Generative AI Consortium (Ltd)

AI/ML Internship: Assignment 1

Name: ABINAYA B

Email: abinayabala484@gmail.com

S.No	Age	Salary (INR)	Loan Approved	Credit Score	Is Outlier
1	28	45000	No	650	No
2	35	85000	Yes	720	No
3	40	95000	Yes	690	No
4	23	40000	No	630	No
5	50	150000	Yes	780	No
6	45	120000	Yes	710	No
7	30	2000000	No	850	Yes

APPLYING TERMINOLOGIES:

Feature: Individual independent variables that act like an input in your system.

Example: `Age`, `Salary`, `Credit Score`.

Label: Identification of raw data.

Example: `Loan Approved`.

Prediction: Project a probable dataset that relates back to original data.

Example: For a new record in the dataset with `Age=32` and `Salary=70000`, the model might predict `No`.

Outlier: Data that is unique/different from other data.

Example: ID=7 where `Is Outlier=Yes`.

Test Data: Ensure that the model works for the given testing data.

Example: Records of ID=6 and ID=7.

Training Data: Data that is used to train the model.

Example: Records from ID=1 to ID=5.

Model: Program that can make decisions from previously unseen datasets.

Example: Logistic Regression, Random Forest.

Validation Data: Uses a sample of data that is withheld from training.

Example: Records of ID=3 and ID=4.

Hyperparameter: Parameters that are set before training a model and controlling the learning process.

Example: The learning rate and number of trees in a Random Forest model.

Epoch: Each time a dataset passes through an algorithm, it is said to have completed one epoch. Therefore, it refers to the one complete passing of training data through the algorithm.

Example: One pass through records of ID=1 to ID=5.

Loss Function: Quantifies the difference between predicted outputs of a machine learning algorithm and actual target values.

Example: Binary Cross-Entropy, Mean Absolute Error.

Learning Rate: Tuning parameter in an optimization algorithm that determines the step size at each iteration while moving towards a minimum of a loss function.

Example: Starting with a learning rate of 0.01 and reducing it by a factor of 0.1 every 5 epochs.

Overfitting: A behavior that occurs when the learning model gives accurate predictions for training data but not for new data.

Example: If the model perfectly predicts loan approval on the training data but performs poorly on test data

Underfitting: When a model is too simple and has not learned the patterns in the training data well and is unable to generalize well on new data.

Example: If a linear model fails to capture the non-linear relationship between salary and loan approval.

Regularization: Set of methods to reduce overfitting.

Example: L2 Regularization, Dropout in neural networks.

Cross-Validation: Technique of resampling different portions of training data for validation on different iterations.

Example: k-Fold Cross-Validation.

Feature Engineering: Technique that leverages data to create new variables that aren't in the training set.

Example: Creating a new feature `Income Level` by binning `Salary` into categories like low, medium, and high.

Dimensionality Reduction: Method of reducing variables in a training dataset used to develop machine learning models.

Example: Principal Component Analysis (PCA).

Bias: Systematic error that occurs in the model itself due to incorrect assumptions on the machine learning process.

Example: If the model assumes a linear relationship between salary and loan approval, causing it to systematically miss the actual pattern.

Variance: Changes in the model when using different portions of the training dataset.

Example: A complex model that changes significantly with small changes in the training data has high variance.