FLIP ROBO

MALIGNANT COMMENTS CLASSIFICATION

Submitted by:

DR RUCHI SHARMA

INTRODUCTION

- ## Business Problem Framing

The proliferation of social media enables people to express their opinions widely online. However, at the same time, this has resulted in the emergence of conflict and hate, making online environments uninviting for users. Although researchers have found that hate is a problem across multiple platforms, there is a lack of models for online hate detection.

Online hate, described as abusive language, aggression, cyberbullying, hatefulness and many others has been identified as a major threat on online social media platforms. Social media platforms are the most prominent grounds for such toxic behaviour.

- ## Conceptual Background of the Domain Problem

There has been a remarkable increase in the cases of cyberbullying and trolls on various social media platforms. Many celebrities and influences are facing backlashes from people and have to come across hateful and offensive comments. This can take a toll on anyone and affect them mentally leading to depression, mental illness, self-hatred and suicidal thoughts.

Internet comments are bastions of hatred and vitriol. While online anonymity has provided a new outlet for aggression and hate speech, machine learning can be used to fight it. The problem we sought to solve was the tagging of internet comments that are aggressive towards other users. This means that insults to third parties such as celebrities will be tagged as un offensive, but “u are an idiot” is clearly offensive.

Analytical Problem Framing

- ## Mathematical/ Analytical Modeling of the Problem

The label can be either 0 or 1, where 0 denotes a NO while 1 denotes a YES. There are various comments which have multiple labels. The first attribute is a unique ID associated with each comment.

The data set includes:

Malignant: It is the Label column, which includes values 0 and 1, denoting if the comment is malignant or not.

Highly Malignant: It denotes comments that are highly malignant and hurtful.

Rude: It denotes comments that are very rude and offensive.

Threat: It contains indication of the comments that are giving any threat to someone.

Abuse: It is for comments that are abusive in nature.

Loathe: It describes the comments which are hateful and loathing in nature.

ID: It includes unique Ids associated with each comment text given.

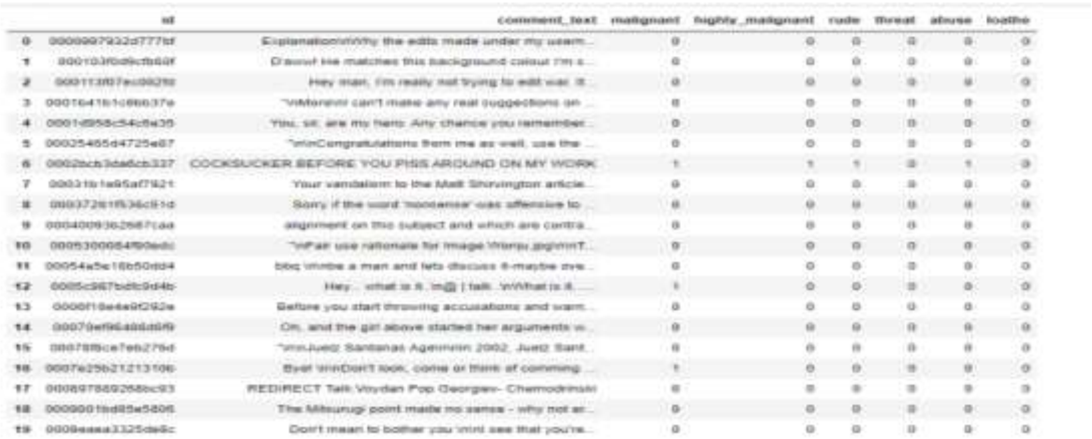
Comment text: This column contains the comments extracted from various social media platforms.

And as I am taking label as a target variable which having 0 and 1 class, identify that its is a binary type classification problem.

- ## Data Sources and their formats

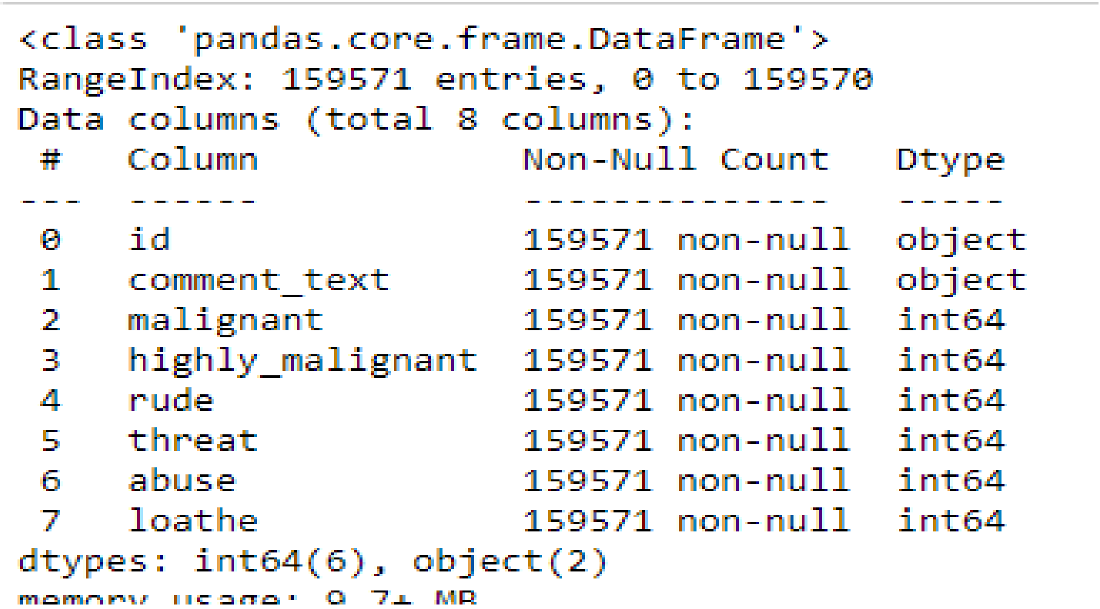
The data set contains the training set, which has approximately 1,59,000 samples and the test set which contains nearly 1,53,000 samples. All the data samples contain 8 fields which includes 'Id', 'Comments', 'Malignant', 'Highly malignant', 'Rude', 'Threat', 'Abuse' and 'Loathe'.

