

Jewellery Price Optimization with Machine Learning

**(A Data Science Project for Gemineye
Emporium)**

By

Kehinde Ogundana



Business Introduction, Problems & Objectives

➤ **Business Company:** *Gemineye Emporium*

- Industry: Luxury goods and jewellery
- Known for craftsmanship, quality, and innovation.
- Expanding operations across the country, increasing costs and operational complexities.
- Currently relies on manual pricing by gemmologists and appraisal experts, which is expensive and time-consuming.

➤ **Business Problem - Challenges in Pricing:**

1. Overpricing risks losing price-sensitive customers.
2. Under-pricing reduces profit margins.
3. Lack of dynamic adjustments based on market trends, preferences, and competition.
4. Inconsistent pricing strategies across regions and product lines.
5. Absence of data-driven demand prediction.

➤ **Objectives:**

1. **Maximized Revenue :** Data-driven pricing to optimize sales volume and profit margins.
2. **Competitive Edge:** Implement dynamic pricing to respond swiftly to market trends.
3. **Improved Customer Retention:** Develop personalized pricing for diverse customer segments.
4. **Efficient Decision-Making:** Automate pricing decisions, reducing reliance on manual interventions.
5. **Actionable Insights:** Use ML models to analyse customer behaviour and demand patterns.

Design Methodology and Tools

➤ Methodology:

Adopt the **CRISP-DM** (Cross-Industry Standard Process for Data Mining) framework:

1. **Business Understanding** – Define goals and challenges.
2. **Data Understanding** – Explore and evaluate data quality.
3. **Data Preprocessing** – Handle missing data, outliers, and transformations.
4. **Modelling** – Train predictive models for pricing optimization.
5. **Evaluation** – Assess model accuracy and performance.
6. **Deployment** – Implement and monitor pricing recommendations.

➤ Tools:

1. **Pandas** – Data manipulation and cleaning.
2. **NumPy** – Numerical computations.
3. **Matplotlib/Seaborn** – Data visualization.
4. **Scikit-learn** – Machine learning models.
5. **MLflow** – Experiment tracking and model management.
6. **Git, VS Code, Jupyter Notebook** – Collaboration and development tools

Data Overview

➤ **Dataset Size:** 95,910 rows and 13 features.

- **Features:** Includes order details, jewellery categories, user demographics, material attributes, and the target variable (*price*).

➤ **Key Observations**

- **Missing Values:** 9 columns with missing values, of most critical missing data were in *Target_Gender* (48,000 rows) and *Main_Gem* (34,000 rows).
- **Duplicates:** 2,589 duplicate rows.
- **Feature Variety:** Low variety in *SKU_Quantity*, *Main_Color*, and *Main_Metal*.
- **Outliers in Price:** Price ranges from \$0.99 to \$34,448.60, indicating possible outliers or premium items.
- **Data Quality Issues:** Incorrect values in *Category* (e.g., '451.10', '283.49') need correction.

➤ **Next Steps for Data Preparation**

1. Data Exploration
2. Address missing values and duplicates
3. Correct inconsistent or corrupt entries.
4. Handle outliers in the *price* feature.
5. Normalize categorical and numerical features for modelling.

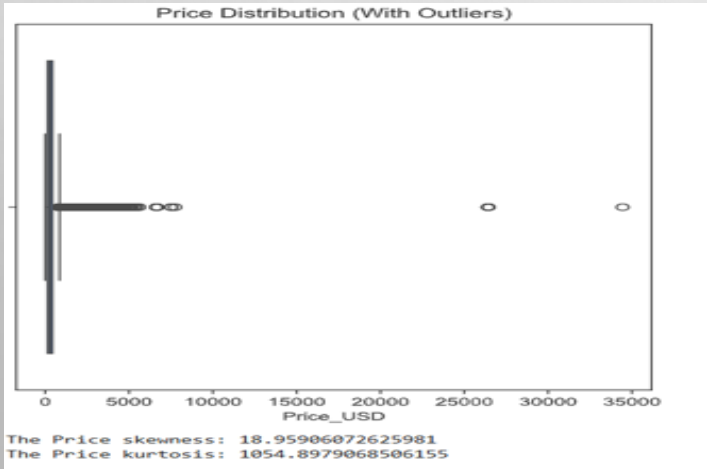
Exploratory Data Analysis (EDA)

