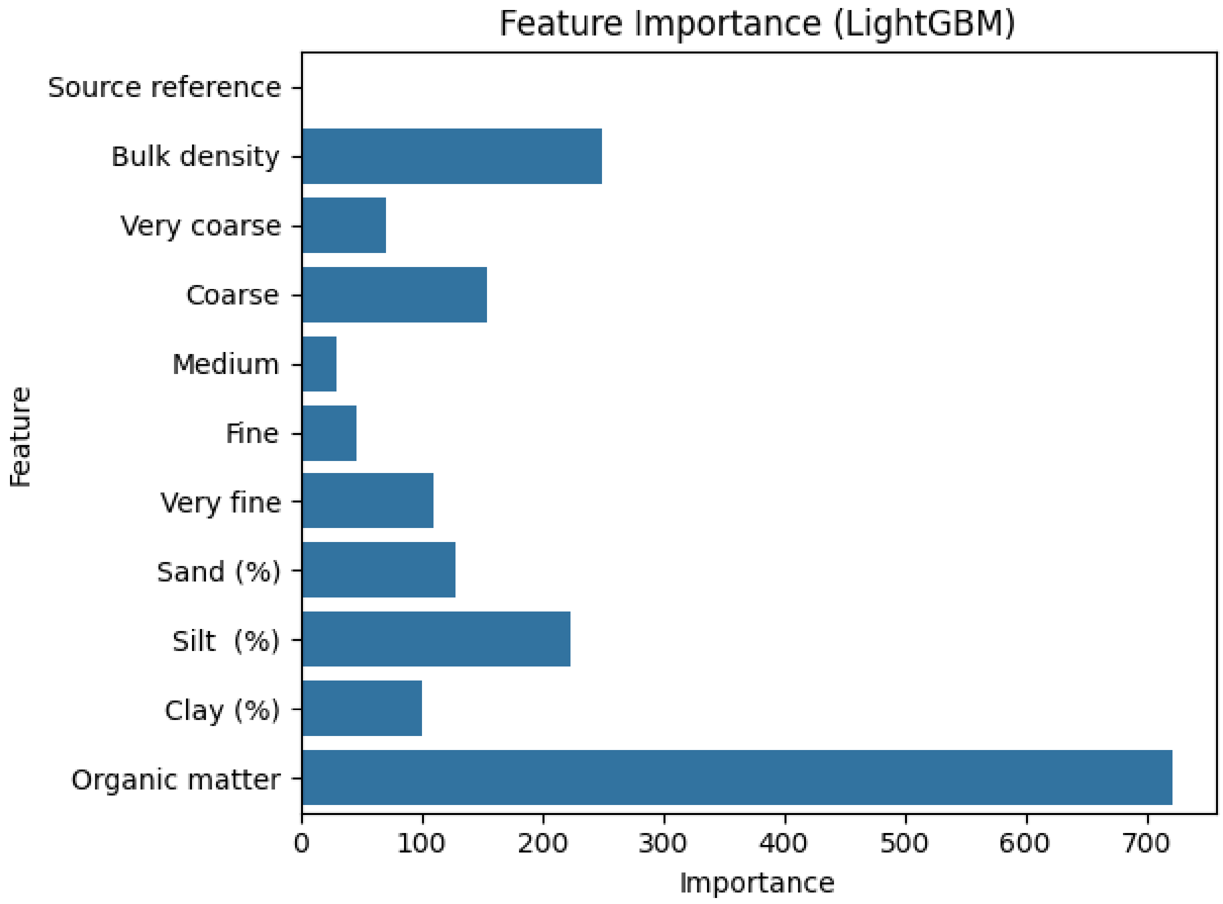
Model Challenge 1: ‘KSAT Quest: Regression Runoff’

Data preprocessing:

The original dataset consisted of 27433 rows and 33 columns. The Excel file included multiple sheets that each referenced a specific soil type. One of the sheets was called ‘All data’. This sheet combined all the reference sheets, but it was highly disorganized because the columns did not align correctly. To fix this, multiple sheets from 'data.xlsx' were parsed (excluding the “All data” and “References” sheets). Code was added to extract relevant features from each sheet to bring together a new dataset that aligned columns correctly (combined_df). Missing values were a big issue. Columns and rows that contained mostly missing values were removed. String values were cleaned by removing unwanted characters (e.g., commas, percentage signs) and converting to float.