The label can be either 0 or 1, where 0 denotes a NO while 1 denotes a YES. There are various comments which have multiple labels. The first attribute is a unique ID associated with each comment.

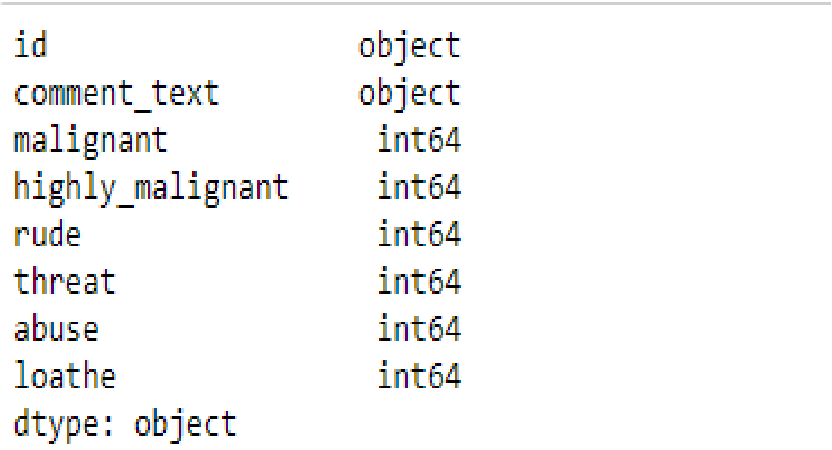


	id	comment_text	malignant	highly_malignant	rude	threat	abuse	loathe
0	0000997932d777bf	Explanation\nWhy the edits made under my userm...	0	0	0	0	0	0
1	000103f0d9cfb60f	D'aww! He matches this background colour I'm s...	0	0	0	0	0	0
2	000113f07ec002fd	Hey man, I'm really not trying to edit war. It...	0	0	0	0	0	0
3	0001b41b1c6bb37e	"\nMore\nI can't make any real suggestions on ...	0	0	0	0	0	0
4	0001d958c54c6e35	You, sir, are my hero. Any chance you remember...	0	0	0	0	0	0
5	00025465d4725e87	"\n\nCongratulations from me as well, use the ...	0	0	0	0	0	0
6	0002bcb3da6cb337	COCKSUCKER BEFORE YOU PISS AROUND ON MY WORK	1	1	1	0	1	0
7	00031b1e95af7921	Your vandalism to the Matt Shirvington article...	0	0	0	0	0	0
8	00037261f536c51d	Sorry if the word 'nonsense' was offensive to ...	0	0	0	0	0	0
9	00040093b2687caa	alignment on this subject and which are contra...	0	0	0	0	0	0
10	0005300084f90edc	"\nFair use rationale for Image:Wonju.jpg\n\nT...	0	0	0	0	0	0
11	00054a5e18b50dd4	bbq \n\nbe a man and lets discuss it-maybe ove...	0	0	0	0	0	0
12	0005c987bdfc9d4b	Hey... what is it..\n@ | talk .\nWhat is it...	1	0	0	0	0	0
13	0006f16e4e9f292e	Before you start throwing accusations and warn...	0	0	0	0	0	0
14	00070ef96486d6f9	Oh, and the girl above started her arguments w...	0	0	0	0	0	0
15	00078f8ce7eb276d	"\nJuelz Santanas Age\n\nIn 2002, Juelz Sant...	0	0	0	0	0	0
16	0007e25b2121310b	Bye! \n\nDon't look, come or think of comming ...	1	0	0	0	0	0
17	000897889268bc93	REDIRECT Talk:Voydan Pop Georgiev- Chernodrinski	0	0	0	0	0	0
18	0009801bd85e5806	The Mitsurugi point made no sense - why not ar...	0	0	0	0	0	0
19	0009eaea3325de8c	Don't mean to bother you \n\nI see that you're...	0	0	0	0	0	0

- Their are 159571 rows and 8 columns in dataset.
- Two types of data type present in dataset. 1. object, 2. integer
- No null values has been observed in dataset



