MACHINE LEARNING GRADED PROJECT

PREPARED FOR
MACHINE LEARNING MENTOR

PREPARED BY
NITIN KUMAR SINGH

INDEX

1.1 Basic data summary, Univariate, Bivariate analysis, graphs, checking correlations, outliers and missing values treatment (if necessary) and check the basic descriptive statistics of the dataset.
1.2 Split the data into train and test in the ratio 70:30. Is scaling necessary or not?
1.3 Build the following models on the 70% training data and check the performance of these models on the Training as well as the 30% Test data using the various inferences from the Confusion Matrix and plotting a AUC-ROC curve along with the AUC values. Tune the models wherever required for optimum performance.: a. Logistic Regression Model b. Linear Discriminant Analysis c. Decision Tree Classifier – CART model d. Naïve Bayes Model e. KNN Model f. Random Forest Model g. Boosting Classifier Model using Gradient boost.
1.4 Which model performs the best?
1.5 What are your business insights?
2.1 Pick out the Deal (Dependent Variable) and Description columns into a separate data frame.
2.2 Create two corpora, one for those who secured a Deal, the other for those who did not secure a deal.
2.3 The following exercise is to be done for both the corpora: a) Find the number of characters for both the corpuses. b) Remove Stop Words from the corpora. (Words like 'also', 'made', 'makes', 'like', 'this', 'even' and 'company' are to be removed) c) What were the top 3 most frequently occurring words in both corpuses (after removing stop words)? d) Plot the Word Cloud for both the corpora.
2.4 Refer to both the word clouds. What do you infer?
2.5 Looking at the word clouds, is it true that the entrepreneurs who introduced devices are less likely to secure a deal based on your analysis?

SUMMARIZE DATA

There are 444 rows × 9 columns present in our dataset, 7 Columns are int64 ('Age', 'Engineer', 'MBA', 'Work Exp', 'Salary', 'Distance', 'license') & 2 is object (Gender,Transport), No null & No missing value present in data

SKEWNESS OF DATA

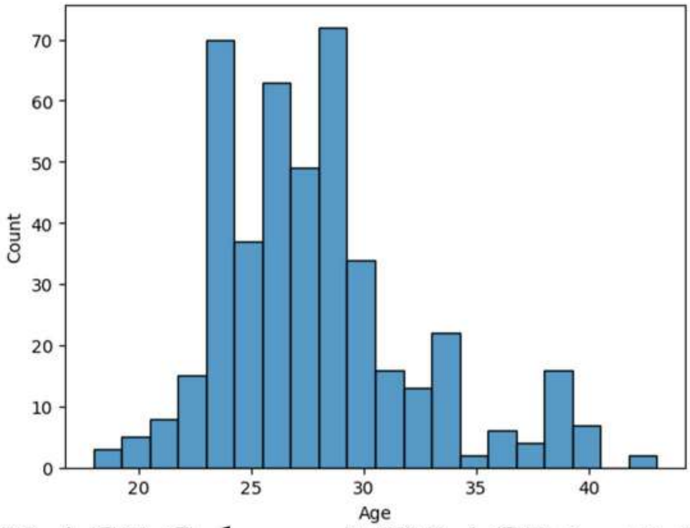
Age	0.955276
Gender	0.937952
Engineer	-1.186708
MBA	1.144763
Work Exp	1.352840
Salary	2.044533
Distance	0.539851
license	1.259293
Transport	0.753102

