Abstract for the WeldRight Competition

Prakriti Shetty, 200020095, Chemical Engg, IIT Bombay

Background:
Welding is an essential and critical activity for the manufacturing activities of Godrej Aerospace. With proper real-time weld monitoring methods, weld defects are expected to be identified quickly and weld parameters can be adjusted accordingly

Goal:
Come up with an ML model to predict welding defects in the materials by developing algorithms

Dataset Description:

Machine Data		Process Parameters	
Employee Code	*(Office id of employee)*	Current	*(In Ampere)*
Machine	*(Welding machine name)*	Humidity	*(Relative humidity in %)*
Production	*(Project order number)*	Temperature	*(in degree celsius)*
Order Operation No	*(This captures the activity to be performed by the operator as per the job description card)*	Flow	*(in liters per min (LPM))*
Date	*(date of activity)*	Job Temp	*(in degree celsius)*
Time	*(Timestamp for the activity)*	Voltage	*(in volts)*
Employee Code	*(Office id of employee)*	Defect	*(if the defect occurred)*

We notice that since we have a column for 'Defect', we are tackling a supervised ML problem.

STEP 1: LITERATURE REVIEW

Zhou, B., Pychynski, T., Reischl, M. et al. Machine learning with domain knowledge for predictive quality monitoring in resistance spot welding. J Intell Manuf 33, 1139–1163 (2022). https://doi.org/10.1007/s10845-021-01892-y

Exactly our use case, to predict defects in resistance spot welding by use of actual production data, with model interpretation interspersed with domain knowledge.

Asif, K., Zhang, L., Derrible, S. et al. Machine learning model to predict welding quality using air-coupled acoustic emission and weld inputs. J Intell Manuf 33, 881–895 (2022). https://doi.org/10.1007/s10845-020-01667-x

We compare two methods for prediction: (1) logistic regression and (2) adversarial sequence tagging. Existing research that leverages machine learning for weld quality prediction tends to use only the features at a specific point in time, without considering the predicted weld quality in preceding time steps. However, when there is a possibility of spurious variations of signals over time, a sequential model that accounts for predictions made at preceding time steps generally offers a more stable model. As a standard, generic, and non-sequential model, we use Logistic Regression (LR), which uses a softmax function to predict weld quality (target label) probabilistically. For the sequential approach, we use Adversarial Sequence Tagging (AST) that uses both instantaneous features and previous predictions. This method uses a minimax game to

find the probability distribution for the prediction that is robust to all possible welds producing a similar sequence of features.

Shuo Feng, Huiyu Zhou, Hongbiao Dong, Using deep neural network with small dataset to predict material defects, Materials & Design, Volume 162, 2019, Pages 300-310, ISSN 0264-1275, https://doi.org/10.1016/j.matdes.2018.11.060.

DNN trained by conventional methods with small datasets commonly shows worse performance than traditional machine learning methods, e.g. shallow neural network and support vector machine. This inherent limitation prevented the wide adoption of DNN in material study because collecting and assembling big dataset in material science is a challenge. In this study, we attempted to predict solidification defects by DNN regression with a small dataset that contains 487 data points. It is found that a pre-trained and fine-tuned DNN shows better generalization performance over shallow neural network, support vector machine, and DNN trained by conventional methods. The trained DNN transforms scattered experimental data points into a map of high accuracy in high-dimensional chemistry and processing parameters space. Though DNN with big datasets is the optimal solution, DNN with small datasets and pre-training can be a reasonable choice when big datasets are unavailable in material study.

Shin S, Jin C, Yu J, Rhee S. Real-Time Detection of Weld Defects for Automated Welding Process Base on Deep Neural Network. Metals. 2020; 10(3):389. https://doi.org/10.3390/met10030389

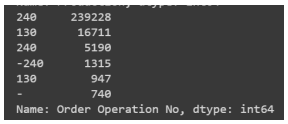
Explains the various process variables

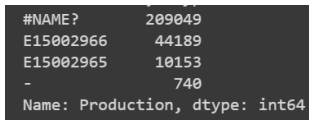
Work done so far:

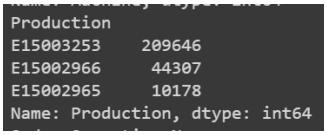
I have started work on Step 2 as follows, completed the literature review, missing value identification and handling, and recategorisation. The correlations and visualisations is where I'm currently at.

