# **CIS 568 Data Mining - Fall 2024**

## **Final Project Report**

**Gemstone Price Prediction**

**Student Names:** Anish Kolaparthi, Rahul Sai Sudeer Vadala, Sri Haritha Deevi

**Department Name:** Department of Computer Science and Engineering

**Responsibilities:**

* **Anish Kolaparthi:**
  + Conducted Exploratory Data Analysis (EDA)
  + Implemented Decision Tree and Random Forest models
  + Evaluated overall performance
* **Rahul Sai Sudeer Vadala:**
  + Defined research questions
  + Implemented LightGBM and XGBoost
  + Contributed to technical documentation
* **Sri Haritha Deevi:**
  + Conducted dataset research
  + Implemented KNN and Linear Regression models
  + Performed a literature review

1. **Introduction**

**Topic:** Forecasting gemstone prices using machine learning techniques.

**Background:**

Gemstones are highly sought-after natural resources, esteemed for their aesthetic appeal, scarcity, and cultural importance. Their pricing is intricate and shaped by a confluence of physical characteristics, market dynamics, and subjective assessments. The 4Cs: carat, cut, color, and clarity are the principal variables influencing a gemstone's price. Yet, these elements alone cannot entirely account for pricing discrepancies. In contrast to precious metals like gold or silver, which possess defined valuation methodologies, gemstone pricing is influenced by market volatility, limited transparency, and expert subjectivity, resulting in value discrepancies.

**Objective:**

This study seeks to tackle the absence of a dependable pricing system by utilizing machine learning methodologies to forecast diamond values. Through the analysis of an extensive gemstone dataset, we employ sophisticated algorithms to discern and measure the correlations among attributes such as carat weight, cut quality, clarity grade, and price.

**Summary:**

We employed various machine learning models, such as Linear Regression, Decision Tree, KNN, Random Forest, LightGBM, and XGBoost, to create a precise pricing prediction model. The dataset experienced comprehensive preparation, encompassing the management of missing values, normalization of features, and encoding of categorical variables. After thorough assessment of criteria such as RMSE and R², LightGBM was identified as the superior model, with a R² score of 0.98 and an RMSE of 534.54. The results underscore the capability of machine learning to enhance transparency and precision in gemstone valuation.

1. **Methods used:**

**Technologies and Tools:**

* **Language**: We employed **Python**, a flexible programming language adept at data analysis and machine learning, to construct a strong forecast model for gemstone pricing.
* Essential **libraries** comprise:
  + **Pandas with NumPy**: For effective data manipulation and preparation.
  + **Seaborn and Matplotlib**: For visualizing feature distributions, correlations, and insights.
  + **Scikit-learn**: For the implementation of foundational models such as Linear Regression, K-Nearest Neighbors, and Decision Trees.
  + **LightGBM and XGBoost**: Sophisticated ensemble learning methodologies selected for their proficiency in managing non-linear connections and extensive datasets, yielding elevated accuracy and diminished mistakes.
* **Insights from Research Papers**:
  + Diamond prices are significantly connected with carat weight and clarity, as emphasized by Mihir et al. We concentrated on these essential elements throughout data preprocessing and analysis due to this comprehension.
  + Mankawade et al. demonstrated that ensemble models such as XGBoost can identify non-linear connections in pricing datasets. Incorporating their findings, we used XGBoost and LightGBM to enhance model performance.
  + Ramírez et al. underscored the importance of processing efficiency in modeling large datasets, advocating for Extreme Learning Machines. This guided our hyperparameter optimization approach to reconcile accuracy with computational cost.
  + All investigations underscored issues such as distorted price distributions and the requirement for robust feature engineering. These elements influenced our data preparation and model evaluation.

**Implementation Steps:**

1. **Data Preprocessing:**

To guarantee data quality, absent values were methodically resolved, and numerical attributes were standardized to standardize their scales on the dataset obtained via Kaggle. Categorical variables such as cut, color, and clarity were converted into numerical formats by one-hot encoding, facilitating interoperability with machine learning techniques.