1. Price Distribution:

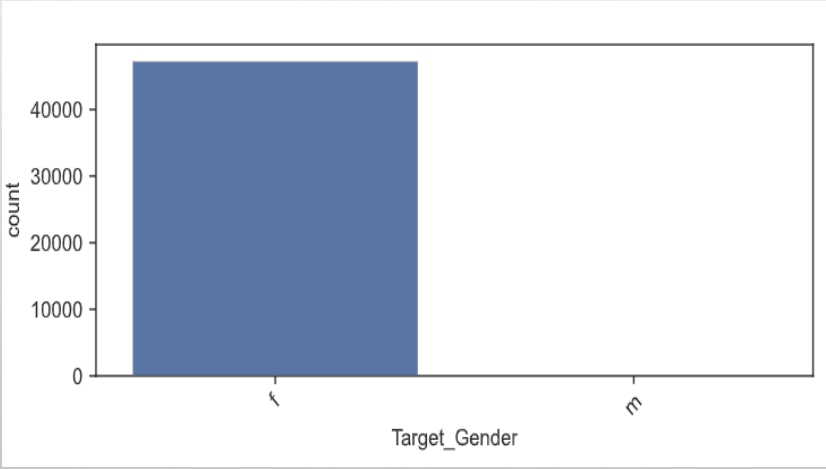
The prices ranges from £0.99 to £34,448.60 with a mean of £362.21 and standard deviation of £444.16. The skewness and kurtosis shows £18.96 and £1054.90 respectively, indicating a highly right-skewed distribution with potential outliers.

2. Categorical Features Distribution:

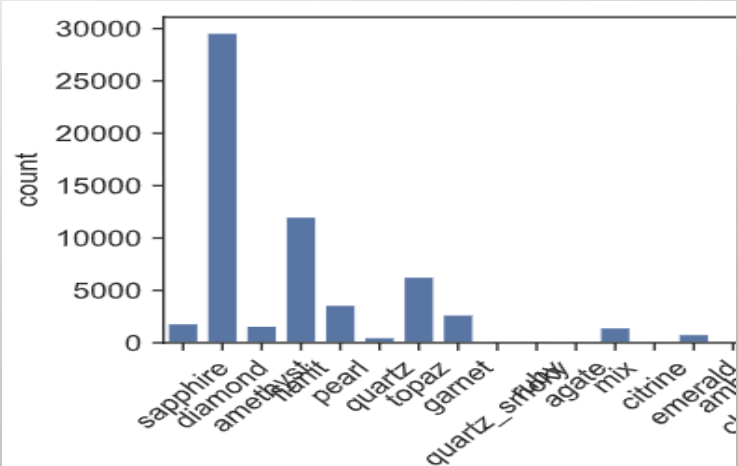
- **Target Gender:** Female Buyers are made of 99.24% and male Buyers at 0.76%.
- **Main Colors:** The jewelleries colour are of 'Yellow', 'White', 'Red', 'Black' but most common is that of Red (69,510 occurrences).
- **Main Metals:** consist of 'Gold', 'Silver', 'Platinum', most common is Gold (89,081 occurrences).
- **Main Gems:** are made of 'Sapphire', 'Diamond', 'Amethyst', 'Fianit', 'Pearl', 'Quartz', 'Topaz', etc., most common is Diamond (29,609 occurrences).
- **Jewellery Categories:** are of eight(8) different types, However, most common category is that of 'Jewelry.Earring' (29,051 occurrences).



