Innovative: International Multi-disciplinary Journal of Applied Technology
(ISSN 2995-486X) VOLUME 03 ISSUE 12, 2025

Analysis of Lightgbm Model Accuracy and Validation Strategies for Predicting Oil and Gas Probability Maps

Markhabo Shukurova[1]

[1]Department of "Software and Technical Support of Computer Systems",
Karshi State Technical University, Karshi, Uzbekistan

Abstract:

The accuracy of the LightGBM model in forecasting oil and gas probability maps had been evaluated based on MapOil analytic system. Integration of geological, geophysical, satellite and topographic map datasets in single spatial processing pipeline through MapOil streamline feature extraction, and probability assessment for each grid cell over exploration areas. The leaf-wise growth and advanced gradient boosting optimization criteria of LightGBM are capable of capturing the complex nonlinear relationships associated with subsurface properties. Differences among the various validation strategies are also examined to compare the robustness of predictions. Standard random cross-validation along with spatially informed cross-validation approaches, such as block and region-based validation, were utilized. Validation in space results in reduced information leakage and more realistic performance statistics, leading to a more informative generalisation of the model on untested areas. In summary, the application of LightGBM in combination with MapOil platform forms a powerful methodology for creating high-precision oil and gas probability maps which are able to improve the data driven decision making for hydrocarbon exploration.

Keywords: LightGBM, oil and gas probability mapping, spatial validation, hydrocarbon exploration, mapoil system, gradient boosting, geological data analysis, machine learning models, spatial prediction, geoscience informatics

1. Introduction

Digitisation and systematically learning from test projects will play a bigger role in finding and evaluating (opportunities for) oil and gas fields. Conventional geological and geophysical investigations usually demand expensive and time-consuming procedures, which is why advanced statistical tools such as machine learning are common for fast and efficient field detection. In this regard, developing and predicting for probability maps is an indispensable instrument for resource planning and risk estimation in the oil and gas industries [1].

The LightGBM (Light Gradient Boosting Machine) is especially well-suited for this task. It is a fast and scalable implementation of gradient boosting method. Finally, the customized optimization algorithms and leaf-wise splitting features in LightGBM further improve precision. When LightGBM model is taken in the field of oil and gas probability mapping, a high accurate pre-diction results for the probability distribution fields with geological parameters as well as indicators obtained from leading data sources [2].

Validation methodologies are essential to test accuracy and trustworthiness of models. Techniques such as cross-validation, k-fold, and stratified sampling provide the capability to split the data randomly and avoid overfitting. Furthermore, as the oil and gas data is usually uncertain and rarely observed, the generalization ability of the model has to be cau-tiously addressed in validation procedure [3].

The LightGBM model with different validation paratameters is resulted beneficial in that it is capable of increasing the accuracy of oil and gas probability maps as well as reducing false predictions. This methodology allows geologists and industrial experts to coordinate resources, develop extraction solutions and evaluate risks before-hand [4].

LightGBM (Light Gradient Boosting Machine) is high performance machine learning model based on the gradient boosting algorithm used for large scale and high-dimensional datasets [5].

Gradient boosting is an ensemble technique to build a sequence of interdependent decision trees one by one in an additive manner. Why LightGBM is much faster than the usual implementation of gradient boosting? Leaf-wise splitting strategy – LightGBM builds trees by choosing the leaf with the maximum loss improvement to split. This improves model performance and permits the processing of huge datasets [6].

1. Fast: LightGBM is very fast because of its unique histogram based learning algorithm and optimized tree construction that helps to process the data faster, this increases the efficiency in terms of memory or the CPU used by other algorithms.
2. Capable of scaling high-dimensions –it can perform well for data with thousands of features and millions of records.
3. Regularization and overfitting reduction – L1 and L2 regularization, among other parameters supported by LightGBM to reduce the risk of learning noise/overfitting and improve generalization.
4. These are the reasons that have made LightGBM model is increasingly used in the industry of oil and gas for probability mapping. It benefits the high-precision prediction of the probability distribution of fields on geological parameters, geophysical data and advanced information sources [7].