1. **Exploratory Data Analysis (EDA)**

Exploratory Data Analysis was performed to reveal patterns and correlations within the dataset. Histograms and box plots were employed to illustrate feature distributions, revealing patterns and probable outliers. The analysis of relationships between attributes, including carat and price, utilized scatter plots and correlation matrices, indicating that carat weight and clarity grade are the primary determinants of gemstone pricing. This phase yielded essential insights on feature significance and data dynamics.

1. **Model Training**

Six prediction models were constructed and assessed: Linear Regression, Decision Tree, Random Forest, KNN, LightGBM, and XGBoost.

**Hyperparameter Tuning:**

* Random Forest: Set to 100 trees with a maximum depth of 10 for enhanced performance.
* XGBoost: Optimized with a learning rate of 0.1, a maximum depth of 6, and 200 estimators to achieve a balance between accuracy and computing efficiency.
* KNN: Executed using five neighbors and the Euclidean distance measure for localized forecasting.

**Results of Model Evaluation:**

Two primary indicators were employed to assess model performance:

1. **R² (Explained volatility)** quantifies the proportion of volatility in the dependent variable (price) elucidated by the model. Elevated figures indicate enhanced predictive capability.
2. **RMSE** quantifies the average size of prediction errors, providing a clear indication of the model's accuracy. A lower RMSE indicates greater forecasting accuracy.

**Model Comparison:**

Ensemble models such as LightGBM and XGBoost outperformed Linear Regression and Decision Trees. Their capacity to amalgamate weak learners into a robust predictor encapsulates non-linear relationships and diminishes prediction errors. The gradient-boosting structure of LightGBM demonstrated superior efficiency and accuracy, rendering it the optimal selection for our project.

A screenshot of a computer

Description automatically generated

**Fig 1. Comparison of Results of Different Machine Learning Models**

The findings in Fig 1. illustrate the significance of employing sophisticated machine learning techniques on complex datasets to enhance predictive accuracy and provide valuable pricing insights.

1. **Experiments**

**Data:**

The information utilized for this project comprises 10 parameters, including carat, cut, clarity, color, depth, table, and pricing. The dependent variable is price, whereas the other qualities serve as independent predictors. Carat weights span from 0.2 to 5.01, with values fluctuating considerably between $326 to $18,823, illustrating the intricacies of gemstone pricing. Attributes including as cut, clarity, and color are categorical, but carat and dimensions (x, y, z) are continuous, necessitating different preprocessing methods to address their specific traits.

**Experiments conducted:**

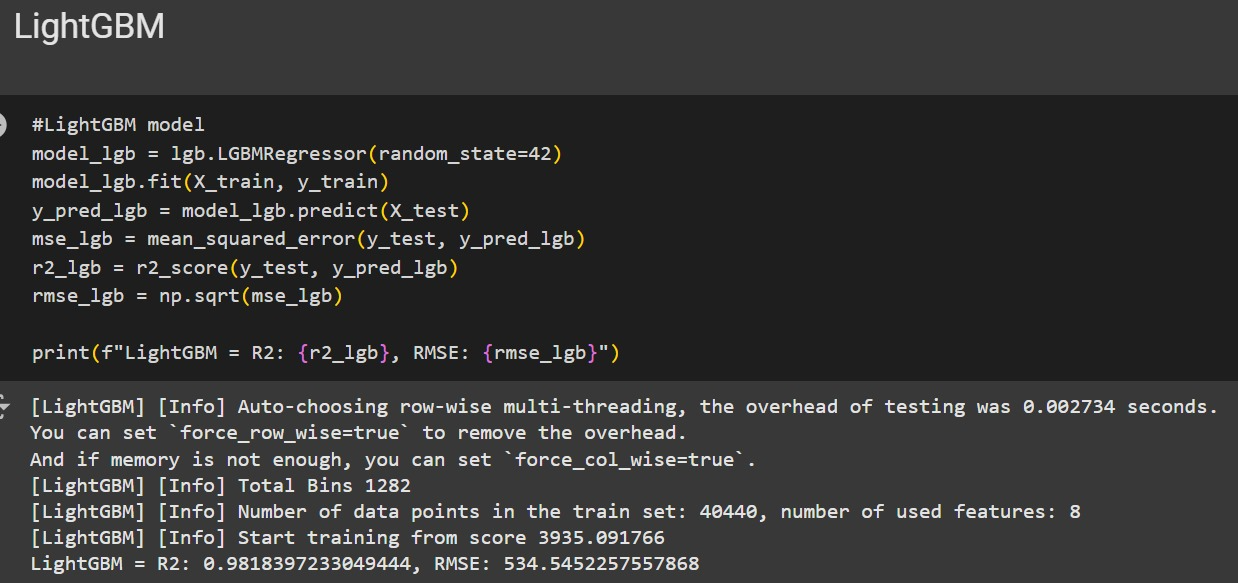
We utilized an 80-20 train-test split to assess the dataset, therefore establishing a rigorous testing framework to verify model generalizability. Various machine learning models, such as Linear Regression, Decision Tree, Random Forest, XGBoost, and LightGBM, were developed and evaluated. Each model was assessed for accuracy and error utilizing R² and RMSE as performance indicators.

