# **Requirements Document/ Scoping Document**

**A note to all groups**

***Overview: you must provide the requirements/scoping document to your client before the due date for their feedback***.

### 

[Requirements Document/ Scoping Document 1](#_heading=h.gjdgxs)

[1. Data Understanding 2](#_heading=h.3znysh7)

[1.1 Initial Data Sources 2](#_heading=h.2et92p0)

[1.2 Data Collection Methods 2](#_heading=h.tyjcwt)

[1.3 Assumed Data Quality, Structure, and Completeness 3](#_heading=h.1t3h5sf)

[1.4 Mechanisms for Evaluation 4](#_heading=h.2s8eyo1)

[2. Data Preparation 4](#_heading=h.17dp8vu)

[2.1. Data Preparation Activities 4](#_heading=h.3rdcrjn)

[2.2. Data Processing Pipelines 5](#_heading=h.lnxbz9)

[2.3. Different Preparation Pipelines 5](#_heading=h.35nkun2)

[2.4. Justification for Different Models 6](#_heading=h.1ksv4uv)

[2.5. Data Storage Mechanisms 6](#_heading=h.44sinio)

[3. Modelling 6](#_heading=h.2jxsxqh)

[Relationship to the Chosen Model: 8](#_heading=h.z337ya)

[3.2. How will the Effectiveness and Validity be checked? 9](#_heading=h.3j2qqm3)

[3.3. Test Data 9](#_heading=h.1y810tw)

[4. Evaluation 10](#_heading=h.2xcytpi)

[4.1. Evaluating Results 10](#_heading=h.1ci93xb)

[4.2. Review Process 11](#_heading=h.3whwml4)

[4.3. Next Steps 11](#_heading=h.2bn6wsx)

[5. Deployment 12](#_heading=h.qsh70q)

[5.1. Deployment Plan: 12](#_heading=h.3as4poj)

[5.2. Monitoring and Maintenance: 12](#_heading=h.49x2ik5)

[5.3. Final Report and Project Review: 13](#_heading=h.147n2zr)

### 

## **1. Data Understanding**

### **1.1 Initial Data Sources**

* **Primary Dataset:**
  + **Data Source:** Kaggle Dataset with 1,000 rows.
  + **Content:** Dataset includes a variety of columns relevant to insurance claims including months\_as\_customer, age, policy\_number, policy\_state, incident\_type, total\_claim\_amount, and a crucial column, fraud\_reported, which indicates whether each entry is fraudulent.
  + **Widespread Availability:** It has been discovered that this dataset is identical across multiple sources, websites, and forums, suggesting that it is widely used and possibly well-vetted.
  + Despite the dataset being a good starting point, it needs to be evaluated and prepared properly to ensure it meets and services the project’s needs and can assist in fraud detection purposes.
* **Additional Data:**
  + **Synthetic Data Generation:** To expand the dataset, approximately 9,000 additional rows will be generated using three different data synthesis methods (Scikit-learn, GAN, and AI-based generation). This will then be combined to give a dataset with a total of ~~about~~ 10,000 rows.

### **1.2 Data Collection Methods**

* **Kaggle Dataset:**
  + The initial dataset is already collected and readily available for analysis.
  + **Columns and Structure:** The dataset includes the following columns:
    - **Customer Information:** months\_as\_customer, age, insured\_sex, insured\_education\_level, insured\_occupation, insured\_hobbies, insured\_relationship.
    - **Policy Information:** policy\_number, policy\_bind\_date, policy\_state, policy\_csl, policy\_deductable, policy\_annual\_premium, umbrella\_limit.
    - **Incident Information:** incident\_date, incident\_type, collision\_type, incident\_severity, authorities\_contacted, incident\_state, incident\_city, incident\_location, incident\_hour\_of\_the\_day, number\_of\_vehicles\_involved, property\_damage, bodily\_injuries, witnesses, police\_report\_available.
    - **Financial Information:** capital-gains, capital-loss, total\_claim\_amount, injury\_claim, property\_claim, vehicle\_claim.
    - **Vehicle Information:** auto\_make, auto\_model, auto\_year.
    - **Outcome Information:** fraud\_reported.
* **Bronze Layer:**
  + **Raw Data Storage:** The unclean original dataset (1,000 rows) is stored in the Bronze layer in its raw form. The bronze layer is the initial repository for all data before any cleaning or processing.
  + **Purpose:** Keeping the unclean data in Bronze ensures that the original dataset is always accessible for reference or reprocessing if needed. It also makes the project traceable and reproducible.
* **Silver Layer:**
  + **Data Cleaning:** The raw dataset from the Bronze layer is cleaned using Azure Functions. This cleaned dataset (1,000 rows) is then stored in the Silver layer as a cleaned version of the original data.
  + **Synthetic Data Combination:** The cleaned dataset from Silver is used as the basis for generating synthetic data (~ 9,000 rows). The combined dataset, now totaling 10,000 rows, is also stored in the Silver layer.
  + Data Organisation: The Silver layer will contain different folders to organise the original cleaned data and the combined dataset.
* **Gold Layer:**
  + **Final Data Preparation:** The data in the Silver layer is further refined and stored in the Gold layer. This final data is fully prepared and ready for use by the AI team. It will contain columns relevant and useful to the AI model and will have a feedback loop from the AI Team which will trickle down to the silver layer if our data quality needs to be improved by more cleaning and refining.

