Modeling

Predictive Model Development for ATM Withdrawal Probabilities

Modeling Techniques

Classification

Selected Algorithms:

- **Random Forest**
  - Ensemble of decision trees
  - Handles non-linear relationships & high-dimensional data
  - Built-in feature importance analysis
- **Multinomial Logistic Regression**
  - Linear model with multinomial link function
  - L1/L2 regularization support
  - Efficient for large-scale datasets
  -

Regression

- linear regression
- desicion tree
- random tree

Key Assumptions:

1. Features treated as distribution-free
2. No perfect multicollinearity between predictors
3. Explicit handling of missing data:
   - Critical nulls: Row removal
   - Residual nulls: Spark's `handleInvalid` parameter

Test Design

Component	Specification
Train-Test Split	80% training / 20% testing
Cross-Validation	3-fold stratified
Evaluation Metrics	Weighted F1 Accuracy (final comparison) - for classification, MAE for regression

Hyperparameter Tuning

Model	Parameters	Search Grid
Random Forest	`numTrees`	[20, 50]
	`maxDepth`	[5, 10]
Logistic Regression	`regParam`	[0.01, 0.1]
	`elasticNetParam`	[0.0, 0.5]

Model Implementation

Random Forest Workflow:

1. Parallel tree training with feature bagging
2. Majority voting across 20-50 trees
3. Depth limitation (5-10 levels) to prevent overfitting

Logistic Regression Setup:

- Multinomial family for >2 classes
- ElasticNet mixing (0.0 = L2, 0.5 = L1/L2 blend)
- 100 iterations maximum