Data Scientist Intern - Take Home Challenge

Instruction

Hello Candidate!

Here are some fun and intellectually interesting questions related to a singular dataset. Since we are the company behind Pokemon Go, we've supplied you with a dataset that focuses on Pokemon. There are four questions included below, and you are expected to spend no more than two to three hours on this challenge. Please submit your write-up in PDF to the Greenhouse link in the email.

A few things to note:

> The dataset is a single csv file attached to this take-home challenge. If you have difficulty opening this csv file or have any questions during the take-home challenge, please reach out to the recruiter.
>
> Tools for analysis: unless specified in the questions, feel free to use whatever analysis software you think makes the most sense.
> Prior knowledge: Having prior knowledge on this topic can be helpful but is not necessary. If you have applicable contextual information, feel free to mention it in your write-up. However, note that we are more interested in the thoroughness of your data analysis skills here.
> Outside help: internet research for generic questions is fine but you are expected to complete this challenge without consulting other individuals for help or collaboration.
> Submission and presentation: The only submission you will need to send to your recruiter is your completed answers in PDF to the four questions below. You can submit any supporting documents in any format you prefer.
>
> If you excel in this take-home challenge, you will have the chance to present your findings to your interviewer in your next interview. Treat that interview as a two-way conversation: you will start with a few minutes of presentation, and then your interviewer may ask you a few follow-up questions and you will respond based on your understanding of the data. You will be evaluated on the quality of your data analysis as well as the clarity and depth of your discussion. Please note, the discussion on the take-home challenge will only be a portion of the next interview - please be prepared to answer a few additional technical questions.

Dataset

The Pokémon Dataset is focused on the stats and features of the Pokémon in the Pokémon RPG games until Generation 6. You'll be using this for Question 2 to Question 4.

This database includes 21 variables for each of the 721 Pokémon of the first six generations, plus the Pokémon ID and its name. These variables are briefly described here:

Number. Pokémon ID in the Pokédex.
Name. Name of the Pokémon.
Type_1. Primary type.
Type_2. Second type, in case the Pokémon has it.
Total. Sum of all the base stats (Health Points, Attack, Defense, Special Attack, Special Defense, and Speed).
HP. Base Health Points.
Attack. Base Attack.
Defense. Base Defense.
Sp_Atk. Base Special Attack.
Sp_Def. Base Special Defense.
Speed. Base Speed.
Generation. Number of the generation when the Pokémon was introduced.
isLegendary. Boolean that indicates whether the Pokémon is Legendary or not.
Color. Color of the Pokémon according to the Pokédex.
hasGender. Boolean that indicates if the Pokémon can be classified as female or male.
Pr_male. In case the Pokémon has Gender, the probability of its being male. The probability of being female is, of course, 1 minus this value.
Egg_Group_1. Egg Group of the Pokémon.
Egg_Group_2. Second Egg Group of the Pokémon, in case it has two.
hasMegaEvolution. Boolean that indicates whether the Pokémon is able to Mega-evolve or not.
Height_m. Height of the Pokémon, in meters.
Weight_kg. Weight of the Pokémon, in kilograms.
Catch_Rate. Catch Rate.
Body_Style. Body Style of the Pokémon according to the Pokédex.

Question 1

In Pokemon Go, players can both catch pokemon and battle in raids. Given the sample data below, what is the number of raid battles per player divided by the number of raid battles per battler?

row_number	event	user_id
1	battle_raid	1
2	battle_raid	1
3	battle_raid	2
4	catch_pokemon	3
5	catch_pokemon	4
6	battle_raid	4
7	catch_pokemon	2
8	catch_pokemon	3

A. 0.5
B. 0.75
C. 1.0
D. 1.33
E. 2.0

Question 2

Suppose the Pokémon Dataset is a SQL table called ‘PokemonStats’. In a SQL dialect you are most comfortable with, find…

1) The number of distinct primary types present across Pokemon,
2) The average Total stats for each Pokemon generation,
3) The white Pokemon with the highest Total stats

For each question, your answer should include both a SQL query and the returned result.

Question 3

Imagine a new Pokemon game where you are only allowed to collect ONE type of Pokemon. Similar to other Pokemon games, your goal is to have the strongest battlers and defenders for battles and raids. Which type will you pick? Why?

Question 4

If you want to predict whether the Pokemon is a legendary Pokemon (a.k.a. predict the field isLegendary using other fields), which models would you use? List your top 3 models.

Pick one model and implement it in a language you are most comfortable with (preferably Python or R). Please do not use the 'Catch_Rate' field (if you're a Pokemon fan, you know why!). What is your in-sample classification accuracy and what fields did you end up using?

Your answer should include:

1) The code implementing the model (including feature processing, model fitting and cross validating);

2) The formula/description of your final model along with the model evaluation;

3) Why did you pick your model over the other two options?

In addition to the code and the model specification, if you choose to submit a presentation / dashboard as part of your writeup, you can present your results in any way you like.