Lab 2 Use Generative AI To Support Advanced Data Analytics

1. Introduction

Data analytics and Large Language Models (LLMs) have distinct yet complementary capabilities, and it's important not to view them as alternatives to one another. Instead, these technologies can work together, combining the reliable predictive power of machine learning-based advanced analytics with the natural language understanding of LLMs. Leveraging these complementary strengths, we see significant potential for generative AI to address challenges during the development and deployment phases of advanced analytics, both in predictive and prescriptive applications.

In practice, LLMs can be particularly useful for integrating unstructured data sources into analyses, translating business problems into analytical models, and interpreting and explaining model results. In this lab, you will learn how to use ChatGPT to enhance advanced data analytics, especially in incorporating unstructured data into usable features for analysis. You'll also practice writing precise prompts that guide an LLM to review a dataset for key themes and return its output as data formatted with standard labels, making it suitable for predictive models.

To assist you in crafting effective prompts, you will be provided with prompt text segments that you can adapt to create your own.

2. Tasks

Here is a dataset of a large number of entries, in this case wines (over 400 000), that contained both pricing information and professional reviews. The data was scraped from a wine website, based in the USA, which describes itself as "the world's largest wine store".

1. Exploratory Data Analysis (EDA)

Please conduct an Exploratory Data Analysis (EDA) to explore the dataset and gain an overall understanding of its structure. Additionally, look for any patterns or anomalies that stand out.

a) To test the hypothesis that people are willing to pay for wines based on subjective value rather than any clear, definable metric, calculate and display the correlation between the numerical fields. For instance, examine whether the average wine rating strongly

influences its price (please present this as Figure 1: Correlation between numerical features in the dataset).

b) Check for any imbalance issues in the data, such as those related to reviews, host sites, or grape varieties. Please include the following visualizations:

Figure 2: Top ten countries with the most wines in the dataset

Figure 3: Top ten grape varieties with the most entries in the dataset

Figure 4: Distribution of unique values among user review scores

Figure 5: Average user rating from the top ten countries

2. Data Cleaning and Feature Engineering

Please perform data cleaning to facilitate the following feature engineering tasks. Begin by checking for any entries that are missing a winemaker note or have null values for the wine's provenance. Additionally, since you will be using natural language processing (NLP) to analyze the sentiment of reviews and understand the winemaker's subjective experience, discard any records that do not contain a review.

To prepare numerical values for regression analysis, remove any non-numeric characters (in this case, "the") from the "Price-Out-of-Stock" and "Price" columns. Convert the values to float types, then combine them into a single feature named "Price_Feature." This ensures that the prices of wines, whether in stock or out of stock, are treated consistently.

You may find that the 'Origin' entries are formatted inconsistently—some only list the country, while others provide more specific details like the state or region. Split any entries containing multiple locations in the 'Origin' column into separate columns, labeling each appropriately as 'Country' or 'Region'. For wines originating from the United States that only list the state, assign "United States" as the country of origin.

Since the year of production may influence both the price and review of wine, it is crucial to extract this information. While the dataset doesn't have a dedicated column for the year, the year is typically included at the end of the wine's name. Extract the year from the wine's name and create a new feature, "Year," to store this information.

By the end of this process, your dataset should contain a final set of features: 2 text features, 7 categorical features, and 6 numerical features (see Table here).

Data Types	Features
Text	Winemaker_notes, Review
Categorical	Name, Variety, Country, Region, Zone, Attr_1, Attr_2
Numerical	Alcohol_percentage, Alcohol_vol, Avg_rating, N_ratings, Price_feature, Year

For the cleaned dataset, please create 6 histograms to display the distribution of each numerical feature. Additionally, please re-draw Figures 2, 3, and 5, highlighting any changes compared to the raw dataset.

3. Lab Report

In the lab report, please describe how you used ChatGPT to complete the two tasks. Be sure to list the prompts, Python code, and results. Include all relevant figures to illustrate how you completed the tasks. Additionally, upload your cleaned dataset and provide a brief evaluation report to confirm that your dataset contains the full set of features: 2 text features, 7 categorical features, and 6 numerical features.