Texas A&M Student Chapter - Society of Petroleum Engineers

A&M Technical Exhibition & Networking Event

CHALLENGE
JAN 23 - FEB 7

Data Science Challenge

Join us and test your data-science skills!

Registration opens:	NOW
Challenge timeline:	Jan. 23- Feb. 7, 2021
Challenge level:	Intermediate, Basic
Prizes:	$400 (Int), $100 (Basic)

Scan the QR Code below or visit this link for more information.
Contact: tamu.spe.te.ne@gmail.com

Sponsored by:

SM ENERGY

TE&NE Challenge

A&M

Saturday, January 23rd, 2021

Technical Exhibition & Networking Event

What is TE & NE?

The technical exhibition and networking event (TE&NE) is an event designed for students to explore the eight pillars of the Society of Petroleum Engineers (SPE). The event is hosted by the Texas A&M University SPE student chapter. As part of the event, we would like to challenge students with one of the currently trending pillar: Data Analytics & Engineering Analysis.

Basic Challenge

The basic challenge is designed to be solved by 1-2 people. The challenge will be a regression exercise to generate artificial sonic logs perfectly from the least number of available logs. You will be provided with a well log data and challenge guide. Winning team will get $100.

Intermediate Challenge

The intermediate challenge is designed for those with relatively advanced knowledge of data analytics. The challenge will be to develop an improved time-series forecasting model using production data. You will then need to compare your implemented data-driven model with the conventional Arps decline model and highlight the differences. You will be provided with a challenge guide and data to be analyzed. Winning team will get $400.

FREE SWAGS!!!

Not wanting you to miss the goodies from the in-person technical exhibition, our team got some swags for you to pick-up. The first 20 participants to submit their completed project will receive some swags, which include (but not limited to): TE&NE tote bag, TE&NE notebook, TE&NE spray sanitizer, NOV swags, Don-Nan swags, and more! So, sign-up now and book your spot!

Many thanks to SM Energy for sponsoring this challenge!

SM | ENERGY

TE&NE Challenge Problem Statements

Saturday, January 23rd, 2021

A&M Technical Exhibition & Networking Event

Problem 1 (Intermediate)

The purpose of this challenge is to benchmark and discover new approaches that may lead to a class of more reliable production forecasts than traditional approaches such as Arps, modified hyperbolic, Duong and logistic DCA models. The quality of the forecasts will be established by using incremental hindcasting every 6 months of available historical data; that is, blind forecast tests using 6, 12 and 18 months of historical data. To compare errors in the training and testing phases use the root mean square error (RMSE). Comparisons should be made against the Arps model (base model).

Deliverables:

- Functional code in Python or Matlab
- Short presentation (5 slides max) highlighting:
    - Description of the methodology employed
    - Comparative performance of your method vs Arps
    - Validation of the results in terms of RMSE
    - Remarks of your learnings and possible improvements

- All of the material should uploaded to the following link in a single zip file for submission titled Problem1_<TeamNumber>_LastNames_Final: https://forms.gle/pEQ9e9ZNQQZyMA7Q8

Problem 2 (Basic)

Build a data-driven model to predict both compressional travel time (DT) and shear travel time (DTS) logs using least number of logs/inputs from the logging-while-drilling dataset. In other words, obtain the best prediction with least number of input logs. Such a data-driven model will help approximate the geomechanical properties in wells where sonic logs cannot be run. Such a data-driven model will require limited number of input logs to generate the desired target logs. Sonic travel time logs are used to quantify the elastic moduli and brittleness of the hydrocarbon-bearing formations.

Deliverables:

- The Deployment sheet with predictions of DT and DTS logs.
- Functional code in Python or Matlab
- Short presentation (5 slides max) highlighting:
    - Description of the methodology employed
    - Comparative performance of your method vs other alternatives you tried
    - Validation of the results in terms of RMSE and other valid metrics
    - Remarks of your learnings and possible improvements

- All of the material should uploaded to the following link in a single zip file for submission titled Problem2_<TeamNumber>_LastNames_Final: https://forms.gle/pEQ9e9ZNQQZyMA7Q8

Problem 1 (Intermediate)

Methodology

Since rate production curves can be seen as a time series object, there are a plethora of methods to perform forecasting out there. For the purpose of the present challenge, we are going to consider the following classes of methods:
1. Ensemble based methods - These are based on the agrupation of weak models to create a strong predictive model.
2. Recurrent neural networks (RNNs) - These are basically neural networks with memory that can be used for predicting time-dependent targets.
3. ARIMA type of methods - These models incorporate Autoregressive (AR) and Moving Average (MA) approaches to build a composite model of the time series.
4. Modal decomposition type of methods - These are methods that reconstruct and predict spatio-temporal patterns hidden in the data based on eigenvalue decompositions.
5. The Facebook Prophet model - This is a robust time series forecasting package based on an additive model where non-linear trends are fit with seasonality effects at different time scales.
We are going to explore a representative set of these approaches in separate teams.

