Approach used:
Supervised Learning based models - This type of ML involves training the algorithm through labelled data. The algorithm receives a set of inputs along with the corresponding correct outputs and the algorithm learns by comparing the output it produces with the correct outputs to find errors. Through methods like classification, regression, prediction, and gradient boosting, supervised learning uses patterns to predict the values of the label on additional unlabelled data. Other supervised learning techniques include Support Vector Machines (SVMs), Decision Trees, Naive Bayes, Random Forest, etc.

Data Cleaning and Modification:
Data cleansing is essential to machine learning. Regardless of how sophisticated the ML algorithm is, we can't obtain good results from bad data. Depending on the dataset, different procedures and methods are used to clean the data.

Having clean data will ultimately increase overall productivity and allow for the highest quality information in our decision-making. The benefits include removal of errors when multiple sources of data are at play. This helps in fewer errors which makes for happier clients and less-frustrated employees.

1. **Redundant Columns:**
   The dataset had many unused empty columns that required deletion. Moreover there were few columns like Date, Employee Code and Machine that weren't directly related to the output and didn't give any useful information that could be used to train or improve the model performance. These columns were deleted.

2. **Incorrect Data Labeling:**
   The target column i.e the 'Defect' column had many incorrect labels. The labels consisted of some extra numbers along with a class name. The issue was solved by mapping all the labels to 0 and 1. The class names that contained 'No Defect' were mapped to 0 while the other labels were mapped to 1.

3. **Drawing Equal Samples:**
   The given dataset had a high class imbalance. Class imbalance directly influences the performance of our model. Machine Learning algorithms used for classification are designed around the assumption of an equal number of samples for each class.The class imbalance results in a model that has a poor predictive performance for the minority class.z
   The majority class had approximately 8,20,000 samples while the minority class had approximately 5700 samples. The issue was solved by taking the same number of samples as that of minority class from the majority class without replacement.

4. **Standard Scaling:**
   Standard Scaling is a highly used data engineering technique in machine learning. It

helps reduce the differences in the scales across input values which is a great factor affecting the difficulty of the problem that is being modeled.

In the given dataset, it can clearly be seen that the values of the different input classes have significant differences. The values range from anywhere in decimal values to large whole number values. The highly spreaded values cause large weight values and result in an unstable model. This makes our model poor in performance and highly sensitive to input values.

The simple technique of Standard Scaling solves this issue. The standard score (z) of a sample x is calculated as following:

$$z = \frac{x - \mu}{\sigma}$$

where μ is the mean of the training samples, and σ is the standard deviation of the training samples.

5. **Data Splitting**
   In order to test the accuracy of our models, we split our data into Train and Test datasets. The dataset was split in the ratio of 60:40 where 60% of the data was used for training and the rest, 40% was used for testing.

Business impact:

1. We have been able to achieve a high model accuracy along with a high F1 Score. This means that we will have really good detection of welding defect cases and very low false detection of welding defect cases.

2. The computational requirements for our model to run are very low. We don't require a high amount of computational power due to which one doesn't need to invest in expensive computers or machines with high end chipsets like Intel i7 or AMD Ryzen 7 and above. Our model can work on a chip of quality as low as a Raspberry Pi 4 or Nvidia Jetson Nano.

3. The low computational power requirements also affect our model prediction time. It is able to give highly accurate outputs in a small amount of time. This helps in saving time as well as conserving human efforts.

4. The above factors also lead to greater advantage, that is, it has equal to no maintenance and running cost. This helps the company in the long run as it only needs to invest

money once for a long term solution.

Personalization of the Customer Experience:
Machine Learning is handy for attracting more clients to any product or service and converting them into repeat customers. It paves way to examine clients' browsing experience and activity on a platform in order to fulfill customers' requirements and desires.

Reasonable Resource Management:
A corporation can estimate the resources needed to fulfill shifting demand for its products or services based on machine learning forecasts. Knowing ahead of time what your consumers anticipate from your firm will aid you in inventory and process management.

