Multimodal Credit Rating Prediction - Machine Learning & Data Analytics Documentation

1. Project Overview

1.1 Objective

The project aims to develop and evaluate multimodal machine learning models to predict corporate credit ratings using a combination of structured financial data and unstructured textual information from annual reports. The focus is on enhancing predictive accuracy through the integration of diverse data types.

1.2 Key Tasks

- **Binary Classification:** Differentiate between Investment Grade (BBB and above) and Below Investment Grade (BB and below).
- **Multi-class Classification:** Predict one of six rating categories (AA+, A, BBB, BB, B, CCC-).

1.3 Input Configurations

Three input configurations were evaluated:

- **Tabular Only:** Utilizes financial ratios and sector information.
- **Tabular + NLP Scores:** Combines financial data with sentiment/risk scores derived from MD&A text.
- **Tabular + NLP Scores + Full Text:** Incorporates financial data, NLP scores, and the complete MD&A text.

1.4 Experimental Scenarios

Each classification task was evaluated under two scenarios:

- **With Ticker:** Includes the company ticker as a feature.
- **Without Ticker:** Excludes the company ticker to prevent potential data leakage.

2. Dataset Information

2.1 Data Summary

- **Total Samples:** 2,029 company filings
- **Clean Samples After Processing:** 1,639
- **Time Period:** August 2005 to December 2016

- **Sectors Encoded:** 12 unique sectors
- **Companies (Tickers):** 593 unique companies

2.2 Class Distribution

Binary Classification

- Investment Grade (1): 978 samples (59.7%)
- Below Investment Grade (0): 661 samples (40.3%)

Multi-class Classification

Rating Category	Samples	Percentage
BBB	548	33.4%
BB	385	23.5%
A	339	20.7%
B	225	13.7%
AA	84	5.1%
CCC	46	2.8%
AA+	7	0.4%
CCC-	5	0.3%

2.3 Feature Engineering

- **Without Ticker:** 37 features
- **With Ticker:** 38 features (including encoded ticker)

3. Model Architecture

3.1 Algorithm Portfolio

Nine machine learning algorithms were implemented:

- Logistic Regression
- K-Nearest Neighbors (KNN)
- Naive Bayes
- Decision Tree

- Random Forest
- XGBoost
- Support Vector Machine (SVM)
- Deep Neural Network (DNN)
- Ensemble Methods (Stacking and Voting)

3.2 Neural Network Architecture

- **Input Layer:** Variable based on feature set (37-38 features)
- **Hidden Layers:** 2-3 dense layers with ReLU activation
- **Output Layer:**
  - **Binary:** Sigmoid activation
  - **Multiclass:** Softmax activation
- **Optimizer:** Adam with learning rate = 0.001
- **Loss Functions:**
  - **Binary:** Binary Crossentropy
  - **Multiclass:** Categorical Crossentropy

3.3 Ensemble Methods

- **Stacking Ensemble:** Meta-learner (Logistic Regression) trained on base model predictions.
- **Voting Ensemble:** Hard voting from best-performing base models.

4. Results Analysis

4.1 Binary Classification Results

Best Performing Models

WITHOUT Ticker Feature:

- **XGBoost:** Accuracy = 82.62%, F1-Score = 85.64%, ROC-AUC = 90.08%
- **Random Forest:** Accuracy = 82.32%, F1-Score = 85.50%, ROC-AUC = 88.97%
- **Stacking Ensemble:** Accuracy = 81.71%, F1-Score = 84.77%, ROC-AUC = 89.83%

WITH Ticker Feature:

- **XGBoost:** Accuracy = 83.54%, F1-Score = 86.36%, ROC-AUC = 90.42%
- **Stacking Ensemble:** Accuracy = 83.23%, F1-Score = 86.01%, ROC-AUC = 90.82%
- **Random Forest:** Accuracy = 79.88%, F1-Score = 83.50%, ROC-AUC = 89.61%

Error Analysis (Without Ticker)

- **False Positives:** 31 (Type I Error: 23.48%)
- **False Negatives:** 26 (Type II Error: 13.27%)
- **Total Errors:** 57 (Misclassification: 17.38%)

4.2 Multi-class Classification Results

Best Performing Models

WITHOUT Ticker Feature:

- **XGBoost:** Accuracy = 59.15%, F1-Score = 58.13%, ROC-AUC = 86.41%
- **Random Forest:** Accuracy = 58.23%, F1-Score = 56.92%, ROC-AUC = 81.15%
- **Stacking Ensemble:** Accuracy = 55.17%, F1-Score = 53.38%, ROC-AUC = N/A

WITH Ticker Feature:

- **XGBoost:** Accuracy = 58.84%, F1-Score = 57.90%, ROC-AUC = 83.75%
- **Random Forest:** Accuracy = 57.32%, F1-Score = 55.39%, ROC-AUC = 81.40%
- **Stacking Ensemble:** Accuracy = 54.93%, F1-Score = 53.56%, ROC-AUC = N/A

4.3 Feature Set Comparison

Binary Classification - Top Models by Feature Set

Rank	Feature Set	Model	Accuracy	F1-Score
1	Tabular + NLP + Ticker	Stacking Ensemble	83.50%	83.46%
2	Tabular + Ticker	Stacking Ensemble	83.25%	83.12%
3	Tabular + NLP	Stacking Ensemble	83.25%	83.24%
4	Tabular + NLP + Ticker	XGBoost	82.76%	82.72%
5	Tabular Only	Stacking Ensemble	82.76%	82.68%

