Production prediction assessment

Diego Babativa
Data Scientist

S&P Global
Commodity Insights

1. Data validation and data understanding

Overview

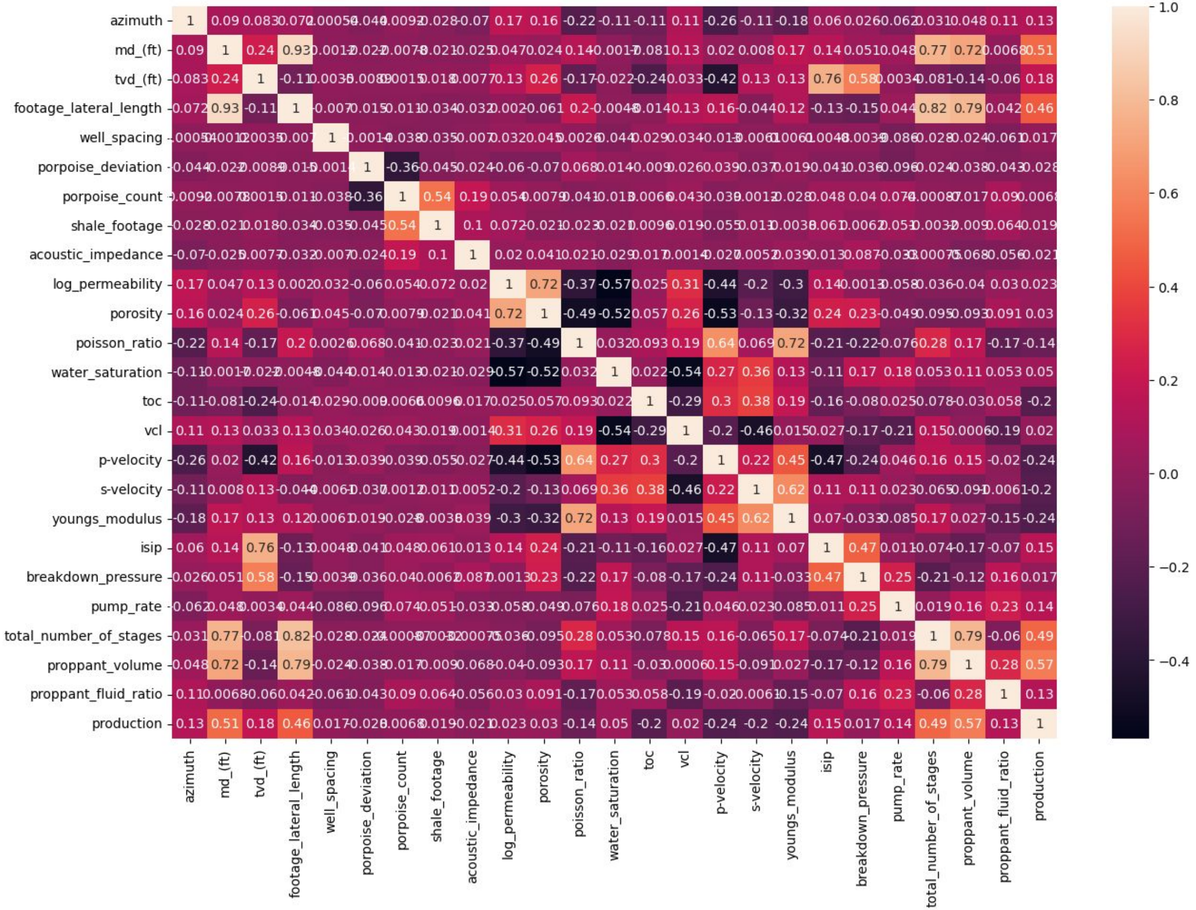
Dataset Statistics

Number of Variables	28
Number of Rows	1000
Missing Cells	1920
Missing Cells (%)	6.9%
Duplicate Rows	0
Duplicate Rows (%)	0.0%
Total Size in Memory	399.5 KB
Average Row Size in Memory	409.0 B
Variable Types	Categorical: 3 Numerical: 25

First insights:

Variable	**Description/Notes**
treatment company	31 different companies
date on production	Data from 01/01/2011 to 03/01/2019
operator	36 different operators
water saturation	has 57.7% missing values
breakdown pressure	has 74.4% missing values
azimuth	has 91.4% negative values
shale footage	has 33% zeros

