Initial Investigation on Models

1) Methodology

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

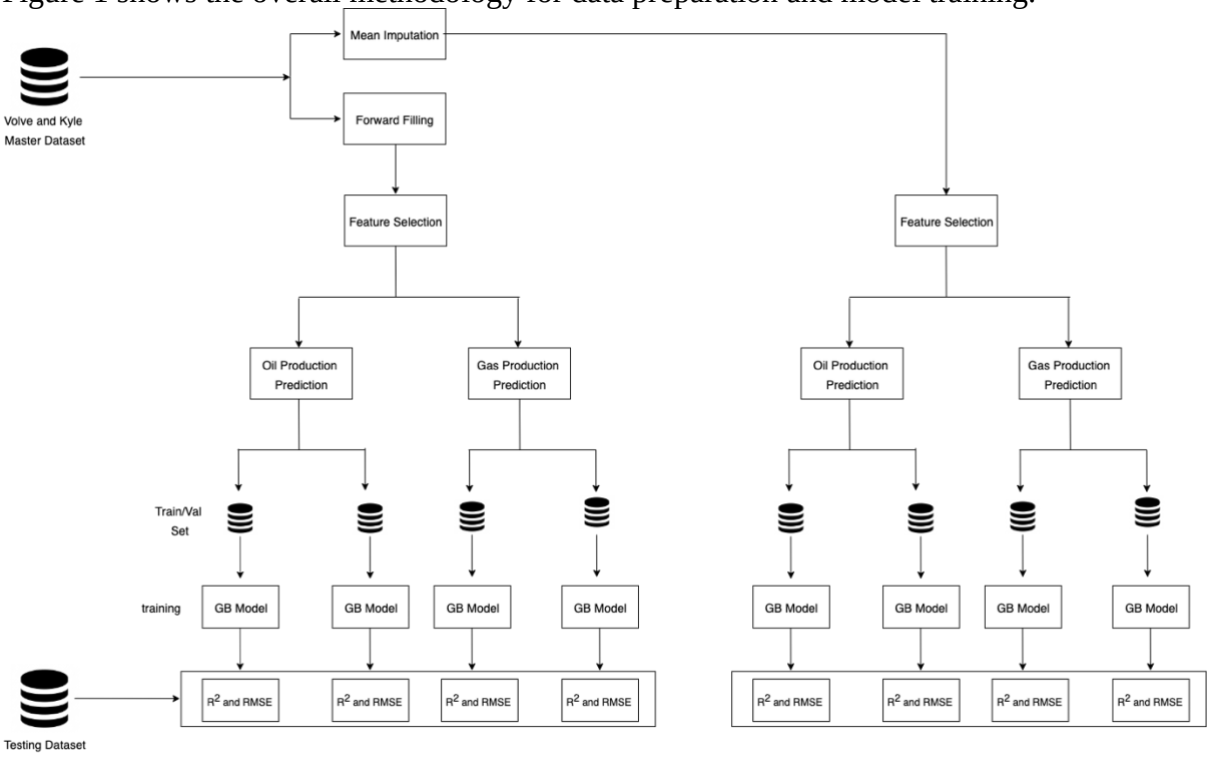

Figure 1: Methodology

2) 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. A gradient boosting model and random forest baseline models were 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.

	Gradient Boosting		Random Forest	
	R^2	RMSE (m^3)	R^2	RMSE (m^3)
Oil prediction	0.77	175	0.53	261
Gas prediction with oil predicted value	0.25	95601	0.72	72488
Gas prediction without oil predicted value	0.80	53121	0.75	76189

Table I: Forward filling method used on gradient boosting and random forest models

	Gradient Boosting		Random Forest	
	R^2	RMSE (m^3)	R^2	RMSE (m^3)
Oil prediction	0.0000887	750	0.15	408
Gas prediction with oil predicted value	0.35	110113	0.60	98575
Gas prediction without oil predicted value	0.77	65050	0.75	88199

Table II: Mean imputation method used on gradient boosting and random forest models

Table I and Table II showed that the models performed worse if the values of the predicted oil production are included in the training. Therefore, it would be better to not include the predicted oil production value for model training. In terms of oil prediction and gas prediction without the oil predicted values, Table I and II showed that the forward filling method is better. This is because the forward filling method gave a lower RMSE and R^2 score in both cases. Furthermore, the R^2 value for the oil prediction is extremely low when the mean imputation method is used. A low R^2 value indicates a poor model whereas a high RMSE value indicates a poor model. Therefore, forward filling is the better method for data imputation.

In the forward filling method, Table I and II showed that the gradient boosting model performed better than the random forest model. The gradient boosting model gave a higher R^2 value and a lower RMSE value compared to the random forest model. Therefore, this makes the gradient boosting model a better choice compared to the random forest model.

3) 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].

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 III and Table IV. 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 III, model A.1 showed the best performance with 94 trees, whereas in Table IV, 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 III: 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	RMSE (m^3)
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 IV: 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	RMSE (m^3)
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

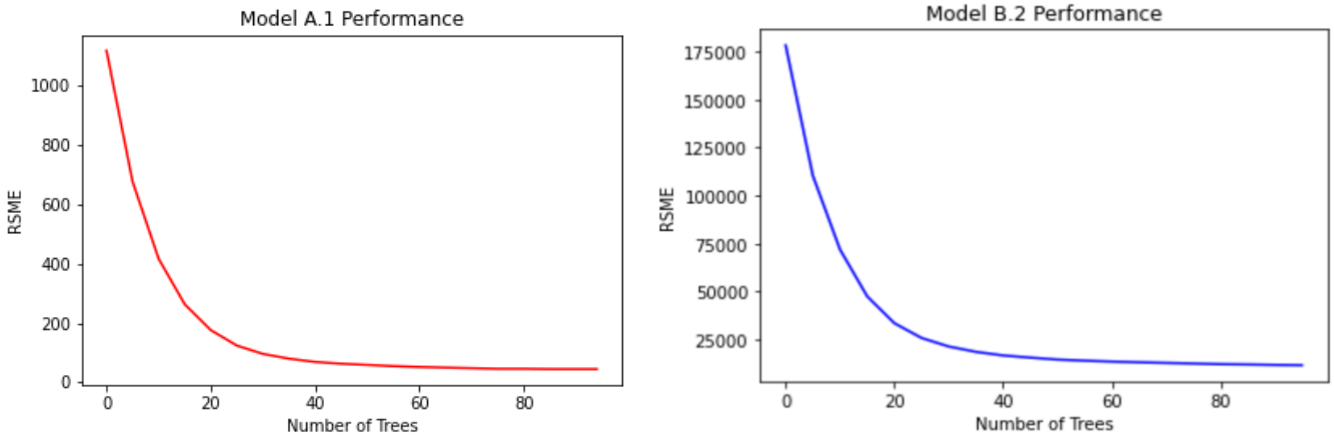

Figure 2: Performance of Model A.1 and B.2 during training

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/.