Multi-class Classification - Top Models by Feature Set

Rank	Feature Set	Model	Accuracy	F1-Score
1	Tabular Only	Random Forest	56.16%	54.46%
2	Tabular Only	Stacking Ensemble	55.17%	53.38%
3	Tabular + Ticker	Stacking Ensemble	54.93%	53.56%
4	Tabular + NLP	Stacking Ensemble	54.68%	53.15%
5	Tabular + NLP + Ticker	Stacking Ensemble	54.19%	52.55%

5. Key Findings

5.1 Ticker Feature Impact

Binary Classification:

- **WITH Ticker:** Average Accuracy = 73.31%
- **WITHOUT Ticker:** Average Accuracy = 72.81%
- **Difference:** +0.50% improvement with ticker

Interpretation: Including the ticker slightly improved binary classification performance, suggesting some company-specific patterns are useful for distinguishing investment grade status.

Multi-class Classification:

- **WITH Ticker:** Average Accuracy = 40.42%
- **WITHOUT Ticker:** Average Accuracy = 40.09%
- **Difference:** +0.33% improvement with ticker

Interpretation: The ticker feature had minimal impact on multi-class classification, indicating that company identity is less critical for granular rating prediction.

5.2 NLP Feature Contribution

- **Binary Classification:** NLP features (sentiment/risk scores) provided marginal improvements when combined with tabular data.
- **Multi-class Classification:** NLP features showed mixed results, with tabular-only models performing best.

- **Full Text Addition:** Including complete MD&A text did not significantly outperform NLP scores alone.

5.3 Model Performance Insights

- **Tree-based Models Dominance:** XGBoost and Random Forest consistently outperformed other algorithms.
- **Ensemble Benefits:** Stacking ensembles improved robustness and generalization.
- **Neural Network Performance:** DNNs performed moderately but were outperformed by ensemble tree methods.
- **Simple Models Competitiveness:** Logistic Regression maintained reasonable performance despite model complexity.

6. Performance Metrics Deep Dive

6.1 Binary Classification Metrics

Best Model (XGBoost with Ticker):

- **Accuracy:** 83.54%
- **Precision:** 85.50%
- **Recall:** 87.24%
- **F1-Score:** 86.36%
- **ROC-AUC:** 90.42%
- **Log Loss:** 0.4474

6.2 Multi-class Classification Metrics

Best Model (XGBoost without Ticker):

- **Accuracy:** 59.15%
- **F1-Score (Macro):** 58.13%
- **ROC-AUC (Macro):** 86.41%
- **Log Loss:** 1.2856
- **Top-K Accuracy (K=2):** 92.68%

7. Business Implications

7.1 Practical Applications

- **Credit Risk Assessment:** Models can assist analysts in initial credit screening.
- **Portfolio Monitoring:** Automated rating prediction for large portfolios.
- **Early Warning System:** Identify companies at risk of downgrade.

7.2 Model Selection Recommendations

- **For Binary Classification:**
  - **Primary Choice:** XGBoost with Tabular + NLP features.

    - **Alternative:** Stacking Ensemble for better generalization.
    - **Scenario:** Include ticker for in-sample analysis, exclude for out-of-sample prediction.
- **For Multi-class Classification:**
    - **Primary Choice:** Random Forest with Tabular-only features.
    - **Alternative:** XGBoost for slightly better accuracy.
    - **Consideration:** Top-2 accuracy of 92.68% suggests models are effective at identifying rating neighborhood.

7.3 Risk Considerations

- **Data Leakage Risk:** Ticker feature may cause overfitting to specific companies.
- **Temporal Bias:** Data spans 2005-2016; may not capture recent market conditions.
- **Class Imbalance:** Limited samples for AA+ and CCC- categories.

8. Technical Implementation

8.1 File Structure

```
ML_Analytics/
├── data_processing.ipynb         # Data preprocessing and feature engineering
├── binary_classification.ipynb   # Binary classification models
├── multiclass_classification.ipynb # Multi-class classification models
├── ensemble_methods.ipynb        # Ensemble model implementation
├── feature_analysis.ipynb        # Feature importance and ablation studies
├── results_analysis.ipynb        # Results visualization and comparison
├── models/                       # Saved model files
├── results/                      # Output CSV files
└── config.py                     # Configuration parameters
```

8.2 Key Dependencies

- Python 3.8+
- Scikit-learn 1.0+
- XGBoost 1.5+
- TensorFlow 2.8+
- Pandas 1.4+
- NumPy 1.21+

8.3 Reproducibility

- Set random seeds for all models.
- Use stratified train-test split (80-20).
- Implement cross-validation for hyperparameter tuning.
- Save all model artifacts with versioning.

9. Conclusion

The project successfully developed multimodal machine learning models for credit rating prediction with the following key achievements:

- **Binary Classification:** Achieved 83.54% accuracy using XGBoost with tabular and NLP features.
- **Multi-class Classification:** Achieved 59.15% accuracy for six rating categories.
- **Feature Analysis:** Demonstrated the value of combining financial ratios with textual analysis.
- **Robust Modeling:** Addressed data leakage concerns through ticker ablation studies.

9.1 Future Work

- **Advanced NLP:** Incorporate transformer models for text analysis.
- **Temporal Modeling:** Implement LSTM/GRU for time-series prediction.
- **Explainability:** Add SHAP/LIME for model interpretability.
- **Real-time Deployment:** Develop API for real-time credit assessment.

9.2 Contact

For questions regarding this implementation, please contact manishkujur05@gmail.com