In The Loupe

Phase 1 Executive Summary

PRODUCT DESCRIPTION & OBJECTIVES

Consumer confidence in the fairness of luxury pricing and sourcing is at an all time high, but price-shopping for gemstones remains a rare endeavor among new 18-25s. ‘In The Loupe’ is a project which intends to bring attention to innovations on FBay’s gemstone marketplace, by offering price predictions for listings and customer searches. This phase of the project will produce a Jupyter Notebook demonstrating the data pipeline, model training, and interactive model predictions.

CUSTOMERS

This data product’s eventual customers will be netizens visiting FBay and searching for diamonds. When these shoppers start a search or visit a listing on FBay, the features of the gem they search for or view will be used to predict a fair market price with the model developed in this phase, which they will compare to listed prices in order to make a purchase at a price they will feel confident in.

EXISTING SYSTEMS

There are currently no site-backed price prediction services available for online luxury goods shoppers.

DATA

FBay has sold nearly 54,000 diamonds through the gemstone marketplace, and those listings contain metadata originally supporting our search tools. These metadata include cut, color, clarity, carat, depth, table (the size of the largest top facet), length in x, y, and z, and sale price. This data was collected into a CSV file and hosted at kaggle.com. Evaluator: This data did not originate from FBay, the fictional company. This data originated from the linked kaggle.com dataset, where it was already formatted in a clean CSV file.

PROJECT METHODOLOGY AND PLAN

Evaluator: This section describes both the methodologies and plan I will follow to develop the data product.

This project's planned implementation will follow the DIKW (Data, Information, Knowledge, Wisdom) process:

1. Data
    a. Diamond sale data will be loaded into a Python Jupyter Notebook.
2. Information
    a. Data will be cleaned based on having zero or empty values.
    b. Outliers will be removed based on having a z-score (num. Standard deviations away from the mean) of >= 10.
    c. Ordinal data (categorical data with an ordering) will be approximated by numerical data maintaining the same order (e.g. Monday->1, Tuesday->2, etc).
    d. Data will be explored to identify artifacts, trends, biases, and distributions.
        i. If any are found which remove the data's integrity, return to step 1.
3. Knowledge
    a. Correlations between features will be investigated.
    b. Additional inspection will be performed to ensure data has been cleaned and organized correctly.
        i. If any issues arise, return to step 2.
    c. A Decision Tree will be fit to the data, using the SKLearn Python library.
4. Wisdom
    a. The model will be used to make predictions about diamond price.
    b. The model will be evaluated for accuracy, throughput, and latency.
    c. These measures will be used to inform whether 'In The Loupe' phase 2 should proceed.
    d. This model's predictions could be used by customers to make decisions about whether to buy a diamond at a particular price.

PROJECT OUTCOMES

This phase of the project focuses on determining if a model can be produced, and the direct customers are the stakeholders and decision makers who will approve or reject the continuation to Phase 2. As such, this project's outcomes are centered around software, reports, and documentation.

Software:

- A Jupyter Python Notebook which demonstrates the data pipeline, model creation, model evaluation, and enables model use.

Reports:

- A report evaluating the model's performance, accepting or rejecting the model's ability to meet requirements.
- A report (embedded in the Jupyter Python Notebook) which evaluates the data's trends, distribution, and structure.
- A report documenting the suggested action to be taken relating to Phase 2.

Documentation

- Documentation (embedded in the Jupyter Python Notebook) describing the data pipeline, model construction, and model evaluation.
- Documentation of how the data product (Jupyter Python Notebook) can be accessed for internal validation and testing.

VALIDATION METHODS

In step 4 of the project plan, the model will be evaluated for its value to customers (accuracy), and its value to FBay (throughput and latency).

The model will be fit to 80% of the dataset, and evaluated for accuracy on the remaining 20%, with a passing accuracy score of 95%.

The model will be timed to measure its throughput and latency, with a passing throughput of 50,000 queries/second and a passing latency of 50ms/query. These will be evaluated by timing the model as it evaluates batched queries of different batch sizes, then determining what batch sizes support both the throughput and latency requirements.

PROJECT DEVELOPMENT TIMELINE

This phase of the project will be developed in a Jupyter Notebook running Python, hosted for free by Google Colab. Based on the timeline below, the anticipated cost to FBay is $2,250 in salary + overhead for a single developer, plus $0 in compute costs.

Project	Start Date	End Date	Hours	Dependencies	Resources	Milestones
Data Preparation	Aug. 18	Aug. 18	2h		Google Colab	
Data Analysis	Aug. 18	Aug. 18	2h	Data Preparation	Google Colab	Training features finalized
Model Fit	Aug. 18	Aug. 18	0h	Data Analysis	Google Colab	Trained model
Model Evaluation	Aug. 19	Aug. 19	1h	Model Fit	Google Colab	Model scores for accuracy, throughput, and latency
Report Development	Aug. 19	Aug. 19	2h	Model Evaluation	Google Colab	Report on model performance and recommendations
Documentation	Aug. 19	Aug. 19	2h	Model Evaluation	Google Colab	Documentation of model structure and use