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: Predicting gemstone prices using machine learning.

Background:

Gemstones rank among the most precious natural materials with their prices influenced by the 4Cs: carat, cut, color, and clarity. In contrast to gold or silver, gemstone valuation is notably intricate and often inconsistent due to market fluctuations, lack of transparency, and subjective expert opinions. This complexity underscores the importance of reliable, data-driven predictive models to promote equitable pricing.

Summary:

This project utilizes advanced machine learning techniques to forecast gemstone prices using a comprehensive dataset. Various algorithms such as Decision Tree, KNN, Linear Regression, Random Forest, LightGBM, and XGBoost are applied to build an accurate pricing model. The workflow involves data preprocessing, analysis, model training, and performance evaluation, with metrics like RMSE and R^2 used to measure accuracy and reliability.

2. Methods used

Technologies and Tools:

- **Language**: Python for implementation and analysis.
- **Libraries**:
  - **Data manipulation:** Pandas, NumPy
  - **Data visualization**: Seaborn, Matplotlib
  - **Training Machine learning algorithms:** Scikit-learn
  - **Advanced ensemble learning:** LightGBM, XGBoost

Implementation Steps:

a) Data Preprocessing:

- Imported a dataset from Kaggle comprising 53,940 rows and 10 attributes.
- Addressed missing data and standardized numerical features.
- Applied one-hot encoding to categorical attributes such as cut, color, and clarity.

b) Exploratory Data Analysis (EDA)

- Used histograms and box plots to visualize feature distributions.
- Examined relationships, such as carat versus price, and identified significant correlations.
- Identified carat and clarity as the most influential features.

c) Model Training

- Trained six models: Linear Regression, Decision Tree, Random Forest, KNN, GBM, and XGBoost.
- Hyperparameter tuning:
  - Random Forest: 100 trees, max depth = 10.
  - XGBoost: learning rate = 0.1, max depth = 6, estimators = 200.
  - KNN: neighbors = 5, Euclidean distance metric.

Results:

- **Model Evaluation:**
  - Metrics: R^2 (explained variance) and RMSE (error).
  - Best Model: LightGBM with $R^2 = 0.98$ and RMSE = 534.54.
  - Ensemble models surpassed traditional methods, with LightGBM achieving the highest accuracy.

3. Experiments

Data:

The dataset comprises 10 attributes, such as carat, cut clarity, color, depth, table and price. Price is the target variable, while the other features serve as predictors. Carat weighs range from 0.2 to 5.01, and prices vary from $326 to $18,823.

Experiments conducted:

- Various models were trained and evaluated using an 80-20 train-test split.

- Model performance was measured in terms of accuracy and error across all algorithms.

Results:

a) Linear Regression: R^2 = 0.91, RMSE = 1178.89
b) Decision Tree: R^2 = 0.96, RMSE = 741.47
c) Random Forest: R^2 = 0.94, RMSE = 900.45
d) XGBoost: R^2 = 0.96, RMSE = 720.50
e) LightGBM: R^2 = 0.98, RMSE = 534.54 (**Best Performance**)

4. Conclusion

Summary:

This project successfully built a precise machine learning model to predict gemstone prices. Key findings highlight the significance of carat and clarity in valuation, with ensemble methods outperforming traditional models.

Challenges:

- Mitigating skewness in price distribution.
- Balancing model interpretability with accuracy.

Learnings:

- Gained expertise in feature engineering, hyperparameter tuning, and evaluation metrics.
- Recognized the effectiveness of ensemble methods like LightGBM and XGBoost in managing complex datasets.

Future Work:

a) **Scalability:**
   - Incorporate real-time pricing data.
   - Adapt the model for other luxury goods.
b) **Feature Enhancements:**
   - Integrate external factors such as market trends and origin.
c) **Applications:**
   - Develop pricing tools for traders, auction houses, and customers.
d) **Optimization:**
   - Further refine hyperparameters.
   - Prepare the model for commercial deployment.

5. References

- Kaggle dataset: 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.