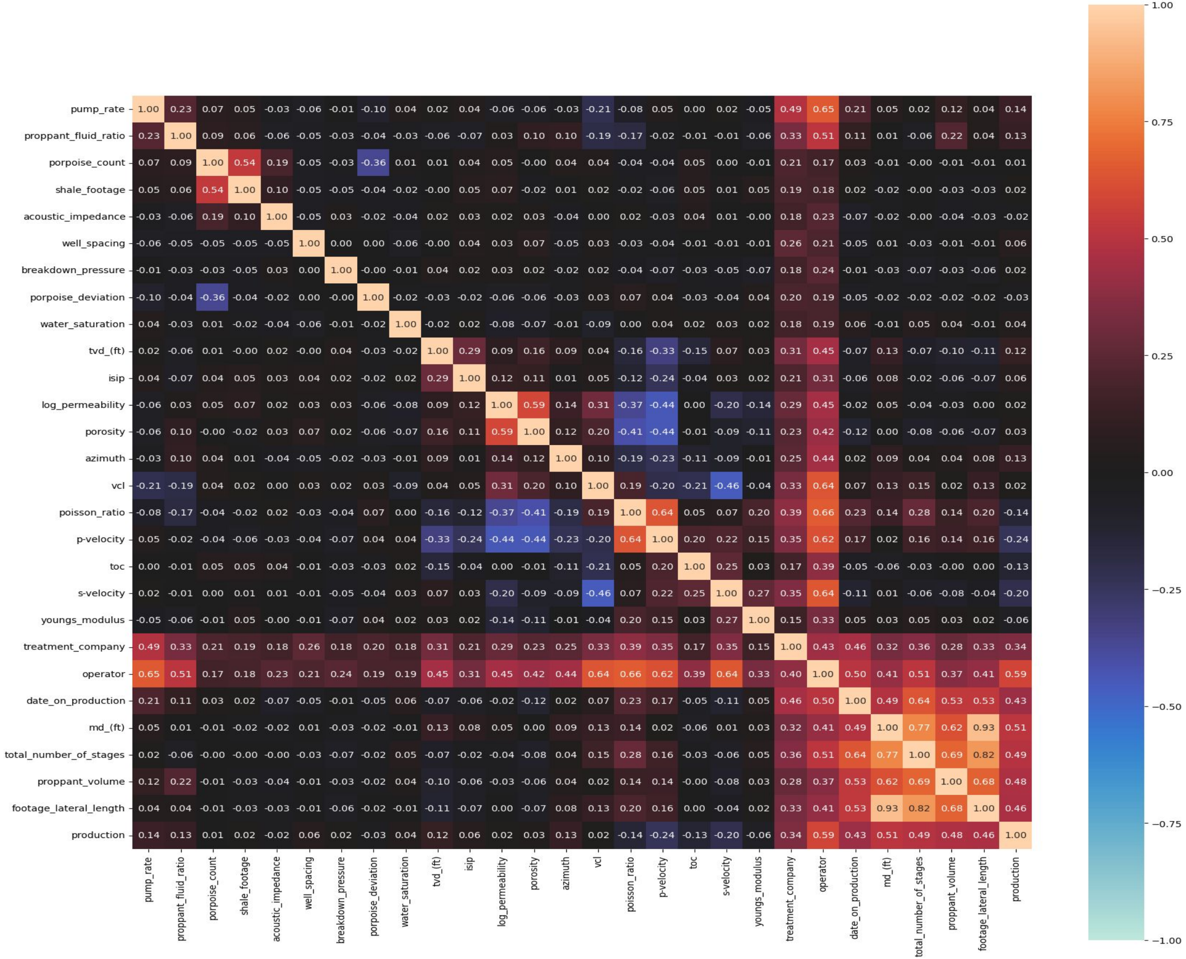
1. Data validation and data understanding - Correlations



Insights:

★ Strong correlation between **measure depth, footage lateral length, and proppant volume** with "**total number of stages"**
★ Good correlation between **youngs_modulus** and poisson_ratio.
★ Good correlation between the target variable (**production**) with **total number of stages** and **proppant volume**.

