Data Science Exercise

Introduction

Mr. Ali owns a perfume manufacturing business which supplies major distributors and retailers in the region. The perfumes are manufactured at two different locations (plants), and both have different processes and recipes.

In the last one year, samples of perfumes from both factories were sent to customers for feedback, and general impressions ("positive" / "like" or "negative" / "dislike") were recorded. This data has been collected and recorded in a tabular CSV format (*assessment_dataset.csv*).

Mr. Ali recently gained an interest in data science, and he wants to know how he can use this data to obtain insights about his products and their appeal to customers. Since he is a novice in this area, he would like you – a qualified and experienced data scientist – to help him out.

Mr. Ali heard a lot about different data science techniques and approaches such as data exploratory analysis, clustering, pattern recognition, anomaly detection, predictive modeling, etc. And he is open to any technique / approach you may suggest. His ultimate goal is to generate value that helps him improve his business.

Data Set Description

In the "*assessment_dataset.csv* " you will find information about perfume samples and their ingredients. There are 808 different perfume samples. Each sample is manufactured using a set of ingredients. Each row in the file represents an ingredient / sample detail. Here is the explanation of the different fields.

1. ***sample_id***: Unique sample ID of the perfume. 808 samples reviewed by customers are present in this dataset.
2. ***ingredient_id***: Each sample has been manufactured using a set of ingredients. This field contains the unique ID of the ingredient used.
3. ***perfume_base***: Each perfume sample has a "base". The base could be (i) alcohol, (ii) oil, or (iii) water base.
4. ***fragrance***: Each perfume sample has a "fragrance type". The fragrance type could be (i) floral, (ii) fresh, (iii) oriental, or (iv) woody.
5. ***manufacture_plant***: The plant where the sample was manufactured from (A or B).
6. ***rating***: Binary field which indicates if the sample was reviewed positively (1) or negatively (0).
7. ***manufacture_date***: Date of manufacture and review of the sample.
8. ***ingredient_type***: Each ingredient has a "type". The ingredient type could be (i) aroma, (ii) solvent, or (iii) fixative.
9. ***ingredient***: Ingredient codes, which are used in place of the actual ingredient names to protect the secret formulas of Mr. Ali's perfumes.

Guidelines

- Explore and analyze the dataset and propose machine learning models based on your findings.
- Demonstrate your skills in data exploration and visualization, data wrangling, pre-processing, feature extraction, model selection and evaluation.
- Give a brief explanation of main steps you perform during this task. Cover your code with comments explaining what each part is responsible for.
- Create a brief presentation (5-6 slides) summarizing the results of your work and your insights: you can show some parts of your exploratory analysis (charts, etc.), your approach, results, recommendations, etc. You can use Google Slides, MS PowerPoint, Keynote, etc.

- Your analysis will be judged based purely on the approach, completeness, validity of insights / conclusions of the analysis, and not based on the precision / accuracy scores of your machine learning model(s).

Tips and Clarifications

- We are not looking for a model that performs perfectly: we are looking to see how you approach a business problem, how you select and evaluate a model, and how you summarize and communicate your findings.
- You can use any programming language you like. Pick a language suited to the task, and one you are comfortable with.
- You can use any free library.
- It would be great if you will use *Jupyter Notebook*, *R Markdown* or similar tool.
- If you do not understand something or have questions, please contact us!

Submitting Your Solution

Please email your presentation and your code, which can be either a plain code, *Jupyter Notebook* or *R Markdown*. If you do any pre-processing to the data, please also include the script you use to do this (or a list of the commands run).