Facebook Comment Prediction

Introduction to the business problem

- In this project, we will identify the dynamic behavior of users towards a post shared on Facebook, we will try to understand how a user comments based on a post or how many comments a post can receive in a certain period.
- We will do this to predict what type of posts receives the highest number of comments or reach. So that the advertisements can be promoted in those posts.
- The User Generated Content (UGC) can be classified as below in Table 1

		Psychological Motivation for Engaging in UGC creation			
		Rational		Emotional	
		Knowledge Sharing	Advocacy	Social Connection	Self-Expression
Platform Base	Group	Wikis (e.g., Wikipedia)	Issue-Centric Communities (e.g., Rachel ray Sucks Community)	Multiplayer online games (e.g., PUBG)	Virtual Presences (e.g., SecondLife)
	Individual	Blogs by experts (e.g., askanexpertblog.com)	Consumers reviews (e.g., Epinions)	Social Networking sites (e.g., Facebook)	Consumer creative inventions (e.g., jumpcut)

Table 1. Typology of UGC Classification and Exemplars

Data Insights

- The data has 43 columns and 32,760 rows
- The data has been collected daily for each posted content
- The columns CC1-CC5 are the count of comments for the post on different timelines like 24 hours after the content is posted
- Column **ID** is a continuous variable, and it is of no importance to us for the prediction
- Out of **32,760** records, only **634** records have values other than **24** in the **"H Local"** column.
- Column **Post Promotion Status** has only values 0 in it and it won't help us in predicting.

Exploratory Data Analysis

- Most of the Independent Variables are **Left Skewed**, most variables have value