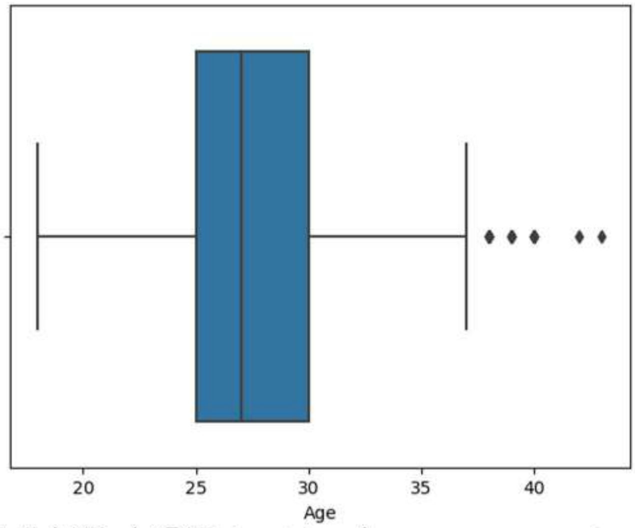
COORELATION OF DATA

	Age	Engineer	MBA	Work Exp	Salary	Distance	license
Age	1.000000	0.091935	-0.029090	0.932236	0.860673	0.352872	0.452311
Engineer	0.091935	1.000000	0.066218	0.085729	0.086762	0.059316	0.018924
MBA	-0.029090	0.066218	1.000000	0.008582	-0.007270	0.036427	-0.027358
Work Exp	0.932236	0.085729	0.008582	1.000000	0.931974	0.372735	0.452867
Salary	0.860673	0.086762	-0.007270	0.931974	1.000000	0.442359	0.508095
Distance	0.352872	0.059316	0.036427	0.372735	0.442359	1.000000	0.290084
license	0.452311	0.018924	-0.027358	0.452867	0.508095	0.290084	1.000000

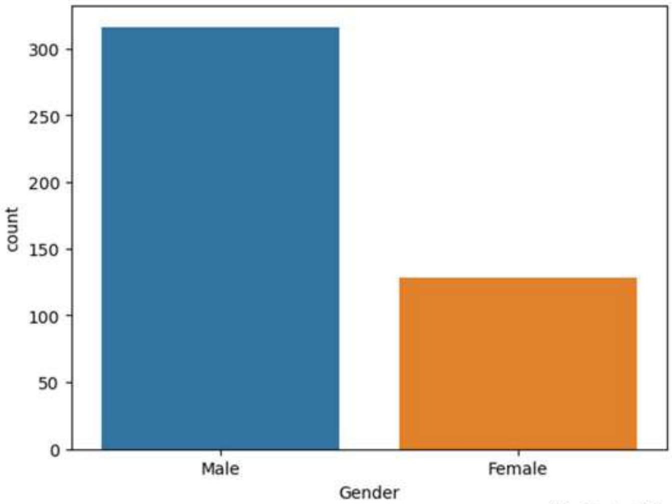
AGE IS HIGHLY COORELATED WITH **WORK EXP AND SALARY**

UNIVARIATE ANALYSIS

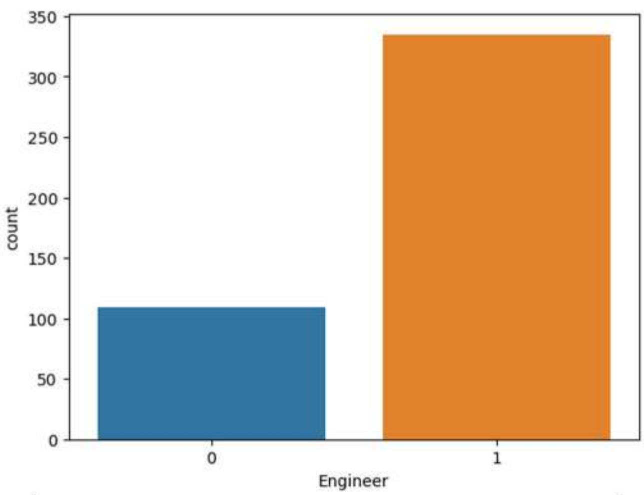




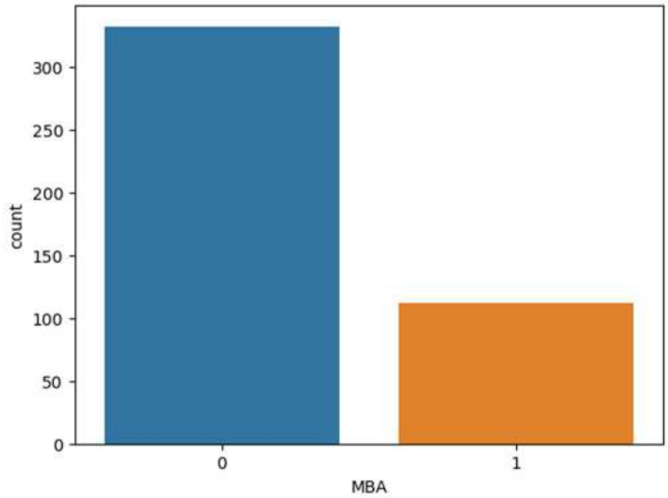
IN AGE Column MIN AGE is 18 & MAX AGE is 43. Aprrox most of data lie b\W 18-37 and Some outliers are present in data



