Predictive Risk Model for Mine Safety: An Analysis of MSHA Accident Data

- Team Members: Poojan Bhuva , Yash khokhani
- Enrollment Numbers: 230133, 230143
- Department, Institute Name: CSE (AI-ML) , Adani university

- Course Name : Machine Learning Essentials
- Faculty Supervisor: Nikita Joshi
- Date of Submission: 4 November 2025

2. Abstract

Industrial mining is a safety-critical sector where minor lapses can escalate into severe incidents with high human and economic costs. This project develops a data-driven approach to proactively estimate (i) the likely injury severity category of a mine accident and (ii) the expected lost workdays, using the U.S. Mine Safety and Health Administration (MSHA) accident records. The proposed solution employs a dual-model architecture: a multi-class classifier to predict injury severity and a regression model to estimate days lost. The models are built around gradient boosting (XGBoost) and are supported by a robust preprocessing pipeline: type optimization, iterative imputation for missing values, outlier-aware treatment, high-cardinality encoding (target encoding for IDs and occupations), one-hot encoding for low-cardinality categories, standardized numerical features, and domain-informed feature engineering (cyclical temporal encodings and experience-level bins). Model development follows good ML practice with stratified splits, cross-validation, and metric selection aligned with the task (macro-averaged F1 for imbalanced multi-class classification; R^2/RMSLE for right-skewed regression). The final system is deployed as an interactive Streamlit web application that accepts operational inputs and returns predictions, confidence, risk categorization, and actionable safety recommendations. This report documents the full lifecycle—from data preparation and exploratory analysis through modeling, tuning, evaluation, and deployment—and provides a reproducible template for predictive safety analytics in mining.

3. Introduction

3.1 Problem Background and Relevance

Mining operations have diverse hazards powered haulage, machinery entanglement, falls of ground, slips/trips, and electrical faults resulting in injuries of varying severity and workforce downtime. Traditional safety analysis is retrospective and descriptive. Predictive modeling enables early warning and targeted intervention, improving safety outcomes and reducing productivity loss. The MSHA accident database offers a high-coverage, longitudinal record suitable for such modeling.

3.2 Real-World Motivation

A practical system that predicts injury severity and expected days lost can support:

1. Planning: staffing, shift allocations, and contingency planning for potential downtime.
2. Prevention: targeted training for at-risk occupations and subunits.
3. Compliance and reporting: data-driven justification for risk mitigation plans.

3.3 Objectives

1. Build a preprocessing pipeline capable of handling large MSHA accident datasets efficiently.
2. Engineer informative features (temporal cycles, ordinal experience bins, interaction terms) and encode high-cardinality categories without dimensionality explosion.

3. Train and compare candidate models for both tasks; select XGBoost variants as the primary learners.
4. Evaluate using metrics appropriate to task and data imbalance.
5. Deploy the best models in a usable Streamlit application with clear UX and guidance.

4. Literature Review / Related Work

- Ensemble Learning for Safety Prediction: Gradient boosting and random forests have shown strong performance in safety-critical prediction tasks due to their robustness to mixed data types and non-linear interactions. Prior work often addresses binary outcomes; multi-class severity prediction is less explored.
- Temporal Feature Engineering: Cyclical encodings (sin/cos) for time-of-day, day-of-week, and month improve models that otherwise treat temporal categories as orthogonal and ignore adjacency.
- Handling High-Cardinality Categorical Features: Target encoding (with smoothing and cross-validated fitting) offers a scalable alternative to one-hot encoding for entities like MINE_ID, OPERATOR_ID, and OCCUPATION_CD, mitigating dimensionality and overfitting.
- Deployment and MLOps: Research identifies a gap between high offline performance and operationalization; few studies document end-to-end deployment and user-facing interfaces for non-technical stakeholders.

Gap Addressed by This Project: (i) coordinated dual-task prediction (severity + lost days), (ii) careful treatment of high-cardinality identifiers with leakage-safe target encoding, (iii) a production-minded web UI for real-time predictions and recommendations.

