CS 6350-001 Fall 2023

Final Report on Prediction of Mineral Grades using Random Forest from Geological Drilling data.

(**Team Members**: Ishmael Anafo & Raviteja Tatikonda)

GitHub Link: <https://github.com/Raviteja-py/CS5350-6350-Machine_Learning_Fall23/tree/main/Final%20Project>

**Problem Definition:**

This project focused on predicting gold mineral grades from geological drilling data. Gold grade refers to the concentration of gold within the ore, and accurately determining this is essential for profitable and sustainable mining operations. However, predicting gold grades is exceptionally challenging due to the complex, nonlinear, and varied nature of geological data.

**Motivation:**

The motivation behind this project is twofold:

* *Economic and Environmental Impact:* Accurately predicting gold grades can significantly impact the profitability and environmental sustainability of mining operations. It allows for more efficient resource extraction and reduces unnecessary exploration, minimizing environmental disruption.
* *Machine Learning Application:* This problem presents a unique challenge due to the complex and varied nature of geological data. Machine learning offers tools and techniques capable of handling such complexity with nonlinear properties.

**Machine Learning Models Used:**

The team developed ensemble-based machine learning solutions for this mineral grade regression task. Ensemble methods combine multiple base models to produce superior predictive performance compared to single models (Chen et al., 2020). They are exceptionally well suited for unstable problems featuring small datasets and significant variance and incoherence like mineral forecasting.

* *Random Forest:* An ensemble method combining multiple de-correlated decision trees trained on random data subsets. It is robust to noise, avoids overfitting and handles nonlinear data well. The integrated predictions of diverse trees improve accuracy (Rani et al., 2022).
* *XGBoost:* A gradient boosting algorithm that incrementally builds an ensemble by focusing on previously mis-predicted instances. This adaptive approach improves prediction accuracy across a compounding collection of shallow decision trees.
* *AdaBoost:* Like XGBoost, AdaBoost is an adaptive boosting technique that trains models sequentially, concentrating more on previously mis-predicted data points. This enables improved accuracy across an ensemble of weak learners.
* *Bagging:* A bootstrap aggregation approach that trains base models like decision trees on random samples of the training data, aggregating their probabilistic predictions to improve stability and accuracy through consensus.

These algorithms provide integrated feature selection and modeling capabilities for handling mixed data types like the combination of categorical and continuous geology data. The feature set contained key parameters identified by domain specialists as influencing gold grades, including depth, spatial coordinates, geological angles and intersections, and mineralization indicators.

Appropriate hyperparameter tuning for factors like tree depth and leaf size was critical to optimize model generalization. The team avoided overfitting through cross-validation and partial model retraining approaches. The modelling pipeline, data splitting and evaluation workflows were coded in Python leveraging common SkLearn libraries.

**Data Description:**

The data is made up of drill data from a mine in Ghana that was collected for over a decade. It was originally structured into several tables, which included Collars, Geology, Survey, and Samples tables. This created the need to manually link these tables using MS SQL, MS Excel and Power BI. Below is the summary of the data used to train the model can be seen in the tables below:

|  |  |  |  |
| --- | --- | --- | --- |
| **Criteria** | **X Coordinate** | **Y Coordinate** | **Z Coordinate** |
| **Max Value** | 2048.76 | 7647.71 | 269.77 |
| **Min Value** | 1651.00 | 7250.91 | 160.45 |
| **Range in Meters** | 397.76 | 396.79 | 109.32 |
| **Pearson Correlation with Gold** | -0.012264 | 0.000115 | -0.013172 |
| **Spearman Rank Correlation with Gold** | -0.008998 | 0.001889 | -0.01341 |

Table 1: Data Description

The table below is the description of the target values, which is the gold grades. This is column is measured in grams per ton or parts per million (ppt).

