Predictive Modeling for Proximity Classification in the Broken Down Lead Deposit

Prepared by: Mohd Zuhair

Date: 03-Feb-2025

1. Executive Summary

The Broken Down lead deposit in Tasmania presents a significant opportunity to improve exploration efficiency by classifying geochemical samples into zones closer to or farther from the orebody. By identifying these zones, drilling efforts can be prioritized, costs reduced, and resource estimation improved.

A predictive model using geochemical assay data (8 key elements: As, Au, Zn, Fe, S, Cu, Mo, Pb) was developed with **XGBoost**, a leading machine learning algorithm. Key takeaways include:

- **Strong Performance**: Achieved 87.0% accuracy on the test set, with robust classification of samples closer to the orebody (Class A).
- **Key Predictors**: Lead (Pb) and Arsenic (As) emerged as the most influential factors for proximity classification.
- **Practical Impact**: Enables exploration teams to focus on high-priority areas, leading to faster and more cost-effective decision-making.

While the model performs well for proximal zones, refinement is needed for distal samples.

2. Problem Statement

The objective is to classify geochemical samples into proximal (Class A) and distal (Class B) zones based on their chemical composition. Addressing this challenge helps to:

- Identify zones closer to the orebody, optimizing drilling campaigns.
- Provide a systematic, data-driven complement to geological methods.
- Enable reliable predictions for 767 unlabeled samples, enhancing the understanding of the resource area.

3. Data Overview

The dataset comprises **4,472 geochemical samples**, measured across 8 elements (As, Au, Zn, Fe, S, Cu, Mo, Pb). Summary:

- **Labeled Data**: 4,004 samples (Class A: 2,861, Class B: 1,143).
- **Unlabeled Data**: 767 samples requiring classification.
- **Metadata**: Drill hole metadata (uniqueID, holeid, from, to) was excluded from modeling per the problem scope.

3.1 QAQC Steps

- **Missing Data**: Replaced sentinel values (-999) with `NaN` and imputed missing values using feature means.
- **Standardization**: Scaled features to ensure consistent ranges.
- **Duplicates**: Checked for duplicate `unique_id` entries (none found).
- **Class Distribution**: Verified imbalance (2.5:1 ratio between Class A and B).

4. Approach & Methodology

XGBoost was chosen for its ability to handle complex data relationships, missing values, and imbalanced classes. Key steps:

1. **Addressing Missing Data**: Imputed missing values with feature averages.
2. **Handling Class Imbalance**: Used `scale_pos_weight` to account for the 2.5:1 class ratio.
3. **Training and Validation**: Split data into training (80%) and test (20%) sets with stratified sampling.
4. **Optimization**: Tuned hyperparameters using GridSearchCV.

5. Model Performance & Insights

5.1 Key Metrics

- **Accuracy**: 87.0% on the test set.
- **Class A (Proximal)**: Precision = 88.9%, Recall = 93.5%.
- **Class B (Distal)**: Precision = 81.4%, Recall = 70.7%.
- **ROC-AUC**: Test set AUC = 0.928.

5.2 Feature Importance

The model identified key predictors:

- **Lead (Pb)**: Highest contribution (0.40).
- **Arsenic (As)**: Second-most influential (0.25).
- **Molybdenum (Mo)**: Moderate impact (0.15).

5.3 Limitations

- **Class Imbalance**: 2.5:1 ratio reduced recall for Class B.
- **Feature Skewness**: Pb and As exhibited skewed distributions.
- **Overfitting Risk**: Train vs. test AUC gap (0.991 vs. 0.928).
- **Unlabeled Data**: Predictions assume similar distributions to labeled data.

6. Recommendations

6.1 Priority Enhancements

1. **Improve Class B Predictions**:
   - Apply SMOTE/ADASYN for class balancing.
   - Collect additional Class B samples.
2. **Feature Engineering**:
   - Log-transform skewed features (Pb, As).
   - Collaborate with geologists to derive domain-specific features.
3. **Model Generalization**:
   - Test simpler models (e.g., Random Forest).
   - Add L1/L2 regularization.

6.2 Next Steps

- **Operational Deployment**: Integrate the model into exploration workflows.
- **Validation**: Partner with geologists to assess unlabeled predictions.
- **Continuous Improvement**: Retrain with new data and feedback.