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 159571 entries, 0 to 159570
Data columns (total 8 columns):
 #   Column            Non-Null Count   Dtype 
---  ------            --------------   ----- 
 0   id                159571 non-null  object
 1   comment_text      159571 non-null  object
 2   malignant         159571 non-null  int64 
 3   highly_malignant  159571 non-null  int64 
 4   rude              159571 non-null  int64 
 5   threat            159571 non-null  int64 
 6   abuse             159571 non-null  int64 
 7   loathe            159571 non-null  int64 
dtypes: int64(6), object(2)
memory usage: 9.7+ MB
```



```
id                  object
comment_text        object
malignant            int64
highly_malignant     int64
rude                 int64
threat               int64
abuse                int64
loathe               int64
dtype: object
```

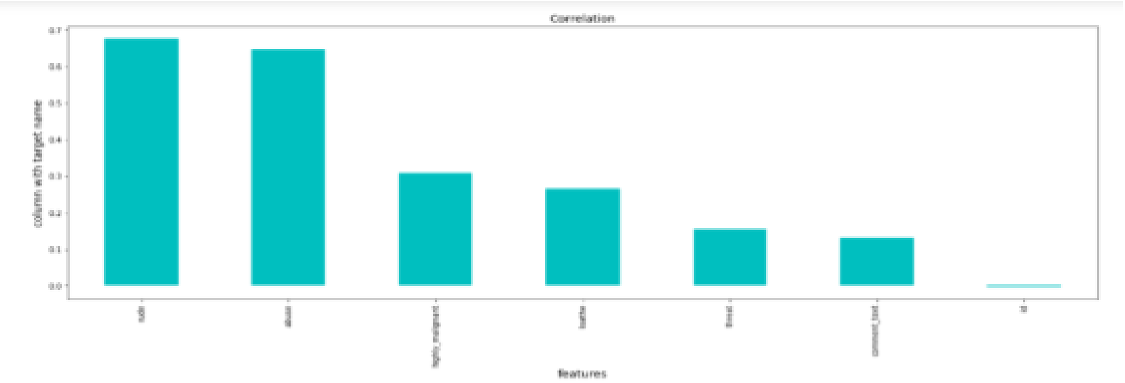
- ## Data Preprocessing

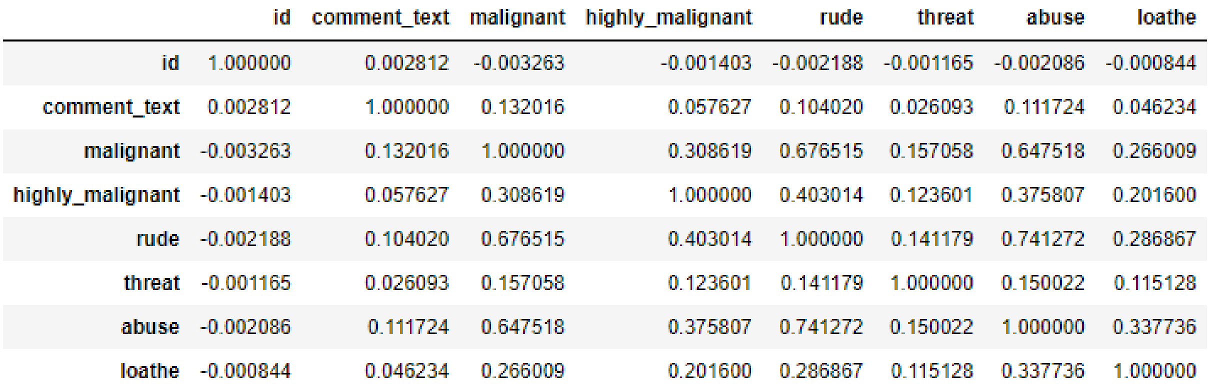
  Column named-Unnamed: 0 having S.No. which is not relevant for loan paying prediction, so decided to drop it

  As it has been observed that two columns have object type data, converted them to intergers as machine learning model

  Dataset observed for checking null values, it has been observed that no null value present in dataset.

- Correlation matrix has been checked with target variable, 6 columns shown +ve correlation with label. 1 column shown -ve correlation with label. Highest correlation is observed with *rude* col- 0.676515.