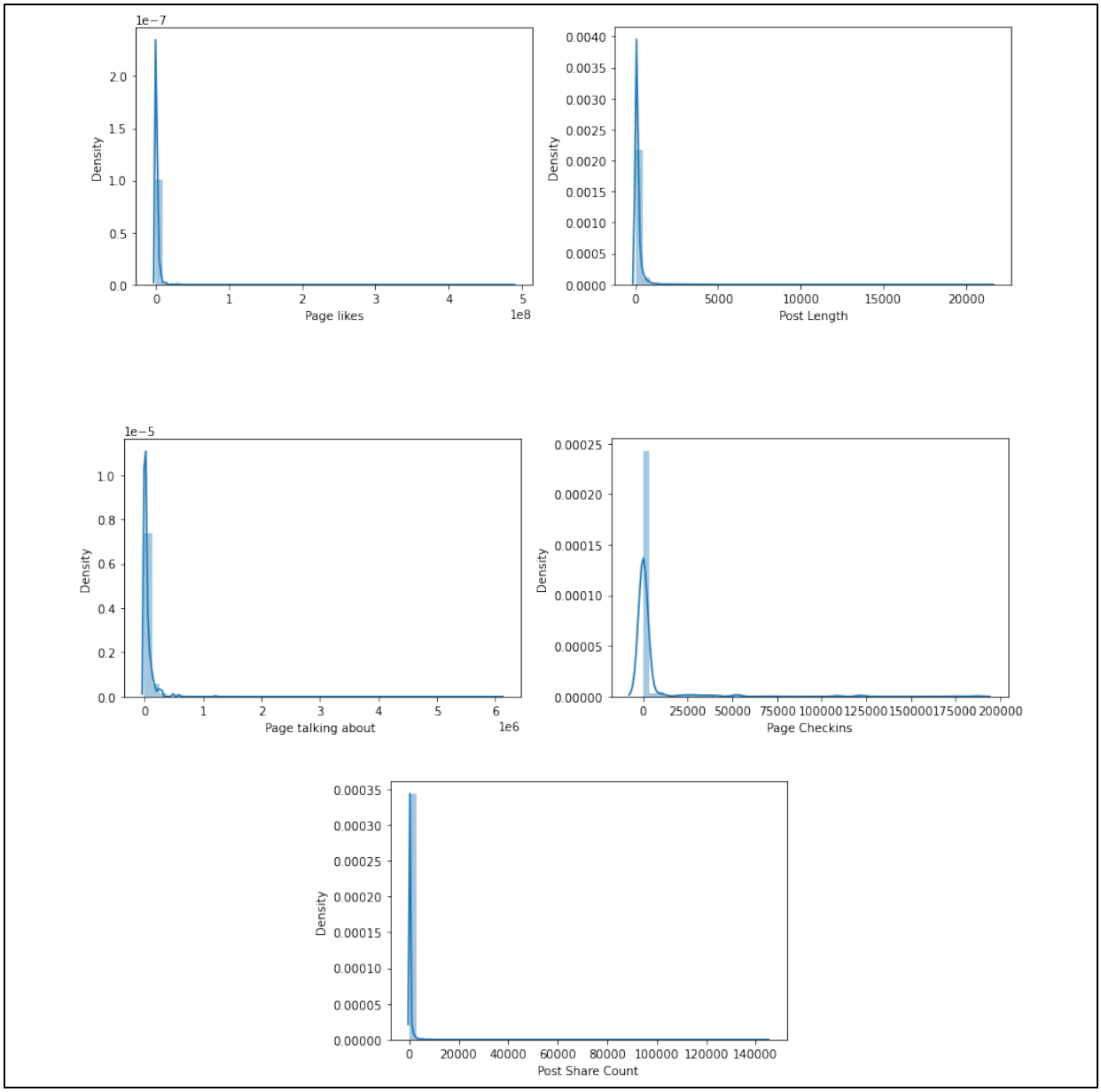

Figure 1: Distribution of Independent Variables

- Column Post Length can be converted into a categorical variable as length less than 100 can be termed as Short Post, post length greater than 100 and less than 500 can be termed as Medium Post and above 500 can be termed as Long post and create a new column postSize.

- The categorical variable Page Category has more density between 0 and 40

Figure 2: Page category Density

- Post posted on Wednesday has received the greatest number of comments, which can be visualized in below figure3

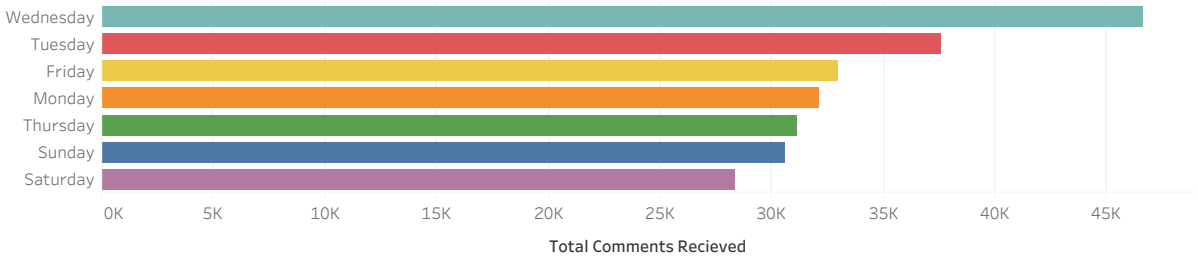

Figure 3 Weekday vs Comment Received

- Correlation on different page features can be understood from the below figure 4