1. Data validation and data understanding - Correlations Cat and Numeric



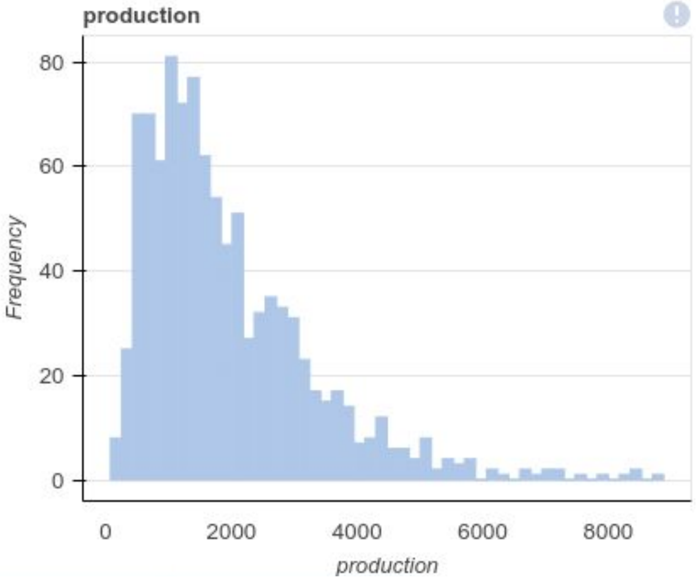
Insights:

★	Good correlation between **operator** and **production**.
★	Regular correlation between **treatment company** and **production**
★	Good correlation between **operator** and **pump rate.**

1. Data validation and data understanding - prediction variable

production
numerical

Approximate Distinct Count	**1000**	Mean	**1949.9195**
Approximate Unique (%)	**100.0%**	Minimum	**76.1072**
Missing	**0**	Maximum	**8880.6712**
Missing (%)	**0.0%**	Zeros	**0**
Infinite	**0**	Zeros (%)	**0.0%**
Infinite (%)	**0.0%**	Negatives	**0**
Memory Size	**16000**	Negatives (%)	**0.0%**



Insights:

- This variable has a left skewed distribution.
- Target variable to predict with trained models

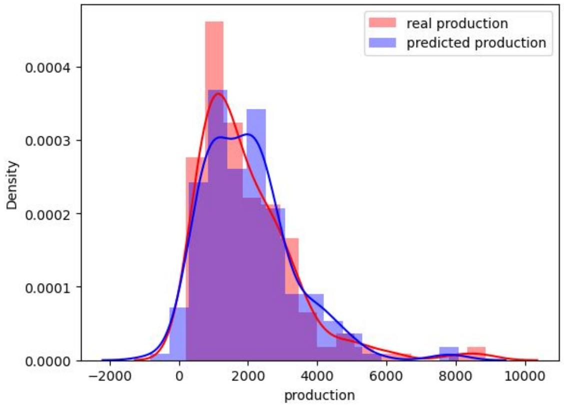
2. Modeling process

Training Test - split	
Training	800 rows
Test	200 rows

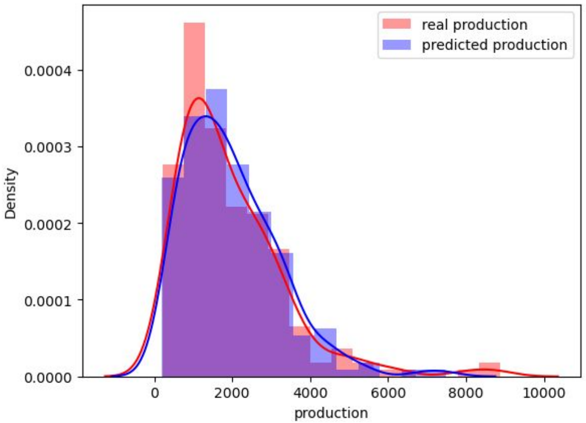
Model	Description/Notes
Linear regression	Sklearn model
Neural Network	Sequential model builded with Keras and tensorflow. Its architecture is composed of 2 Dense layers with 64 neurons each one and finally a Dense layer with 1 neuron to compute the output. The early stopping applied is useful to avoid overfitting. There aren't a lot of data for deep complex models, our proposal architecture with a small net with a few hidden layers to avoid overfitting.
XGBoost Regressor v1	Model optimized with Bayesian Optimization. Logged with Weights and Biases (https://wandb.ai/dbabativa/spg/sweeps/z8dmz4o5/table?workspace=user-dbabativa)
XGBoost Regressor v2	Version 2 with optimizer technique.

