Model Card: Synthetic Data Generation for Imbalanced Classification

1. Problem Framing

1.1 Problem Statement

This model addresses class imbalance in (your classification task) through synthetic data generation. The process aims to improve minority class detection for (your specific domain) without degrading overall performance, providing a systematic approach to mitigate imbalanced data effects in ML models.

1.2 Performance Metrics and Success Criteria

- **Primary**: F1-score for minority class(es)
  - Target: Increase F1-score by at least X compared to baseline
- **Secondary**:
  - Precision (Target: X%)
  - Recall (Target: Y%)
  - AUC-ROC (Target: $\geq$ Z)
  - Other metrics as needed

Success is meeting the target improvement in primary metric while maintaining acceptable secondary metrics.

1.3 Requirements and Constraints

ID	Category	Description	Success Criteria
R1	Performance	Improve minority class detection	F1-score increase $\geq$ X over baseline
R2	Efficiency	Resource constraints	Complete in Y hrs; handle Z samples
R3	Reproducibility	Reproducible, modular process	Fixed seeds; versioned transforms
R4	Data Quality	Statistical integrity	JS divergence $\leq$ X; coverage $\geq$ Y%
R5	Additional	Additional requirement if needed	Success criteria for this requirement

Table 1: Requirements for Synthetic Data Generation

1.4 Imbalance Analysis

- **Ratio**: X:Y (majority:minority)
- **Minority Count**: X instances (Y% of total)
- **Type**: Binary or multi-class classification
- **Patterns**: Description of minority class patterns (clustered, scattered, etc.)
- **Features**:
  - Dimensions: X features
  - Types: Numerical, categorical, or mixed
  - High-dim: Yes/No
  - Missing: Percentage of missing values if applicable

1.5 Ethical Considerations

- **Fairness**: Describe specific ethical concerns related to the minority classes and any vulnerable populations they might represent
- **Bias**: Detail specific bias detection and mitigation methods to be used
- **Metrics**: List specific fairness metrics that will be tracked
- **Privacy**: Note relevant privacy regulations and compliance measures

- **Impact**: Describe how improvements may affect stakeholders and potential outcomes

---

2. Data Preparation and Quality Assurance

2.1 Data Cleaning

The following data quality issues were addressed to ensure high-quality input for synthetic data generation:

- **Missing Values**: (Describe imputation strategy used, with special attention to minority class)
- **Duplicates**: (Detail how duplicate records were identified and handled)
- **Outliers**: (Explain outlier detection method and treatment, especially for minority class instances)
- **Data Quality Metrics**:
  - Initial data completeness: X%
  - After cleaning completeness: Y%
  - Number of records removed: Z

2.2 Feature Engineering and Transformations

- **Categorical Encoding**: (Specify encoding methods used for categorical features)
- **Numerical Scaling**: (Detail normalization/standardization approaches)
- **Transformations Applied**:
  - Feature 1: (transformation applied and rationale)
  - Feature 2: (transformation applied and rationale)
  - Additional features as needed
- **New Features Created**: (List and justify any engineered features)

2.3 Dataset Splitting

- **Split Ratio**: Training X% / Test Y%
- **Stratification**: (Describe stratification approach to maintain class distribution)
- **Class Distribution**:
  - Original dataset: Majority X% / Minority Y%
  - Training set: Majority X% / Minority Y%
  - Test set: Majority X% / Minority Y%
- **Random Seed**: (Seed value used for reproducibility)

2.4 Data Quality Verification

Quality Check	Status	Notes
Missing values addressed	Complete/Partial	Details of approach
Duplicates handled	Complete/Partial	Method and impact
Outliers processed	Complete/Partial	Strategy used
Features appropriately encoded	Complete/Partial	Special considerations
Data properly scaled	Complete/Partial	Scale parameters preserved
Train/test split performed correctly	Complete/Partial	Isolation verification
Preprocessing steps documented	Complete/Partial	Location of documentation

Table 2: Data Quality Verification Checklist

2.5 Ethical Considerations

- **Data Removal Impact**: (Assess if data cleaning disproportionately affected minority class)
- **Privacy Protection**: (Describe anonymization or removal of sensitive attributes)
- **Fairness Verification**: (Detail checks performed to ensure preprocessing didn't introduce bias)
- **Representation**: (Explain how preprocessing maintained representation of important subgroups)

By the end of this stage, the dataset is cleaned, properly formatted, and split into appropriate sets for synthetic data generation and evaluation. All preprocessing steps are documented

to ensure reproducibility, and ethical considerations have been addressed to prevent introducing bias during data preparation.