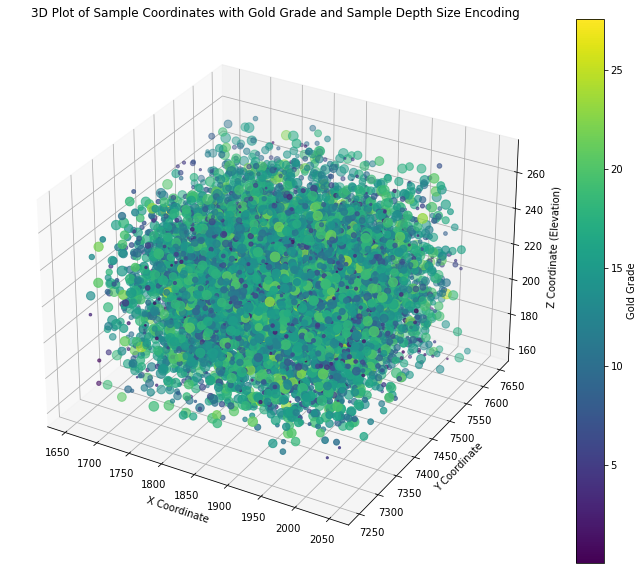
|  |  |
| --- | --- |
| **Criteria** | **Gold** |
| **Max Value** | 27.594205 |
| **Min Value** | 0.036032 |
| **Range in PPM** | 27.558173 |
| **Mean** | 12.782876 |
| **STD** | 4.713851 |
| **25%** | 9.445494 |
| **50%** | 13.137053 |
| **75%** | 16.31472 |

Table 2: Gold Data Summary

**Data Visualization and Interpretation:**

It can be seen from the graph below how incoherent the data is with low and high grades distributed evenly.

* Figure 1 suggests a relatively uniform distribution of samples across geographical space.
* The size of the points indicates that samples have been taken at varying depths, providing a comprehensive vertical profile of the geological site.
* The color gradient allows for the immediate visual identification of areas with higher gold concentrations, which appear to be sporadically distributed, without a clear pattern indicating a dependency on the location or depth.
* The lack of a distinct pattern in gold grade distribution relative to collar coordinates could imply a complex geological structure or varying mineralization processes affecting gold deposition.



**Fig. 1 Spatial Plot of Coordinates and Grade values**

**Ensemble Modeling:**

For the production-grade solution, the team chose Random Forest (RF) and XGBoost (XGB) algorithms to construct an ensemble Voting Regressor that combined their individual strengths. The integrated predictions of these robust and accurate tree-based ensembles outperformed any single model.

Additionally, the team built AdaBoost and Bagging ensembles with Decision Tree base learners to provide comparative benchmarks. Tuning and evaluation was conducted across all modeling approaches. While AdaBoost and Bagging showed respectable performance, the Random Forest and XGBoost ensemble delivered superior predictive accuracy on key metrics like R2 score. This validated the approach of blending two state-of-the-art algorithms to produce optimal real-world results.

The comparisons further demonstrated the power of advanced ensembles for handling complex geological data. But the final Voting Regressor integrating Random Forest and XGBoost represents an optimal solution tailored via rigorous experimentation.

**Model Development:**

This report presents drill data, survey data and assay data from a gold mining operation in Ghana. The data is collected from various sources and stored in the Geology, Survey, Collars, and Samples tables. A relational database is to be created from the above data to create one table that contains the Hole ID, X, Y, Z, Rock Type and Gold Grade columns. Once this table is created, its columns are then used in various machine-learning algorithms.

* *Data Collection:* The team obtained real-world geological drilling data in CSV format containing relevant features identified by domain experts that influence gold mineral grades, including spatial coordinates, depth, dip, azimuth, and geological intersections.
* *Data Exploration:* Performed initial investigation of the dataset using pandas, matplotlib and seaborn to understand the distribution of features, correlations, and relationships within the data.
* *Feature Engineering:* Key inputs like drill hole angles, mineralization indicators and gold grade target variable were prepped for modeling based on geological knowledge.
* *Hyperparameter Tuning:* Extensive hyperparameter tuning using randomized search cross-validation was implemented to optimize model generalization capability.