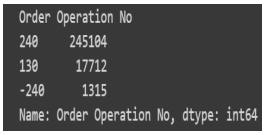
<u>STEP 2:</u> DATA PREPROCESSING: DATA WRANGLING AND EXPLORATORY DATA ANALYSIS

1. Handling missing values:
    a. The 'Current' column had 3 empty values, since this is insignificant with respect to the ~2.5lakh entries, I just imputed it with the most frequent value
    b. The 'Production' column and 'Order Operation No' had about 740 missing values(denoted with a hyphen). The 'Employee code' had ~66K unfilled values(denoted by 0). All these values are significant but fall within the range of imputation being a reasonable choice. Hence, I took the top 10 most frequent values, judged their probabilities of occurring again, and then carried out the imputation.



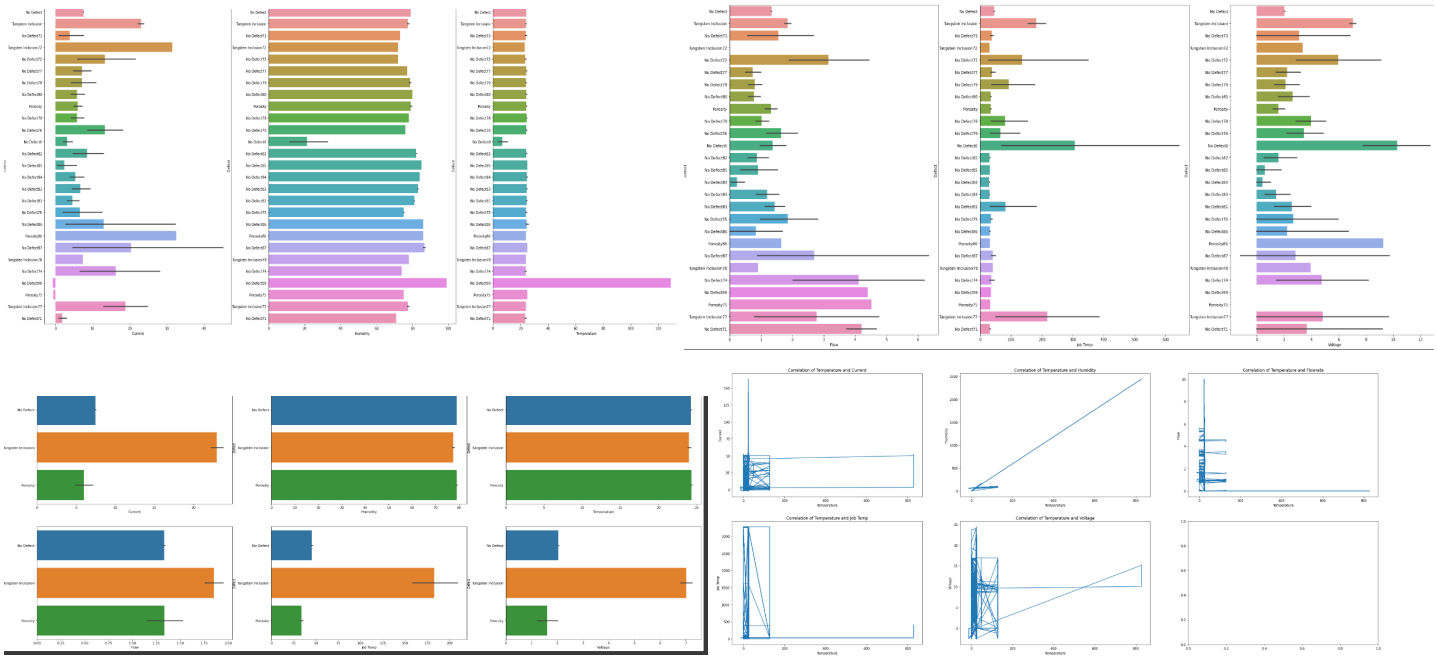


before => after




2. Recategorisation
    a. The 'Production' column had a little inconsistency (~25%) in the values of the column, to be specific: "E15003253" was being considered as an unrecognised label because it was entered in the formula format "=e15003253".
    b. The defects column had certain suffixes on some labels Eg NoDefect59, NoDefect60 etc. On probing, I realised they were depicting the humidity levels for the examination. Since creating a different label with just a different subscript doesn't help me in classifying any better, I decided to get rid of those suffixes and group it under more broad labels of 'Tungsten Inclusion', 'Porosity' and 'No Defect'.
3. Correlation between variables (Questions and hypothesis testing)
    a. Question: Does each variable, taken separately, affect the prediction of a a particular type of defect?

i. Current, Voltage: Yes, found an almost uncontested probability of a Tungsten Inclusion defect in the case of very high currents/ voltages.
ii. Humidity, Temperature, Flow: No substantial causation observed
iii. Job temp: There is a slight inclination towards tungsten inclusion at higher job temperatures, but nothing definitive can be said



b. Question: Is there a correlation between any of the process variables?
   i. Yes, there is a direct relationship between temperature and humidity, so I plan to drop one of these during model building
c. How is the variation between voltage and current, is it ohmic?
   i. No, it is not a linear relationship as would be expected if it was an ohmic device

4. Data Visualization and Statistical Analysis
   a. I analysed correlations between variables and causation effects of all the process variables on the prediction of particular defects through barplots, lineplots.
   b. I also extracted the statistics for each of the numerical variables.