---

3. Method Selection for Synthetic Data Generation

3.1 Technique Selection

- **Chosen Method**: (Specify the synthetic data generation method selected)
- **Method Category**:
  - ☐ Interpolation-based (e.g., SMOTE and variants)
  - ☐ Generative Adversarial Networks (GANs)
  - ☐ Variational Autoencoders (VAEs)
  - ☐ Diffusion Models
  - ☐ Other: (Specify)

3.2 Selection Rationale

- **Data Characteristics Match**: (Explain why this method is appropriate for the dataset's characteristics)
- **Alignment with Requirements**:
  - R1 (Performance): (How the method supports the target metric improvement)
  - R2 (Efficiency): (How the method meets time/resource constraints)
  - R3 (Reproducibility): (How reproducibility is ensured with this method)
  - R4 (Data Quality): (How the method preserves data integrity and realism)
- **Trade-offs Considered**: (Discuss trade-offs between complexity, quality, and resources)

3.3 Method Configuration

- **Key Parameters**:
  - (Parameter 1): (Value and rationale)
  - (Parameter 2): (Value and rationale)
  - Additional parameters as needed
- **Implementation Details**:
  - Library/Framework: (Specify implementation used)
  - Version: (Version number for reproducibility)
  - Custom modifications: (If any adaptations were made to standard implementation)

3.4 Method Evaluation

Evaluation Criteria	Assessment	Justification
Computational complexity	Low/Medium/High	Resource impact assessment
Expected sample quality	Low/Medium/High	Based on literature/prior experience
Training data requirements	Sufficient/Insufficient	Assessment of available minority samples
Modularity	Good/Limited	Ability to integrate/replace in pipeline
Tuning complexity	Simple/Moderate/Complex	Effort required for optimization

Table 3: Method Evaluation Matrix

3.5 Alternative Methods Considered

- **Alternative 1**: (Method name)
  - Advantages: (Potential benefits of this alternative)
  - Limitations: (Why this method wasn't selected)
- **Alternative 2**: (Method name)
  - Advantages: (Potential benefits of this alternative)

  - Limitations: (Why this method wasn't selected)

By the end of this stage, a specific synthetic data generation method has been selected based on analysis of the dataset characteristics, project requirements, and available resources. The method configuration has been defined and documented to ensure reproducibility and optimal performance.

---

4. Synthetic Data Generation

4.1 Implementation Details

- **Generation Framework**: (Specify the framework/library used)
- **Hardware Configuration**: (Detail the hardware used for generation)
- **Execution Time**: (Total time taken for generation process)
- **Random Seed**: (Seed value used for reproducibility)

4.2 Parameter Configuration

Parameter	Value	Rationale
(Parameter 1)	(Value)	(Reason for selection)
(Parameter 2)	(Value)	(Reason for selection)
(Parameter 3)	(Value)	(Reason for selection)
...	...	...

Table 4: Generation Method Parameter Configuration

4.3 Generation Volume

- **Original Class Distribution**:
  - Majority class: X samples (Y%)
  - Minority class: X samples (Y%)
- **Generation Strategy**:

  - Target ratio: (Describe target majority:minority ratio)
  - Generation rule applied: (Explain rule used to determine synthetic sample count)
- **Synthetic Data Generated**:
  - Number of synthetic samples: X
  - Percentage increase of minority class: Y%
- **Resulting Class Distribution**:
  - Majority class: X samples (Y%)
  - Minority class: X original + Y synthetic = Z total (W%)

4.4 Integration Process

- **Data Labeling**: (Describe how synthetic samples were labeled)
- **Integration Method**: (Explain how synthetic samples were combined with original data)
- **Verification Steps**:
  - Format consistency check: (Confirm synthetic data matches original format)
  - Distribution verification: (Verify class distribution meets target)
  - Test set isolation: (Confirm no test data was used in generation)