Feature Selection:

The target variable was chosen to be Ksat because this environmental data has the information to investigate what features would predict Ksat (Soil hydraulic conductivity). These predictions would be meaningful because social conductivity is a critical factor governing the movement of water and dissolved substances. It is crucial for effective soil management and monitoring. Feature selection was done by removing metadata and features that had a lot of missing data. We then picked features that were relevant to the size of the grain in soil (important in determining Ksat) and features like organic matter because they are also relevant to Ksat (inorganic matter may poorly affect soil conductivity). Key features included: 'Bulk density', 'Very coarse', 'Coarse', 'Medium', 'Fine', 'Very fine', 'Sand (%)', 'Silt (%)', 'Clay (%)', and 'Organic matter'. However, organic matter eventually had to be dropped due to an excessive number of missing values.



Methods:
The LightGBM model was selected due to its efficiency and ability to work with large and messy data. Data was split 80/20. Key Parameters were:

- Objective: regression
- Metric: RMSE
- Learning rate: 0.05
- Number of leaves: 31
- Boost rounds: up to 1000 with early stopping (50 rounds)

Hyperparameters were set based on domain experience. Early stopping was employed to dynamically determine the optimal number of boosting rounds. Advanced tuning (e.g., grid search or Bayesian optimization) was not applied but is recommended for future work with more time.

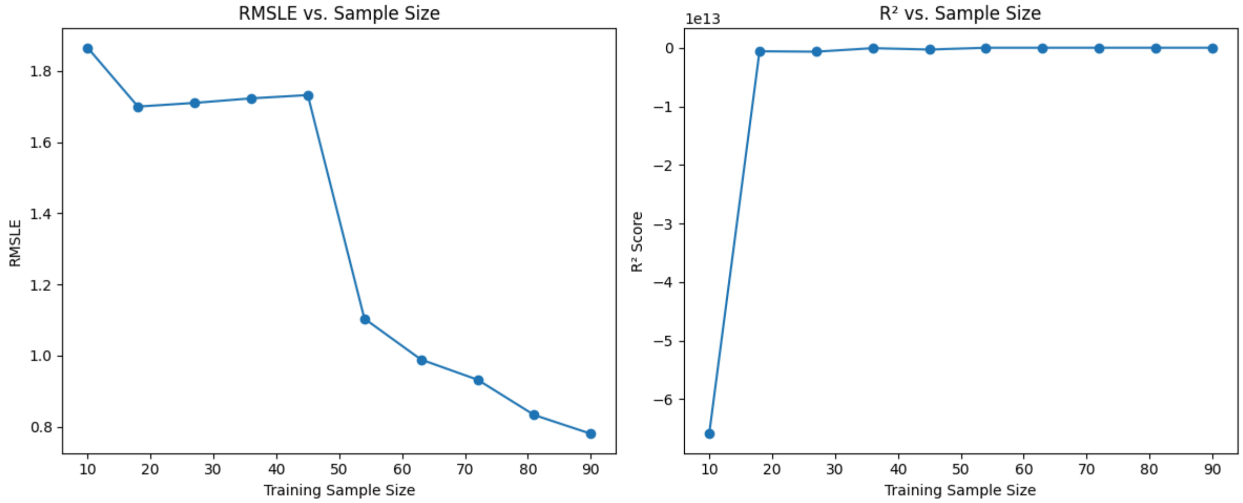
Data subsets were obtained from separate reference sheets (ex. Ref #1, Ref #3). Each sheet had slightly different available features and needed custom header parsing. These subsets were merged into a unified dataset to improve generalization and reduce site-specific overfitting. The subsets were less than 2,000 due to our cleaned dataset being smaller than 2,000 rows. After running the model the results were

R² Score: 0.9427800935362117

MAE: 2.0834327890727202

MSE: 8.062342276716313

These are very good results. R² explains a notable amount of variance in the data. MAE might be worth looking into deeper because it may not be ideal, considering the range of the Ksat numeric values. MSE is higher than MAE, hinting that outlier treatment may be useful in the future.



The results indicate a significant relationship between training data sample size and model performance. Specifically, at smaller sample sizes, the model exhibited higher Root Mean Squared Logarithmic Error (RMSLE) and lower R-squared values. This means reduced accuracy and explanatory power. As the sample size increased, both RMSLE and R-squared showed substantial improvement, reflecting enhanced model performance. However, the magnitude of these improvements diminished beyond a sample size of approximately 60. This implies that the model's performance gains diminish returns with increasing data volume beyond this threshold.

Results:

The LightGBM model demonstrated strong predictive performance, with an R² score of 0.94 showing that it captured a significant portion of the variance in Ksat values. While the MAE and MSE values suggest the model was generally accurate, the higher MSE compared to MAE signals potential sensitivity to outliers. Importantly, an analysis of model performance across different training sample sizes revealed that while larger datasets improved performance metrics (lower RMSLE and higher R²), the benefits of increasing sample size began to plateau around 60 observations. This suggests that even modestly sized, well-cleaned datasets can yield robust models, but improvements taper off as more data is added. Future work could benefit from more advanced hyperparameter tuning and targeted outlier treatment to further refine accuracy.