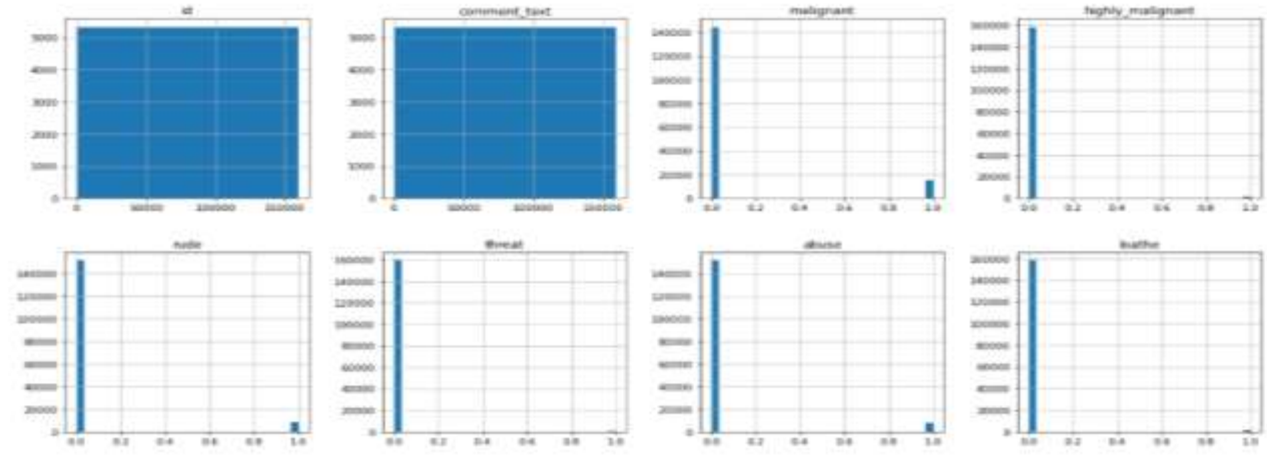


	id	comment_text	malignant	highly_malignant	rude	threat	abuse	loathe
id	1.000000	0.002812	-0.003263	-0.001403	-0.002188	-0.001165	-0.002086	-0.000844
comment_text	0.002812	1.000000	0.132016	0.057627	0.104020	0.026093	0.111724	0.046234
malignant	-0.003263	0.132016	1.000000	0.308619	0.676515	0.157058	0.647518	0.266009
highly_malignant	-0.001403	0.057627	0.308619	1.000000	0.403014	0.123601	0.375807	0.201600
rude	-0.002188	0.104020	0.676515	0.403014	1.000000	0.141179	0.741272	0.286867
threat	-0.001165	0.026093	0.157058	0.123601	0.141179	1.000000	0.150022	0.115128
abuse	-0.002086	0.111724	0.647518	0.375807	0.741272	0.150022	1.000000	0.337736
loathe	-0.000844	0.046234	0.266009	0.201600	0.286867	0.115128	0.337736	1.000000

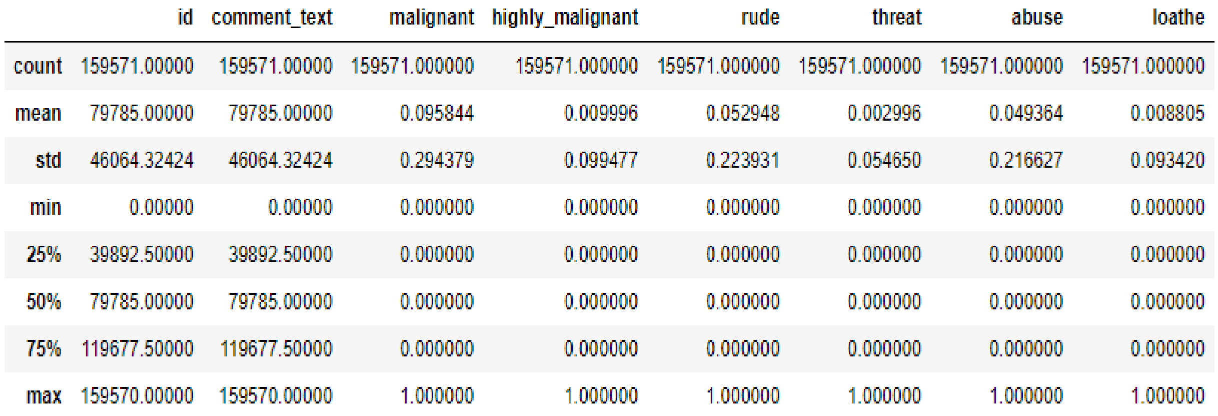
Correlation summary to target variable is shown in above table

No Outliers present in dataset

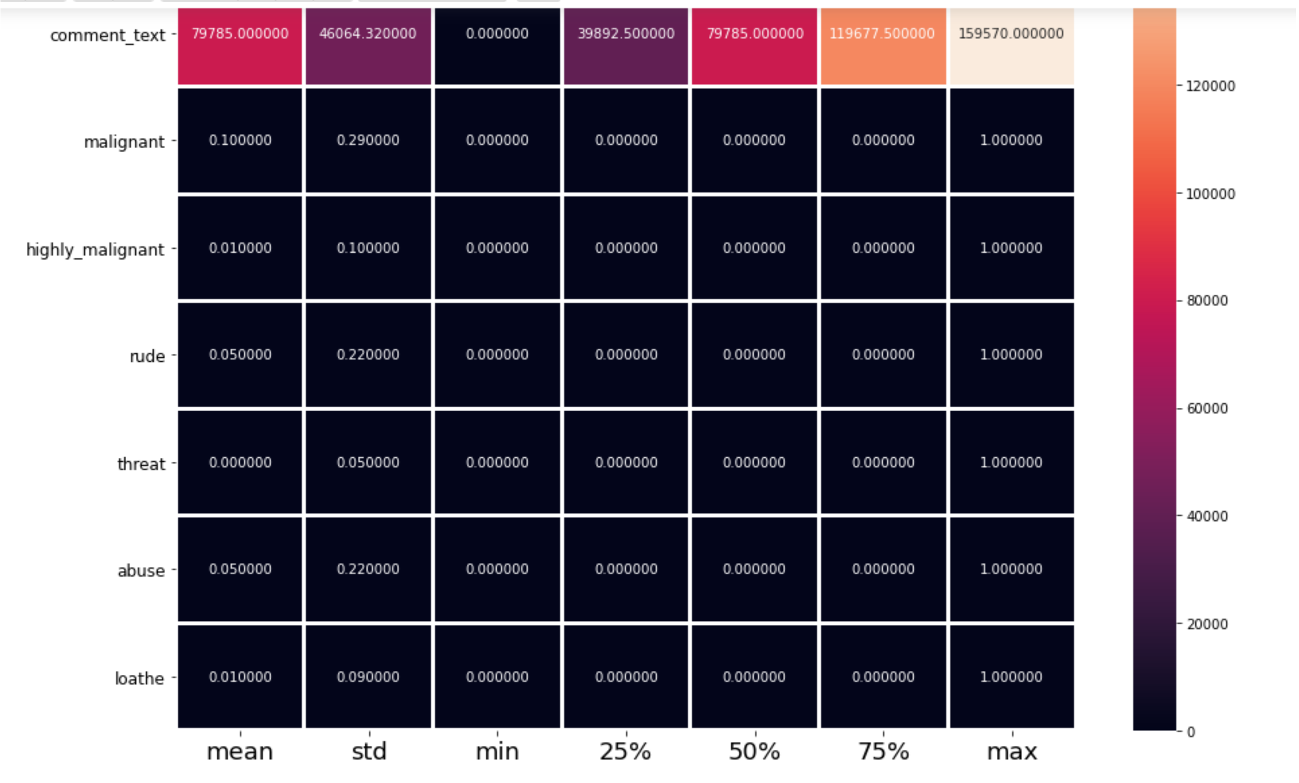
Histogram for checking distribution



Describing dataset



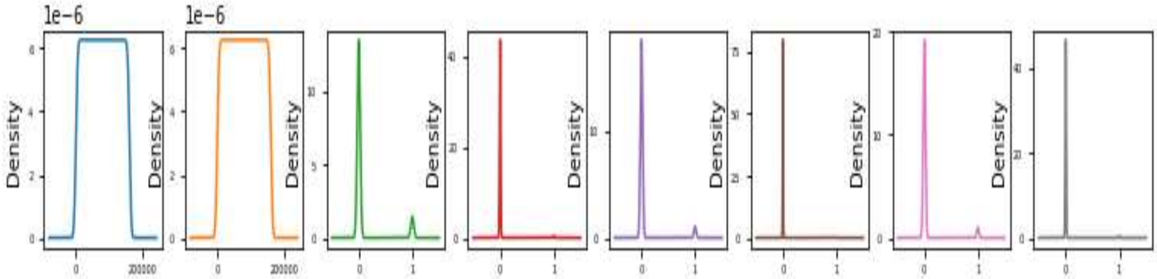
	id	comment_text	malignant	highly_malignant	rude	threat	abuse	loathe
count	159571.00000	159571.00000	159571.000000	159571.000000	159571.000000	159571.000000	159571.000000	159571.000000
mean	79785.00000	79785.00000	0.095844	0.009996	0.052948	0.002996	0.049364	0.008805
std	46064.32424	46064.32424	0.294379	0.099477	0.223931	0.054650	0.216627	0.093420