4.5 Quality Control

- **Initial Quality Assessment**: (Describe preliminary checks performed on synthetic data)
- **Issues Encountered**:
  - (Issue 1): (How it was addressed)
  - (Issue 2): (How it was addressed)
  - Add more if applicable
- **Iterations Required**: (Number of generation attempts before acceptable quality)

4.6 Reproducibility Measures

- **Generation Code**: (Location of stored code)
- **Model Artifacts**: (Location of saved generator model if applicable)
- **Generated Data**: (Location of saved synthetic data)
- **Augmented Dataset**: (Location of final combined dataset)

By the end of this stage, synthetic minority class samples have been generated using the selected method and integrated with the original training data. The process has been fully documented to ensure reproducibility, and the augmented dataset is ready for validation and model training.

---

5. Synthetic Data Validation

5.1 Distribution Analysis

- **Statistical Comparison**:
  - Feature-wise comparison: (Summary of statistical tests performed)
  - Central tendencies: (How means/medians compare between real and synthetic data)
  - Dispersion measures: (How variances/standard deviations compare)
  - Distribution tests: (Results of distribution similarity tests)
- **Visualization Methods**:
  - (List visualization techniques used to compare distributions)
  - Key observations: (Insights gained from visualizations)

5.2 Quality Metrics

Metric	Value	Threshold	Interpretation
Jensen-Shannon divergence	(Value)	(Threshold)	(What this value indicates)
Coverage score	(Value)	(Threshold)	(What this value indicates)
Novelty measure	(Value)	(Threshold)	(What this value indicates)
(Additional metrics)	(Value)	(Threshold)	(What this value indicates)

Table 5: Synthetic Data Quality Metrics

5.3 Coverage and Diversity Assessment

- **Feature Space Coverage**:
  - Dimensionality reduction method: (PCA/t-SNE/UMAP)
  - Coverage analysis: (How well synthetic data covers minority class regions)
  - Gap filling: (Assessment of how synthetic data fills sparse regions)
- **Diversity Measurement**:
  - Novelty score: (Measure of how different synthetic samples are from real ones)
  - Variety assessment: (How diverse the synthetic samples are from each other)

5.4 Duplication and Memorization Check

- **Duplicate Detection**:
  - Method used: (Describe approach for finding duplicates)
  - Results: (Number/percentage of exact or near duplicates found)
  - Action taken: (How duplicates were handled, if any)
- **Memorization Risk**:
  - Evaluation approach: (How memorization was assessed)
  - Findings: (Level of memorization detected)

5.5 Class Representation Verification

- **Semantic Validity**:
  - Verification method: (How class representation was verified)
  - Results: (Confirmation that synthetic samples represent minority class)
- **Feature Coherence**:
  - Relationship preservation: (Assessment of feature interdependencies)
  - Implausible combinations: (Check for unrealistic feature combinations)

5.6 Ethical and Privacy Assessment

- **Bias Evaluation**:
  - Bias metrics: (Measures used to detect introduced bias)
  - Findings: (Any bias detected in synthetic data)
- **Privacy Protection**:
  - Identifiability risk: (Assessment of re-identification potential)
  - Sensitive attribute handling: (How privacy of sensitive data was preserved)
- **Fairness Verification**:
  - Fairness metrics: (Measures used to verify fairness)
  - Compliance: (How synthetic data aligns with fairness requirements)

5.7 Validation Outcome

- **Overall Assessment**: (Summary evaluation of synthetic data quality)
- **Decision**:
  - ☐ Approved for model training

  - ☐ Requires refinement (specify issues)
  - ☐ Rejected (specify reasons)
- **Documentation**: (Location of detailed validation report and visualizations)

By the end of this stage, the synthetic data has been thoroughly validated to ensure it closely resembles the real minority class data distribution, adds useful diversity without introducing noise or bias, and maintains privacy and fairness requirements. The validation outcome determines whether the synthetic data is ready for use in model training or requires further refinement.

