Name: Sushant Patil

1. Sources

We will use Pandas, Numpy, Scikit Learn, eli5, re libraries to work the given csv file dataset. Much of the data cleaning process and model fitting is referred from Introduction to Data Science in Python (Coursera) and Kaggle Micro Courses on Pandas, Intermediate Machine Learning,etc and Google of course.

2. Data Cleaning

Load the dataframe as train and test and observe the table
Filter the training dataset for success probability between [0,100]
Filter the testing dataset for Internal rating between [1,5]

Survey the categorical columns such as ‘Resource’ to reduce the cardinality for both the datasets. Similar action the ‘Designation’ column as well

Look at all the columns containing null(Nan) values. Note down the column type as well.

a) Filter the ‘Deal value’ and ‘Weighted amount’ column to remove the $ sign and convert the columns as float datatype. About dealing with Nan values, fill those with mean of their respective column
b) Fill Nan values for ‘Industry’ , ‘POC name’ , ‘Location’ as ‘Unknown’ thus, we induce another categorical datatype into these columns
c) Fill Nan values for ‘Geography’, ‘Last lead update’ , ‘Resource’ with backward/forward filling whichever results in eradication of null values

Repeat process a) b) c) for testing dataset as well. At the end, recheck if any null values are present.

Now, we process our categorical datatypes to convert them to numerical datatypes

Filter out the good label columns from categorical columns as their cardinality matches. This means we can easily encode these columns for conversion using LabelEncoder.

3. Model Fitting

We will be using the RandomForestRegressor model and tune the hyperparameters to reduce overfitting on the training model. This is done so that the model performs better on predicting the unseen test dataset.