min	0.00000	0.00000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	39892.50000	39892.50000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	79785.00000	79785.00000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
75%	119677.50000	119677.50000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
max	159570.00000	159570.00000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000



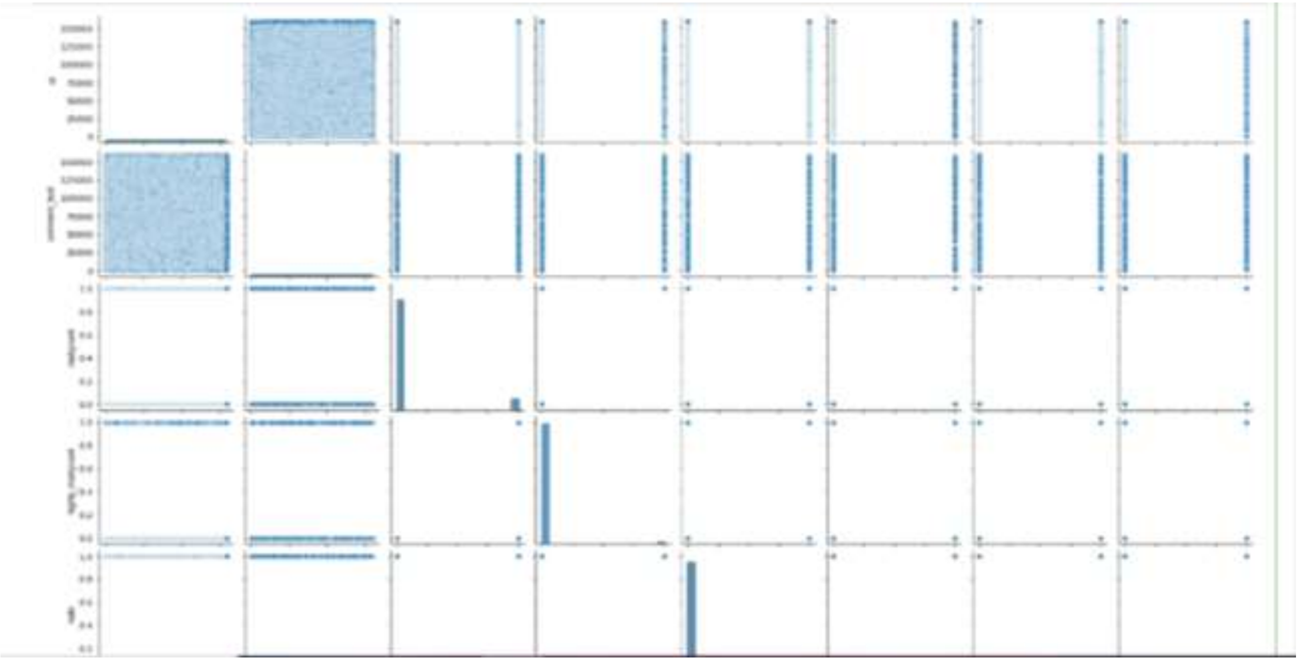
	mean	std	min	25%	50%	75%	max
comment_text	79785.000000	46064.320000	0.000000	39892.500000	79785.000000	119677.500000	159570.000000
malignant	0.100000	0.290000	0.000000	0.000000	0.000000	0.000000	1.000000
highly_malignant	0.010000	0.100000	0.000000	0.000000	0.000000	0.000000	1.000000
rude	0.050000	0.220000	0.000000	0.000000	0.000000	0.000000	1.000000
threat	0.000000	0.050000	0.000000	0.000000	0.000000	0.000000	1.000000
abuse	0.050000	0.220000	0.000000	0.000000	0.000000	0.000000	1.000000
loathe	0.010000	0.090000	0.000000	0.000000	0.000000	0.000000	1.000000

Variable summary related to target variable is shown in above heatmap

Dataset also checked for skewness in columns, andtreated for skewness

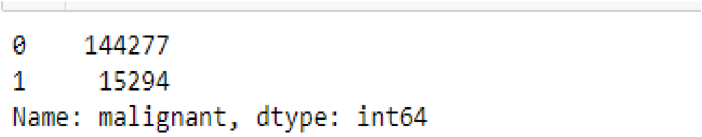


Plot for seeing relation between columns-using pairplot

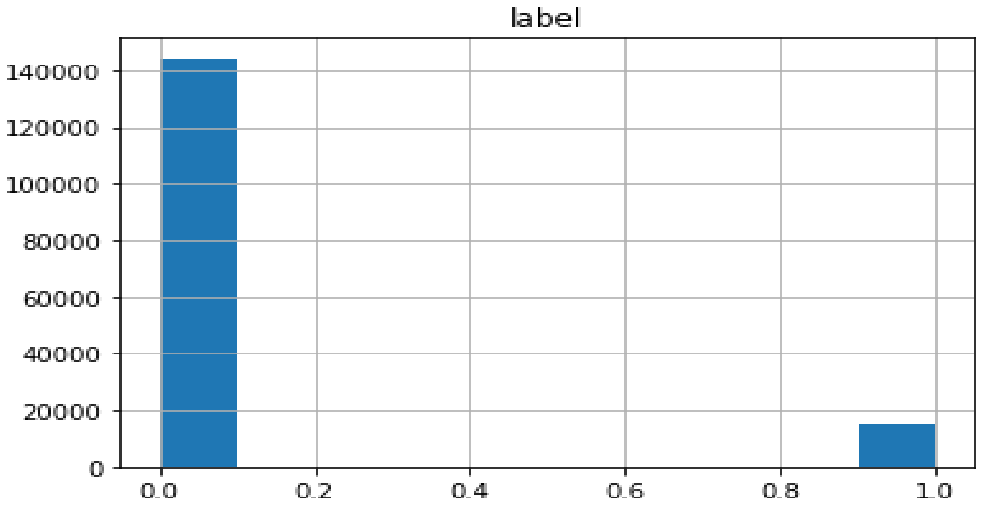


Their is imbalance in classes of target variable/label, decided to treat them with sampling technique(SMOTE function for over sampling)

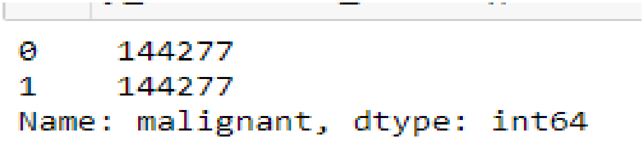
Data info before oversampling