---

6. Model Training

6.1 Classification Algorithm Selection

- **Algorithm**: (Specify the classification algorithm used)
- **Selection Rationale**: (Explain why this algorithm was chosen for the task)
- **Implementation**: (Library/framework and version used)

6.2 Training Configuration

Hyperparameter	**Value**
(Hyperparameter 1)	(Value)
(Hyperparameter 2)	(Value)
(Hyperparameter 3)	(Value)
...	...

Table 6: Model Hyperparameters

- **Training Hardware**: (Specify the hardware used for training)
- **Training Duration**:
  - Baseline model: (Time taken)
  - Augmented model: (Time taken)

- **Convergence Criteria**: (Specify how training completion was determined)

6.3 Baseline Model Training

- **Training Data**:
  - Size: (Number of samples)
  - Class distribution: (Majority:Minority ratio)
- **Training Process**:
  - Iterations/epochs: (Number completed)
  - Learning curve: (Description of convergence pattern)
  - Stopping criteria: (When and why training was stopped)
- **Regularization**: (Methods used to prevent overfitting)

6.4 Augmented Model Training

- **Training Data**:
  - Size: (Number of samples including synthetic data)
  - Class distribution: (Adjusted Majority:Minority ratio)
- **Training Process**:
  - Iterations/epochs: (Number completed)
  - Learning curve: (Description of convergence pattern)
  - Differences from baseline: (Any notable training behavior differences)
- **Overfitting Mitigation**: (Special measures taken due to synthetic data)

6.5 Model Evaluation on Test Set

Metric	Baseline	Augmented	Improvement
F1-score (minority)	(Value)	(Value)	(Difference)
Precision (minority)	(Value)	(Value)	(Difference)
Recall (minority)	(Value)	(Value)	(Difference)
AUC-ROC	(Value)	(Value)	(Difference)
(Additional metrics)	(Value)	(Value)	(Difference)

Table 7: Performance Comparison on Test Set

6.6 Success Assessment

- **Target Achievement**:
  - R1 (Performance): (Whether performance improvement targets were met)
  - R2 (Efficiency): (Whether efficiency requirements were satisfied)
  - R3 (Reproducibility): (Whether the process is documented for reproducibility)
  - R4 (Data Quality): (Whether data quality metrics were maintained)
- **Overall Assessment**:
  - ☐ Successful - synthetic data improved model performance
  - ☐ Partial success - some improvements with trade-offs
  - ☐ Unsuccessful - needs revision in earlier stages
- **Next Steps**: (Recommendations based on results)

By the end of this stage, two models have been trained under identical conditions: a baseline model using only original data and an augmented model using original plus synthetic data. The performance comparison provides direct evidence of the impact of synthetic data augmentation on addressing the class imbalance problem.

---

7. Performance Evaluation

7.1 Test Set Evaluation

- **Test Data Characteristics**:
  - Size: (Number of test samples)
  - Class distribution: (Majority:Minority ratio in test set)
  - Confirmation of isolation: (Verification that test data was not used in any previous stage)
- **Evaluation Procedure**: (Description of evaluation methodology)

7.2 Detailed Performance Metrics

Metric	**Baseline**	**Augmented**	**Change**	**Target Met?**
Minority Class Performance				
F1-score	(Value)	(Value)	(Value)	Yes/No
Precision	(Value)	(Value)	(Value)	Yes/No
Recall	(Value)	(Value)	(Value)	Yes/No
Majority Class Performance				
F1-score	(Value)	(Value)	(Value)	N/A
Precision	(Value)	(Value)	(Value)	N/A
Recall	(Value)	(Value)	(Value)	N/A
Overall Performance				
Accuracy	(Value)	(Value)	(Value)	N/A
AUC-ROC	(Value)	(Value)	(Value)	Yes/No
(Other metrics)	(Value)	(Value)	(Value)	Yes/No

Table 8: Comprehensive Performance Evaluation

7.3 Confusion Matrix Analysis

- **Baseline Model**:
  - True Positives (minority): (Count)
  - False Positives: (Count)
  - False Negatives: (Count)
  - True Negatives: (Count)

- **Augmented Model**:
  - True Positives (minority): (Count)
  - False Positives: (Count)
  - False Negatives: (Count)
  - True Negatives: (Count)
- **Error Analysis**: (Key insights from confusion matrix comparison)