The table below proposes 6 different ideas on how to go about exploring these approaches

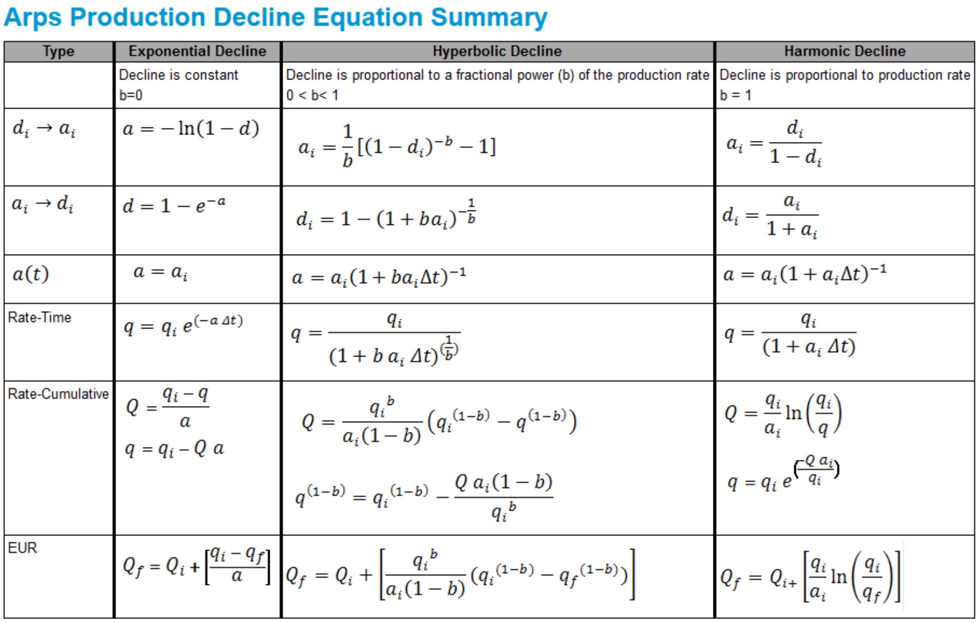
#	Teams	Methods	Source
1	Ensemble	Random forests, LightGBM, CatBoost,...	https://machinelearningmastery.com/gradient-boosting-with-scikit-learn-xgboost-lightgbm-and-catboost/
2	Recurrent	LSTM, GRU, ...	https://towardsdatascience.com/predictive-analysis-rnn-lstm-and-gru-to-predict-water-consumption-e6bb3c2b4b02
3	Moving Average	ARIMA, SARIMA, ...	https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/
4	Modal	DMD, EDMD, ...	https://github.com/mathLab/PyDMD
5	Prophet	Facebook Prophet API	https://facebook.github.io/prophet/

Additional Tips

Keep in mind that good data science practice includes understanding of the requirements, the data and the performance of the model. Make sure to check them to achieve the desired deliverables of the present challenge. Coordinate carefully you work and unleash your collective creativity. Feel free to preprocess the data to improve the predictability of your approach. You may consider filtering, smoothing, transforming or clustering the production curves to that purpose.

You may also want to consider reviewing the basic building blocks of a traditional decline curve theory. **The data repository also has piece of code that gets you started on building the traditional decline curves on the production data.** While that provides a great baseline to start (and finish) constructing traditional decline curves on the production data, please feel free to modify the same based on your idea of a better industry standard.

