Data Science Exercise

For the position: "Data Scientist – Industry Projects"

You've been contacted by Caroline Ducrex, the owner of a small flour trading business, who would like to automate some of the tasks she carries out on a regular basis. Each week, Caroline buys 1000 packages of flour, evaluates its quality and sells it back on the market. Caroline already has the tools needed to automatically measure some characteristics of flour, while other information is shared directly by suppliers. In particular, for each package, it is possible to trace this information:

- Gluten Content (%): % of gluten in the flour
- Dough Elasticity Index: measure of the ability of the dough to regain its original shape, in a scale from 0% to 100%
- Dampening Time: time in hours during which the wheat is left in water before the milling
- Package Weight (g): weight of the flour package, including the bag
- Ash Content: mineral content of a flour, in %
- Production Recipe: recipe process used in the milling process
- Moisture (%): % of water in the flour
- Starch content (%): % of starch in the flour
- Production Mill: name of the factory where the flour was produced
- Package Volume: volume of the bag
- Proteins Content (%): % of proteins in the flour
- Color: flour color tone, provided on a discrete scale from 1 (lighter) to 10 (darker)

Caroline divides the flour into three quality levels: High, Average and Low. The quality of the flour determines the selling price:

- High quality flour is sold at 5 CHF/package
- Average quality flour is sold at 2.5 CHF/package
- Low quality flour is sold at 1.2 CHF/package

Here is additional information Caroline shared:

- Caroline must refund all customers who receive flour of a different quality with respect to the purchased one
- Caroline claims she never shipped a wrong order. We are skeptical, but don't want to upset her
- All flour purchased by Caroline is sold every week, and you can assume that the sales have the same proportions in terms of quality as in the dataset (roughly 39% Low, 59% Average, 2% High)
- Caroline spends about 10 hours/week sorting flour packages, but she would really like to spend that time to grow her business. She is considering delegating the task to her assistant, whom she pays 50 CHF/h. She estimates her assistant to be as good as her in classifying the quality of the flour
- All flour, regardless of quality, is purchased at 1 CHF/package
- We can assume there is no shipping cost

Caroline is open to any suggestion and insight you might have and would be curious to know how much money she can save. In a few days' time, she expects you to present her your analysis, conclusions and recommendations.

However, she understands that this is not a lot of time and she is fine with a proof-of-concept solution, to understand how you would tackle the problem.

Disclaimer: the dataset for this exercise has been created from synthetic data, and there is a very limited connection between the data and the typical characteristics of commercial flour. Therefore, we don't encourage candidates to spend time in understanding what makes a good flour in real commercial applications, because this will have no impact on the resolution of the exercise.