**Link to Code:** [**https://github.com/Forecasting-Gemstone-Prices-using-Machine-Learning-Techniques/**](https://github.com/anishkolaparthi/Forecasting-Gemstone-Prices-using-Machine-Learning-Techniques/blob/main/Data_Mining_Project_Group11_Kolaparthi%2C_Deevi%2C_Vadala.ipynb)

**Discussion of Results:**

1. Linear Regression: Exhibited fundamental performance with R² = 0.91 and RMSE = 1178.89.
2. Decision Tree: Enhanced accuracy (R² = 0.96) with decreased RMSE (741.47).
3. Random Forest: Demonstrated competitive outcomes (R² = 0.94, RMSE = 900.45).
4. KNN: Achieved an R² of 0.9674 and an RMSE of 716.09, demonstrating strong predictive performance with minimal error.
5. XGBoost: Achieved superior accuracy (R² = 0.96, RMSE = 720.50).
6. LightGBM: (**Optimal Model**) Had the highest efficacy, with a R² score of 0.98 and RMSE of 534.54.

These results demonstrate its exceptional ability to comprehend intricate data relationships and generate precise predictions.



**Fig 2. Results of LightGBM**

The findings from Fig 2. underscore LightGBM's proficiency in accurately capturing intricate linkages, positioning it as the most dependable model for gemstone price forecasting.

1. **Conclusion**

This research developed a robust machine learning model to predict diamond prices. Principal findings indicated that carat and clarity influence value, however ensemble techniques such as LightGBM and XGBoost surpassed conventional approaches in managing the dataset's intricacy.

**Challenges and Learnings:**

Addressing the skewness in price distribution and reconciling model interpretability with accuracy were significant problems. We acquired knowledge in feature engineering, hyperparameter optimization, and evaluation metrics, showcasing the effectiveness of advanced ensemble methods in handling intricate datasets.

**Future Work:**

The model may be expanded by incorporating real-time pricing data and luxury items in addition to jewelry. External factors such as market trends and gemstone provenance might enhance pricing estimations. This system provides merchants, auction houses, and clients with reliable pricing instruments. Furthermore, refining hyperparameters and commercializing the model can enhance its practicality and industrial impact. This research demonstrates how machine learning may transform valuation procedures to enhance transparency and accuracy.

1. **References**
   * Kaggle dataset: [Diamonds](https://www.kaggle.com/datasets/shivam2503/diamonds).
   * H. Mihir et al., "Diamond Price Prediction using Machine Learning," 2021 2nd International Conference on Communication, Computing and Industry 4.0 (C2I4), 2021.
   * A. Mankawade et al., "Diamond Price Prediction Using Machine Learning Algorithms," International Journal for Research in Applied Science and Engineering Technology, 2023.
   * J. Ramírez et al., "Extreme Learning Machines for Predicting Diamond Prices," 2023 IEEE CHILECON, 2023.