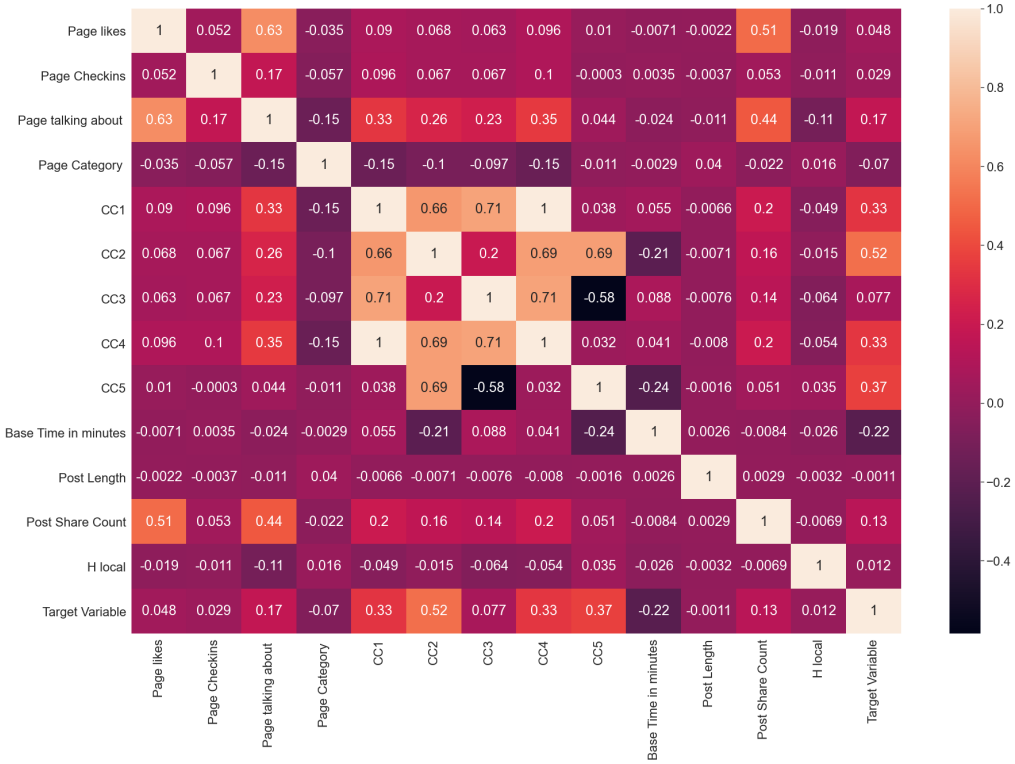

Figure 4: Page Features Correlation

- From figure 4 we can interrupt that Variable CC1 and CC4 are highly correlated
- Also, we can see there is an inverse correlation between CC3 and CC5
- And CC1 is closely related to CC2 and CC3. CC2 is closely related to CC4 and CC5. CC3 is closely related to CC4.
- To Support our correlation identification, refer to the below figure 5 how data are distributed.

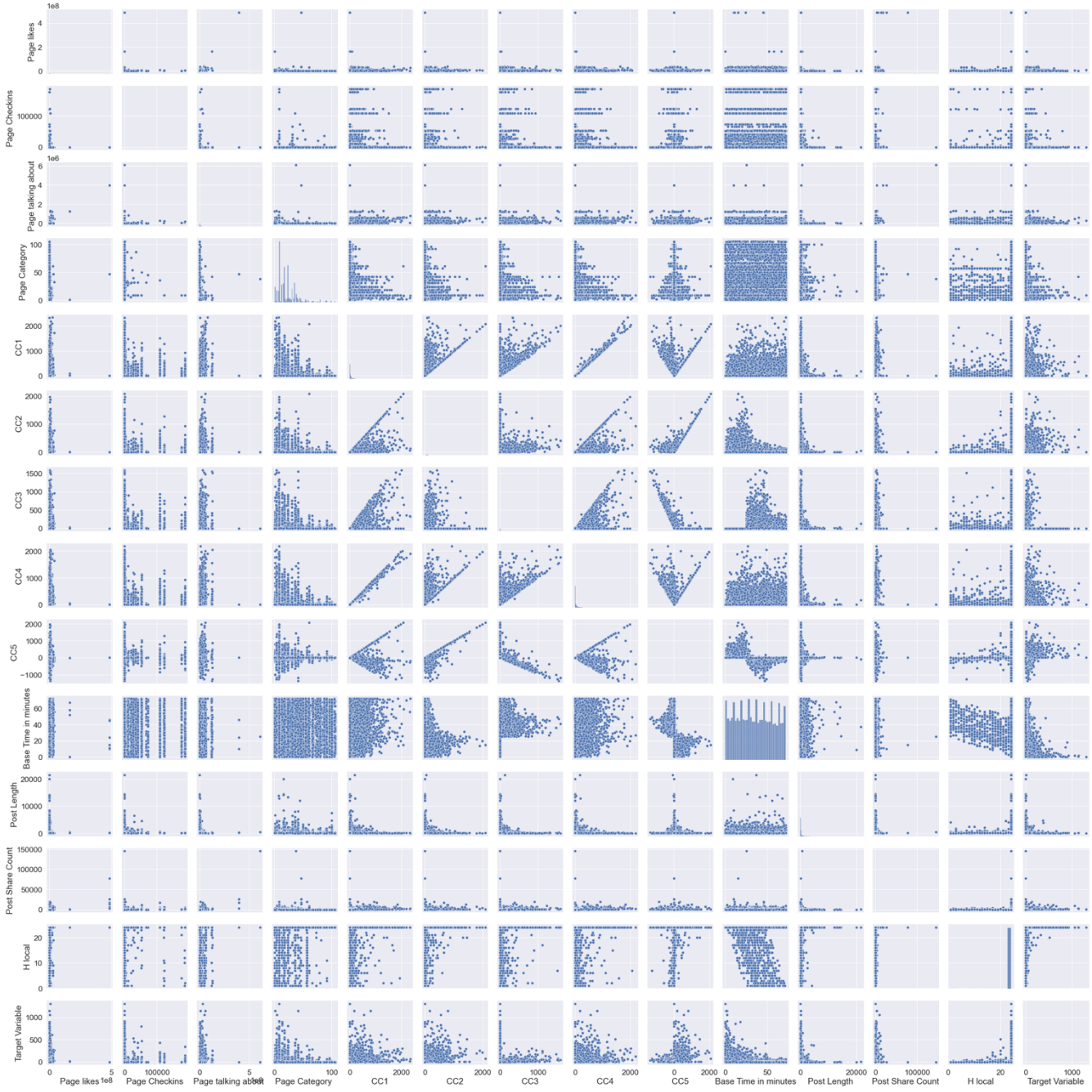

Figure 5 Distribution of all Page Features

- The relationship between post length and number of comments received are detailed in below figure 6