7.4 Requirements Assessment

- **R1 (Performance)**:
  - Target: (Original target)
  - Achieved: (Actual improvement)
  - Status: (Met/Partially Met/Not Met)
- **R2 (Efficiency)**:
  - Target: (Original constraint)
  - Achieved: (Actual resource usage)
  - Status: (Met/Partially Met/Not Met)
- **R3 (Reproducibility)**:
  - Target: (Original requirement)
  - Achieved: (Documentation and versioning status)
  - Status: (Met/Partially Met/Not Met)
- **R4 (Data Quality)**:
  - Target: (Original quality thresholds)
  - Achieved: (Actual quality measures)
  - Status: (Met/Partially Met/Not Met)

- **Additional Requirements**: (Assessment of any other project-specific requirements)

7.5 Fairness and Ethical Assessment

- **Bias Analysis**:
  - Performance across subgroups: (Analysis of model fairness across demographic or other sensitive attributes)
  - Error distribution: (How errors are distributed across groups)
- **Fairness Metrics**:
  - Metric 1: (Value and interpretation)
  - Metric 2: (Value and interpretation)
  - Additional metrics as needed
- **Ethical Considerations**: (Assessment of whether any ethical concerns or unintended consequences emerged in final model)

7.6 Final Project Status

- **Overall Assessment**: (Summary of project success or failure)
- **Decision**:
  - ☐ Ready for deployment
  - ☐ Requires iteration (specify stages to revisit)
  - ☐ Project goals not achievable with current approach
- **Next Steps**: (Recommendations for deployment or future improvements)
- **Lessons Learned**: (Key insights from the process for future synthetic data projects)

This final evaluation stage provides a comprehensive assessment of whether the synthetic data generation approach successfully addressed the class imbalance problem. By comparing the augmented model's performance against both the baseline model and the predefined

success criteria, the evaluation determines if the project goals were met and the model is ready for deployment.

---

8. Deployment and Monitoring

8.1 Deployment Strategy

- **Target Environment**:
  - Infrastructure: (Describe the production environment)
  - Integration approach: (Explain how the model is integrated into existing systems)
- **Model Packaging**:
  - Format: (Specify the model's export format)
  - Dependencies: (List required libraries and versions)
  - Preprocessing components: (Detail how preprocessing steps are included)
- **Deployment Validation**:
  - Testing methodology: (How consistency was verified)
  - Results: (Summary of validation tests)

8.2 Documentation and Knowledge Transfer

- **Deployed Documentation**:
  - User documentation: (Describe guides for end users)
  - Technical documentation: (Detail documents for technical stakeholders)
  - Model limitations: (List known constraints and assumptions)
- **Stakeholder Communication**:
  - Knowledge transfer: (How information was shared with operational teams)
  - Training provided: (Any training for users or maintainers)

8.3 Monitoring Framework

- **Performance Metrics**:
  - Tracked metrics: (List metrics being monitored in production)
  - Collection frequency: (How often metrics are gathered)
  - Reporting mechanism: (How metrics are reported and to whom)
- **Drift Detection**:
  - Data drift monitoring: (Methods for detecting input distribution changes)
  - Concept drift monitoring: (Methods for detecting changes in relationships)
  - Minority class prediction tracking: (How minority class prediction rates are monitored)
- **Alert Thresholds**:
  - Performance thresholds: (When alerts are triggered for performance drops)
  - Drift thresholds: (When alerts are triggered for drift detection)
  - Response protocol: (Actions to take when thresholds are crossed)

8.4 Compliance and Security

- **Privacy Considerations**:
  - Synthetic data implications: (Ensuring no sensitive information leakage)
  - Production data handling: (How new data is securely processed)
- **Regulatory Compliance**:
  - Applicable regulations: (Industry or regional requirements)
  - Compliance measures: (Steps taken to ensure compliance)
- **Security Implementation**:
  - Access controls: (How model access is restricted)

  - Data encryption: (Protection measures for data in transit and at rest)

