Weldright Round 2 Submission - Documentation- IIT Bombay TechFest 2022

Team ID: HA-226956
Team Members:
Sejal Kotian, Safdar Inamdar, Saad Ahmed, Akshita Mittal

Problem Statement: -

Considering the importance of the operation, flawless welds are very much desired. Flawless welds largely depend on the procedure. Thus, the process parameters must be monitored to achieve the desired accuracy in welds. Therefore, we are required to use machine learning to predict the kind of defect that could get infested during the procedure, given the process parameters.

About the Dataset:-

The dataset provided consists of **827534** instances, with each having 11 features, namely *'Employee code', 'Machine', 'Production', 'Order Operation No', 'Date', 'Time', 'Current', 'Humidity', 'Temperature', 'Flow', 'Job Temp', 'Voltage', '**Defect'.***
Of these, the target column is the 'Defect' column which consists of the three classes: 'No Defect', 'Tungsten Inclusion', and 'Porosity'.
Since the data is taken from industrial applications, it is a highly imbalanced dataset with **99.3% data consisting of No Defects** while **0.56% Tungsten Inclusion** and **0.14% Porosity**.

This is a **Multiclass Imbalanced Data Classification problem.**

Data Analysis and Cleaning:

An extensive EDA (Exploratory Data Analysis) was performed on the dataset:
- Anomalous readings in the Defect column were rectified
- Rows with missing Order Operation Number, and Production Number were deleted (740 rows).
- The features 'Employee code', 'Machine', 'Production', 'Date', and 'Time' were dropped from the feature set.

On plotting the scatter plot for various features in the binary data frame obtained from the pair of each of the classes, it was noticed that the data points of Porosity and No Defect were quite overlapping. Thus it was not possible to discern between the points. This was unlike the case with No defect and Tungsten Inclusion, whose feature points were quite distinct.

Work Done till Round1 Submission:-

Objective of the classification- high recalls for the defective classes.
Observing the imbalance in the dataset, it was concluded that a sampling operation is to be performed for proper fitting of the models.
Training the models- data was split into train and test sets in a ratio of 0.25; got the validation results on the test set.
Several models were implemented initially, considering the classification to be multiclass. However, due to the imbalance, the results were disappointing. Without any form of sampling, the models were being overfitted on the 'No Defects' class due to the vast number of instances for such.

Dealing with the imbalance:
1. SMOTE technique was used for upsampling the minority classes with various sampling ratios.
2. RandomUnderSampler technique was used for downsampling.
3. Subsampling with ensembling was also implemented, but the results were not up to the mark.

We realized that training the model with multiple classes in the dataset was not giving the desired results.

Thus, we switched to binary classifications using the OvO (One vs. One strategy). It was noted that working on a binary dataset is easier and meticulous tunings could be done to achieve the optimum results.

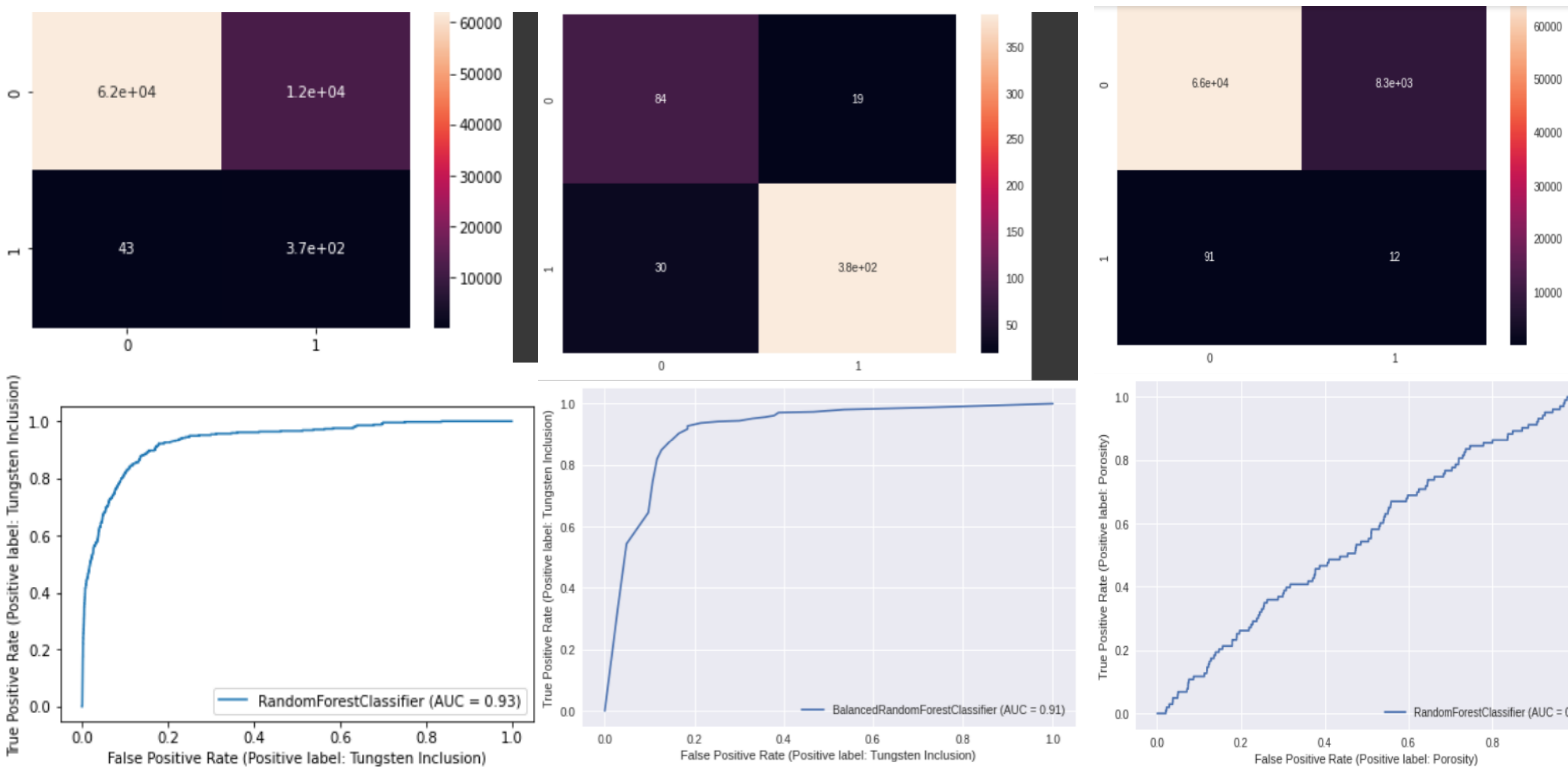
The dataset was divided into 3 data frames with {('No Defect', 'Porosity') ('No Defect', 'Tungsten Inclusion') ('Tungsten Inclusion', 'Porosity')} combinations without any sampling. Several algorithms were run, like SVM, RFC, GaussianNB, and KNN, along with subsampling and max voting ensemble classification, and binary classification was performed for training with these combos.