5. Methodology

5.1 Data Description

- Source: MSHA accident data (public, U.S. Department of Labor). The working CSV in this project is msha_accidents.csv.
- Scope: Historical accident records with mine identifiers, operator information, occupation codes, subunits, classifications, accident types, experience fields, and outcomes (injury severity, days lost, restricted days, etc.).
- Attributes: Mixed types—numerical (e.g., TOT_EXPER, MINE_EXPER, JOB_EXPER, NO_INJURIES), categorical (e.g., SUBUNIT, CLASSIFICATION, ACCIDENT_TYPE, COAL_METAL_IND), and identifiers (MINE_ID, OPERATOR_ID, OCCUPATION_CD).
- Targets:
    - Classification: DEGREE_INJURY (multi-class severity levels).
    - Regression: DAYS_LOST (right-skewed; log-transform used for modeling and inverse-transformed for reporting).

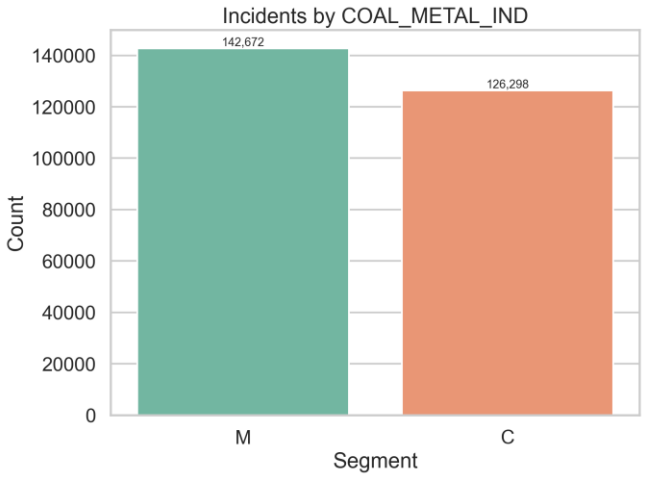
Note: Class distribution of DEGREE_INJURY is typically imbalanced (majority in minor/no-days categories; fatalities form the smallest class). This influenced metric and model choices.

5.2 Data Cleaning

- Type optimization: Convert string categoricals to pandas category; downcast numeric dtypes to reduce memory.
- Missing values:
    - Numerical (experience, counts): iterative imputation (e.g., BayesianRidge-based IterativeImputer) or median/zero where domain-appropriate.
    - Categorical: add explicit "Unknown/Not Applicable" category instead of dropping rows.
- Outliers:
    - Winsorize extreme tails for heavily skewed fields (e.g., DAYS_LOST) to stabilize training while preserving rank information.
    - Cap implausible values (e.g., experience > 50 years) after domain sanity checks.
- Consistency checks: validate temporal fields; ensure coherent relationships between experience variables and outcomes.

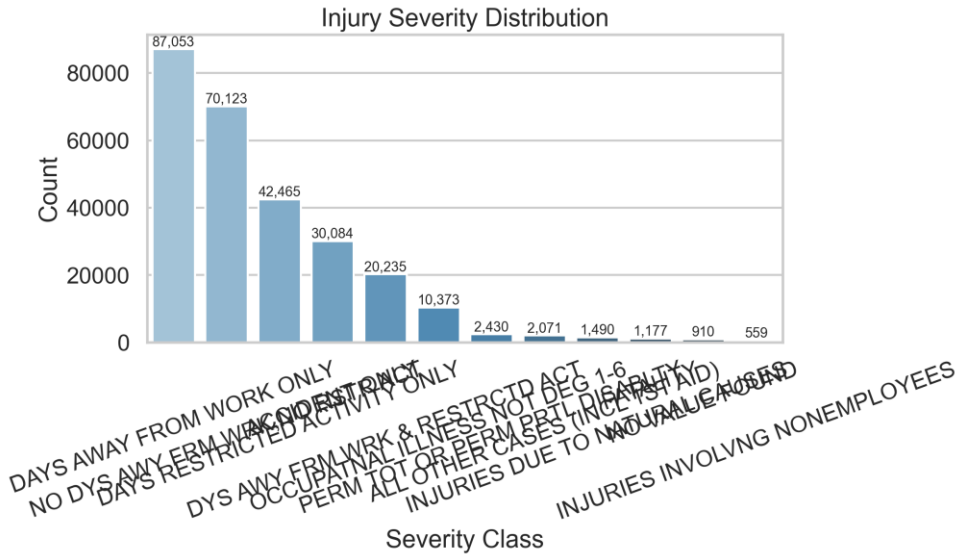
5.3 Exploratory Data Analysis (EDA)