Enhanced Data Security:
The effective growth of a corporation depends greatly on the security of the firm and its clients. Numerous firms worldwide use machine learning to guarantee payment security. The

Models used:

1) Random Forest Classifier:
Random forest classifier consists of a large number of individual decision trees that operate as an ensemble. It uses bagging and feature randomness when building each individual tree to create an uncorrelated forest of trees. Each individual tree in the random forest splits out a class prediction and the class with the most votes becomes the model's prediction.

Classification Table

	precision	recall	f1-score	support
0	0.772497	0.838377	0.804090	1454.000000
1	0.819092	0.747716	0.781778	1423.000000
accuracy	0.793535	0.793535	0.793535	0.793535
macro avg	0.795794	0.793046	0.792934	2877.000000
weighted avg	0.795543	0.793535	0.793054	2877.000000

Fig - ROC Curve

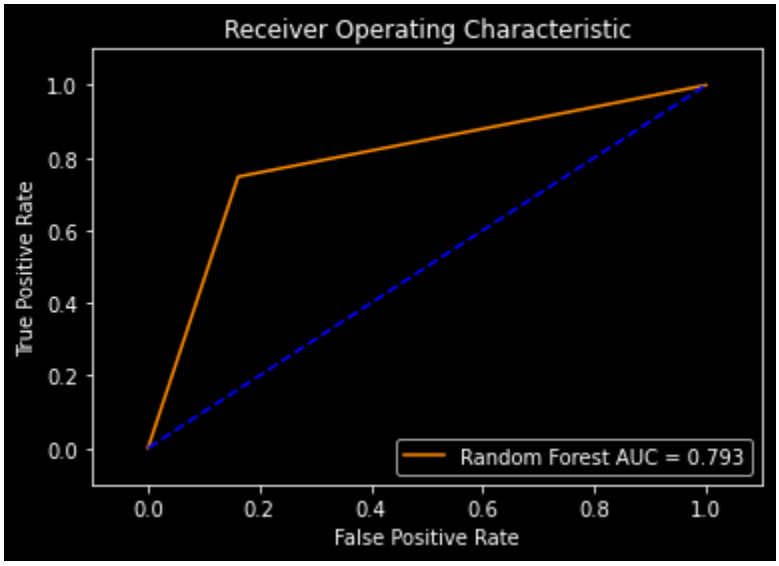
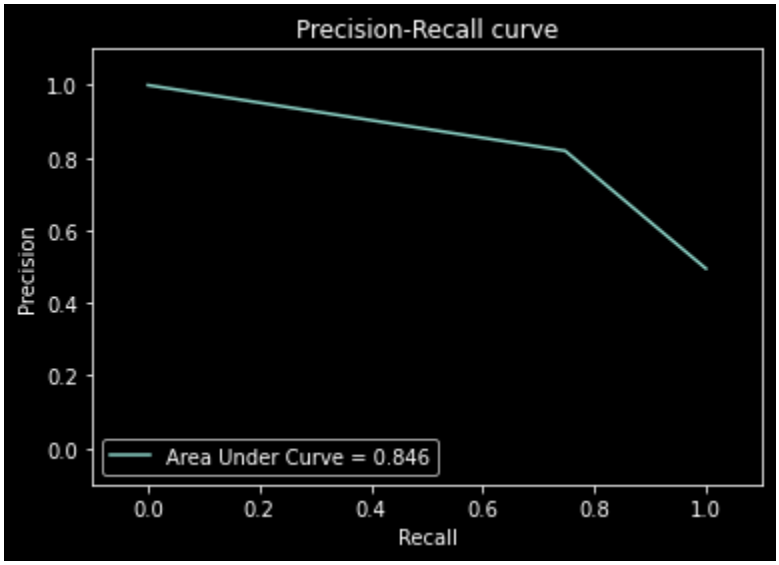


Fig - Precision Recall Curve



2) Gradient Boosting:
A supervised learning technique which builds models sequentially, each one having less errors compared to the previous one. It relies on the intuition that when prior models are coupled with the best feasible upcoming model, the overall prediction error is minimized. The key idea is to set the target outcomes for the next model in order to minimize the error.