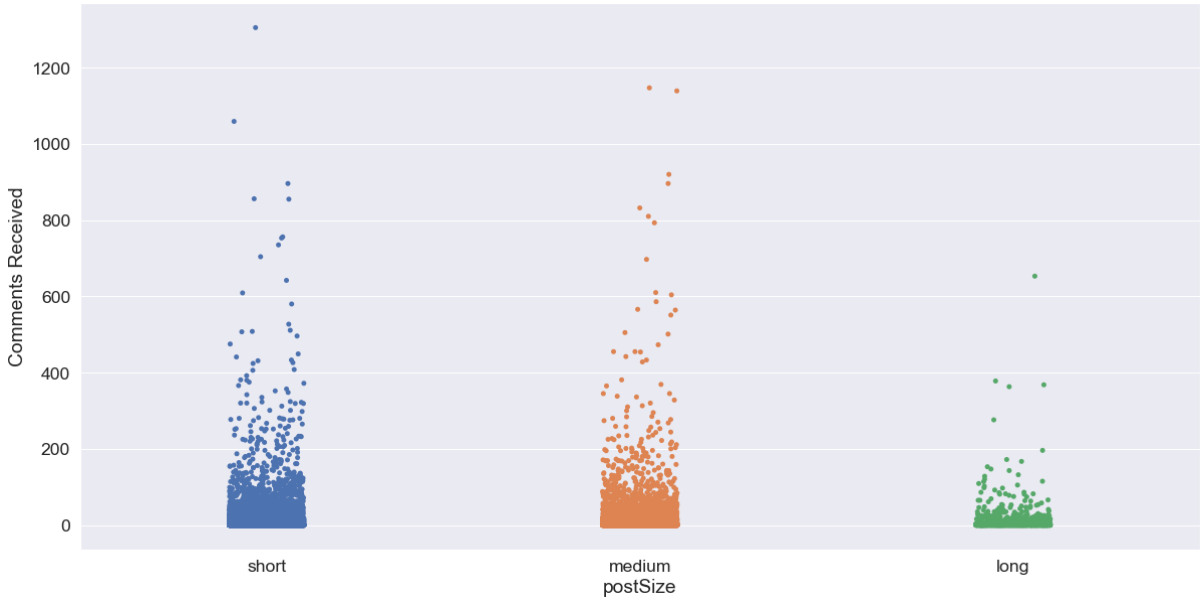

Figure 6 Post Length vs Comments Received

- We can see comments are high for a post with less than 500 characters
- We can also see in figure 7 the content posted at 24 hours received more comments

