| **Document Field** | **Document Value** |
| --- | --- |
| Title | Responsible Data Science Policy -  [USE CASE NAME] Sub-Procedure (“Sub-Procedure”) |
| Document ID |  |
| Version |  |
| Last Updated | [DATE] |
| Last Reviewed | [DATE] |
| Responsible Body | [SUB-PROCEDURE OWNER] |

# 1. Introduction

## 1.1 PURPOSE

This Sub-Procedure supports the Organization’s Responsible Data Science Parent Procedure (“Parent Procedure”) by defining clear processes for Data Science projects that meet the scope description below.

## 1.2 **SCOPE**

In order for a Project to be approved for use under this Sub-Procedure, the following conditions must be met:

* [list or expression of logical conditions]
* [[EXAMPLE: Use case is targeted direct mail credit card offer for US non-veteran households.]]
* [[EXAMPLE: Use case is employee retention salary adjustment for salaried US and UK staff.]]

Additional requirements related to the Project’s Dataset(s) or Model(s) may be listed in the respective sections below.

## 1.3 DATASET REQUIREMENTS

The Project’s Dataset(s) must meet the following requirements:

* [list of expressions or logical conditions]
* [[EXAMPLE: Source: ExperifaxUnion PeopleBuyData™,
* [[EXAMPLE: Filter: Pre-filtered by ExperifaxUnion to remove all veteran households]]
* [[EXAMPLE: Source: CSV export from Organization’s HumanSoft HR system]]
* [[EXAMPLE: Source Records: CSV export from tbl\_PeoplePerformanceReview from ERP HR module]]

## 1.4 MODEL REQUIREMENTS

If any pre-existing models are utilized in this Project, they must meet the following requirements:

* [list of expressions or logical conditions]
* [[EXAMPLE: Must not have been developed using data that includes EU residents]
* [[EXAMPLE: Must have undergone prior internal screening for disparate impact or bias on race or gender]]
* [[EXAMPLE: Must not use ReallyDeepThinking, smell2tensor embeddings, or related techniques.]]

# 2. Procedures - Research and Development

Personnel must review the technical specifications below to confirm their understanding of the requirements for the Research and Development phase of a Project. Where no specific requirements are provided, Personnel should exercise their best judgement. If Personnel have questions, comments, or concerns related to deviations from or ambiguities in the specifications, they should contact [SUB-PROCEDURE OWNER] at [SUB-PROCEDURE OWNER CONTACT].

## 2.1 NOTICE

If Personnel believe their use case is valid under this sub-procedure, they should provide notice to [SUB-PROCEDURE OWNER CONTACT] that they are initiating a new Project.

## 2.2 PSEUDO-ANONYMIZATION

Project Datasets must be transformed for pseudo-anonymization as specified below. Transformations and their acceptable software implementations, when not otherwise defined, are specified in the Concepts and Techniques reference available at [Responsible Data Science Policy - Concepts and Techniques Inventory].

### 2.2.1 Attribute Transformations

Attributes, which may typically be present as fields or columns of tabular data, must be transformed as specified below:

* [FIELD NAME OR DESCRIPTION : TRANSFORMATION TYPE, TRANSFORMATION PARAMETERS]
* [[EXAMPLE: Household Salary: Convert to quantile and round to the nearest multiple of 0.025.]
* [[EXAMPLE: Employee Name: Hash with SHA-256 using a project-specific 32-byte salt.]]
* [[EXAMPLE: Employee Zip: Truncate the last two digits; (see also l-diversity constraint below).]]

### 2.2.2 Dataset Transformations

In addition to Attribute-level transformations, the Dataset as a whole must be transformed as specified below:

* [TRANSFORMATION TYPE, TRANSFORMATION PARAMETERS]
* [Truncate ZIP-5+ to ZIP-3. Group by (ZIP-3, transformed Household Salary). Calculate mode, median, or mean for all other fields.]

### 2.2.3 Acceptance Criteria

The Transformations above must result in the Project Dataset meeting or exceeding the criteria below:

* [list of criteria or constraints for dataset as a whole or specific classes/attributes of interest]
* [[EXAMPLE: k-anonymity and l-diversity must both be at least 5.]
* [[EXAMPLE: Each aggregated (ZIP-3, Salary) record must contain at least 3 households.]]

## 2.3 MODELING

### 2.3.1 Systems

Research and development [may/must] be conducted using systems as described below:

* [list of approved or prohibited source control or research systems]
* [[EXAMPLE: All Project source code must be stored and managed in Git.]]
* [[EXAMPLE: All model training artifacts must be stored in MLFlow.]]
* [[EXAMPLE: Final feature and training data must be stored in the following NFS path: /nfs123/compliance/modeling/project-id/123/]]

### 2.3.2 Cleaning

Prior to augmentation, synthetic data generation, or feature engineering, the following cleaning steps [may/must] be performed:

* [ATTRIBUTE NAME OR TYPE, TRANSFORMATION TYPE, TRANSFORMATION PARAMETER]
* [[EXAMPLE: Convert all diacritical marks or non-ASCII characters to ASCII.]]
* [[EXAMPLE: Drop underexposed images by removing the bottom 5% of samples by total intensity.]]

