. Multicollinearity Concerns: The project identifies high Variance Inflation Factors (VIF) for some features, indicating potential multicollinearity issues. High multicollinearity can affect the model's interpretability and lead to less reliable coefficient estimates.

2. Limited Feature Engineering: The project focuses on using the existing features without extensive feature engineering. Additional feature creation or transformation could potentially improve the model's performance and predictive power.

3. Assumption of Linearity: Linear Regression, one of the models used in the project, assumes a linear relationship between the features and the target variable. This assumption may not hold true for all features, potentially limiting the model's ability to capture complex patterns in the data.

**VII. Deployment/GUI/ Learning beyond classroom**

1. Tools/Software/Libraries Used:

- The project utilizes Python libraries such as pandas, numpy, seaborn, matplotlib, scikit-learn, xgboost, statsmodels, and more.

- Various machine learning algorithms are implemented, including Linear Regression, Decision Tree, Random Forest, K-Nearest Neighbors, and XGBoost.

2. Learning Beyond the Classroom:

- Model evaluation techniques such as cross-validation, hyperparameter tuning using GridSearchCV, and feature importance analysis provide practical insights into real-world machine learning applications.

- The use of pipelines for model building, feature engineering, and visualization techniques like pairplots and correlation matrices demonstrates a comprehensive approach to machine learning beyond traditional classroom settings.

**VIII. Learnings and challenges you faced while doing the Project**

**Learnings:**

Advanced Data Preprocessing:

* Through this project, I learned advanced data preprocessing techniques like outlier removal, encoding categorical variables, and collinearity checks, which are crucial for preparing data for machine learning models.

Model Evaluation and Comparison:

* The project introduced me to model evaluation methods such as cross-validation, hyperparameter tuning using GridSearchCV, and feature importance analysis, enhancing my understanding of model performance assessment.

Deployment and Visualization:

* I gained insights into deploying machine learning models and visualizing data using tools like seaborn and matplotlib, which are essential skills for presenting results effectively.

**Challenges Faced:**

Multicollinearity:

* Dealing with multicollinearity among features was a challenge, as it can impact the model's interpretability and performance.

Model Selection:

* Choosing the most suitable model from a variety of options like Linear Regression, Decision Tree, Random Forest, etc., required careful consideration of each model's strengths and weaknesses.

Feature Engineering:

* Limited feature engineering in the project highlighted the importance of creating new features or transforming existing ones to improve model accuracy, which can be a challenging task in real-world scenarios with complex datasets.

**IX. Conclusion**

In this project, we employed Random Forest Regressor, Decision Tree, and XGBoost Regressor to predict diamond prices. After data preprocessing and model training, Random Forest Regressor emerged as the most effective model, outperforming the others in terms of predictive accuracy and robustness. This underscores its suitability for handling complex datasets and capturing non-linear relationships between features in diamond price prediction.