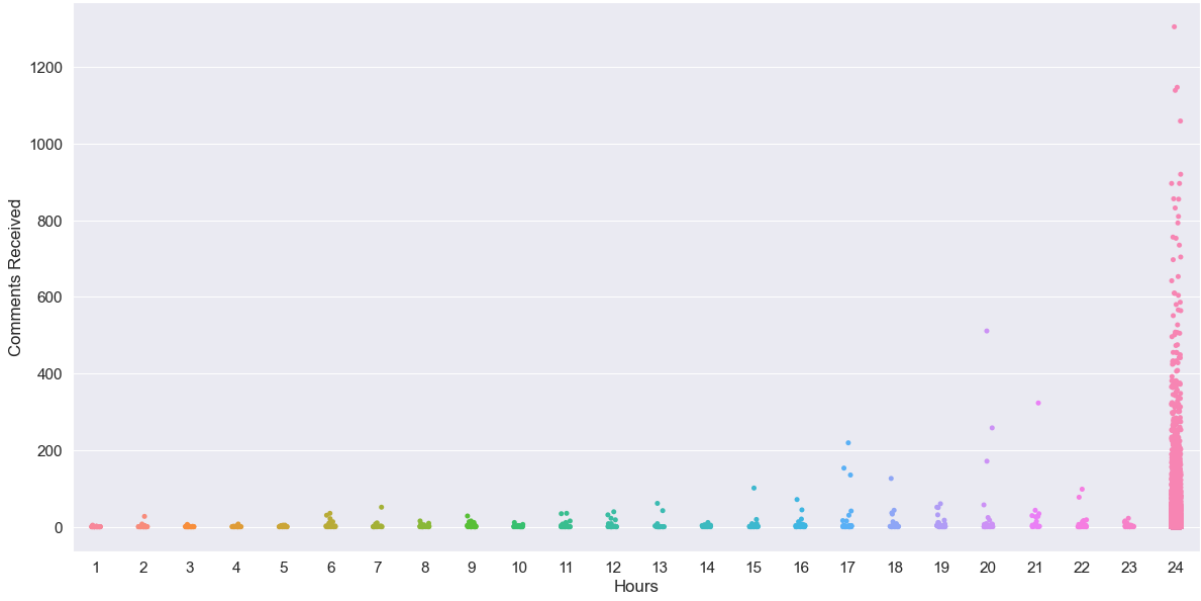

Figure 7 Hours vs Comments Received

- The variable ID is a continuous variable and we do not have any use with it so we can remove it.
- The variable Post promotion Status has only one value 0, so we can remove this column as well.
- There are a lot of values missing in most of the variables

- We will find mean and impute these missing values in figure 8

ID	0
Page likes	3208
Page Checkins	3255
Page talking about	3255
Page Category	3024
Feature 5	0
Feature 6	0
Feature 7	1679
Feature 8	0
Feature 9	0
Feature 10	1632
Feature 11	0
Feature 12	0
Feature 13	1643
Feature 14	0
Feature 15	1692
Feature 16	0
Feature 17	0
Feature 18	1605
Feature 19	0
Feature 20	1600
Feature 21	0
Feature 22	1601
Feature 23	0
Feature 24	0
Feature 25	1600
Feature 26	0
Feature 27	1598
Feature 28	0
Feature 29	1600
CC1	3199
CC2	0
CC3	0
CC4	3198
CC5	3200
Base Time in minutes	0
Post Length	0
Post Share Count	0
Post Promotion Status	0
H local	0
Post published weekday	0
Base DateTime weekday	0
Target Variable	0

Figure 8 Missing Values

- There are few Outliers in our data as seen in figure 9. We will find the interquartile range and bring them the accepted range.

Figure 9 Outlier on Variables

- We can see even our Target variable has some outliers we will not treat it.
- "Page likes" variable has most outliers followed "Page talking about"
- We, Will, transform the "Post Length" variable into a categorical variable by putting them in bins of post length less than 100 as Short Post, greater than 100, and less than 500 as Medium post and length greater than 500 as long post.

Business Insights

- The data looks imbalanced as the Target variable is more skewed to the left as seen in figure 10

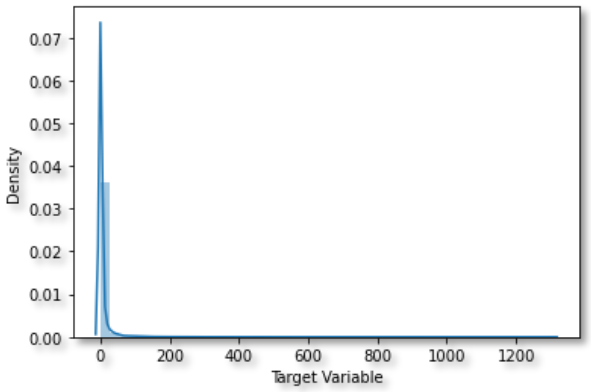

Figure 10 Skewness on Target Variable

- Even the data is dense more towards the bottom as seen in figure 11

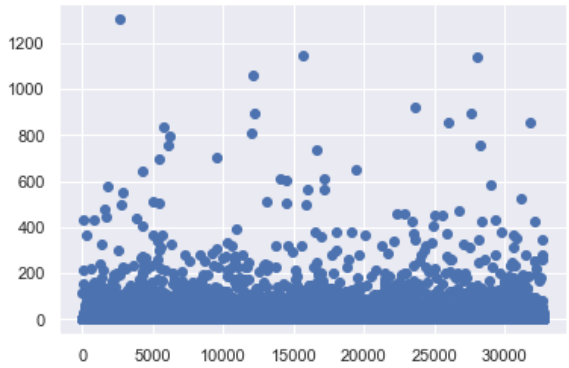

Figure 11 Scattering of Target Variable

- We can solve the Imbalance in continuous variables using Synthetic Minority Over-Sampling Technique for Regression (SMOTE for Regression).

Appendix

UGC:
https://www.researchgate.net/profile/Sandeep_Krishnamurthy/publication/274176873_Note_from_Special_Issue_Editors/links/5553be2c08aeaaff3bf19cc5.pdf

SMOTE for Regression: https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/
https://pypi.org/project/smogn/