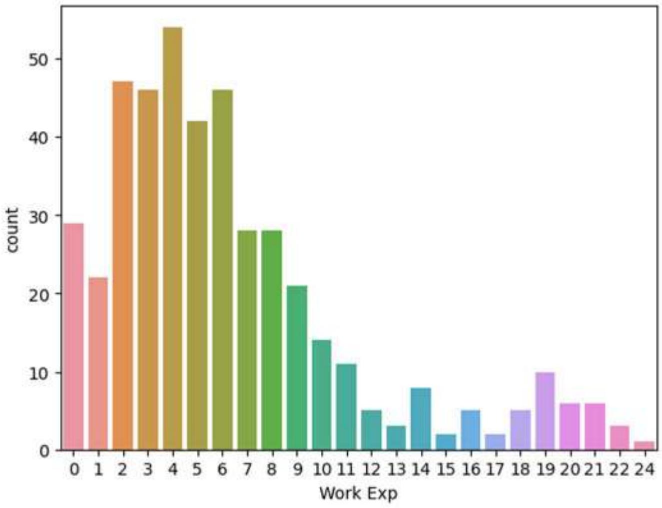
IN GENDER Column two category are present (MALE:316,FEMALE:128) Values present in dataset

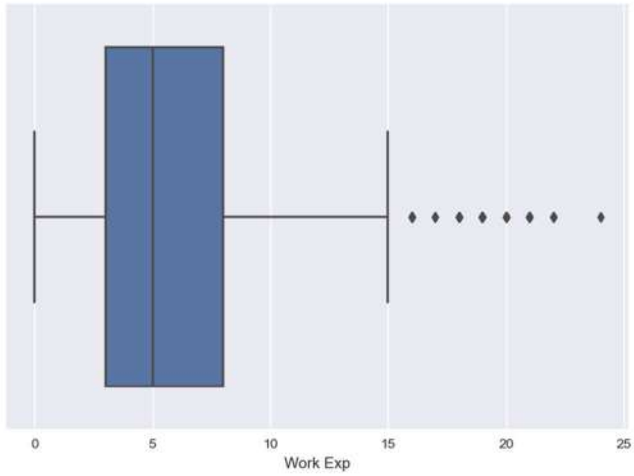


IN ENGINEER Column two category are present (0:335,1:109) Values present in dataset(0 is non engineer, 1 is engineer)

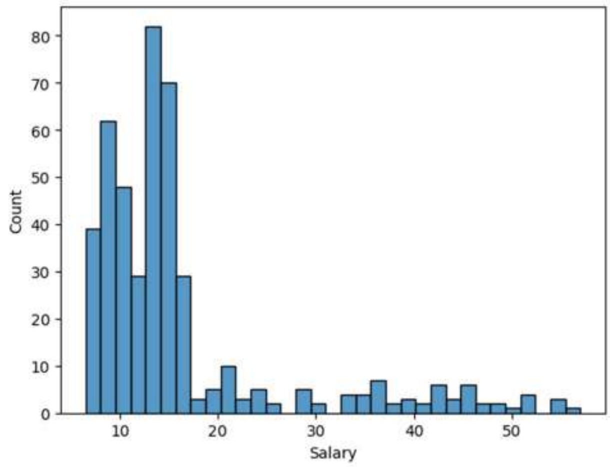


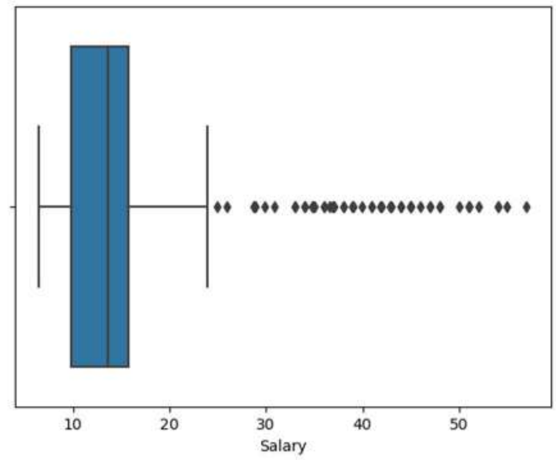
IN MBA Column two category are present (0:322,1:112) Values present in dataset (0 is NON MBA, 1 is MBA)