Classification Table

	precision	recall	f1-score	support
0	0.773256	0.823246	0.797468	1454.000000
1	0.806622	0.753338	0.779070	1423.000000
accuracy	0.788669	0.788669	0.788669	0.788669
macro avg	0.789939	0.788292	0.788269	2877.000000
weighted avg	0.789759	0.788669	0.788368	2877.000000

Fig - ROC Curve

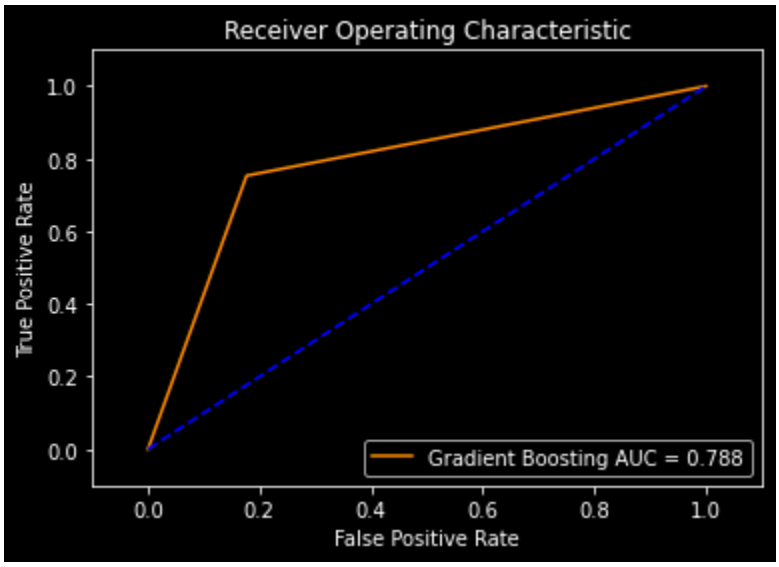
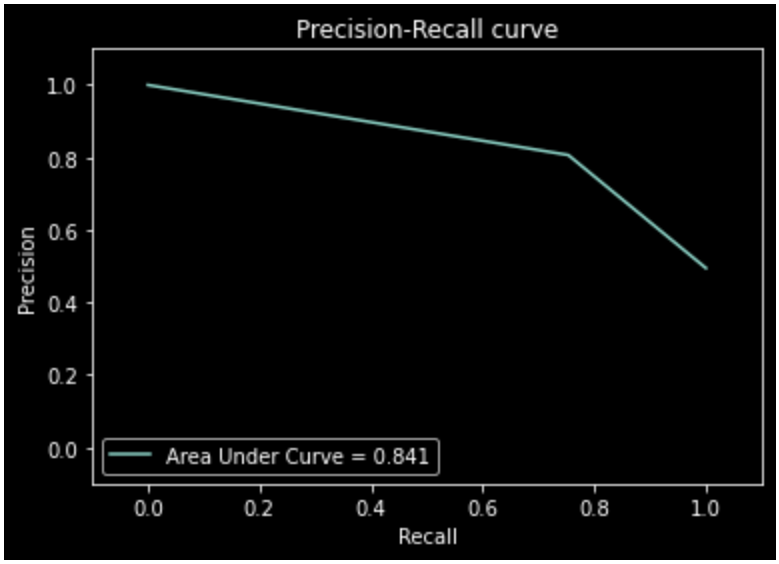


Fig - Precision-Recall Curve



3) Extra Trees:
A supervised learning technique that fits a number of randomized decision trees on various samples of the dataset and uses averaging to improve the accuracy and control overfitting of the model. It combines the predictions from many decision trees.

Classification Table

	precision	recall	f1-score	support
0	0.777704	0.810867	0.793939	1454.000000
1	0.797943	0.763176	0.780172	1423.000000
accuracy	0.787278	0.787278	0.787278	0.787278
macro avg	0.787824	0.787021	0.787056	2877.000000
weighted avg	0.787715	0.787278	0.787130	2877.000000

4) Decision Tree:
A supervised learning approach and a non parametric algorithm which best fits the training data in constructing the mapping function. We used a regressive decision tree keeping in mind the nature of the dataset and other requirements.

Maximum F1 score achieved: 0.804090 (Random Forest Classifier)