KGI 04 CASE 3 MACHINE LEARNING CANVAS

Designed for: Microsoft KGI Hackathon Designed by: KGI 04 Date: 26/04/2025 Iteration: 8

PREDICTION TASK

What is the type of task?
Which entity are predictions made on?
What are the possible outcomes to predict?
When are outcomes observed?

- Regression task to predict the percentage of silica in the final product of an iron ore flotation process.

- Process measurements from iron ore flotation operations at specific timestamps.

- Continuous values representing silica percentage in the concentrate

- After the flotation process is completed and quality testing is performed on the final product

DECISIONS

How are predictions turned into actionable recommendations or decisions for the end-user? (Mention parameters of the process / application for this.)

- Process engineers can make real-time adjustments to operational parameters (reagent flow rates, pH levels, air flow) when the model predicts silica content will exceed acceptable thresholds.

- Starch flow (depressant), Amina flow (collector), ore pulp pH, ore pulp density, and air flow rates in flotation columns can all be adjusted to optimize the process.

VALUE PROPOSITION

Who is the end beneficiary, and what specific pain points are addressed?
How will the ML solution integrate with their workflow, and through which user interfaces?

- Mining companies and process engineers benefit by gaining the ability to proactively maintain product quality instead of reactive quality control. This addresses critical challenges including: inconsistent product quality, excessive reagent consumption, production of off-spec material requiring rework, and missed quality targets leading to customer penalties.
- The model integrates with existing process control systems, providing real-time silica predictions based on current operational parameters. It enhances decision-making by offering predictive insights that complement traditional process monitoring, enabling operators to make proactive adjustments before quality issues occur.

- Control room dashboards with real-time prediction displays, alert systems for predicted quality deviations, mobile applications for engineers on the move, and integration with existing SCADA systems to overlay predictions with current process values

DATA COLLECTION

How is the initial set of entities and outcomes sourced (e.g., database extracts, API pulls, manual labeling)?
What strategies are in place to update data continuously while controlling cost and maintaining freshness?

- Historical operational data from plant control systems including process parameters and laboratory analysis of silica content in the final product.

- Automated data pipeline from process control systems and laboratory information management systems (LIMS); regular model retraining on new batches of production data; anomaly detection to flag unusual process conditions.

DATA SOURCES

Where can we get data on entities and observed outcomes? (Mention internal and external database tables or API methods.)

- Plant SCADA/DCS systems for real-time process parameters; laboratory analysis for silica content measurements; historical databases of plant operation; quality control databases.

- Direct integration with plant control systems through APIs; automated extraction from laboratory information management systems; batch uploads of historical data for initial training.

IMPACT SIMULATION

What are the cost/gain values for (in)correct decisions?
Which data is used to simulate pre-deployment impact?
What are the criteria for deployment?
Are there fairness constraints?

- Correct predictions enable process optimization that reduces reagent costs while maintaining quality specifications; incorrect predictions could lead to off-spec product (requiring rework at ~$1000/ton) or excessive reagent usage.

- Test dataset (20% split) used to validate model performance; historical process interventions analyzed to estimate potential savings.

- RMSE below 0.5, R^2 above 0.80, and MAE below 0.4 with consistent performance across different operational conditions.

- Model should perform well across different ore types, operating shifts, and seasonal variations.

MAKING PREDICTIONS

Are predictions made in batch or in real time?
How frequently?
How much time is available for this (including featurization and decisions)?
Which computational resources are used?

- Real-time predictions for process control; batch predictions for quality forecasting and shift planning.
- Continuous real-time predictions (every few minutes) to match process control cycles; batch predictions at the start of each shift for planning.
- Predictions must be available within seconds to support timely process control decisions.
- Lightweight inference pipeline (scaler + trained model) deployed on edge devices or integrated with control systems.

BUILDING MODELS

How many models are needed in production?
When should they be updated?
How much time is available for this (including featurization and analysis)?
Which computation resources are used?

- One optimized model after comparing Linear Regression, Random Forest, XGBoost, and Gradient Boosting
- Monthly or when process conditions significantly change
- Initial model training requires several hours for feature engineering, cross-validation, and hyperparameter optimization with Optuna; subsequent updates can be performed overnight.
- Parallel processing (-1 jobs) for cross-validation and hyperparameter tuning; moderate computing resources for model training.

FEATURES

What representations are used for entities at prediction time?
What aggregations or transformations are applied to raw data sources?

- Raw process variables (iron and silica feed content, reagent flows, pH, density, air flows) plus engineered features including temporal information (year, month, day, hour), material ratios (Iron-to-Silica feed ratio, Starch-to-Amina ratio), process ratios (Air Flow Ratio), and lagged variables (Iron_Feed_Lag1, Silica_Feed_Lag1, pH_Lag1).

- Decimal normalization (comma to period conversion); robust scaling to handle varying units and magnitudes; outlier handling using IQR method with 3*IQR threshold; forward and backward filling for missing values caused by lags.

MONITORING

Which metrics and KPIs are used to track the ML solution's impact once deployed, both for end-users and for the business?
How often should they be reviewed?

- Technical metrics: prediction accuracy (RMSE, R^2, MAE), prediction vs. actual silica content; Process KPIs: percentage of product meeting specifications, reagent consumption per ton, production rate; Business KPIs: cost savings from reduced reagent usage, reduced rework costs, increased customer satisfaction

- Technical metrics: daily monitoring; Process KPIs: weekly review; Business impact: monthly evaluation.