Price Distribution



Target Gender



Main Gems

Data Preprocessing

1. Filter Relevant Categories and Conduct EDA

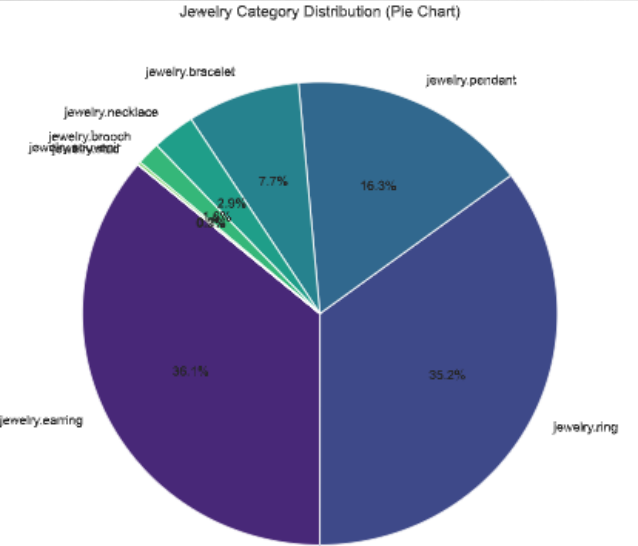
- **Corrupt Data Removed:** 15,000 rows (16% of the dataset) from the 'Category' feature were removed due to errors. Further Insights Post-Cleanup reveals that Earrings, rings, and pendants account for 87% of sales. Female buyers(over 90%), primarily purchases earrings, rings, and pendants while male buyers represent <1% of the sales.

2. Handled Missing Values using ‘SimpleImputer’.

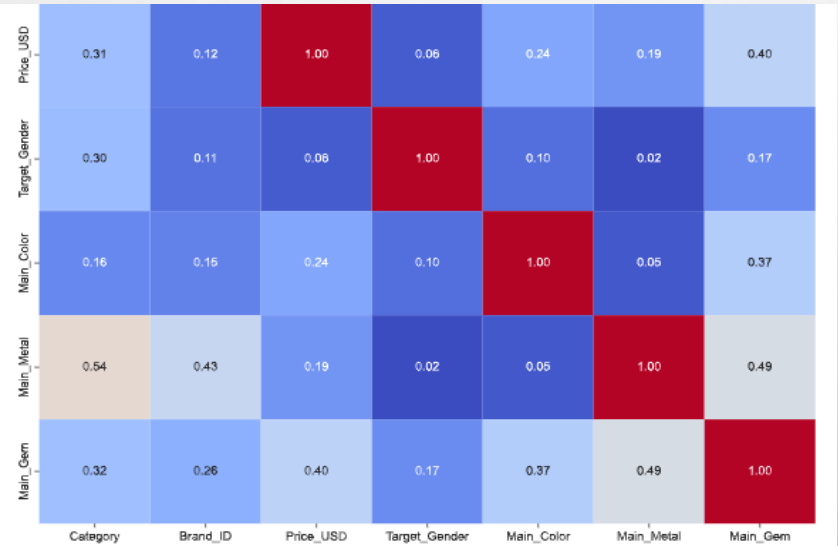
3. Drop Irrelevant Features such as Order_datetime, Order_ID, Product_ID, SKU_Quantity, as they do not contribute meaningful information to price prediction and retained features like Category, Brand, Price_USD, Target_Gender.

4. Outlier Detection and Removal: Detected 804 outliers using the Isolation Forest algorithm.

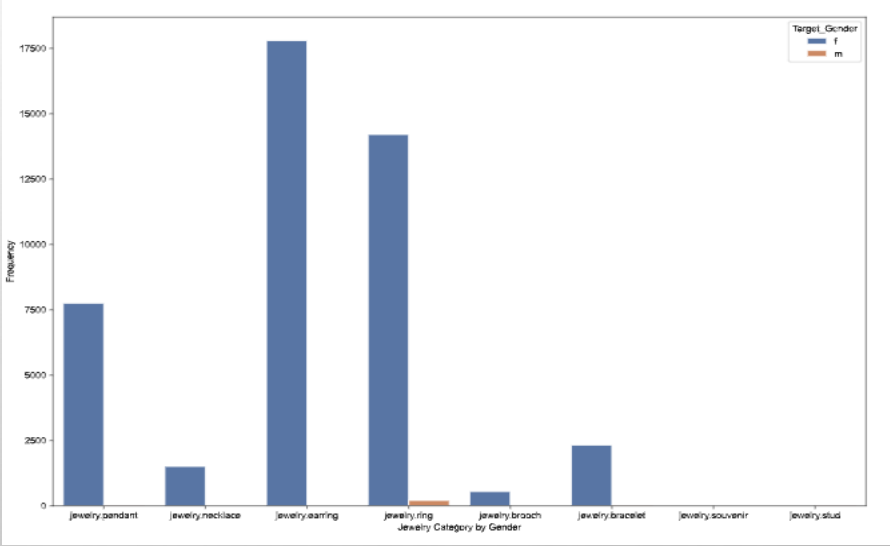
5. Feature Selection(Advanced Correlation Analysis - Phik Heatmap): The results shows strong associations of Main_Metal and Category(0.541), Main_Metal and Main_Gem(0.487) and moderate associations between Price_USD with Category (0.315) and Price_USD with Main_Gem (0.401).



