**Overview**

Independent variables in the regression model are highly correlated to each other. It makes it hard to interpret the model and also creates an overfitting problem. Change in one variable would cause change to another and so the model results fluctuate significantly.

Due to features Derived5 - Derived29 and CC1 - CC5 are very highly correlated to each other. Hence breaking the assumption of collinearity.

Most features have a large number of zero values, as is the Target Variable. This leads to a severe imbalance in the data set. Many statistical tests and machine learning models rely on normality assumptions, severe imbalance means that the data is abnormal and may affect the statistical testing or machine learning predictive ability.

**Missing Value Ratio**

Graphical user interface, application

Description automatically generated

No missing value. We don’t need to anything.

**Dimensionality Reduction**

**Low Variance Filter**

Table

Description automatically generatedTable

Description automatically generated

Except for Binary Encoding value, all the variance of all features is >= 5. We don’t need to anything.

**Method1: High Correlation Filter (****Multicollinearity)**

A picture containing text

Description automatically generatedChart

Description automatically generated

|r|<0.3 means low linear correlation

0.3=<|r|<=0.5, which means low-to-medium linear correlation

0.5=<|r|<0.8, indicating a medium-to-high linear correlation

0.8=<|r|<1, which means high linear correlation

From the heatmap, we can see the features below are high multicollinearity.

**Drop:** Page talking about, CC1, CC4, CC5, Derived6, Derived7, Derived8, Derived9, Derived10, Derived11, Derived13, Derived14, Derived16, Derived17, Derived18, Derived19, Derived20, Derived21, Derived22, Derived23, Derived24, Derived26, Derived27, Derived29

Chart, scatter chart

Description automatically generatedChart

Description automatically generated

This is a constant data.

**Drop:** Post Promotion Status

**Table

Description automatically generated**

**The filtered features are:**

Page Popularity/likes,

Page Checkinsâ€™s,

Page Category,

Derived5

Derived12

Derived15

Derived25

Derived28

CC2

CC3

Base time,

Post length,

Post Share Count,

H Local,

Post published weekday-Sunday,

Post published weekday-Monday,

Post published weekday-Tuesday,

Post published weekday-Wednesday,

Post published weekday-Thursday,

Post published weekday-Friday,

Post published weekday-Saturday,

Post published weekday,

Post published weekend,

Base DateTime weekday-Sunday,

Base DateTime weekday-Monday,

Base DateTime weekday-Tuesday,

Base DateTime weekday-Wednesday,

Base DateTime weekday-Thursday,

Base DateTime weekday-Friday,

Base DateTime weekday-Saturday,

Base DateTime weekday,

Base DateTime weekend

Then we can keep select the most important features with the 3 methods below to reduce dimensionality further.

**Method1: Random Forest (for FCN and other distance based algorithms)**

<https://towardsdatascience.com/feature-selection-using-random-forest-26d7b747597f>

Random forests are one the most popular machine learning algorithms. They are so successful because they provide in general a good predictive performance, low overfitting, and easy interpretability. This interpretability is given by the fact that it is straightforward to derive the importance of each variable on the tree decision. In other words, it is easy to compute how much each variable is contributing to the decision.

**Table

Description automatically generated with medium confidence**

**The filtered features are:**

Page Popularity/likes,

Page Checkinsâ€™s,

Page Category,

Derived5

Derived12

Derived25

Derived28

CC2

CC3

**Method2: Backward Feature Elimination**

Page Popularity/likes,

Page Checkinsâ€™s,

Page Category,

Derived5,

Derived12,

Derived15,

Derived25,

Derived28,

CC2,

CC3,

Base time,

Post length,

Post Share Count,

H Local,

Post published weekday-Wednesday,

Post published weekday,

Base DateTime weekday

**Method3: Forward Feature Selection**

Set F >= 1000

**The filtered features are:**

Page Popularity/likes,

Page Checkinsâ€™s,

Page Category,

Derived5,

Derived12,

Derived15,

Derived25,

Derived28,

CC2,

CC3,

Base time,

Post length,

Post Share Count,

H Local,

**We can see that the filtered features with red color are the most important features that contains the most information.**

**Encoding**

Graphical user interface

Description automatically generated with medium confidence

<https://www.analyticsvidhya.com/blog/2020/08/types-of-categorical-data-encoding/>

All data are Value Encoding, Decimal Encoding and Binary Encoding. We don’t need to anything.

**Imbalanced Data**

**Box Cox Transformation of high skewed features**

<https://towardsdatascience.com/top-3-methods-for-handling-skewed-data-1334e0debf45>

**Before**

**Table

Description automatically generated**

Total 9 skewed features

**After**

Graphical user interface

Description automatically generated with medium confidence

Only 1 skew feature

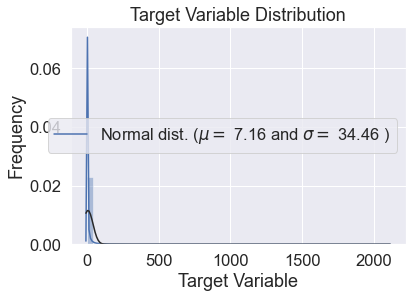
We can see that almost all the skewness are disappeared.

**The target variable is right skewed, we apply Log Transformation here.**

Graphical user interface, application

Description automatically generated

**Before**

 Chart

Description automatically generated

**After**

Chart, histogram

Description automatically generated Chart, line chart

Description automatically generated

We will not use sampling because: <https://stats.stackexchange.com/questions/30162/sampling-for-imbalanced-data-in-regression>