DATA SCIENCE ASSIGNMENT

The goal of this assignment is to provide you with an opportunity to demonstrate your data science skills in practice. Besides highlighting your technical skills and displaying your creativity as a data scientist, we are also interested in how you approach the assignment as a consultant, managing your time, effectiveness and communication. Being a consultant at Vantage AI requires that you can easily explain the results to our customers and demonstrate the quality of your product. In the next meeting, we invite you to give a presentation of your solution where you need to be able to convince technical experts, as well as laymen stakeholders of your expertise.

USE CASE & REQUIREMENTS

CASE

The case you have to solve is about water pumps in Tanzania. The data is stored in `[water_pump_set.csv]` and `[water_pump_labels.csv]`, and is sent to you together with this document. The data set originates from the Tanzanian Ministry of Water. Currently, this ministry maintains its pumps based on a maintenance schedule or, of course, when they break down. We feel that the maintenance of the Tanzanian water pumps could improve in both the cost of maintenance and the prevention of break downs by introducing machine learning to predict if a water pump is in need of repair or even the moment of failure of each water pump.

OBJECTIVE

Your objective is to develop a reproducible model that can predict which pumps will fail in the future, either the moment or just whether they fail. The purpose of your model is to be used by the Tanzanian government to effectively maintain their water pumps. Your solution can be a classification model or something else, as long as it can be effectively used to perform predictive maintenance.

PROGRAMMING LANGUAGE

Any language of choice. Examples are languages like R, MATLAB, Scala, Java, Python. You can also use other machine learning tools, or platforms, or combine them.

APPROACH

You are free to approach the case in any way you think is best. It is recommended that you use a model which classifies the status of a pump given the features. You should be able to show the (most important) steps to complete a data science project (e.g. data exploration, feature engineering, modeling, evaluation, …). Include the different techniques and approaches that you have used in your presentation, and the rationale why you used them. A clean and working solution that you can explain is preferred over an almost-amazing-but-not-working version. We advise you to start simple and pragmatic, and build from there.

ENGINEERING

The focus of this assignment is on data science. However, to be effective as a data scientist, we also expect some proficiency in and affinity with IT and engineering. Significant bonus points can be earned by including engineering elements in your solution, such as clean and reliable source code, databases, a serving application, demo frontend, reproducible environments, distributed processing frameworks, etc.

TIME

Use the time you need for this assignment. From our experience, we know that it can easily take 8-10 hours in total and we advise you not to underestimate the investment. Due to time restraints you will most likely not be able to do everything you have thought of. It is important that you use the presentation to show what else you would have done if you had more time.

DELIVERABLE

The deliverable is a zip file containing your code. Send this deliverable to **assignments@vantage-ai.com** one week after receiving the assignment at the latest. For any questions regarding this assignment, also use this mail address. During the next meeting, we invite you to give a presentation of your solution. You are not required to send your presentation slides/material to us. Good luck!