Jewellery Category Distribution



Correlation Analysis



Jewellery Category by Gender

Model Selection, Training & Evaluation

1. Models Chosen(ML):

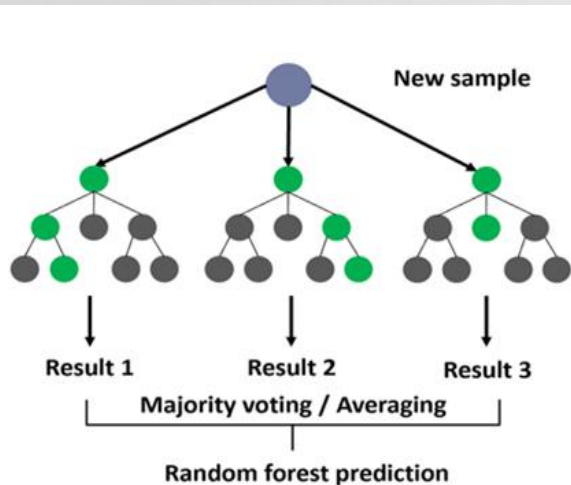
- **Random Forest Regressor, XGB Regressor, Gradient Boosting Regressor and LGBM Regressor:** Selected for their strong performance in regression tasks and ability to handle nonlinear relationships. These models are known for their robustness, scalability, and hyperparameter tuning flexibility.

2. Model Training:

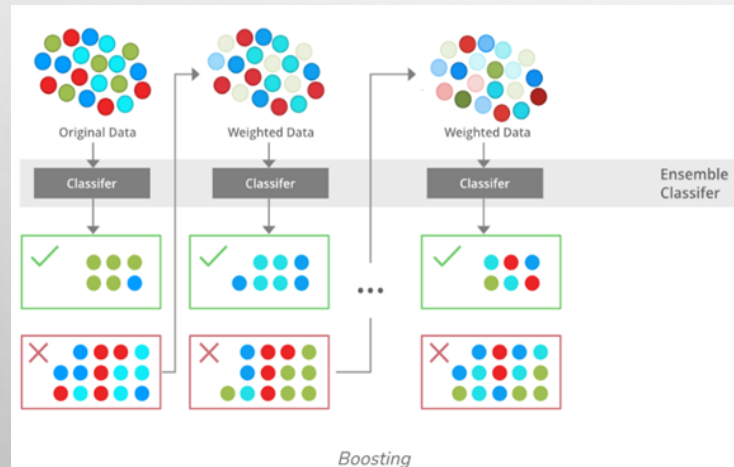
Hyperparameter tuning using GridSearchCV with 5-fold cross-validation was applied to ML models.

3. Best parameters identified::

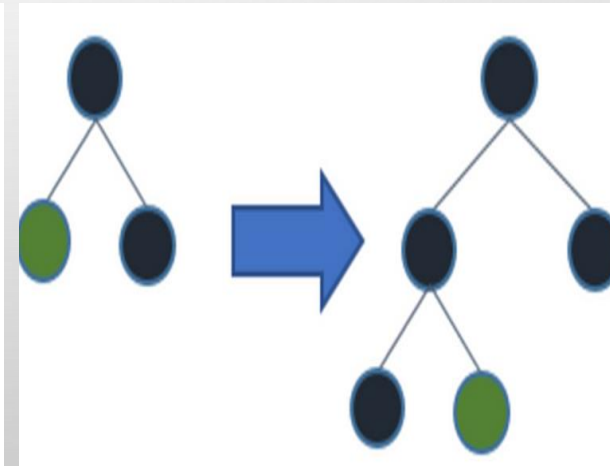
- **RandomForestRegressor:** max_depth=30, min_samples_split=2, n_estimators=100.
- **XGBRegressor:** learning_rate=0.2, max_depth=10, n_estimators=300.
- **GradientBoostingRegressor:** learning_rate=0.2, max_depth=7, n_estimators=300.
- **LGBMRegressor:** learning_rate=0.3, num_leaves=31, n_estimators=300.



