Final project idea proposal

Predicting the price of brilliants

Group: IT-2202

Students:

1. Alua Orynbassarova
2. Zhanmadi Toleugazyuly
3. Sultan Kudaybergenov

Teacher: Sultanmurat Yeleu

Idea Description:

The project focuses on developing a machine learning model to predict the cost of brilliant-cut gemstones. This predictive tool will benefit jewelers, gemstone enthusiasts, and industry professionals by providing accurate and efficient estimations based on various gemstone characteristics.

Background Information on the Problem:

Estimating the cost of gemstones, especially those with brilliant cuts, is a complex task influenced by multiple factors such as carat weight, cut quality, color, clarity, and other attributes. Traditional methods may lack precision, and a machine learning approach can enhance the accuracy of cost predictions.

Available Solutions with Links:

While there are existing methods for gemstone cost estimation, integrating machine learning provides a more data-driven and adaptive approach. Relevant research and articles on gemstone pricing include:

- "Gemstone by value": ["Gemstones By Value - A Full Comparison Of Gem Prices And Value | Chroma Gems & Co"]
- "Gemresearch Swiss Lab": ["Dramatic rise in prices for fine gemstones as high-end ruby, emerald and sapphire sales soar - GRS (gemresearch.ch)"]

How to Get the Data:

The dataset for this project can be obtained from gemstone databases, industry-specific sources, or reputable gemstone vendors. Dataset includes features such as carat weight, cut quality, color, clarity, and other relevant attributes. In our work we will use *seaborn* library, which has different datasets and our will be "diamonds".

https://github.com/mwaskom/seaborn-data

Brief Description of Your Solution:

Our solution involves building a machine learning model that leverages features such as carat weight, cut quality, color,depth, and clarity to predict the cost of brilliant-cut gemstones. The model will be trained on a curated dataset, preprocessed to handle outliers and missing

values, also will be evaluated using regression metrics to ensure accurate predictions. We will use LinearRegression, DecisionTreeRegressor and Random Forest Regressor, in order to compare which one will be better and at the end choose the model, that will have higher accuracy. In the document also will be represented graphs, histograms to show correlation between parameters.

Website:

We are planning to create a website, to predict the price of a diamond based on its characteristics (e.g., carat weight, cut, clarity, color).

Tech Stack That Will Be Used:

- Programming Language: Python
- Machine Learning Libraries: scikit-learn, matplotlib,seaborn,pandas
- Data Manipulation and Analysis: Pandas, NumPy
- Data Visualization: Matplotlib, Seaborn
- Development Environment: Jupyter Notebooks
- Model will be trained in Google Colab
- Libraries for website: Flask, Bootstrap, Joblib, Pandas, Scikit-learn.

Conclusion:

The success of this project relies on the quality of the dataset, careful feature selection, and thorough model evaluation. Regularization techniques, such as L1 or L2 regularization, will be explored to prevent overfitting and ensure the generalization of the model.

We anticipate that the predictive model developed through this project will significantly contribute to the gemstone industry, providing a reliable and efficient tool for estimating the cost of brilliant-cut gemstones.