```
0    144277
1     15294
Name: malignant, dtype: int64
```



After oversampling



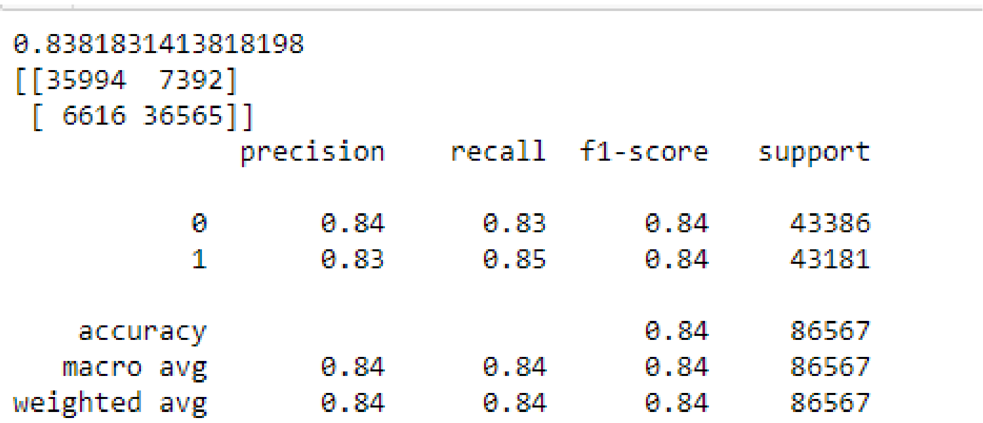
```
0     144277
1     144277
Name: malignant, dtype: int64
```

Model/s Development and Evaluation

Four different classification model has been build for micro-credit loan prediction

1. Linear Regression
2. Decision Tree Classifier
3. Random Forest Classifier
4. SVC model
5. KNN Classifier

- ✓ Model is selected on the basis of accuracy and cross-validation report
- ✓ Best Accuracy % obtained in Decision Tree Classifier

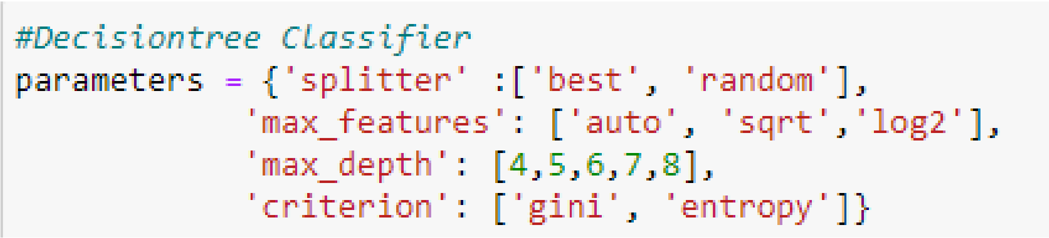