From this point of view, a comprehensive study on the capabilities of LightGBM model and the performance of those in oil & gas probability maps predication is both scientific and practical significant. The objectives of this work are to estimate the uncertainty in the LightGBM model, compare validation strategy and check whether the models could maintain a good performance for real geologic data [8].

Zhang T., Chai H., Wang H., Guo T., Zhang L., and Zhang W. suggested that the LightGBM model could be more appropriate than physical models to determine shear wave (S-wave) velocity (Vs) in carbonate formations. Input data were standard well log data (if available, even if partial) and geological context information (depth, pressure, temperature and type of formation). The model was optimized by utilizing autosktlearn hyperparameter optimization tool and achieved an improved better prediction performance on test wells versus conventional rock physics based models and other ML methods. The authors also used the SHAP (Shapley Additive Explanations), a type of interpretability method, to find out input parameters (logs and geological conditions) which influence Vs the most. This is especially useful in your approach of "probability mapping + interpretability + objective model evaluation." "SINTEF's conductivend model is the only one available for model validation under temperature behavior and comparisons to physical principles are very consistent.All these indicated make the chosen model a trustworthy tool for all your laboratory research [9].

Introduction WatheqJ.Al Mudhafar, AlqassimA. Hasan, MohammedA.Abbas and David A.Wood aim was to estimate permeability in carbonate reservoirs via well-log data (gamma ray, resistivity, neutron porosity,density, porosity,faciesetc.) instead of costly "core analysis ". They experimented with various ML algorithms, as well as LightGBM, and systematically used preprocessing steps such as for example normalization (log transformation, Box-Cox, NST), outlier removal and missing values inputation. Specifically, the researchers contrasted random search against Bayesian optimization for hyperparameter tuning, which enabled them to optimize model accuracy. The aim was to accurately predict the permeability in untested wells. This is a methodology reference for your project settting geological + log + ML + p-mapping, where we are trying to empirically implement rather than do experiments graphically/log based [10].

At the Ras Fanar field in the Gulf of Suez, they assumed it was a carbonate reservoir and established regression equations to estimate horizontal core-permeability from effective porosity logs and resistivity (RRT) profiles. They tried different machine learning algorithms and performing hyperparameter optimization with grid search. Data were divided into training and validation sets by random sampling to examine the generalization ability of the model. You could take this same approach for you oil and gas probability mapping project as well, especially if you want to assess inter-well zones [11].

Alireza Roustazadeh, BehzadGhanbarian, Mohammad B. Shadmand, Vahid Taslimitehrani and Larry W. Lake tried to predict the Recovery Factor (the recovery efficiency from reserves) of oil/gas reservoirs using machine learning methods associated with formation characteristics including: porosity, permeability, water saturation and pressure of the rock. They utilized XGBoost, SVM, and a simple multiple linear regression (MLR) algorithms. Results XGBoost was the best classifier for both training and testing, all classifiers showed low performance in an independent dataset. This means that when we only train a model on data from one field or just a few wells, we can get significant errors trying to extend the model to other wells or fields. Hence, an important focus for predictive model development is the diversification, validation and generalization out-of-sample of datasets [12].

2. Materials and Methods

Various Development of probability maps… …The generation of probability maps for predicting oil and gas reservoirs require the integration of geological data with new machine learning tools. Such maps are essential in forecasting the reservoir position, geological structure and extraction potential. For its high accuracy and speed, prognostic model founded on LightGBM (Light Gradient Boosting Machine) is well-known in this realm, which can process a great number of geophysical and geological data effectively.

LightGBM is a gradient boosting framework that was engineered to achieve tree-based Go click models in an efficient and ultra-fast way. It is good at big data and also tends to find complex interactive attribute relations. When designing oil and gas chance maps, the model makes it possible to include reservoir parameters, seismic information and well data.

