**Test assignment for a Data Scientist interview position.**

**Business Task:** Make an accurate forecast for each field and a weighted average (by field area) forecast for the cluster (group of fields) for the year 2020.

**Data:** The dataset includes observations for several years from different fields

* - information about the field (Field, year, yield, belonging to a geographic zone\*, belonging to a cluster\*, FAO hybrid\*, field area, applied fertilizers\*, previous crop type on the field)
* - weather information (weekly weather in the current year, average weather for the previous 18 years in different months)
* - vegetation information (weekly NDVI\* index)

*geographic zone* - an area with similar weather and soil conditions

*cluster* - a group of fields with similar climatic conditions and field processing technology

*FAO hybrid* - hybrid characteristic; the higher the FAO value, the higher the yield of the hybrid

*applied fertilizers* - N- nitrogen, P- phosphorus, K- potassium

*NDVI* - vegetation index [0, 1]. The higher the value, the more likely that there are no diseases in the field, and the crop is healthy

**Task:** Build a model that will be robust for each year of observations. Choose a metric to evaluate the algorithm's performance independently.

**Expected steps:**

* - EDA (Exploratory Data Analysis),
* - data cleaning (some data is inaccurate and scaled),
* - feature engineering,
* - model training and evaluation,
* - feature importance analysis,
* - future steps

**Additional Task:** With help of SHAP lib, describe what is most important features for the model and how they interact.

**Submission:** Push the solution to a repository on GitHub; we will review and discuss the results together.

Feature naming

'Field' – Field ID

'Year' - observation year

'Cluster' – admin cluster ID

'Area' – field area in hectares

'Yield' – yield in ton per hectare

'Geozone' – ID of geozone

'Predecessor' – ID of crop from previous year for this field

'FAO' – hybrid FAO

'N\_kg/ga', 'P\_kg/ga', 'K\_kg/ga' – the amount of applied fertilisers in kg per hectare

Weather features

'5\_YEARLY\_median\_AVERAGE\_TEMPERATURE'

5 - the month in which the average value for 18 years of observations was calculated

median/sum – aggregation

AVERAGE\_TEMPERATURE - feature type

'22\_WEEK\_AVERAGE\_TEMPERATURE'

22\_WEEK – week id

AVERAGE\_TEMPERATURE – feature type

The weighted average forecast by area per cluster can be calculated using the following formula



where

a - area ,

y - yield ,

i – through field id from one cluster

5\_YEARLY\_median\_AVERAGE\_TEMPERATURE