3. Modeling selection

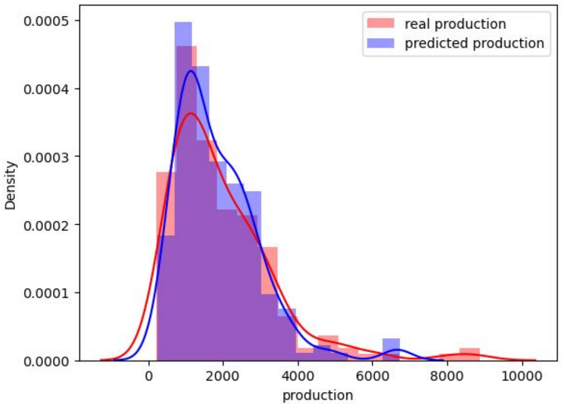
Linear regression



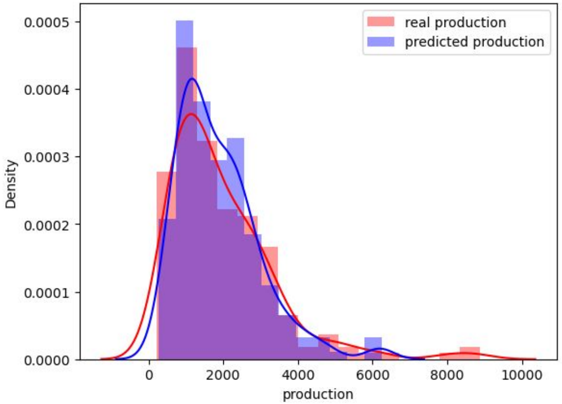
Neural Network



XGBoost Regressor v1



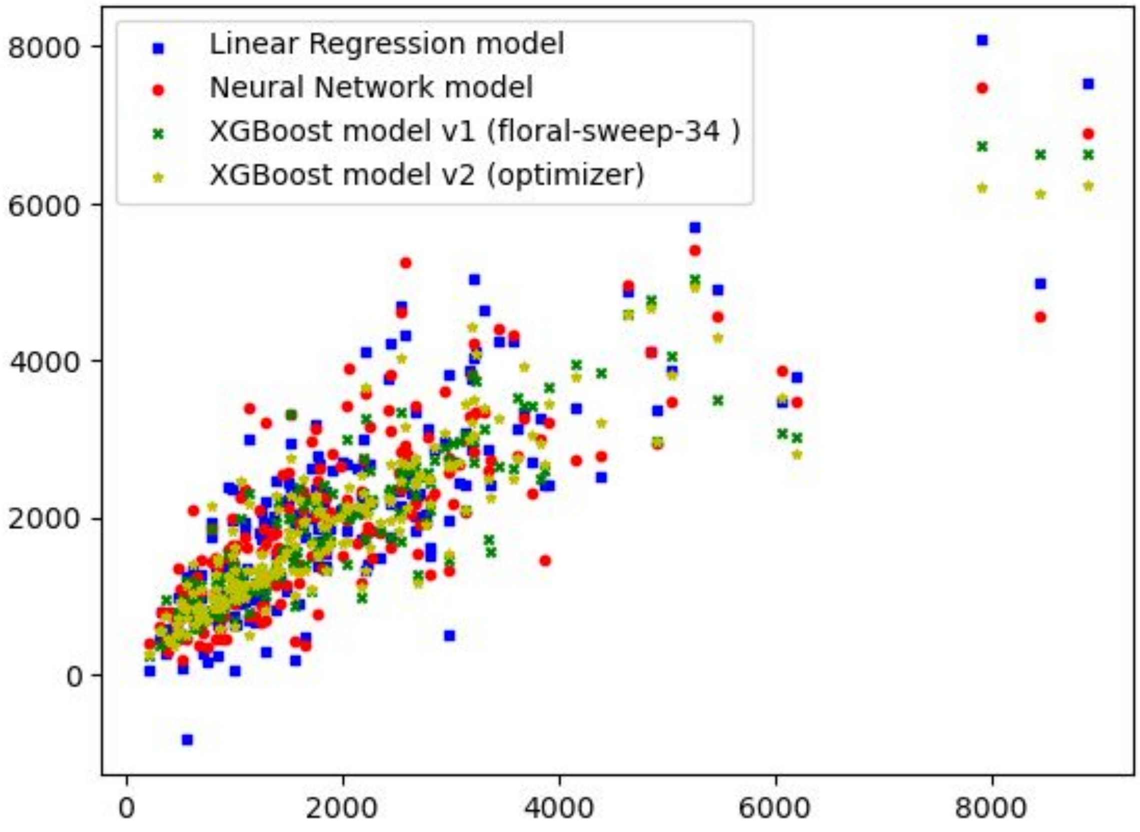
XGBoost Regressor v2



Insights:

★ Very similar distributions (real vs predicted) with XGBoost models
★ The Linear regression and neural network models missing a high density from 0 to 2000 mmcf.