<u>STEP 3</u>: MODEL BUILDING

Before model building: train- test split, and a transformers pipeline for encoding of categorical variables and scaling of numerical columns.

Models that can be used:
Linear Regression,
Multi class Classification algos like- Multi Layer Perceptron with 1 hidden layer, Logistic regression, KNNs, CART, random Forest, gradient boosting etc and support vector machines

Approach: I plan to build an extensive model pipeline, that judges the best model that can be used, alongwith the most optimal hyperparameters. The final model will be selected based on model metrics like the confusion matrix etc and the graphs I plot depicting model performance with changing hyperparameters.

Reasons for choosing :
Linear Regression- Selected as a representative of classical ML, Industry's favorite, primarily because of the transparency involved in the working
KNNs- although they are affected by outliers, easy to implement and responds quickly to changes in input data, which is vital in an industry setup.
SVMs - Provides good scaling of high dimensional data and the non constant weights of each variable adds more entropy to the prediction of the output.
XGBoost - provides optimisation through functions rather than parameters, making it easy to use custom functions. Also does a cross validation at each iteration of the process, thus optimizing the number of iterations
MLP- Although training is time consuming, it can be applied to solve complex non linear problems with large input datasets.
CART - Unlike MLP, it assigns a specific value to the inputs and output of each decision of the problem thus every probability can be evaluated.
Reasons for discarding :
Logistic regression - doesn't work well with correlated variables
Random Forest- makes sense to do much harder computations than decision trees only when the benefit of working with sparse datasets is to be availed- not required here.

A final thing to note is that till now, I've been talking about tackling the process variables and their judgement of whether there is an impending defect or not.
In a future plan of work, I'll also start analysing the machine data and perform advanced analytics on the welder performance, machine utilisation, return on investment calculations, and total cost of ownership calculations. I've not quite figured out a plan of action for that as yet, but I'll definitely make that in the successive weeks.

That's a whole overview of my game plan, I really hope it covers the ground you were looking for, cheers!

Note: The entire code is available at the following link: TechfestWeldRight