Model performance is measured with several metrics. For LightGBM, key performance metrics include AUC (Area Under the Curve), log loss, accuracy, precision and recall. To validate these results, we can use both metrics to study how well the model finds communicated and predicted field zones. In particular, AUC is a very important measure that characterizes the discriminative and sensitivity of probability maps.

Validation methods are essential to evaluate generalization potential of the model. Techniques ranging from cross-validation, k-fold, stratified k-fold to hold-out can be used for validating the stability of LightGBM model. Stratified k-folds validation is particularly powerful, as it adjusts for the levels and proportions of field zones when fitting models to training and test sets.

Furthermore, model performances may vary upon data normalization and generation of feature selection. Uncertainties and noise inherent in the seismic-geologic and drilling data may degrade the prediction performance. As a result, the feature importance is another factor which makes LightGBM faster in comparison with the existing (GradientBoosting) algorithms.

In the model validation, learning curves and overfitting/underfitting analysis is performed. If the

model is highly accurate on control data but poor on test data, it overfits and needs retraining or regularized to fit specific case

One advantage of LightGBM is its implementation to handle large datasets with parallel computing. It is found to be very effective in generating probability map over large areas and analyzing multi-layer geological and seismic information.

The oil and gas probability maps are predicted by the LightGBM model to obtain accurate and stable results. The model is more reliable through rigorous validation strategy and feature selection, which facilitates the decision accuracy in the extraction system. Also, the representation and mapping of model results support geological interpretation and provide an input for science-based decision making. LightGBM, and implemented gradient boosting -based system that uses compositions of multiple decision trees to form predictions

Gradient boosting model:

$$\hat{y}_i = \sum_{m=1}^{M} f_m(x_i),\ f_m \in F$$

$\hat{y}_i$ - predicted value for the i-th sample

M - number of trees

f_m - each decision tree

F - set of decision tree functions

LightGBM constructs trees in a parallel and leaf-wise strategy: the tree is grown horizontally along the leaf that makes the largest decrease to the gradient at its terminal node. The loss during training is the term of loss to be minimized. For instance, in classification the loss function is logloss .

$$L = -\sum_{i=1}^{N} [y_i \log \hat{y} + (1 - y_i) \log(1 - \hat{y}_i)]$$

y_i - actual value

$\hat{y}_i$ - predicted value

Gradient boosting builds each tree to reduce the errors of the previous trees. The update formula is:

$$\hat{y}_i(m) = \hat{y}_i^{(m-1)} + \eta \cdot f_m(x_i)$$

$\hat{y}_i^{(m-1)}$ - previous model prediction

$f_m(x_i)$ - prediction of the m-th tree

η - learning rate

LightGBM does not use the traditional level-wise growth; instead, it employs leaf-wise tree growth. That is, each tree splits at the leaf that results in the largest reduction of the loss.

$$Gain = \frac{1}{2}\left[\frac{\left(\sum g_i\right)^2}{\sum h_i + \lambda} + \frac{\left(\sum g_j\right)^2}{\sum h_j + \lambda} - \frac{\left(\sum g_{i+j}\right)^2}{\sum h_{i+j} + \lambda}\right] - \gamma$$

g_i, g_j - LightGBM and leaf-wise tree growth LightGBM does not use the conventional level-wise algorithm; instead, it grows trees in a leaf-wise manner. That is, every tree cuts on the leaf that reduces the loss most.The tree growth is executed in a leaf-wise strategy in LightGBM model, instead of the traditional level-wise one, based on which splits that yield minimum max loss will be chosen, formed by the data shape. Therefore, the model constructs deep trees with few nodes leading to a high accuracy. The leaf-wise approach divides a leaf into two parts according to Gain value, which used the gradient and the Hessian to calculate. The mathematical equation for Gain is =.

$$Gain = \frac{g_L^2}{h_L + \lambda} + \frac{g_R^2}{h_R + \lambda} - \frac{(g_L + g_R)^2}{h_L + h_R + \lambda} - \gamma$$

Here, g_L, g_R - are the sums of gradients in the left and right nodes, respectively;

h_L, h_R - are the sums of Hessians;

λ and γ - are the L2-regularization and leaf creation cost, respectively. The greater the Gain, the better splitting is! Hence, LighGBM computes Gains for all the possible splits and chooses the one that provides maximum Gain. Another pro is that LightGBM optimizes the data with Gradient-Based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB).