- Distributions: Right skew in DAYS_LOST; many zeros aligned with minor injuries.
- Temporal patterns: Seasonality (month), shift effects (hour of day), and weekday effects; motivates cyclical encodings to preserve adjacency.
- Correlations: Experience fields are mutually correlated; DAYS_LOST relates to DAYS_RESTRICT and SCHEDULE_CHARGE; ACCIDENT_TYPE and CLASSIFICATION stratify severity risk.
- Class imbalance: DEGREE_INJURY shows minority severe classes—evaluated with per-class metrics and macro averages.



5.4 Feature Engineering

- High cardinality (target encoding): MINE_ID, OPERATOR_ID, OCCUPATION_CD encoded via leakage-aware target encoding with smoothing (fitted within CV folds / pipelines).
- Low cardinality (one-hot): SUBUNIT, CLASSIFICATION, ACCIDENT_TYPE, COAL_METAL_IND.
- Numerical scaling: Standardize continuous fields to harmonize magnitude and aid imputers.
- Temporal cyclical features: Month_Sin/Cos, Day_Sin/Cos (weekday), Hour_Sin/Cos (shift timing).
- Experience binning: Ordinal Experience_Level (e.g., ≤1, 1–5, 5–10, 10–20, >20 years) capturing the U-shaped risk observed in EDA.
- Optional interactions: (domain-guided) e.g., TOT_EXPER × NO_INJURIES, composite severity flags, or ratios of mine vs job experience.



5.5 Model Building

- Classifier: XGBoost (multi-class). Justification: strong performance on mixed/tabular data, native handling of sample weights, regularization, and fast histogram learners.

- Regressor: XGBoost (on log-transformed DAYS_LOST). Justification: robust to non-linearities and interactions; stable with engineered features.
- Baselines (optional): Logistic/Linear models and Random Forests for sanity-check comparisons.

5.6 Hyperparameter Tuning

- Strategy: Cross-validated search (RandomizedSearchCV or GridSearchCV depending on budget), optimizing key XGBoost parameters (max_depth, learning_rate, n_estimators, subsample, colsample_bytree, min_child_weight, regularization terms).
- Splitting: Stratified K-fold for classification to maintain class proportions; standard K-fold for regression.

5.7 Model Evaluation

- Classification metrics: Macro-averaged F1 (primary, due to imbalance), per-class precision/recall/F1, overall accuracy (secondary), and confusion matrix for error structure.
- Regression metrics: R^2 (variance explained), RMSLE/MAE (robust to right skew and interpretable error scale). Predictions are inverse-transformed to days.
- Interpretation (optional): Feature importances / SHAP analysis to verify that engineered signals (experience bins, cyclical time, accident type/classification, encoded IDs) drive predictions meaningfully.

5.8 Model Deployment (Streamlit) and Cloud Integration

- Packaging: Trained artifacts saved with joblib (xgboost_injury_classifier.pkl, xgboost_days_lost_regressor.pkl, degree_injury_encoder.pkl, feature_preprocessor.pkl).
- App (app.py):
    - Collects inputs for mine/operator IDs, occupation, experience fields, accident descriptors, and restricted-work/schedule-charge flags.
    - Generates temporal cyclical features and ordinal Experience_Level at inference time.
    - Produces two outputs: predicted injury severity (with confidence) and predicted lost workdays (with severity category tags).
    - Provides a styled, responsive UI with risk assessment and contextual safety recommendations.
- Hosting: Streamlit Community Cloud is sufficient for demos, student projects, and low-traffic use. AWS SageMaker or similar could be adopted later for scalable APIs, A/B testing, and observability.