1. For the Random Forest ensemble, key parameters tuned included:

n\_estimators: Number of decision trees (values: 100, 500, 1000)

max\_depth: Maximum tree depth (values: 5, 10, 20)

min\_samples\_split: Minimum samples at a leaf node (values: 5, 10)

max\_features: Number of features considered per split (values: auto, log2)

1. For the XGBoost model, important hyperparameters optimized:

n\_estimators: Number of gradient boosted trees (values: 100, 500, 1000)

max\_depth: Maximum tree depth (values: 3, 5, 7)

learning\_rate: Shrinkage rate (values: 0.01, 0.05, 0.1)

subsample: Subsample ratio of training instances (values: 0.7, 0.8, 0.9)

The randomized search method performed 10 iterations evaluating 5 cross-validation splits per iteration for a broad search with statistical rigor to find the optimal parameter combinations.

**General Comments**

* The best performing RF and XGB models identified via this extensive tuning process were combined in the ensemble for further performance gains.
* Standardization scaling using StandardScaler preprocesses data and improves training.
* train/validation/test splits created for robust evaluation.
* Gridsearch based hyperparameter tuning implemented for Random Forest and XGBoost using cross-validation.
* An ensemble Voting Regressor combining the best RF and XGB models is constructed and evaluated on the test set.

**Experimental Results:**

Performance Metrics

* R2 Score: Measures the proportion of variance in the dependent variable that is predictable from the independent variables.
* MSE (Mean Squared Error): Indicates the average of the squares of the errors.
* RMSE (Root Mean Squared Error): Shows the square root of MSE.
* MAE (Mean Absolute Error): Represents the average of the absolute errors.

**Results**

Below is the summary of the performance of the models.

|  |  |  |
| --- | --- | --- |
| Model | R2 Score | Mean Square Error |
| Random Forest Ensemble Model | 0.6969 | 6.4295 |
| Adaboost | 0.6869 | 6.6410 |
| Bagging | 0.6517 | 7.3886 |

Cross-Validation Scores were [0.70170965, 0.73589627, 0.73868408, 0.70542278, 0.70432414]

**Conclusion:**

The models all worked well with the ensemble model being the best. They all had an over 30% error and this is understandable since mineral deposition and their related grades vary in ways that cannot be perfectly modelled. This performance are high enough to give a higher than chance prediction of grades within the search area. This could lead to significant savings since zones that record zero grades can be avoided during drilling. This will lead to selective drilling in sites with high grades as confirmation. The cost savings can be high in the process.

**Potential Extensions:**

Further research directions that could be pursued include:

*Real-world operational data testing:* Applying models using actual geological survey data from active mining sites to evaluate performance in ongoing production scenarios for potential operationalization.

*Advanced deep learning models:* Exploring more sophisticated neural network architectures tailored for spatial data like CNNs or graph neural networks to capture additional data relationships.

*Expanded feature engineering:* Creating and evaluating derived features in conjunction with geology experts that better expose orebody indicators to improve model accuracy.

*Knowledge integration:* Collaborating with veteran geology scientists and engineers to incorporate physics-based first principles and domain heuristics into a hybrid ML approach.

*Geo-statistical techniques:* Incorporating established geo-stat methods like Kriging that can factor in spatial correlations. Integrating these with ML predictions may yield further improvements.

**REFERENCES**

Chen, CH., Tanaka, K., Kotera, M. et al. Comparison and improvement of the predictability and interpretability with ensemble learning models in QSPR applications. J Cheminform 12, 19 (2020).

Rani D, Gill NS, Gulia P, Chatterjee JM. An Ensemble-Based Multiclass Classifier for Intrusion Detection Using Internet of Things. Comput Intell Neurosci. 2022 May 20;2022:1668676. doi: 10.1155/2022/1668676. PMID: 35634069; PMCID: PMC9142322.