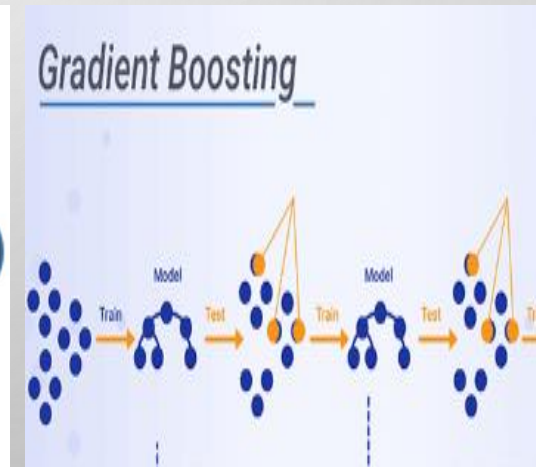
Random Forest Regressor



XGB Regressor



LGBM Regressor



Gradient Boosting Regressor

Evaluation Metrics and Results

❖ Chosen Metrics:

- 1. **R² (Coefficient of Determination):** Measures goodness of fit.
- 2. **MAE (Mean Absolute Error):** Evaluates average prediction error.
- 3. **MSE (Mean Squared Error):** Penalizes larger errors to assess overall variance.

Metrics align with regression task objectives and provide a holistic evaluation.

❖ Model Performance Summary:

- **Best algorithm: Gradient Boosting Regressor with the highest R² (0.294).**
- **Other models had comparable but slightly lower R² scores (~0.29).**
- **Results indicate limited predictive power, suggesting that the current dataset lacks sufficient features and data quality to model jewellery prices effectively.**

Model	MAE (USD)	MSE (USD ²)	R ² Score
Gradient Boosting	152.86	52225.38	0.2944
Random Forest	152.97	52246.09	0.2941
XGBoost	152.93	52283.85	0.2936
LightGBM	153.42	52369.99	0.2925

Model Results

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

R² (Coefficient of Determination)

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

MAE (Mean Absolute Error)

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

MSE (Mean Squared Error)

Conclusion

Recommendations:

1. Invest in Data Quality and Expansion

The dataset faced a 16% reduction due to errors, highlighting significant limitations such as:

- Lack of diverse and detailed features (e.g., customer demographics, promotional campaigns, and purchasing patterns).
- Insufficient data samples for certain categories (e.g., male buyers and less common gems/metals).

➤ Action Plan:

- **Expand Customer Data:** Collect information on age, location, and income level.
- **Enrich Product Features:** Include design specifications, customization options, and seasonal trends.
- **Enhance Sales History:** Track discounts, bundling strategies, and cross-sale patterns..

2. Improve Feature Engineering: Introduce additional variables to enhance model performance;

- **Jewellery-specific attributes:** Weight, purity levels, and certifications.
- **Marketing data:** Ad campaigns, promotions, and customer engagement metrics.
- **External factors:** Seasonality and regional trends impacting demand.

3. Leverage Insights for Business Strategy: Utilize data-driven insights to refine business operations;

- **Target Female Customers:** Focus on earrings, rings, and pendants, which dominate sales.
- **Optimize Inventory and Promotions:** Prioritize popular items like gold jewellery with diamonds.
- **Engage Male Buyers:** Develop targeted campaigns to drive purchases in niche categories like rings.

In conclusion, investing in data quality, expanding feature diversity, and leveraging insights for strategic decision-making will empower the company to optimize pricing strategies, improve operational efficiency, and enhance profitability.

THANK YOU