```
0.8381831413818198
[[35994  7392]
 [ 6616 36565]]
              precision    recall  f1-score   support

           0       0.84      0.83      0.84     43386
           1       0.83      0.85      0.84     43181

    accuracy                           0.84     86567
   macro avg       0.84      0.84      0.84     86567
weighted avg       0.84      0.84      0.84     86567
```

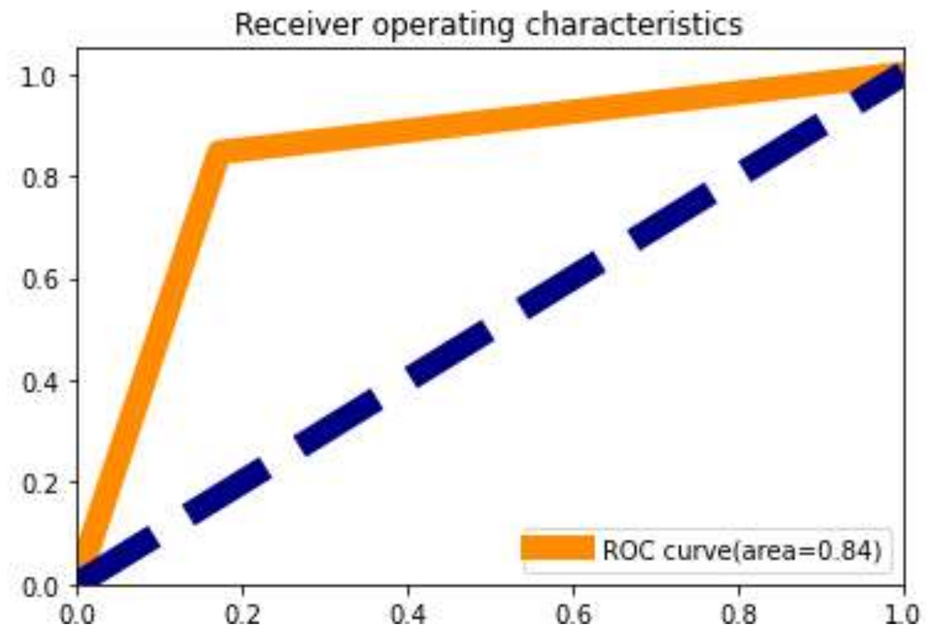
Followed by best model hypertunning-using Gridsearch CV