A summary of key equations can be given by



Arps Production Decline Equation Summary

Type	Exponential Decline	Hyperbolic Decline	Harmonic Decline
	Decline is constant b=0	Decline is proportional to a fractional power (b) of the production rate 0 < b< 1	Decline is proportional to production rate b = 1
$d_i \rightarrow a_i$	$a = -\ln(1-d)$	$a_i = \frac{1}{b}[(1-d_i)^{-b} - 1]$	$a_i = \frac{d_i}{1-d_i}$
$a_i \rightarrow d_i$	$d = 1 - e^{-a}$	$d_i = 1 - (1+ba_i)^{-\frac{1}{b}}$	$d_i = \frac{a_i}{1+a_i}$
$a(t)$	$a = a_i$	$a = a_i(1+ba_i\Delta t)^{-1}$	$a = a_i(1+a_i\Delta t)^{-1}$
Rate-Time	$q = q_i\, e^{(-a\,\Delta t)}$	$q = \frac{q_i}{(1+b\,a_i\,\Delta t)^{(\frac{1}{b})}}$	$q = \frac{q_i}{(1+a_i\,\Delta t)}$
Rate-Cumulative	$Q = \frac{q_i - q}{a}$ $q = q_i - Q\,a$	$Q = \frac{{q_i}^b}{a_i(1-b)}\left({q_i}^{(1-b)} - q^{(1-b)}\right)$ $q^{(1-b)} = {q_i}^{(1-b)} - \frac{Q\,a_i(1-b)}{{q_i}^b}$	$Q = \frac{q_i}{a_i}\ln\left(\frac{q_i}{q}\right)$ $q = q_i\, e^{(\frac{-Q\,a_i}{q_i})}$
EUR	$Q_f = Q_i + \left[\frac{q_i - q_f}{a}\right]$	$Q_f = Q_i + \left[\frac{{q_i}^b}{a_i(1-b)}({q_i}^{(1-b)} - {q_f}^{(1-b)})\right]$	$Q_f = Q_{i+}\left[\frac{q_i}{a_i}\ln\left(\frac{q_i}{q_f}\right)\right]$

Data

We are going to use public production data available from the Vaca Muerta formation in Argentina. The production and well dataset for Problem 1 is provided in the email along with this document.

Additional Reading

1. Dabakoglu, C. (2019). Series Forecasting — ARIMA, LSTM, Prophet with Python

2. Vlachanov, I., (2018). KDNuggets. Data Science Predicting the Future.

3. Cao, Q et al. (2016). Data Driven Production Forecasting Using Machine Learning. SPE N. 180984-MS

Regarding Traditional Decline Curve Analysis:

4.http://www.fekete.com/san/webhelp/feketeharmony/harmony_webhelp/content/html_files/reference_material/analysis_method_theory/traditional_decline_theory.htm

5.https://petrowiki.spe.org/Production_forecasting_decline_curve_analysis

Problem 2 (Basic)

Dataset and Tasks

Regression models need to be developed and evaluated using the data in the **Training-Testing** sheet. **DON'T USE THE DATA IN "DEPLOYMENT" SHEET TO DEVELOP/EVALUATE THE DATA-DRIVEN MODEL.**

After the data-driven model is fully developed and thoroughly evaluated on the data in the **Training-Testing** sheet, the data-driven model needs to be deployed on the data in the **Deployment** sheet to generate the DT and DTS logs for those depths without the sonic log information.

The winning team will be the one that achieves highest accuracy in predicting the DT and DTS logs in the Deployment Sheet. The judges for this contest will run a code to read the team's DT and DTS predictions and compare it with the known DT and DTS values accessible only to the judging panel. The judges will also check the inner workings of the wining team's code and model for reliability and robustness.

Suggestions

- Perform outlier detection and fix missing data if any.
- Perform data scaling/standardization.
- Perform dimensionality reduction (feature selection/feature extraction). Identify the most relevant features. Quantify the importance of features.
- Develop the regression models. Perform cross validation with hyperparameter optimization to ensure the model is well generalizable. Find the best hyperparameters.
- Perform good evaluation using good metrics during the training and testing stages.
- Visualize the predictions and errors.

Data

We are going to use log data available for a well in the Volve Dataset distributed by Equinor. Use the training and testing sheet for building the model and use the deployment sheet to lock in results of your best model. Dataset for Problem 2 is provided in the email along with this document.

Reading Material

He, J., Li, H., & Misra, S. (2019). Data-driven in-situ sonic-log synthesis in shale reservoirs for geomechanical characterization. *SPE Reservoir Evaluation & Engineering*.

Osogba, O., Misra, S., & Xu, C. (2020). Machine learning workflow to predict multi-target subsurface signals for the exploration of hydrocarbon and water. *Fuel*, *278*, 118357.

He, J., & Misra, S. (2019). Generation of synthetic dielectric dispersion logs in organic-rich shale formations using neural-network models. Geophysics, 84(3), D117-D129.