#### 

### **1.3 Assumed Data Quality, Structure, and Completeness**

* **Kaggle Dataset:**
  + **Quality:** The dataset is widely available, suggesting that it is well-vetted.
  + **Structure**: The dataset’s standardised structure simplifies integration but requires careful examination to ensure that no nuances are lost, and all necessary elements for fraud detection are present.
  + **Completeness:** While the dataset is likely complete in terms of the data points it intends to capture, it may not cover all scenarios or edge cases specific to the project, making the generation of synthetic data crucial.
* **Synthetic Data:**
  + **Quality:** The synthetic data is expected to be clean, as it is generated from the cleaned source data. It will be evaluated for quality, ensuring it introduces valuable variance and addresses potential gaps in the original dataset.
  + **Structure:** The synthetic data will match the cleaned original dataset’s structure, ensuring seamless integration.
  + **Purposeful Errors:** Some synthetic data might include purposeful errors or variations to test and validate the robustness of the AI models, helping to identify and rule out potential biases.
  + **Use of SMOTE Elements:** To address any imbalance in fraud versus non-fraud cases, some synthetic data generation methods may incorporate elements from SMOTE (Synthetic Minority Over-sampling Technique).
  + **Completeness:** The goal is to use the synthetic data to enhance the dataset’s completeness by covering underrepresented cases and scenarios in the original data.

#### 

### **1.4 Mechanisms for Evaluation**

* **Data Quality Assessment:**
  + **Initial Dataset:** Perform detailed exploratory data analysis on the initial dataset to identify biases, potential overfitting risks, and any embedded assumptions that might affect the project’s outcomes.
  + **Synthetic Data:** Evaluate the synthetic data against the original dataset to ensure it adds diversity and robustness. The synthetic data should not merely replicate common patterns but should introduce valuable variance. Any purposeful errors will also be evaluated to assess their impact on model training.
* **Data Structure Verification:**
  + Confirm that the dataset’s structure fully supports the analyses planned. Look for any structural limitations in the dataset and address these with synthetic data as needed.
* **Completeness Check:**
  + Ensure that the dataset captures all relevant scenarios, especially for fraud detection. The incorporation of SMOTE elements in the synthetic data generation will help ensure a balanced representation of fraud cases.
* **Additional Data Collection Requirements:**
  + As the project progresses, regularly assess whether more unique or tailored data is needed. Stay flexible to incorporate supplementary data that could provide additional insights or cover overlooked scenarios.

## **2. Data Preparation**

### **2.1. Data Preparation Activities**

* **Data Selection**: In order to select the data, we look into various data sources and identify the relevant features to include in our dataset. In our case, we went ahead with one dataset and created 100 rows of data using information from that dataset. We plan to create 1000 rows of data using this dataset.
* **Data Cleaning:** For data cleaning, we plan to fill in any missing data. We also plan to drop the columns which we will not need and possibly convert the categorical values into numerical ones for example: True will be converted to 1 and False will be converted to 0. We plan to clean the data and store it in the silver layer of our data structure.
* **Data Synthesization:** We plan to synthesise around 9000 rows of data using data zsynthesization techniques. We plan to use 3 different methods which will generate 3000 rows of data each. The methods are the “make\_classification” function in the scikit-learn library in python, GAN model, and AI based generation.
* **Data Integration:** While we are not currently planning to integrate different data sources, we can still do that if we find another valid data source. If we had to merge, we would identify the common keys such as the ClaimID and merge the datasets. We would have to be consistent with the data. If we are using datasets with different formats, we would have to convert the separate formats into a unified structure. Therefore, JSON files will be converted to tables, XML files will be converted to CSV files.
* **Data Formatting:** Data formatting ensures one particular format for our dataset. For our case we will have a CSV file. We select a CSV file as our dataset is relatively small and CSV is useful for small datasets. We have to ensure that all of the rows in the columns are consistent. For example, the date and time format should be the same for all the rows.