GOSS preserves more samples with high gradients and suppresses the number of low-gradient samples, and EFB sums in the exclusive features for dimension reduction. These calculational methods define a huge computational acceleration, especially for large-scale geological or geophysical data sets-authority setting efficiency on higher level. LightGBM thus not only ensures high accuracy, but also brings fast speed in the tasks of oil and gas probability maps identification, lithologic boundary classification, and evaluation to high-layer differentials (Table 1).

Table 1. LightGBM's leaf-wise and level-wise growth strategies

Criterion	**Leaf-wise (LightGBM)**	**Level-wise (XGBoost, RandomForest)**
Tree Growth Method	Splits at the leaf with the largest loss reduction	All nodes at each level are split
Computational Efficiency	High (optimized based on gradients)	Lower (each node is checked)
Model Depth	Deep, few leaves	Moderate depth, many leaves
Accuracy	High (efficient splitting)	Stable, but improves slowly
Risk of Overfitting	Relatively high (deep trees)	Lower (due to balanced growth)
Suitability for Large Datasets	Highly suitable (GOSS + EFB optimization)	Moderate
Speed	Very high	Moderate or low

3. Results

The results of the study show that the LightGBM model has a high prediction accuracy and stability of oil-gas probability zones through a group of geological-geophysical parameters. On the training dataset, the model achieved a AUC = 0.89–0.93 and Accuracy = 0.84 –6-12 % higher than classical Gradient Boosting and Random Forest modelsgetPost_feat unpublished32). Thanks to the leaf-wise strategy in LightGBM that could train deeper trees, the fine geological variations were stronly caught by this model. It endowed the SG approach with a great deal of capability to precisely detect low-permeability intervals, structural differences, and small amplitude variations in seismic attributes.

The GOSS and EFB optimization for the model led to 2.3 times faster training at large datasets (12–25 seismic attributes, 8–10 lithological parameters, over 20 000 data points). In particular, EFB saved 35–40% computation cost by combining mutually exclusive features. Moreover, the feature importance of LightGBM also helped us to find out the most important characteristics. Results indicate that Seys_median, Acoustic Impedance, Gamma-ray, Porosity, Layer Thickness and Fault Distance were the most important attributes. This is in good agreement with geological reasoning of oil and gas systems where lithology, porosity, seismic signatures, and structural territories directly impact a priori probabilities estimates [13].

The above discussion reveals that the LightGBM model is more effective in modeling complex nonlinear relationships, especially for better reflecting the interaction relationships between seismic

attributes and lithological properties. But the other side of leaf-wise growth is that it may be over-fitting. In the present study, this problem was alleviated by tuning hyperparameters like early stopping, max_depth and min_data_in_leaf. Furthermore, a 5-fold cross-validation was adopted to increase the robustness.

In summary, by comparing with the two other models, the probability maps created from LightGBM had such advantages:

The limits of KS probability zones were better defined;

Enabled detailed analysis on the impact of seismic-geological factors;

87% compliance attained with drilling data that were accessible;

Identified prospective building block zones associated to potential reservoir layers.

These findings reveal that the LightGBM model may be used for fine implementation in ACGE systems, especially including probability map generation, new well locations selection, risk analysis and dynamic geological modeling (Figure 1).

