Models Explanation Thesis

1) Methodology

Figure 1 shows the overall methodology for data preparation and model training.

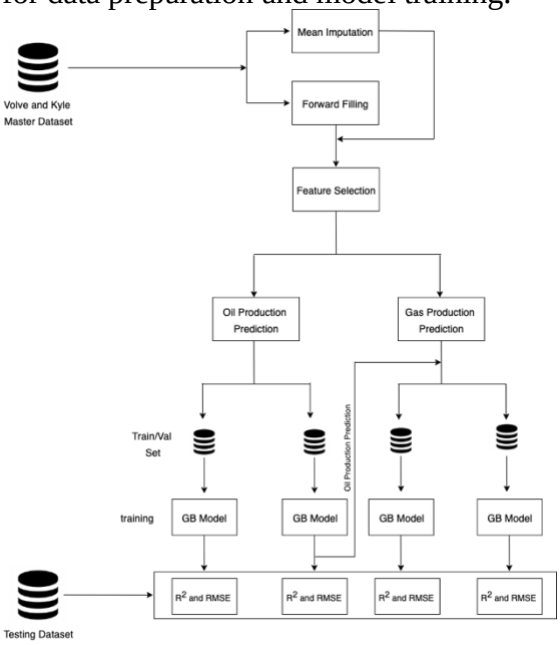

Figure 1: Methodology

2) Automatic Machine Learning

In order to gain a better insight into the possible models that could be used on this dataset, the automatic machine learning (AutoML) algorithm was used. AutoML allows data analysts and scientists to make machine learning models more efficiently [1]. An open-source platform that can be used for AutoML is H2O. The AutoML function in H2O functions by building several models and attempts to discover which is the most ideal model [2]. The models could include Gradient Boosting and Random Forest [2].

3) Forward Filling and Mean Imputation

As shown in Figure 1**,** two methods were used to fill in the missing values in the dataset, namely forward filling and mean imputation. The AutoML model from H2O was trained with both forward filling and mean imputation datasets. The R^2 and RMSE metrics were used to evaluate the performances of the model on the test dataset.

Table I: Evaluation for oil production prediction

Dataset	R^2	RMSE
Forward Filling	0.72	162
Mean Imputation	0.06	485

Table II: Evaluation for gas production prediction

Dataset	R^2	RMSE
Forward Filling	0.81	54354
Mean Imputation	0.07	114976

Table III: Evaluation for gas production prediction while adding oil production

Dataset	R^2	RMSE
Forward Filling	0.78	59033
Mean Imputation	0.06	114975

Table I, Table II, and Table III show that the model performs better when the missing data is filled in with the forward filling method as it has a higher R^2 value and a lower RSME value. Furthermore, Table II and Table III showed that the model performed better if the value of the predicted oil production is included in the training,

4) Model

The *Automatic Machine Learning* section explained that the H2O helps data analysts and scientists by discovering the most ideal model. It returns a leaderboard that ranks all the models that it built. Using H2O showed that for oil production prediction, the ideal model would be gradient boosting. In addition to this, it also showed that the ideal model for predicting gas production would be gradient boosting as well. Different types of the gradient boosting models were made with different parameters to determine the ideal model. The summary of these models is shown in Table IV and Table V. These tables show that the number of trees plays a significant role in the

model's performance. For a gradient boosting model, too many trees would lead to overfitting whereas too few trees would result in underfitting [3]. Therefore, it is vital to find the right number of trees. In Table IV, model A.1 showed the best performance with 94 trees, whereas in Table V, model B.2 showed the best performance with 95 trees. Figure 2 shows the performance of model A.1 and model B.2 on the test dataset.

Table IV: Gradient Boosting Models for Oil Production Prediction

Model	Number of trees	Min depth	Max depth	Mean depth	Min leaves	Max leaves	Mean leaves	R^2	RSME
A.1	94	0	16	13	1	1327	651	0.72	162
B.1	231	10	10	10	28	347	124	0.55	243
C.1	406	11	15	14	35	89	69	0.63	200

Table V: Gradient Boosting Models for Gas Production Prediction

Model	Number of trees	Min depth	Max depth	Mean depth	Min leaves	Max leaves	Mean leaves	R^2	RSME
A.2	82	13	13	13	1	2217	1357	0.75	69247
B.2	95	17	17	17	142	1333	679	0.81	54354
C.2	73	15	15	14	1525	3580	2851	0.80	59146

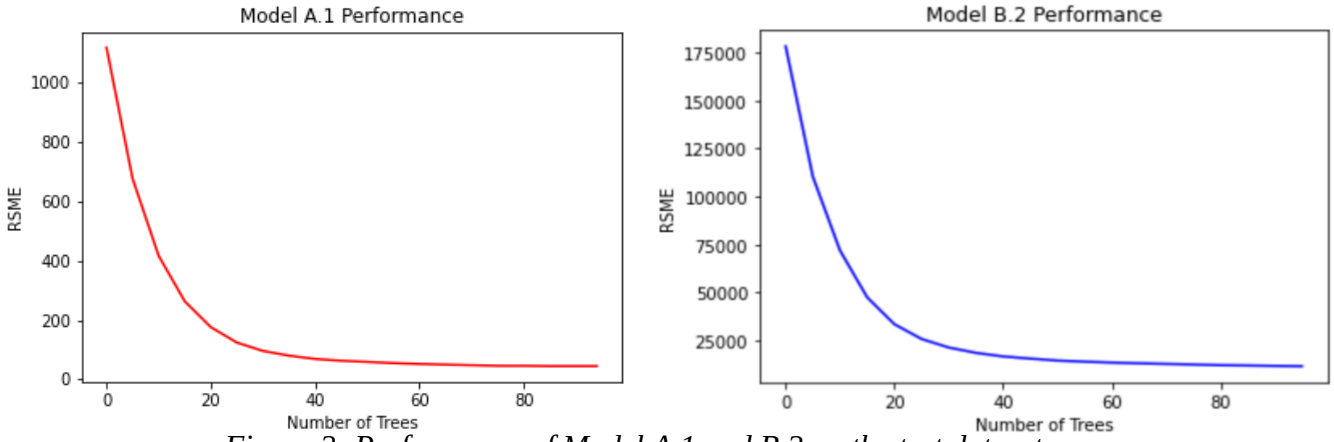

Figure 2: Performance of Model A.1 and B.2 on the test dataset

5) References

[1] Microsoft, "What is automated machine learning (AutoML)?," Microsoft, 17 March 2022. [Online]. Available: https://docs.microsoft.com/en-us/azure/machine-learning/concept-automated-ml.

[2]H2O, "H2O AutoML," H2O, [Online]. Available: https://h2o.ai/platform/h2o-automl/.

[3]S. Dash, "Gradient Boosting – A Concise Introduction from Scratch," 21 October 2020. [Online]. Available: https://www.machinelearningplus.com/machine-learning/gradient-boosting/.