8.5 Maintenance Plan

- **Retraining Strategy**:
  - Retraining frequency: (Schedule or trigger-based approach)
  - Data collection mechanism: (How new training data is gathered)
  - Pipeline reuse: (How components are reused for retraining)
- **Version Control**:
  - Model versioning: (How model versions are tracked)
  - Rollback procedure: (Process for reverting to previous versions)
- **Improvement Cycle**:
  - Feedback mechanism: (How production insights inform improvements)
  - Iteration planning: (Schedule for revisiting framework stages)

8.6 Initial Production Results

- **Early Performance**:
  - Initial metrics: (First results from production environment)
  - Comparison to test results: (How production performance compares to test expectations)
- **Operational Assessment**:
  - Resource utilization: (CPU, memory, storage usage)
  - Latency: (Response time in production)
  - Throughput: (Processing capacity)
- **User Feedback**: (Initial reactions from stakeholders or users)

By completing this stage, the synthetic data-enhanced model has been successfully deployed to production, with comprehensive monitoring and maintenance systems in place. The framework has come full circle, and the model is now actively addressing the class imbalance problem in real-world conditions.

---

9. Documentation and Ethics Review

9.1 Comprehensive Documentation

- **Process Documentation**:
  - Framework stages: (Summary of all stages and key decisions)
  - Code repository: (Location of versioned code)
  - Parameter logs: (Record of all parameters and settings)
- **Artifact Preservation**:
  - Original datasets: (Storage location and access protocol)
  - Synthetic datasets: (Storage location and access protocol)
  - Model artifacts: (Storage location of trained models)
  - Validation reports: (Location of detailed validation documents)
- **Replication Guide**:
  - Step-by-step reproduction: (Instructions for reproducing results)
  - Environment specifications: (Software and hardware requirements)

9.2 Lessons Learned

- **Technical Insights**:
  - Generation method efficacy: (Observations about synthetic data quality)
  - Challenges encountered: (Technical difficulties and solutions)

  - Unexpected findings: (Surprising results or insights)

- **Process Improvements**:
  - Framework refinements: (Suggestions for improving the framework)
  - Efficiency gains: (Methods to improve future iterations)

- **Knowledge Transfer**: (How insights will be shared with the organization)

9.3 Ethics Audit

- **Review Process**:
  - Audit methodology: (Approach to ethics review)
  - Reviewers: (Who conducted the ethics review)
  - Scope: (What aspects were examined)

- **Ethical Considerations Addressed**:

Consideration	Mitigation Approach	Verification Method
Bias in synthetic data	(Approach taken)	(How verified)
Privacy protection	(Approach taken)	(How verified)
Fairness across groups	(Approach taken)	(How verified)
(Additional considerations)	(Approach taken)	(How verified)

Table 9: Ethics Considerations and Verification

9.4 Limitations and Residual Risks

- **Known Limitations**:
  - Model constraints: (Situations where model performance may degrade)
  - Data representation gaps: (Areas where data may still be insufficient)
  - Framework constraints: (Limitations of the current approach)

- **Residual Risks**:

  - Potential issues: (Identified remaining risks)
  - Mitigation plans: (How each risk will be monitored and addressed)
- **Future Considerations**: (Areas requiring further research or development)

9.5 Compliance Verification

- **Organizational Requirements**:
  - Internal policies: (Compliance with organizational standards)
  - Approval status: (Sign-offs and approvals obtained)
- **Regulatory Compliance**:
  - Applicable regulations: (Specific requirements met)
  - Documentation submitted: (Any required regulatory filings)
- **Industry Standards**: (Adherence to relevant industry guidelines)

9.6 Project Conclusion

- **Final Status**:
  - Project completion: (Final state of the implementation)
  - Handover details: (Information about ownership transfer)
- **Success Evaluation**:
  - Original objectives: (Review of problem statement and goals)
  - Achievements: (Summary of what was accomplished)
  - Business impact: (Actual or expected impact on operations)
- **Future Work**: (Recommendations for future iterations or extensions)

This final section completes the documentation of the synthetic data generation framework implementation. By thoroughly documenting the entire process and conducting a compre-

hensive ethics review, this model card ensures transparency, reproducibility, and accountability. The project is now properly archived, with all decisions, code, and results preserved for future reference or audit.