Klym Technical Test

Data Science

17 de agosto de 2023

1. Objective

Hello! Welcome to our Data Science technical challenge. We are very glad to know you are interested in joining us.

Our goal with this challenge is to get a glimpse of your creativity and problem solving skills when faced with a specific business oriented problem from a data-science perspective. We are not only interested in getting to know your technical capabilities but also how you can explain your solution and results to non-technical people or business stakeholders.

2. Process Conditions

The process is very simple, you will be given a problem that you will have to solve and present. This exercise will be given to you 2 days before a 1 hour long interview with one of our Data Scientists in which you will explain what you have done. The objective is to dive deeper into your thinking process and engage in a technical discussion.

You will have a lot of freedom during the test, however we will ask you to focus on specific questions while you analyze your data sets and build your model. Not all of these need to be coded, you will see the **(No Code)** sign next to them. These questions will be discussed during the interview.

The only condition is that you do it in Python as that is the programming language we use and hand the challenge over in a self contained project that we can access. You can upload it to your personal GitHub, send a Dockerfile, a zip file with a setup.py, anything we can access and replicate. As far as libraries, you can use whichever you like as long as we can run the code.

3. Context and problem

Imagine that you work for a big car dealership in the USA. The Innovation and Sales departments are trying to modernize the way the dealership works by introducing less friction in your sales process. To do this you propose to build a model to automatically price cars so that potential customers can opt for a self service buying process either online or in the physical store.

Your job is then to build such a model given the data sets that are available to you (*vehicles.csv*, *crashes poverty.csv* and *counties.csv*).

Once your model is done you will need to present your results to the Innovations and Sales teams, trying to convince them on why your model works in the real world. In order to do this you can use the following assumptions to your convenience:

- Every time your model predicts a price below the market value of a car the sale is always finalized but the dealership loses the difference between the predicted price and the real price. That is:

$$Losses = min(Price_{Predicted} - Price_{Real}, 0)$$

- On the contrary, when your price is above the market value of the car you make a profit but the probability that you sell the car decreases proportionally to the overcharge. That is:

$$P(Buying|Price_{Real}) = 1 - min(1, \frac{Price_{Predicted}}{Price_{Real}} - 1)$$

  Your profits are therefore:

$$Profits = max(Price_{Predicted} - Price_{Real}, 0)$$

4. Data Analysis

During your analysis we would like you to answer the following questions:

- Take each of the following 3 variables:
  - *odometer*
  - *condition*
  - *vin*

  For each of them analyze the following points:

  1. Evaluate the completeness of the variable and highlight any findings you think relevant for a discussion.
  2. Evaluate how well the value discriminates your dependent variable and determine if you would use it in your model.
  3. If the previous answer was affirmative then how would you propose to include it in your model. That is, what transformations, techniques, or tools would you use to include it.

- Are there any duplicate, redundant or irrelevant variables in your data set? If there are, how would you deal with them?

5. Model building

When building your model we would like you to focus on the following points to discuss:

- What model would you propose to solve the problem and why did you choose it?
- What performance metrics would you use and discuss whether they are valid and their results
- How would you present your model to the business stakeholder and ensure them that it works?
- Discuss how would you monitor your model in 3 months time to determine its validity. (No Code)

Good luck and enjoy!