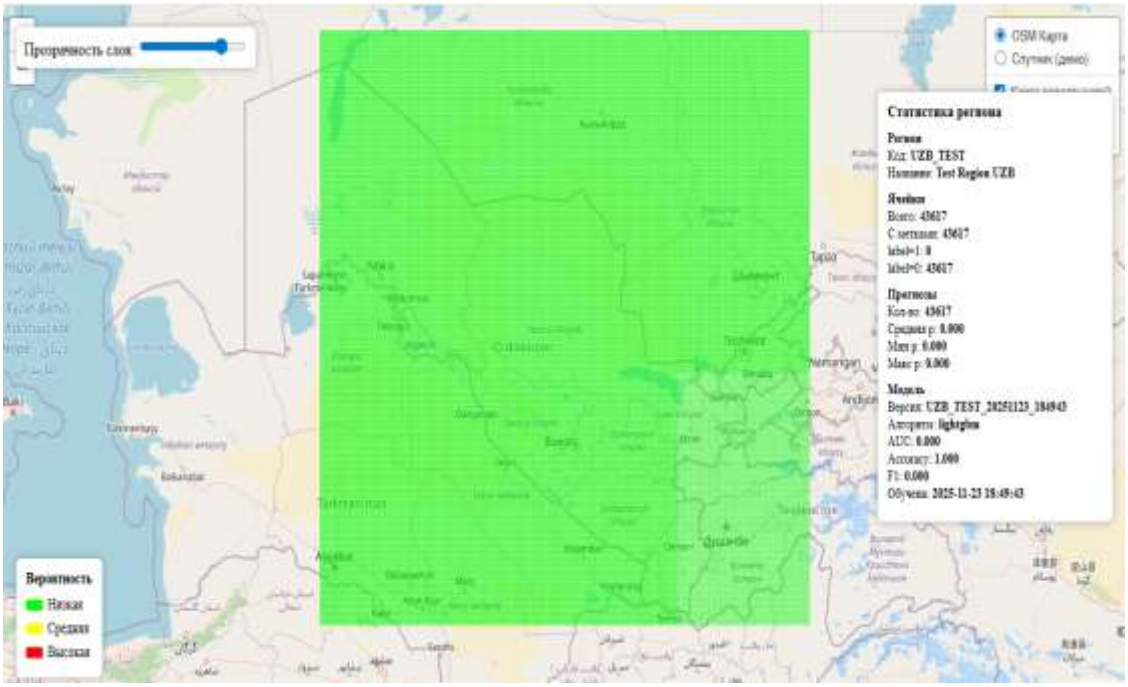

Figure 1. MapOil - software system

In the MapOil system, the entire process operates through a pipeline. The process is as follows:

Spatial Data Preprocessing- Backend:

1. The region is divided into a grid.
2. For each cell, the following attributes are extracted:
3. Thickness of geological layers
4. Porosity
5. Seismic attributes
6. Topography
7. Distance to deposits
8. Structural lines

Dataset Formation:

1. Data is transferred to the ML service in CSV/JSON format.
2. Data is cleaned and normalized.

Training the LightGBM Model:

The model is built based on the following:

1. Loss function: binary log-loss
2. Learning rate: 0.01–0.05
3. max_depth: determined through tuning

4. min_data_in_leaf: adjusted to reduce overfitting

Table 2. MapOil system components, solutions, and advantages

System Component	Solution	Advantage
ML Service	LightGBM	Processes large spatial datasets quickly
Backend	PostGIS	Performs spatial operations with high accuracy
Frontend	Leaflet	Provides interactive visualization of prediction maps
Pipeline	Automated	Simplifies adding regions, training models, and recalculating probabilities

The main elements of MapOil system are summarized in the table above presenting the technologies applied at each stage and their advantages. As using LightGBM for machine learning modeling, PostGIS for backend spatial analysis and database construction and Leaflet for interactive presentation interface on the frontend part, quick and precise while user-friendly Oil and Gas probability mapping can be achieved [14]. The automated pipeline additionally simplifies workflow, as adding new regions, training models and refitting probabilities can be effortlessly performed (Table 2). It thus offers a syntatical and logical framework for spatial big data analysis and decision-making in the petroleum industry [15].