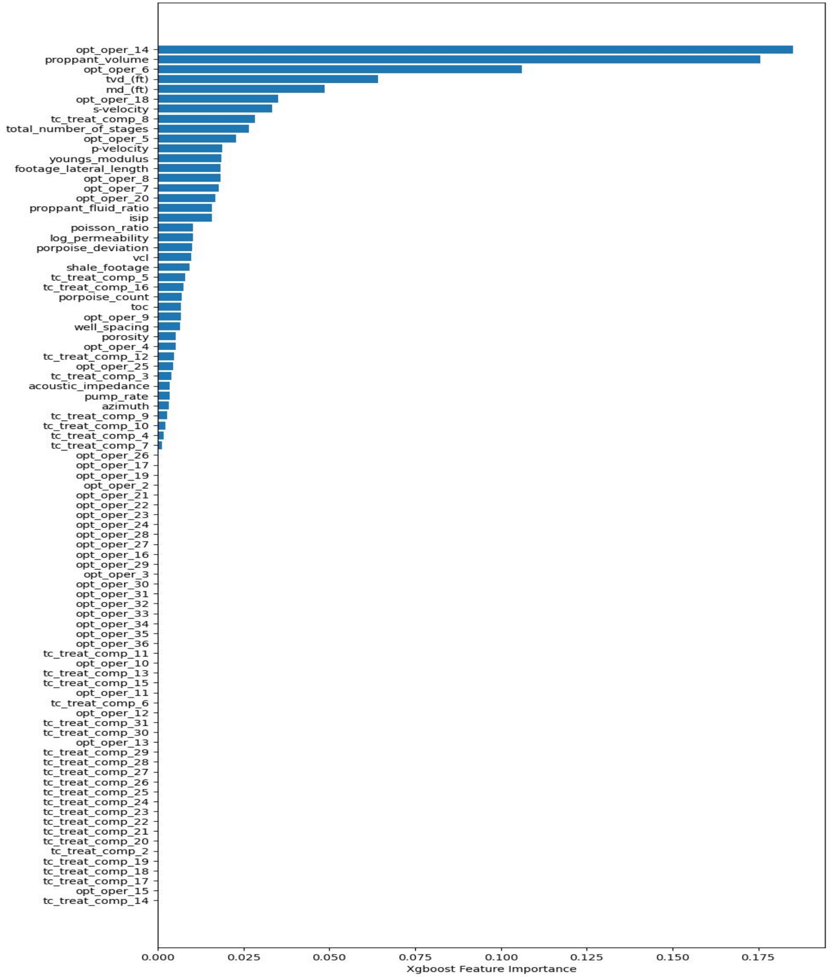
3. Modeling selection - Scatter visualization



Insights:

★ This plot concatenates the result of y_true and y_pred values for all trained models in a scatter plot.
★ Although all models have external points, one more time XGBoost models have a good fitting

3. Modeling selection - metrics results



model	MAE	MSE	RMSE	RMSLE	R2
Linear Regression model	598.217705	663173.173157	814.354452	6.702396	0.669309
Neural Network model	611.814959	719874.629228	848.454259	6.743416	0.641034
XGBoost model v1 (floral-sweep-34)	376.539161	399816.404012	632.310370	6.449380	0.800631
XGBoost model v2 (optimizer)	407.586408	427512.000198	653.844018	6.482869	0.786821

Insights:

- ★ The best results were achieved by **XGBoost model v1** optimized with Bayesian Optimization.
- ★ The another XGBoost model can be a candidate to do a blind test with other data (more updated for example)
- ★ If we inspect over the feature importance of XGboost model v1, variables such as **operator, proppant volume** and **md (ft)** are taking into account for the prediction. This make sense with the correlation analysis done in Data Understanding stage.

XGBoost model v1 Feature importance