Results:
We focused on improving the evaluation metric, **Recall** rather than Accuracy since our goal is to reduce the False Negatives i.e. the model must not falsely predict a defective weld as 'No defect'.

Method	**Precision**	**Recall**	**F1 score**	**Support**
RFC Model (Without Sampling)				
No Defect	0.99	**1.00**	1.00	205440
Porosity	0.00	0.00	0.00	297
Tungsten Inclusion	0.76	0.25	0.38	1147
RFC with RUS(Ratio =1)				
No Defect	1.00	0.48	0.65	82164
Porosity	0.00	0.44	0.00	114
Tungsten Inclusion	0.03	**0.83**	0.07	461
RFC with SMOTETomek Link				
No Defect	1.00	**0.98**	0.99	205422
Porosity	0.01	0.03	0.01	266
Tungsten Inclusion	0.17	0.40	0.23	1196
Binary Classification Dataframes				
RFC with RUS(ratio=0.2), SMOTE(ratio=0.85)				
No Defect	1.00	**0.84**	0.91	73895
Tungsten Inclusion	0.03	**0.90**	0.06	415
Balanced RFC with RUS(0.75)				
Porosity	0.74	**0.82**	0.77	103
Tungsten Inclusion	0.95	**0.93**	0.94	415
RFC with RUS(ratio =1)				
No Defect	1.00	0.45	0.62	73894
Porosity	0.00	0.54	0.00	103
RFC with RUS(ratio =0.9) and SMOTE(ratio=1)				
No Defect	1.00	**0.89**	0.94	73894
Porosity	0.00	0.12	0.00	103
SVM with RUS(ratio=1)				
No Defect	1.00	0.14	0.25	73894
Porosity	0.00	**0.91**	0.00	103

SVM with RUS(ratio=0.9) and SMOTE(ratio=1) No Defect Porosity	1.00 0.00	**0.97** 0.04	0.98 0.00	73894 103

Confusion matrix heatmap and ROC_AUC curve for Nd(0)-Ti(1), P(0)-Ti(1), Nd(0)-P(1) pairs respectively



Result and Conclusive remarks:-

Based on the rigorous work done to minimize the False Negatives, it was found that for Nd-Ti and Ti-P pairs satisfying results were obtained but unfortunately with whichever technique put to use, appreciable results are not achieved for Nd-P pairs. **From an in-depth analysis of relation b/w the various features for Nd-P pair were performed using scatter plotting, we reached the conclusion that data points for this pair have a very great extent of overlapping due to which obtaining a high recall for both is getting quite arduous**. Along with it tuning the ratios for sampling lead to either high overfitting of one or the other or gaining 0.5-0.5 recall which signifies that the model is performing random guesses. So, our main challenge is to build a model that can learn the difference, especially between these two classes of data, and distinguish between the process parameters of the two classes. If so happens, a MaxVoting classification could be performed on the predictions from the respective models of the three pairs.

