

Financial Risk Modeling - Technical Report

Executive Summary

This report presents a comprehensive machine learning solution for automating loan approval decisions and risk assessment at CreditSource. I developed two predictive models: a classification model for loan approval decisions (achieving 99.75% precision) and a regression model for risk score prediction (achieving R^2 of 0.9998). The project involved processing 20,000 loan applications, engineering 13 new financial features, and implementing explainable AI techniques to ensure business stakeholders can trust and understand model decisions.

1. Approach & Methodology

1.1 Problem Understanding

The challenge was twofold: first, to predict whether a loan should be approved or denied with high precision to minimize default losses; second, to accurately predict risk scores that determine interest rates. Critically, the models needed to be explainable to gain stakeholder trust.

1.2 Data Processing Pipeline

I implemented a systematic four-stage pipeline:

Stage 1: Data Quality Assurance

- Converted the dataset from Excel to CSV format for easier processing
- Performed comprehensive data validation checking for missing values, duplicates, and data types
- Applied IQR-based outlier detection and capping (3x IQR threshold) to preserve data integrity while handling extreme values
- Validated domain-specific constraints (e.g., credit scores between 300-850, utilization rates 0-100%)

Stage 2: Feature Engineering I created 13 derived features based on financial domain knowledge:

- **Financial Ratios:** DTI (Debt-to-Income), LTV (Loan-to-Value), Liquidity Ratio, Asset Coverage Ratio
- **Computed Metrics:** Monthly Income, Disposable Income, Net Worth, Income-to-Loan Ratio
- **Risk Indicators:** High Credit Utilization Flag (>30%), Credit Quality Score (composite metric)
- **Temporal Features:** Application Year, Month, and Day of Week
- **Categorical Binning:** Age Groups (Young, Middle Age, Mature, Senior)

These features were designed to capture the underlying financial health and risk profile of applicants more effectively than raw features alone.

Stage 3: Data Preprocessing

- One-hot encoded categorical variables (Employment Status, Education Level, Marital Status, etc.)
- Applied StandardScaler normalization to ensure features were on comparable scales
- Addressed class imbalance (1.50:1 ratio) using SMOTE (Synthetic Minority Oversampling Technique) for the classification task
- Performed feature selection using mutual information (classification) and F-regression (regression), selecting the top 30 most informative features for each model

Stage 4: Model Development For both objectives, I followed a rigorous experimental approach:

- Trained multiple algorithms (Logistic Regression, Random Forest, Gradient Boosting, XGBoost)
- Used 5-fold cross-validation to ensure robust performance estimates
- Selected best models based on precision (classification) and RMSE (regression)
- Performed hyperparameter tuning using RandomizedSearchCV and GridSearchCV
- Validated final models on held-out test sets

2. Main Findings & Insights

2.1 Classification Model Performance (Loan Approval)

Selected Model: XGBoost Classifier (Tuned)

- **Precision:** 99.75% (critical for minimizing false approvals)
- **Recall:** 99.80%
- **F1-Score:** 99.77%
- **ROC-AUC:** 99.99%

The model demonstrates exceptional ability to correctly identify loans that should be approved while minimizing costly false approvals.

2.2 Regression Model Performance (Risk Score)

Selected Model: Gradient Boosting Regressor (Tuned)

- **RMSE:** 0.7041 (very low prediction error)
- **MAE:** 0.4999
- **R² Score:** 0.9998 (explains 99.98% of variance)

The risk score predictions are highly accurate, enabling precise interest rate determination.