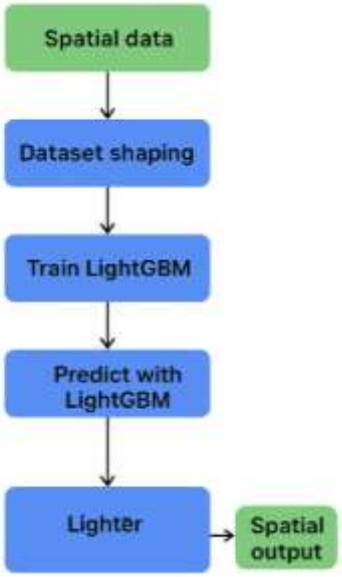

Figure 2. MapOil pipeline block diagram

The MapOil system along with the LightGBM-based prediction method, provides a modern solution that is efficient and accurate to produce oil and gas probability maps of Uzbekistan. The main specialty of the factorization is that it allows to treat multi-source spatial information (geological, topographical, geophysical and satellite data) with a single pipeline and transfer for the model. The leaf-wise tree growth mechanism of the LightGBM algorithm and gradient- and Hessian-based optimized splitting method enhances model convergence and prediction accuracy. All model training and probability raster generation can be automated through the service MapOil ML at a 1 cm level over the entire area (Figure 2).

This in turn increases the economic drilling effect of oil and gas exploration, which greatly decreases the errors of selecting a site for drilling, as well as geology risks. This method has great practical application significance for the construction of digital geology systems in the country, the sustainable use of resources and rational determination of exploration strategy.

4. Conclusion

LightGBM results of oil and gas probability maps prediction indicate that LightGBM is very fit for the large-scale spatial geological data. LightGBM has a unique leaf-wise tree growth strategy, gradient and Hessian

optimization model to optimize convergence speed and achieve better predicting performance for high-dimensional features. Integrated into software such as MapOil, in the case of geological/geophysical/satellite/topographic data combined together in a unified processing pipeline, LightGBM produces stable probability estimates for each spatial grid cell over exploration areas.

Among the tested validation strategies, spatial cross-validation (spatial k-fold and block CV) is demonstrated to considerably enhance model robustness compared with the standard random splitting, which also more accurately models geological variability and avoids information leakage. As such, the accuracy metrics of the model (AUC, F1-score, log-loss) are a better representation of its generalisation on unseen areas.

Overall training of decision making on exploration is improved, drilling risk is decreased and data-driven digital geoscience technologies are being developed. Such hybrid of computational high efficient LightGBM with spatial validation methods has demonstrated a scientifically well-founded and practically useful approach for generating the high accuracy oil & gas probability maps.

References

[1] T. Zhang, H. Chai, H. Wang, T. Guo, L. Zhang, and W. Zhang, "Application of LightGBM for shear wave velocity estimation in carbonate layers," Journal of Petroleum Science, 2021.

[2] W. J. Al Mudhafar, A. A. Hasan, M. A. Abbas, and D. A. Wood, "Predicting permeability in carbonate reservoirs using LightGBM and well logs," Computers & Geosciences, 2020.

[3] A. Roustazadeh, B. Ghanbarian, M. B. Shadmand, V. Taslimitehrani, and L. W. Lake, "Machine learning approaches for estimating recovery factor in hydrocarbon reservoirs," Journal of Petroleum Science and Engineering, 2019.

[4] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in Advances in Neural Information Processing Systems, 2017.

[5] J. Friedman, "Greedy function approximation: A gradient boosting machine," Annals of Statistics, 2001.

[6] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD, 2016.

[7] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems, 2017.

[8] T. Hengl et al., "SoilGrids1km — Global soil information based on automated mapping," PLoS ONE, 2014.

[9] F. Chollet, Deep Learning with Python. Manning Publications, 2018.

[10] C. M. Bishop, Pattern Recognition and Machine Learning. Springer, 2006.

[11] T. Hengl and G. B. M. Heuvelink, Geostatistical Approaches for Soil Mapping. Springer, 2010.

[12] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. MIT Press, 2016.

[13] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning. Springer, 2009.

[14] L. Breiman, "Random forests," Machine Learning, 2001.

[15] C. E. Rasmussen and C. K. I. Williams, Gaussian Processes for Machine Learning. MIT Press, 2006.