Data Scientist Task

Introduction

One of the typical problems we solve in our daily work at Raisa is predicting oil and gas production for a certain well. This can happen at any time in the well's lifetime, for this task you will be trying to infer and analyse the effect of certain features (available at the time of drilling) on the oil and gas production of wells in an area of interest using a method of your choice. Predictive performance is not the key here, our aim is to **capture complex relationships between the features and the targets** and **feature interactions' effect on the targets,** however building a performant predictive model might be a strong tool towards that main goal.

Dataset

You are provided with a dataset that contains 15 columns. The 2 target columns are: NormalizedOilEUR and NormalizedGasEUR

Feature / Target	Data Type	Definition
WellID	str	A unique 10-digit identifier for each well
BVHH	float	A positive number indicating rock quality at the area a certain well is drilled
FormationAlias	str	The zone/rock in the earth's crust the well extracts oil and gas from, more on Formations here
NioGOR	float	A positive number indicating the Gas-to-Oil ratio in the Niobrara formation at the well's coordinates
CodGOR	float	A positive number indicating the Gas-to-Oil ratio in the Codell formation at the well's coordinates
LateralLength	float	The length of the drilled well in feet
ProppantPerFoot	float	The amount (in pounds) of proppant (a chemical used to frac a well) per foot of the well's length, more on proppant here
FluidPerFoot	float	The amount (in gallons) of fluid (used to frac the well) per foot of the well's length
{Left/Right}Distance	float	The distance away from the {left/right} nearest neighbouring well
{Left/Right}NeighbourType	str	Whether this well's {left/right} neighbouring well was drilled prior to the well (parent), drilled with the well (co-developed), or no existing neighbour
TVD	float	The depth of the well (True Vertical Depth)
NormalizedOilEUR	float	The amount of oil (in bbl/ft) produced by the well in its lifetime normalized by its lateral length
NormalizedGasEUR	float	The amount of gas (in mcf/ft) produced by the well in its lifetime normalized by its lateral length

Requirement Specification

You are required to perform analysis on the dataset above that achieves the following:

- Identify the effect of each feature. Possible effects include but are not limited to: positive, negative, no effect, non-linear, dependant on other feature
- Communicate the identified effects in the most efficient manner. Communication media include but are not limited to: plots, metrics, interactive simulations, estimated functions

To achieve the above requirements, you might need to (probably won't need all):

- Perform data cleaning on the dataset
- Choose and use methods of data filling for missing values
- Drop unsuitable features
- Extract a feature from feature(s)
- Split dataset and perform individual analysis on each subset of data
- Build a model to predict the targets
- Pre-process features for modelling/analysis, ex: encode, scale, apply complex transformations, etc.

You are expected to:

- Write all the code needed for this task in python
- Back the decisions you take at all steps of the analysis
- Properly report the performance of any developed model

Deliverables

A Zip file containing:

- All the code written to fulfil the task (notebooks, scripts, modules, etc…)
- All the plots generated in all the analysis steps
- A final presentation communicating all the insights regarding feature effect on the 2 targets (this presentation will be presented by you later in the interview, 15 mins)

You will receive shortly a SharePoint invitation link where you can upload your zip file.

Submission Deadline

Deadline is Saturday Dec. 17, 2022, 11:59PM

Good luck 😊
Data Science team