7. Conclusion

The model successfully classifies geochemical samples into proximal/distal zones, achieving 87.0% accuracy. Pb and As were identified as the strongest predictors, though geological validation is recommended. Addressing class imbalance and feature engineering will further enhance performance.

Call to Action:

- Review predictions for unlabeled samples to prioritize drilling.
- Validate model insights with geological expertise.
- Integrate the model into exploration workflows.

8. Visualizations

8.1 Feature Distributions

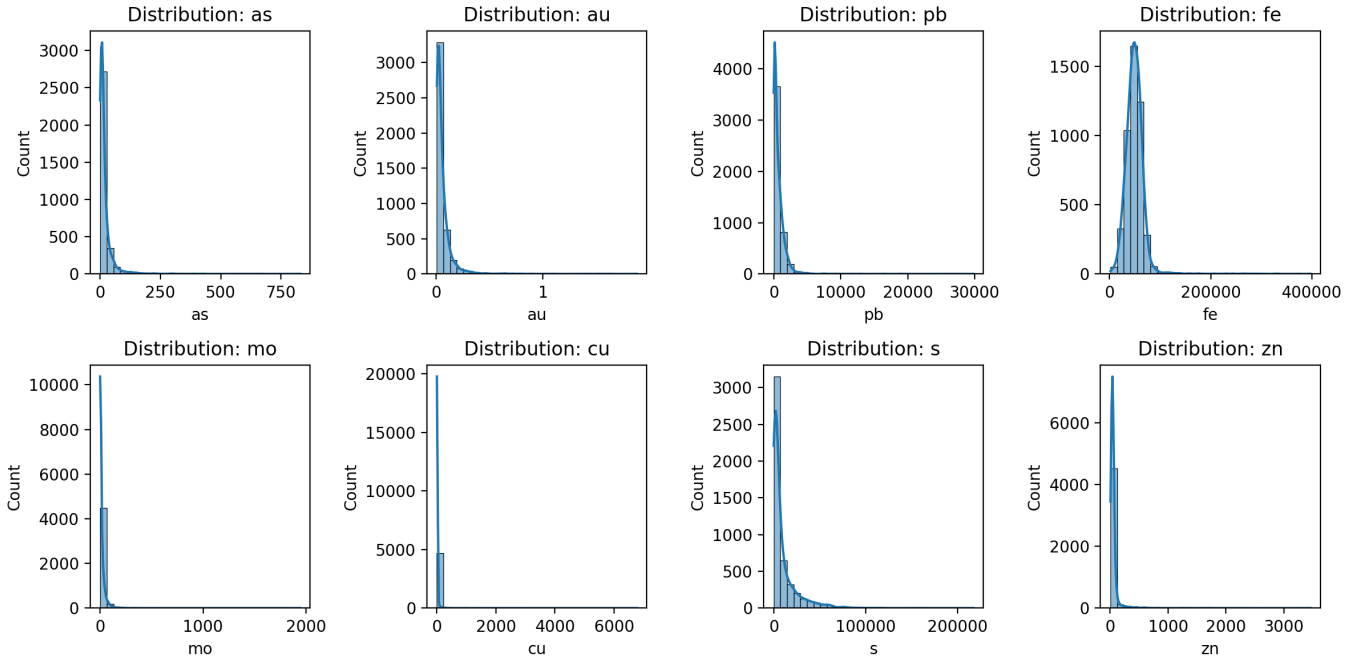

Figure 1: Distributions of Geochemical Features

8.2 Feature Importance

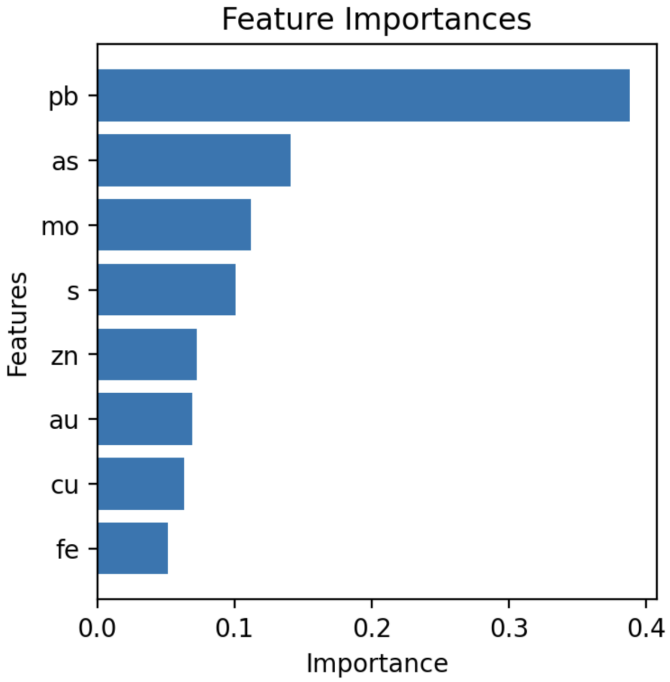

Figure 2: Feature Importance (XGBoost)

8.3 Class Imbalance

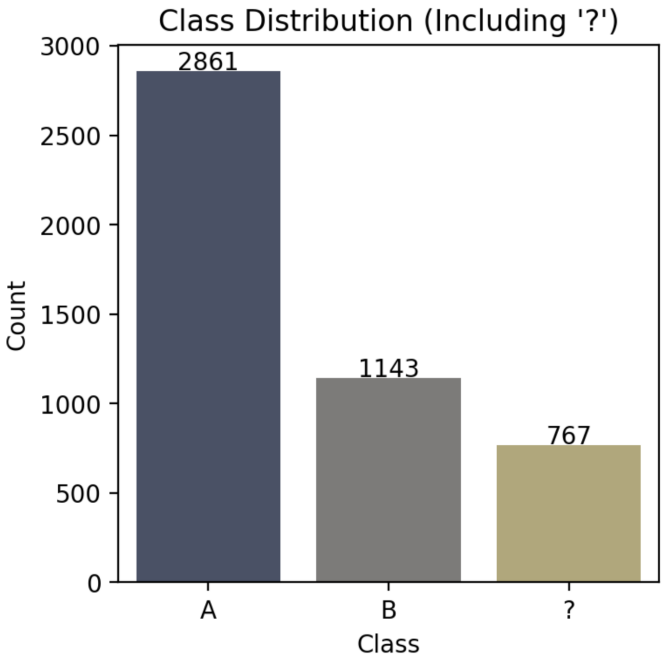

Figure 3: Class Distribution (Including Unlabeled Samples)

8.4 ROC Curve

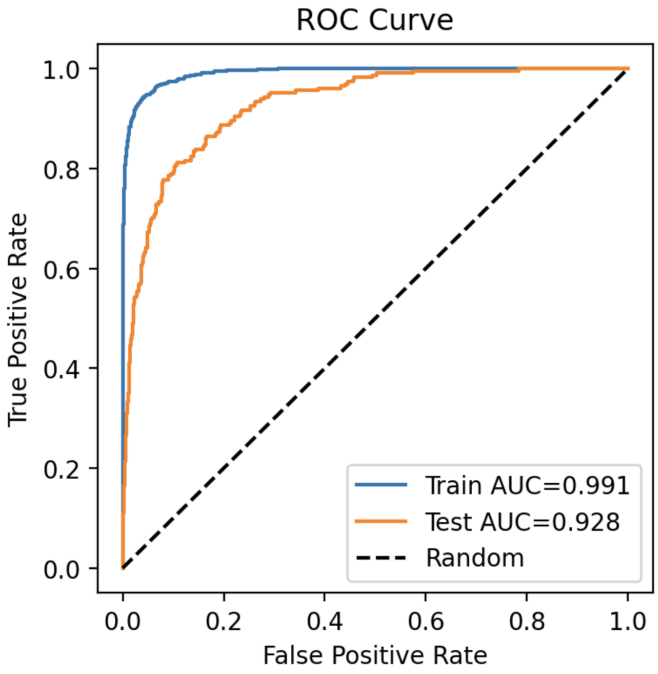

Figure 4: ROC Curve: Train and Test Sets

Contact Information

- Name: Mohd Zuhair
- Email: zuhair.alig31@gmail.com