So for the time being, the biggest issue to tackle is the proper classification of Nd-P pairs.

From a metallurgist point of view, after performing an exhaustive exploration of various algorithms with different sampling ratios, proper classification of Nd-P pairs has still not been achieved. This led us to believe that **maybe the given process parameters might just not be enough for the successful classification of Nd-P pairs** which means that there can or maybe some other process parameters that if taken into account could lead to promising results, leading to a robust classification for the three classes.

Further improvisation will be done by rigorous feature engineering and deploying various feature selection techniques, and in-depth analysis of the Nd-P pairs for their successful classification.

Colab Notebook:https://colab.research.google.com/drive/1_fO_HTxl-id13SwjH1Feyc0pl4phSFQh?usp=sharing

Further Work Carried Out:

Dealing with Overlapping Data Points between Porosity and No-Defect:
In order to deal with the overlapping data points to classify porous weld from the non-porous weld, we implemented

- Linear Discriminant Analysis
- SHAP method
- Feature-wise prediction

While carrying out feature-wise predictions to specifically discern between **No Defect and Porosity Pairs,** we observed that the **best parameters** to classify them are **Flow** and **Job temperature** with a **recall** of **0.52 and 0.59** respectively using EasyEnsembleClassifier with KNN as base estimator.
This algorithm involves undersampling based on overlap. A recall of more than 0.5 for both simultaneously suggests some level of predictability for such a high level of overlapping data.

Analysis:
Effect of Gas Flow Rate: The more powerful the flow of gas, the more turbulence in the air, leading to contaminants mixing with the weld puddle, which in turn causes a porous weld. Thus, an optimum flow rate is required to obtain a non-porous weld.

Effect of Job Temperature:
The fundamental problem in welding titanium alloys is the elimination of atmospheric contamination. Contamination of the weld metal and the adjacent HAZs (Heat Affected Zones) will increase tensile strength and hardness but may reduce ductility to an unacceptably low value. The most likely contaminants are oxygen and nitrogen, picked up from air entrained in the gas shield or from impure shield gas, and hydrogen from moisture or surface contamination.

During welding those parts of the weld exposed to **temperatures above 520°C** will absorb oxygen and nitrogen and must therefore be protected until they have cooled below this critical temperature.

Thus, temperature affects the amount of moisture in the weld. A preheated job would have less moisture and thus less porosity.

Business Perspective

TCO:
The Cost of Ownership of the Project will be the cost of deploying the project onto an existing production environment which, on feeding the model, returns an output that can be used in the industry for decision making. One such Production Environment is Amazon Web Services (AWS)
Considering the current AWS SageMaker Pricing for 6GB RAM and 64 GiB Memory for Amazon EC2 running on a Linux system, the cost is $0.72 per hour.

Specifications for the above lowest cost-EC2 instance are:

Server: r5d.2xlarge
Operating System - Linux
$0.72 - On-demand hourly cost
vCPUs - 8
GPU - NA
RAM- 6GB
Memory(GiB) - 64GiB

Network Performance - Upto 10GigaBit

Thus, TCO for the Model is 11285.76 USD p.a. which is 9,19,112.29 Rs per annum (considering 100% utilization/month)

Measuring the Impact of the model

The dataset has a *99.3: 0.14*: *0.56* imbalance.
Output of the model: *Classification as a defect or no defect*
Motive: to predict a defect before it occurs as per the input parameters
Performance Metric chosen: Due to the heavy losses incurred due to a product being falsely predicted as Not Defective, the performance metric chosen here is *Recall,* as we are striving to minimize the False Negatives. The reason recall was chosen over accuracy is because it is safer to classify a non-defective weld as defective instead of classifying a defective weld as non-defective. The latter would incur huge losses for the company.

RFC with SMOTETomek Link	**Precision**	**Recall**	**F1 Scores**	**Support**	**Support proportion**
No Defect	1.00	**0.98**	0.99	205422	0.9929332
Porosity	0.01	0.03	0.01	266	0.0012857
Tungsten Inclusion	0.17	0.40	0.23	1196	0.0057810
		Weighted Average of Recall = **0.9754255**	**Weighted average of F1 Score =** **0.9843464**		

Interpretation of Performance Metric:
We obtained a 0.9754255 Weighted Recall Average.
This essentially means that out of a 100 instances of Defects, Porosities and TIs, the model can correctly predict 97 of these correctly.

Calculating the ROI:

If a defective weld has been correctly predicted as defective, the cost incurred otherwise would be greatly reduced.
On the correct prediction of defective weld parameters, the cost incurred due to the production of defective weld products can be mitigated.
Thus, a rough estimate of the ROI can be calculated as follows, based on the industrial costs and standards:
ROI = (profit - cost) / cost
Profit = Amount saved by avoided Defective Products (this includes defects + porosities + Tungsten Inclusions)
Cost = Cost of deployment of model + Cost incurred due to wrongly predicted weld defects