### 2.3.3 Augmentation and Synthetic Data Generation

Prior to feature engineering, the following data augmentation or synthetic data generation steps [may/must] be performed:

* [AUGMENTATION TYPE, AUGMENTATION PARAMETERS]
* [SYNTHETIC DATA GENERATION DESCRIPTION]
* [[EXAMPLE: Rotate all images randomly between -15 and +15 degrees.]]
* [[EXAMPLE: Fit a multivariate lognormal-Gamma distribution between income and age. Generate synthetic records by resampling from this hypothetical distribution and selecting attributes from the sample of 5 closest records.]]

### 2.3.4 Feature Engineering

Prior to training, the following feature engineering steps [may/must] be performed:

* [[EXAMPLE: Convert Age to percentile or rank.]]
* [[EXAMPLE: Convert all images to grayscale via perceptual luminance-preserving transform).]]
* [[EXAMPLE: Calculate the day-of-week or business day indicator from the hiring interview date.]]

### 2.3.5 Model Types

The following model types are [allowed/prohibited]:

* [list of allowed or prohibited model types or descriptions]
* [[EXAMPLE: Random forest or related tree ensemble technique (extra trees, xgboost, etc.)]]
* [[EXAMPLE: Support vector machine or related technique (e.g., LibSVM method, Bayesian SVM)]]

### 2.3.6 Training and Testing

Model training, testing, and cross-validation [may/must] be performed as follows:

* [list of acceptable target attributes or annotations]
* [list of acceptable loss functions or metrics]
* [list of requirements or constraints for train/test split percent/ratio, etc.]
* [[EXAMPLE: The test sample should be at least 20% of the total sample.]]
* [[EXAMPLE: The model target should be one of attributes X, Y, or Z.]]
* [[EXAMPLE: SVMs should be trained using L1 loss and Crammer-Singer.]]

### 2.3.7 Cross-Validation and Hyperparameter Optimization

Model cross-validation and hyperparameter optimization [may/must] be performed as follows:

* [list of requirements or constraints for folds, fold construction strategy, etc.]
* [list of requirements for hyperparameter optimization]
* [[EXAMPLE: The model should be selected and optimized with either macro average precision or ROC AUC.]]
* [[EXAMPLE: Time-ordered k-fold cross-validation with k>=3 should be performed.]]

# 3. Procedures - Release

Once the Research and Development phase of a Project is completed, Personnel must complete the Release procedure defined below. Contemporaneous documentation supporting each requirement should be maintained and made available to [SUB-PROCEDURE OWNER] upon request. Personnel should review the technical specifications below to confirm their understanding of the requirements. Where no specific requirements are provided, Personnel should exercise their best judgement. If Personnel have questions, comments, or concerns related to deviations from or ambiguities in the specifications, they should contact [SUB-PROCEDURE OWNER] AT [SUB-PROCEDURE OWNER CONTACT].

## 3.1 FAIRNESS TESTING

Prior to distributing or released Data Science results such as recommendations or models, Personnel must complete the following fairness testing:

* [list of fairness tests and acceptance criteria]
* [[EXAMPLE: Average odds difference for underwriting approval between protected and non-protected classes must be less than or equal to 10%.]]
* [[EXAMPLE: The Thiel index for proposed salary increase must less than X.]]

## 3.2 PRODUCTIONIZATION TESTING

Personnel [may/must] complete the following tests as part of the productionization process:

* [list of productionization requirements]
* [[EXAMPLE: If the Research and Development model is exported via PMML, the Production model must be tested to produce identical results for all training samples.]]

## 3.3 ONLINE OR PERIODIC RETRAINING

If the Project involves a Model that will be run online or periodically retrained, then the conditions below must be met:

* [list of conditions]
* [[EXAMPLE: Streaming Spark models must be serialized, including all training data, at least once per day.]]
* [[EXAMPLE: Periodically-retrained models must not exhibit (hyper)parameter change per refit in excess of 5% in Euclidean space. If such change occurs, the production system should continue use of the prior model until the Fairness Testing requirements above have been completed again.]]

# 4. Related Policies and Procedures

## 4.1 **DATA PROCESSING AND HANDLING**

This Sub-Procedure is not intended to address all aspects of data processing. [The Organization’s [DATA HANDLING POLICY AND PROCEDURES], [CHOICE AND CONSENT POLICY AND PROCEDURES], [DATA CLASSIFICATION POLICY AND PROCEDURES], [ANY OTHER APPLICABLE POLICIES AND PROCEDURES] must be followed when applicable.

## 4.2 **RETENTION**

Documentation required under this Sub-Procedure must follow the Organization’s [RETENTION POLICY AND PROCEDURES].

[LIST OTHER APPLICABLE POLICIES AND PROCEDURES]

# 5. Guidance

Personnel with questions about how to follow these procedures should first consult the Organization’s internal resources:

[LIST RELEVANT RESOURCES, SUCH AS HANDBOOKS, WIKI PAGES, ETC.]

Questions or comments related to this document can be directed to [SUB-PROCEDURE OWNER] at [SUB-PROCEDURE OWNER CONTACT].