#### 

### **2.2. Data Processing Pipelines**

The data processing pipeline is mentioned above. We will have three different layers which are bronze, silver and gold. The bronze layer will have 1000 rows of data which will not be cleaned. While we are going from the bronze to the silver layer, the data will be cleaned using azure functions. While cleaning the data, we want to drop any columns we will not need and fill in any missing values. In the silver layer, we will also synthesise data using three different methods. Each method will synthesise 3000 rows of data. The methods will be Scikit-Learn, GAN(Generative Adversarial Networks), and AI models. This will be stored back in the silver layer and then it will be stored in the gold layer where there will be a feedback loop. The feedback loop will also move to the silver layer where the changes will be made. The AI team (Team 13) will review and give suggestions on the dataset and the Data team (Team 14) will adjust the dataset accordingly.

### **2.3. Different Preparation Pipelines**

Different models will ~~need~~ require different preparation pipelines. The pipeline will ensure that the data is not mishandled and is optimised. This will lead to an improved performance of the model. If a Decision Tree model was chosen, the data would not need to be cleaned that much as that model can handle raw data well. Outliers would need to be removed if they impact the model negatively. However, if a Neural Network model was chosen, the data would need to be normalised and high quality data would need to be provided. Neural Networks are more sensitive therefore, transformations need to be performed to remove outliers and categorical values would need to be converted to numerical values.

### **2.4. Justification for Different Models**

Having different pipelines/models will enhance efficiency. Unnecessary steps can be avoided if there are separate pipelines, therefore, less resources and time will be required. Moreover, separate pipelines allow for easy adjustments if new methods are introduced later on. A tailored pipeline will be beneficial as it will optimise the performance of the model. This will improve the accuracy of the model and reduce the risk of biases.

### **2.5. Data Storage Mechanisms**

Our team has decided on using Azure Data Lake Storagoe Gen2 (ADSL2) to store all of our data. The data will be stored in layers as mentioned above. Azure Data Lake Storage Gen2 is being used as it is cost effective, secure, and scalable. Our dataset will not be that large therefore, the cost for ADSL2 will be cheap. Even if our dataset increases, that will not be a problem as ADSL2 can handle large volumes of data. Therefore, this setup ensures high performance and provides a good solution for data storage.

## **3. Modelling**

For the project, we are working in collaboration with Group 13 for EY. In this collaboration, we are split into two teams, Team 14 who focuses on the data side with our main deliverable being a comprehensive dataset and cloud infrastructure, and Team 13 which focuses on the AI side and will be using our dataset to train their Generative AI model. Therefore, modelling ~~is not entirely in~~ does not fall within the scope of our project with EY however, below we will be discussing some aspects of modelling that will be taking place in relation to the AI team's requirements.

**3.1. Modelling Technique**

The main outcome of using these AI modelling techniques is to find out if an insurance claim being made to NRMA is fraud or not fraud. This is a classification problem as it aims to categorise the data into predefined classes of fraud and not a fraud. Therefore, as a classification issue, we will be using the following models to predict classes: logistic regression model, neural networks, and decision tree.

**Neural Networks**:

* **Why Chosen:** Neural Networks are selected for their ability to model complex relationships within data through layers of interconnected nodes. This technique is particularly useful for capturing non-linear patterns that might be indicative of fraud.
* **Intended Outcome:** The aim is to explore the model's ability to identify subtle, non-linear patterns in transactional data that may not be evident with more traditional methods like Random Forest or GBM.

~~Pros~~ Benefits of Neural Networks

* They form the basis of state-of-the-art models and can be formed into advanced architectures that effectively capture complex features given enough data and computation.