```
#Decisiontree Classifier
parameters = {'splitter' :['best', 'random'],
              'max_features': ['auto', 'sqrt','log2'],
              'max_depth': [4,5,6,7,8],
              'criterion': ['gini', 'entropy']}
```

Checking Accuracy-AUC_ROC Curve

- ✓ getting ROC curve area 0.84, AUC score is 84%
- ✓ Model performance is good(84%) for predicting loan defaulter, hence saving a model

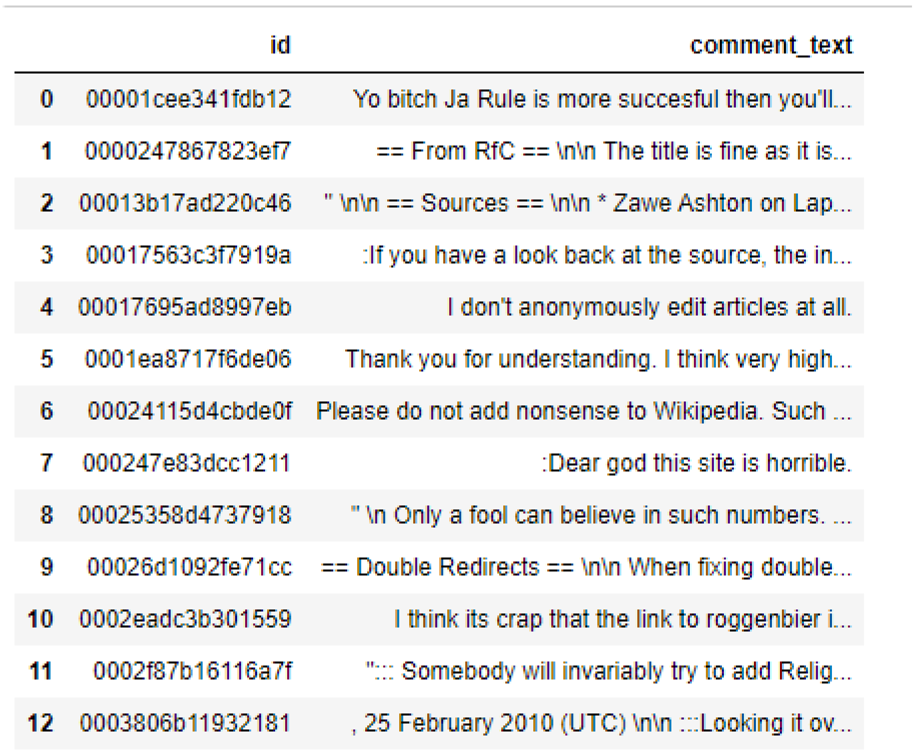


TEST DATASET

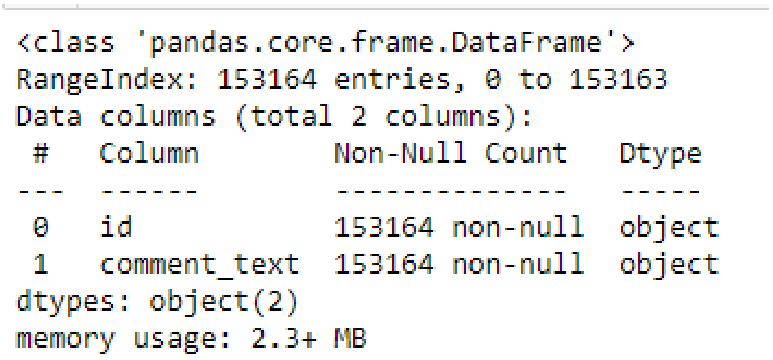
Their are 153164 rows and 2 columns.

Only one type of dataset is observed in test dataset i.e. object

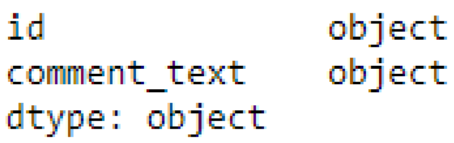
No null values has been observed in test dataset



	id	comment_text
0	00001cee341fdb12	Yo bitch Ja Rule is more succesful then you'll...
1	0000247867823ef7	== From RfC == \n\n The title is fine as it is...
2	00013b17ad220c46	" \n\n == Sources == \n\n * Zawe Ashton on Lap...
3	00017563c3f7919a	:If you have a look back at the source, the in...
4	00017695ad8997eb	I don't anonymously edit articles at all.
5	0001ea8717f6de06	Thank you for understanding. I think very high...
6	00024115d4cbde0f	Please do not add nonsense to Wikipedia. Such ...
7	000247e83dcc1211	:Dear god this site is horrible.
8	00025358d4737918	" \n Only a fool can believe in such numbers. ...
9	00026d1092fe71cc	== Double Redirects == \n\n When fixing double...
10	0002eadc3b301559	I think its crap that the link to roggenbier i...
11	0002f87b16116a7f	":::: Somebody will invariably try to add Relig...
12	0003806b11932181	, 25 February 2010 (UTC) \n\n :::Looking it ov...



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 153164 entries, 0 to 153163
Data columns (total 2 columns):
 #   Column        Non-Null Count   Dtype 
---  ------        --------------   ----- 
 0   id            153164 non-null  object
 1   comment_text  153164 non-null  object
dtypes: object(2)
memory usage: 2.3+ MB
```



```
id              object
comment_text    object
dtype: object
```

Conclusion:

- The data set contains the training set, which has approximately 1,59,000 samples
- Malignant comment classifier-- Model performance is good (84%) for predicting comments.
- Test Dataset-containing 153164 rows and 2 columns