Project Proposal

1. Project title:

Determine the type of fault that occurs on steel plate.

2. Data set:

Data could be got from http://archive.ics.uci.edu/ml/datasets/steel+plates+faults. There are totally 1941 number of instances in the data. It has 34 columns, among which 27 are attributes and the rest 7 are targets.

The dataset is already been one-hot encoded. All the values inside are numeric, either integer or float . The last 7 columns stands for the targets, which are the 7 types of Faults occur along the production process of steel plate.

Attributes:

0	X_Minimum		
1	X_Maximum	14	Edges_Index
2	Y_Minimum	15	Empty_Index
3	Y_Maximum	16	Square_Index
4	Pixels_Areas	17	Outside_X_Index
5	X_Perimeter	18	Edges_X_Index
6	Y_Perimeter	19	Edges_Y_Index
7	Sum_of_Luminosity	20	Outside_Global_Index
8	Minimum_of_Luminosity	21	LogOfAreas
9	Maximum_of_Luminosity	22	Log_X_Index
10	Length_of_Conveyer	23	Log_Y_Index
11	TypeOfSteel_A300	24	Orientation_Index
12	TypeOfSteel_A400	25	Luminosity_Index
13	Steel_Plate_Thickness	26	SigmoidOfAreas

Targets:

27	Pastry	
28	Z_Scratch	14
29	K_Scatch	15
30	Stains	16
31	Dirtiness	17
32	Bumps	18
33	Other_Fault	19

3. Project idea:

1) What problem are you going to be tackling on your project?
In this project, I will first study the relationship between the features and the target, explore the patterns inside, it can tell me each type of faults is more relevant to which features.

Next, I will do some data preprocessing jobs like fulfilling the missing data, getting rid of outliers and reducing data dimensionality by adopting PCA. Then I will train models with the data using different algorithms, such as Logistic Regression, Decision Tree, Naive Bayes, K-Nearest Neighbors and Random Forest that I learned in this course. After that I will compare the performance of the different models and find out the best of them. With the model at hand, I could better distinguish between different fault types based on the observation of all the features.

2) Why is that an interesting/useful application of data science?
The dataset is obtained from the actual production of the steel plant, and there is need to tell apart different type of faults, since this could help the plant find out the root cause of each type of faults, then they could come out with solutions and finally produce good steel plate.

3) What models are you envisioning training to address that (e.g., classification, regression, clustering)?
Classification, and I could also try Clustering with the data.

4) What will a user-facing service that packages your model(s) look like and how will you make it user-friendly for someone to leverage your work?
After a better trained model is got, I could export it and make it into an interface, such as create a function with the model and make it receiving new data as input. In the way, the other users who would like to check up the fault types could leverage my work by calling this function, then the predict result could be given back to the user as output.