~~Cons~~ limitations of Neural Network

* Larger, complex models require significant training time, data, and customisation.
* Careful preprocessing of the data is needed.
* A good choice when the features are of similar types but less so when the features are of very different types.

**Logistic Regression Model**:

* **Why Chosen**: Logistic regression is a powerful yet straightforward method for binary classification problems. It estimates the probability that a given instance belongs to a particular class, making it ideal for predicting whether a claim is fraudulent (class 1) or not (class 0).
* **Intended Outcome**: The model will generate probabilities that an insurance claim falls into the "fraud" category. If the probability is above a certain threshold, the claim is classified as fraud.

Pros of Logistic Regression:

* Simplicity: Easy to implement and interpret.
* Efficiency: Works well with smaller datasets and provides quick results.
* Probabilistic Output: Offers clear probabilities for each class, aiding in decision-making.

Cons of Logistic Regression:

* Linearity: Assumes a linear relationship between the features and the log odds, which may not capture complex patterns.
* Limited Complexity: Less effective with complex or non-linear data compared to more advanced models like neural networks.

**Decision Tree**:

* **Why Chosen**: Decision trees offer an interpretable approach to classification, where decisions are made by traversing a tree structure based on feature values. This makes it easy to understand and justify the model's predictions.
* **Intended Outcome**: The decision tree will provide clear decision rules for classifying claims as fraud or not fraud, making it useful for explaining the reasoning behind each classification.

Pros of Decision Tree

* Easy to Understand
* Useful in Data exploration
* Less data cleaning required
  + It is not influenced by outliers and missing values to a fair degree.
* Data type is not a constraint
  + It can handle both numerical and categorical variables.

Cons of Decision Tree

* Over-fitting
* Not fit for continuous variables
  + While working with continuous numerical variables, decision tree loses information when it categorises variables in different categories

#### **Relationship to the Chosen Model:**

The modelling techniques were chosen based on their ability to handle large, complex datasets and their proven track record in fraud detection scenarios. The intended outcome is to maximise the accuracy, precision, and recall of fraudulent transaction detection while minimising false positives. Each model is expected to contribute unique strengths:

* **Neural Networks:** For capturing complex patterns, non-linear patterns in data, making them highly effective for detecting subtle indicators of fraud.
* **Logistic regression Model:** Provides a simple and interpretable model that effectively handles binary classification tasks, offering clear probability scores for fraud detection.
* **Decision tree:** Offers a transparent, rule-based classification approach, allowing stakeholders to easily understand and trust the decision-making process in fraud detection

### **3.2. How will the Effectiveness and Validity be checked?**

Some issues occur when training models:

* **Overfitting/Underfitting** - accuracy of testing and training data
  + Good in training but poor on tests then overfitting
  + Poor on both training and test then underfitting
* **Curse of Dimensionality** - when you collect features, especially irrelevant features you cannot get any useful information

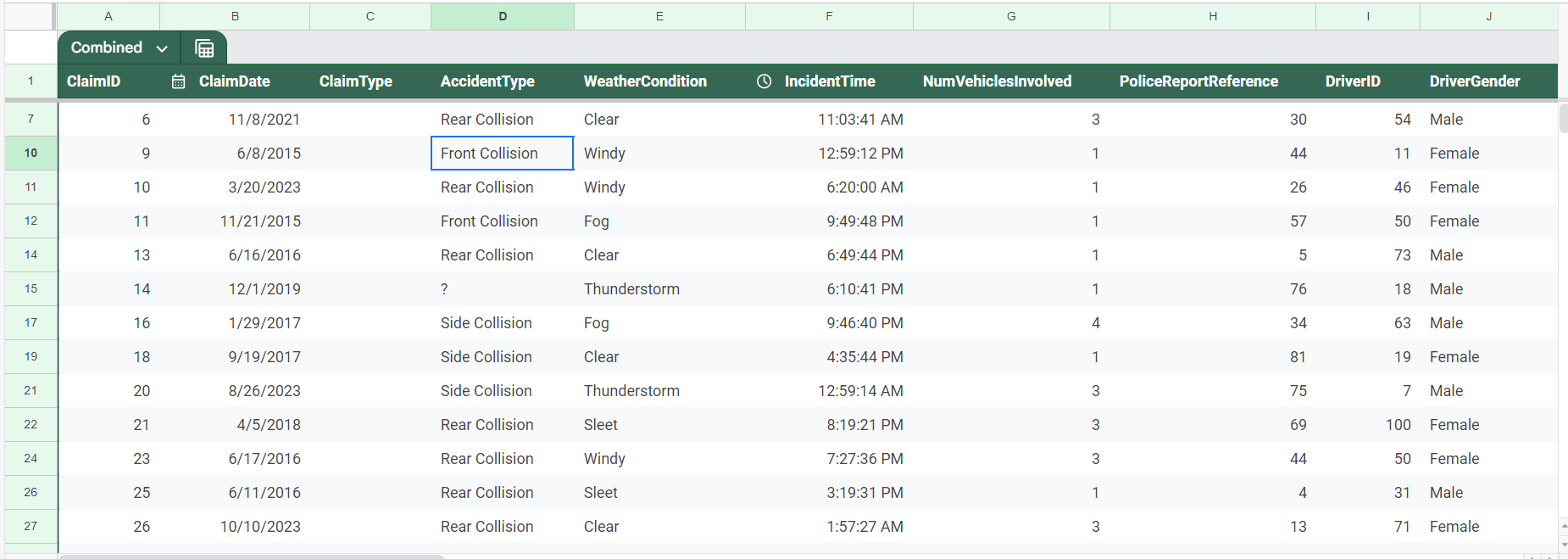
To overcome these issues and to check the validity and effectiveness of the model the following tests should be done; cross-validation, evaluation metrics, hyperparameter tuning, and external validation:

* **Cross-validation:** widely used approach for estimating test error. Estimates can be used to select the best model, and to give an idea of the test error of the final chosen model. The idea is to randomly divide the data into 𝑘 equal-sized parts. We leave out one part, fit the model to the other 𝑘 − 1 parts (combined), and then obtain predictions for the left-out part. This is done in turn for each part 1,2,. . . 𝑘, and then the results are combined. 10-fold or 5-fold is most commonly used. This allows for the assessment of the model's performance across various data sunsets resulting in the minimisation of overfitting.
* **Evaluation metrics**: Accuracy, Precision, Recall, F1-score will be used to evaluate the performance metrics of each model focusing on reducing false positives and negatives.
* **Hyperparameter-tuning:** tuning K in the algorithms we are controlling the complexity of the model thus, avoiding overfitting. This will be done using grid search or random search to fine-tune the hyperparameters of each model for optimal performance.

### **3.3. Test Data**

The models; the logistic regression model, neural network model and the decision tree model all need test data with the following requirements.

* **Balanced Dataset:** A dataset containing a balanced mix of fraudulent and non-fraudulent claims to ensure the model can accurately distinguish between the two classes.
* **Representative Features:** The test data should include the same features used during model training, such as claim amount, claimant history, claim type, and other relevant variables that could indicate fraud.
* **Labelled Data:** Each instance in the test dataset should be labelled as "fraud" or "not fraud" to evaluate the model's classification accuracy, precision, recall, and overall performance.
* **Recent Data:** The test data should be recent enough to reflect current fraud patterns, ensuring the model remains relevant and effective.



***Figure 1: Example of test data given to AI Team***

## 

## **4. Evaluation**

### **4.1. Evaluating Results**

**Null values:**

While reviewing our results, we should ensure that any rows missing any data should be null and the number of null values should also be counted. This will provide a clear understanding about the context of missing data. To solve the case of the null values, we can drop the observations with missing values. We can also use other data imputation methods such as mean, median, KNN to fill in the null values.

**Volume of Data:**

When we evaluate the volume of data, several details would need to be considered. The overall volume of data can impact the performance of the system. However, for our case, the volume of data is not extensive enough so performance should not be affected. While considering the volume of data, we have to ensure that the infrastructure is scalable. The number of rows can be increased, therefore, it has to be ensured that our system is scalable which it is in our case.

**Data Diversity:**

We have to check and analyse the variety of data. If there is a vast range of data, we can conclude that the data is diverse which will help while training the model as there is more to consider and make a decision. A low diversity of data will not be ideal as it will not consider most situations. A diverse dataset will help to avoid bias that can result in inaccurate outcomes while training the models. It is critical that the data is diverse as diverse data will provide a clearer picture which will lead to accurate outcomes.

