High Level Assignment

EDA

Feature	Total Count	Unique Count	Null %
Title	138724	138724	2.00%
Description	138724	116931	9.00%
Authors	138724	97801	0.00%
Publisher	138724	12855	0.00%
Publisher Date	138724	10819	0.25%
Categories	138724	100	0.00%

Target Variable “**IMPACT**” is Normally Distributed with **Mean 786 and SD 63**
A Model giving only Mean as Output will have ~ **8% MAPE** (Baseline Value)

Modelling

Large Data with Multiple Dimensions - **Random Forest** is the most Robust Model

Version 1 - Predicted Impact of books with straight forward Features-

- Number of Characters , Number of Words , Avg Word Length in Title
- Number of Characters , Number of Words , Avg Word Length in Description
- Number of Authors
- Published Year

Version 2 - Added Features Based on EDA-

- Author Type based on Number of Books written (One Hot +Rule based)
- Publisher Type based on Number of Books Published (One Hot + Bins)
- Category Type based on Number of Books written (One Hot + Quantiles)

FUTURE ADDITIONS TO IMPROVE MODEL

Version 3 - Add NLP Features based on Title and Description-

- TF-IDF
- Word Embedding
- N-Grams
- NER

Performance Analysis

No. Of Worker	Version 1		Version 2	
	Cross Val MAPE	Training Time	Cross Val MAPE	Training Time
1	0.0606984	247.3584790	0.0595706	235.5348732
2	0.0606704	149.6760671	0.0595065	224.7799189
4	0.0606687	137.2030070	0.0595308	137.9298708