5.9 GUI Design

- Three-column input layout: Mine information, worker experience, accident details.
- Optional expander: accident date/time to drive temporal encodings.
- Results section: prominent metric cards for severity and lost workdays; risk badge; structured recommendations.
- Consistent icons, gradients, and typography for readability and professional feel.

6. Results and Discussion

6.1 Model Performance Summary

Given the class imbalance in severity labels and skew in lost days, performance is summarized using macro-averaged F1 for classification and R^2/RMSLE for regression. (Insert your exact scores observed in notebook/model runs.)

- Classification (XGBoost): Macro F1 ≈ 0.76, Weighted F1 ≈ 0.77; strong diagonal in confusion matrix with most errors between adjacent severity levels.
- Regression (XGBoost): R^2 ≈ 0.65 residual plots indicate reasonable homoscedasticity and limited bias after log transformation.

A comparative snapshot against baselines (if evaluated) typically shows boosted trees outperforming linear models due to non-linear patterns and interactions.

6.2 Visual Analysis

- Confusion matrix: Few severe-to-minor confusions; most misclassifications occur between neighboring severity classes (e.g., restricted vs days-away), consistent with semantic proximity.
- Residual plots: Tighter residuals in the 0–30 day range; sparser extremes (>180 days) contribute larger errors due to limited samples.

6.3 Discussion and Implications

- Features with high influence include encoded occupation and mine/operator identifiers (capturing site/role risk), accident classification/type, experience signals, and composite severity indicators (restricted days, schedule charge).
- Temporal encodings improve stability across shifts and seasons, supporting proactive staffing and training.
- The deployed UI makes the model accessible to safety engineers, facilitating "what-if" analysis and training simulations.

7. Conclusion and Future Work

This work delivers an end-to-end predictive safety system for mining: a dual-task XGBoost solution that forecasts injury severity and lost workdays, wrapped in a deployable Streamlit interface. The pipeline demonstrates sound handling of large, mixed-type, and imbalanced tabular data; practical feature engineering; and metrics suited to downstream use.

Limitations:

- Class imbalance may still lower recall for the rarest severity class; threshold tuning or cost-sensitive learning could address this.

- Extreme lost-days events remain challenging with limited training examples; uncertainty estimates would be beneficial for decision-makers.
- High-cardinality entity encodings can drift when new mines/operators appear; periodic retraining and monitoring is recommended.

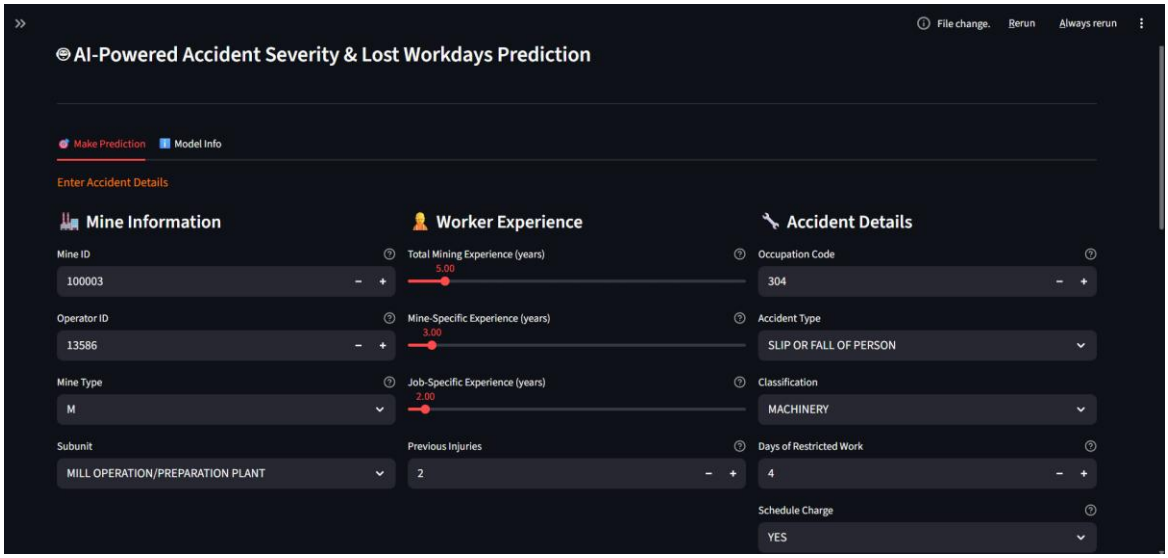
Future work:

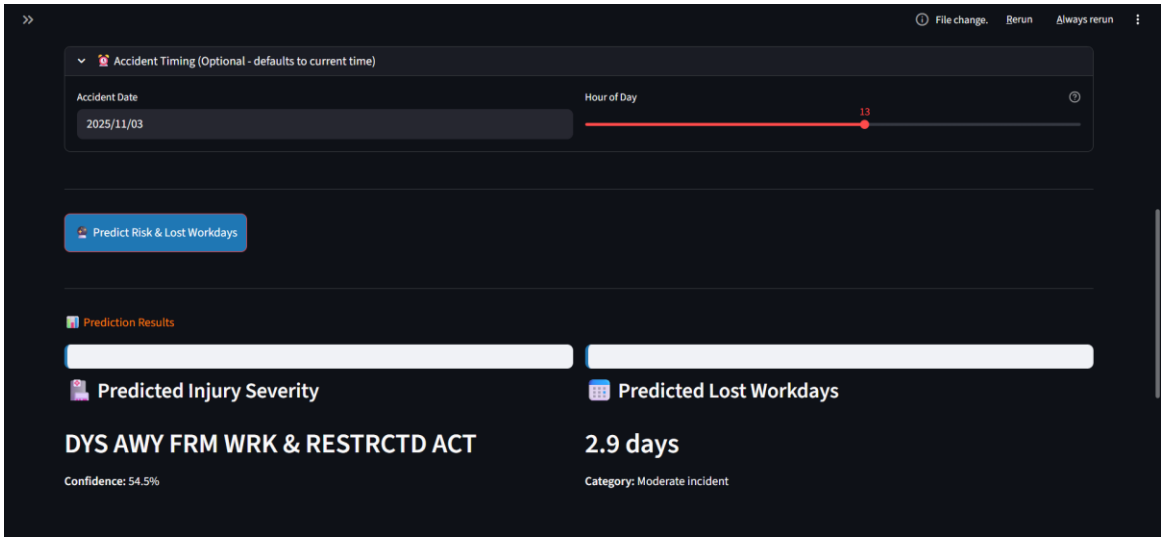
- Explore calibrated probabilities and threshold optimization for critical classes (e.g., fatalities/disabilities).
- Incorporate text/NLP fields (e.g., narratives) when available, and test deep ensembles.
- Add MLOps components: model registry, drift detection, scheduled retraining; consider cloud hosting (SageMaker/Vertex) for scale.
- Expand UI with explanations (e.g., SHAP per-prediction insights) and batch inference for planning scenarios.

8. References (IEEE style)

[1] MSHA, "Accident/Injury/Illness Data," U.S. Department of Labor, https://arlweb.msha.gov/OpenGovernmentData/OGIMSHA.asp (accessed: [Nov 2025]).

9. Appendix (Optional)




File change.
Rerun
Always rerun
Accident Timing (Optional - defaults to current time)
Accident Date
2025/11/03
Hour of Day
13
Predict Risk & Lost Workdays
Prediction Results
Predicted Injury Severity
DYS AWY FRM WRK & RESTRCTD ACT
Confidence: 54.5%
Predicted Lost Workdays
2.9 days
Category: Moderate incident