AI Oilers

Phase 1: Data Processing and Feature Engineering

Initial Data Processing

During the preprocessing stage, we analyzed the dataset to ensure consistency and correctness. The dataset was examined for missing values and duplicates, but none were found. The data was organized in ascending order with a depth interval of 1 ft. All values were of type `float64`, except for the `Date` column, which was transformed from an object format to `DateTime` for proper handling.

Feature Engineering

Feature engineering was performed to enhance data representation. Although it is typically done after data cleaning, we observed that `Depth(ft)` represents measured depth (MD). However, when analyzing temperature, it is crucial to account for true vertical depth (TVD), which was not explicitly available. To address this, we calculated a new feature, `TVD`, preserving the TVD/MD ratio to avoid multicollinearity, especially in vertical wells. Aslo this feature helps model to understand when wellbore started deviation from vertical profile.

Choose necessary features

Before starting analyzing data, we cut off features which are not affecting temperature value from physical point of view.

Exploratory Data Analysis (EDA) and Preprocessing

EDA Preprocessing

Based on expert knowledge, we selected key features for further analysis. Exploratory Data Analysis (EDA) was conducted using scatter plots, focusing on `Bottom Pipe Temp (°F)`. We analyzed three datasets from both training and testing data. The results revealed the need for data cleaning, as outliers significantly impacted predictions.

Data Cleaning and Outlier Treatment

To improve data quality, outlier replacement was performed using the last valid non-null value. Specific threshold-based rules were applied:

- **Bottom Pipe Temp (°F)**:
    - Above 150°F in the first 2000 ft treated as outlier
    - Above 200°F at depths exceeding 6000 ft treated as outlier
- **Annular Velocity**: Negative values were corrected
- **Bit Weight**: Values above 100 klb were flagged as anomalies
- **Co. Man G/L**: Converted into a categorical feature (0 for gains, 1 for losses)
- **Gain Loss**: Gains exceeding 200 bbl were replaced with 0

- **Mud Temp In/Out**:
    - Below 60°F or above 80°F within the first 100 ft considered outliers
- **Mud Gradient Count**: Temperature differences beyond 100°F or below 0°F flagged as outliers
- **Washout Factor**: Negative values treated as outliers
- **ROP Average**: Values exceeding 800 ft/hr flagged as anomalies
- **Differential Pressure**: Values below -600 psi treated as anomalies

EDA Post-Processing

After cleaning, we refined feature selection.

Following features are defined as an important:

`AD ROP SP (ft/hr); Annular Velocity (ft/min); Bit Size (in); Bit Time (hr); Circulating Hrs (hr); Flow In Rate (galUS/min); Mud Volume (bbl); Pump Pressure (psi); WC Bit Weight (klb); tvd`; `ML_mud_temp_grad`; `Mud_temp_grad`

Box plots verified the absence of residual outliers, and a correlation matrix ensured no multicollinearity issues.

ML model building

Before training, it is necessary to separate the target variable from the prepared files. As a result, four datasets with features and four datasets with target values are created, which will be used for model training and validation.

Training is conducted using cross-validation methods, allowing for an objective assessment of the model's ability to adapt to different datasets. Hyperparameter tuning is also performed for each model to avoid overfitting. Since the datasets have different scales, normalization must be applied to bring all variables, except for the target variable, to a common scale. The variable 'Co. Man G/L (bbl)' remains unchanged, as it was converted from a quantitative to a categorical variable.

To automate the process of configuring and selecting the model with optimal hyperparameters, an approach based on transformers and pipelines is utilized. Hyperparameters are tuned using GridSearchCV. The models considered include Random Forest, CatBoost, and LightGBM. According to the cross-validation results, CatBoost was found to be the best performing model. Subsequently, an evaluation of the results on the validation set, which contains the last 500 values of known temperature data, was conducted.

Validation Set Evaluation

During the analysis of the validation data, the models were evaluated both individually and in combination. The testing results are presented in the table below:

№	RF	LGBM	CB	LGBM_CB	RF_LGBM	RF_LGBM_CB
1	5.863437	11.072949	21.348707	15.833827	6.460767	10.880703
2	7.130240	7.892581	5.093314	4.135019	7.269895	4.652773
3	12.543456	10.269936	3.078135	5.952459	10.970682	7.980123
Mean	8.512378	9.745155	9.840052	8.640435	8.233781	7.837867

As a result, it was determined that the best solution is to construct an ensemble of three models for the following reasons:

- Minimal average prediction error on the validation data.
- Lowest variance of errors, making the model more robust. This should provide more stable results when predicting on the blind test.

Additionally, an analysis of feature importance for each model was conducted. It was found that each of the three models uses different key variables. This confirms that the models do not duplicate each other but rather complement one another, enhancing the overall ensemble.

Prediction and Interpolation

In the final stage, predictions need to be made on the blind data, and temperature values within specified intervals need to be determined with a step of approximately 0.5 feet. Before making predictions, the model is retrained using all available data. This is necessary to account for potential changes in patterns at new depths, ensuring maximum prediction accuracy on the blind data.

Approach to the Solution:

- Perform predictions at each point of the blind dataset.
- Interpolate data between neighboring values with a step of 0.001 feet.
- Record the prediction results in the final file.

The classical linear interpolation method was used for interpolation. As a result, three graphs (one for each dataset) were created, where:

- The Y-axis represents 'Bttm Pipe Temp (°F) - predicted',
- The X-axis represents 'Depth (ft)'.

The graphs compare data before and after interpolation. Since the graphs duplicate each other but have different frequency values on the X-axis, this confirms the correctness of the interpolation execution.

Phase 2: Automation and Adaptive Modeling

In the second phase, an automated approach, building on insights from the previous part, was implemented to identify relevant data. A thorough feature analysis was conducted to ensure physical consistency, which involved removing anomalies and replacing outliers based on the logic established in Phase 1. The cleaned dataset was then utilized for automatic fine-tuning of the machine learning model. This process adjusted

hyperparameters according to the new dataset features to develop a more effective model for a different well or area.

№	RF	LGBM	CB	LGBM_CB	RF_LGBM	RF_LGBM_CB
1	12.63	15.17	13.41	14.1	13.37	13.23
2	5.77	6.76	6.81	4.01	5.19	3.59
3	2.65	3.82	2.94	2.91	3.01	2.65
4	13.58	9.46	9.75	8.35	11.44	9.47
5	5.35	6.02	9.19	7.52	5.67	6.75
Mean	7.99	8.24	8.42	7.39	7.73	7.14

The tuning process involved systematically adjusting the input data to ensure the preselected model was well-suited for the new conditions it would encounter. This optimization aimed to enhance the model's predictive capabilities, allowing it to better capture the underlying patterns within the data. After fine-tuning, the final model's predictions were thoroughly analyzed. This analysis confirmed the model's effectiveness in generating reliable temperature estimates for previously unseen data, demonstrating its robustness and adaptability in different contexts. The results provided confidence in the model's ability to perform well in real-world applications, ensuring predictions in a variety of scenarios.