**Gold Layer Standards:**

The Gold Layer is our most important layer. In the gold layer, the dataset will be well refined so that it can be trained by the AI team’s model. So, while evaluating the results, it is important to ensure that the data in the gold layer is consistent and organised. So, the data in the gold layer should be cleaned and there should be any biases. The gold layer data should be ready for analytical processes as the machine learning models will be trained on this dataset. So, we have to assess the performance of the gold layer by checking the data retrieval time and query speed. Moreover, the gold layer will also store the output from the AI team. So, it is crucial that the gold layer is capable of handling that process.

**Quality of Each Column:**

It should be ensured that each column represents legitimate data. For example, the dates should be actual calendar dates, times should be in AM or PM according to the clock. Most importantly, the columns should represent relevant data. While cleaning the data, irrelevant columns should be dropped. Moreover, the data should be consistent. The rows in the columns should follow the same method. For example, the dates should be in DD/MM/YYYY format. Each column should be documented which can explain the data types, ranges, transformations, etc and the columns should be clearly explained so that it is easy to understand.

### **4.2. Review Process**

While reviewing the process, it is important to ensure that the process follows the business requirements set by the client. The business rules set by our client were using a GenAI Solution, using Azure Services, Anonymise personal data, a front-end user interface, and Agile Delivery. The dataset will be reviewed by the Data team and then by the AI team who will give feedback and then a feedback loop will form to enhance the dataset. There should be data validation checks to ensure that the data is accurate. The whole process should also be documented so that each and every step in the process can be explained. The code used should also be reviewed to ensure that the process is not error prone. Finally, there should be clear communication between the Data team and the AI team and also with the client to clear any confusion and streamline the process.

### **4.3. Next Steps**

If the results are positive after the review process, we can scale up and add more rows to the dataset. Adding more rows will provide a more accurate solution as there are more rows of data to consider. As our current infrastructure is made to handle larger volumes of data, this should not be a problem and the process will be smooth. However, if the results are not positive after the review process, it may be worth revisiting the strategy and try to fix it or use a different methodology. Collaborating with the AI team and the clients will be crucial to discuss the next steps and setting up a plan for the methodology will be essential.

## **5. Deployment**

### **5.1. Deployment Plan:**

* **Deliverables to the Client:**
  + **Gold Layer Dataset:**
    - A fully processed and refined dataset stored in the Gold layer, ready for AI model training.
  + **Documentation**:
    - Comprehensive documentation detailing the data preparation process, including the steps taken to clean, synthesise, and integrate the data. This will include guidelines for accessing and using the datasets.
  + **Training:** 
    - Training sessions for the client’s team on how to interact with the Gold layer dataset, how to use the provided documentation, and how to collaborate effectively with the AI team during model development.

#### 

### **5.2. Monitoring and Maintenance:**

* **AI Model Performance Monitoring:**
  + The AI model, expected to be trained using OpenAI’s GPT-4o mini, will be closely monitored for performance metrics such as test accuracy. If the model's accuracy falls below an acceptable threshold (e.g., 85-90%), a collaborative effort will be initiated to diagnose the issue.
  + This process will involve examining both the AI model’s logs and parameters as well as evaluating the quality of the underlying data.
* **Data Re-evaluation and Refinement:**
  + **Inaccuracy Handling:**
    - If the AI model’s accuracy is found to be suboptimal, the data team will revisit the Silver layer, where a version of the combined dataset can be pulled for further cleaning and refinement. This refined dataset will then be stored in a new folder within the Silver layer and subsequently moved to the Gold layer for retraining the AI model.
  + **Overfitting Prevention:**
    - In cases where the AI model appears to be overfitting or performing too accurately, indicating a lack of diversity in the data, the data team will generate additional synthetic data. This may involve deliberately introducing errors or variations during the synthesis process to enhance the dataset's diversity.
* **Data Expansion:**
  + To improve the robustness of the AI model, the dataset may need to be expanded. This could involve generating more synthetic data or employing different synthesis techniques to create a larger, more varied dataset. If a synthesis method is found to be insufficient, alternative techniques will be explored to ensure the highest quality of synthetic data.

#### 

### **5.3. Final Report and Project Review:**

* **Final Report:**
  + A comprehensive final report will be produced, summarising the entire project, including data preparation, synthesis processes, AI model training results, and any refinements made during the deployment phase.
* **Project Review:**
  + A thorough review of the project will be conducted, assessing the effectiveness of the deployment, the quality of the data, and the performance of the AI model. Lessons learned will be documented to inform future projects.