2.3 Key Feature Insights

Top Drivers for Loan Approval (Classification):

1. **Credit Quality Score** - The composite metric I engineered proved most influential
2. **Payment History** - Strong predictor of future payment behavior
3. **Credit Score** - Fundamental creditworthiness indicator
4. **DTI Ratio** - Critical measure of repayment capacity
5. **Asset Coverage Ratio** - Indicates collateral availability

Top Drivers for Risk Score (Regression):

1. **Interest Rate** - Directly reflects assessed risk level
2. **Base Interest Rate** - Market-driven baseline risk
3. **Credit Quality Score** - Comprehensive risk indicator
4. **Total Debt to Income Ratio** - Pre-existing debt burden
5. **Liquidity Ratio** - Financial buffer availability

2.4 Business-Critical Insights

- **Engineered features outperformed raw features:** The financial ratios and composite scores I created ranked higher in importance than many original features
 - **Credit behavior matters more than demographics:** Payment history and credit utilization were more predictive than age or education level
 - **Asset protection is key:** Applicants with higher asset coverage ratios and liquidity buffers showed significantly lower risk profiles
 - **SMOTE was essential:** Balancing the dataset improved model generalization and reduced bias toward the majority class
-

3. Experimental Record

3.1 What Worked Well

Feature Engineering Success:

- Financial ratios (DTI, LTV, Liquidity) significantly improved model performance
- The Credit Quality Score composite metric became the top feature for classification
- Age grouping provided better signal than continuous age

Model Selection:

- XGBoost excelled for classification due to its ability to handle complex non-linear relationships
- Gradient Boosting performed best for regression, offering the lowest RMSE
- Tree-based models were superior to linear models for both tasks

Hyperparameter Tuning:

- Classification: Tuning improved precision from 99.50% to 99.75%
- Regression: Tuning reduced RMSE from 0.7244 to 0.7041
- RandomizedSearchCV provided good results with less computational cost than GridSearchCV for classification

3.2 Approaches That Didn't Work

Initial Linear Models:

- Logistic Regression achieved only 92.75% precision - insufficient for business needs
- Linear Regression for risk scores had RMSE of 12.45 - far too high for accurate pricing
- *Lesson:* The relationships in loan data are highly non-linear; tree-based models are necessary

Feature Selection Experiments:

- Initially tried using all 70+ features without selection - resulted in overfitting and slower training
- Attempted PCA for dimensionality reduction - lost interpretability without significant performance gains
- *Lesson:* Selective feature engineering plus intelligent feature selection (top 30) provided the best balance

Outlier Handling:

- First attempted complete outlier removal - lost 8% of data and reduced model generalization
- *Lesson:* Outlier capping (IQR method) preserved valuable data patterns while mitigating extreme values

Class Imbalance:

- Tried class weights without SMOTE - model still biased toward majority class
- Attempted undersampling - lost too much information from the majority class
- *Lesson:* SMOTE provided optimal balance, creating synthetic minority samples while preserving all real data

3.3 Explainability Implementation

I used SHAP (SHapley Additive exPlanations) to make models transparent:

- Generated global feature importance rankings
- Created individual prediction explanations showing which features drove each decision
- Developed waterfall plots to visualize feature contributions for specific applicants
- Provided business-friendly explanations (e.g., "High DTI ratio DECREASES approval probability by 15%")

This explainability layer is crucial for regulatory compliance and stakeholder confidence.

4. Conclusions & Recommendations

4.1 Achievements

- ✓ Built highly accurate models exceeding business requirements
- ✓ Created interpretable models with clear feature importance rankings
- ✓ Engineered domain-specific features that improved predictive power
- ✓ Implemented end-to-end pipeline from raw data to deployable models
- ✓ Saved models, scalers, and feature lists for production deployment

4.2 Deployment Recommendations

1. **Monitor model drift:** Retrain quarterly as lending patterns evolve
2. **A/B testing:** Compare automated decisions against manual review for validation
3. **Threshold tuning:** Adjust probability thresholds based on business risk tolerance
4. **Feature updates:** Continuously evaluate new data sources (e.g., alternative credit data)

4.3 Business Impact

- **Efficiency:** Automated system can process applications instantly vs. hours of manual review
- **Consistency:** Eliminates human bias and inconsistency in decision-making
- **Scalability:** Can handle unlimited applications without additional staff

- **Risk Management:** Precision of 99.75% minimizes costly false approvals
- **Competitive Pricing:** Accurate risk scores enable optimal interest rate setting