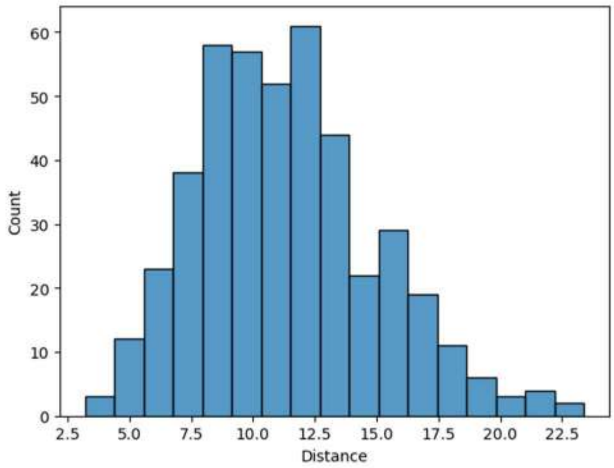


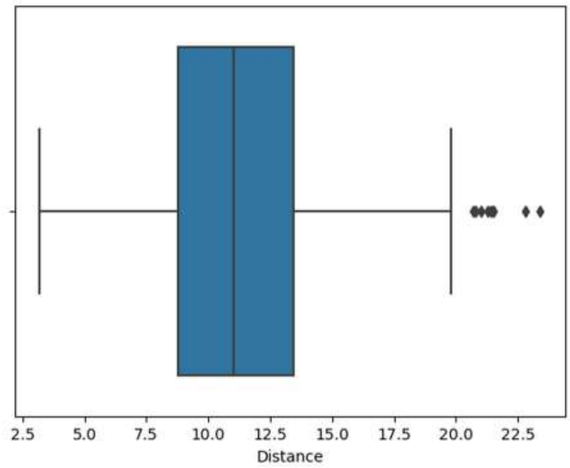
IN WORK EXP Column MIN EXP is 0 is MAX EXP is 24 most of the EXP is B\W 0-11 and some outliers are present in this column



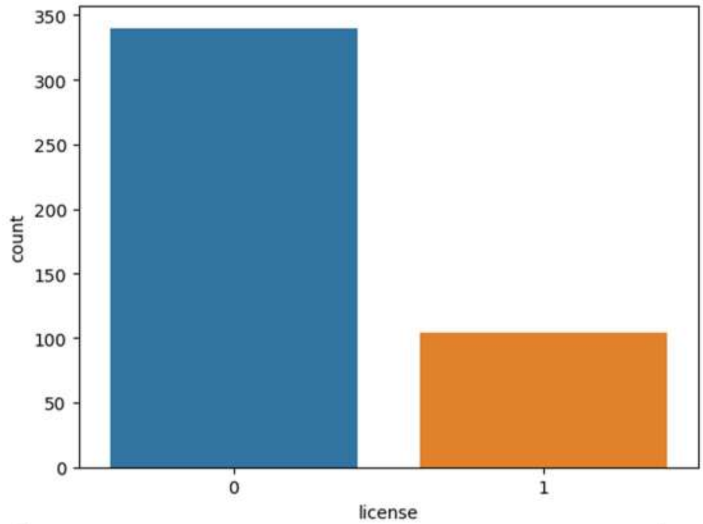


IN SALARY Column MIN SALARY(LPA) is 6.5 & MAX SALARY is 57 most of the SALARY is B\W Approx 6.5-18 and outliers are present in this column

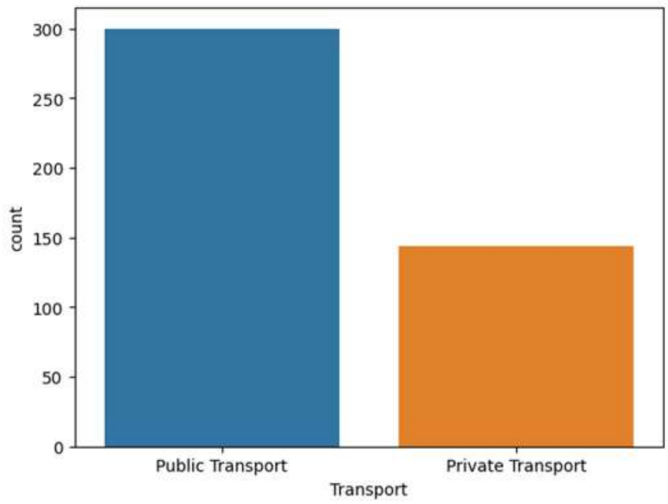




IN DISTANCE Column MIN DIST. is 3.2 & MAX DIST. is 23.4
most of the DISTANCE is B\W Approx 5-18 and outliers are present in this column

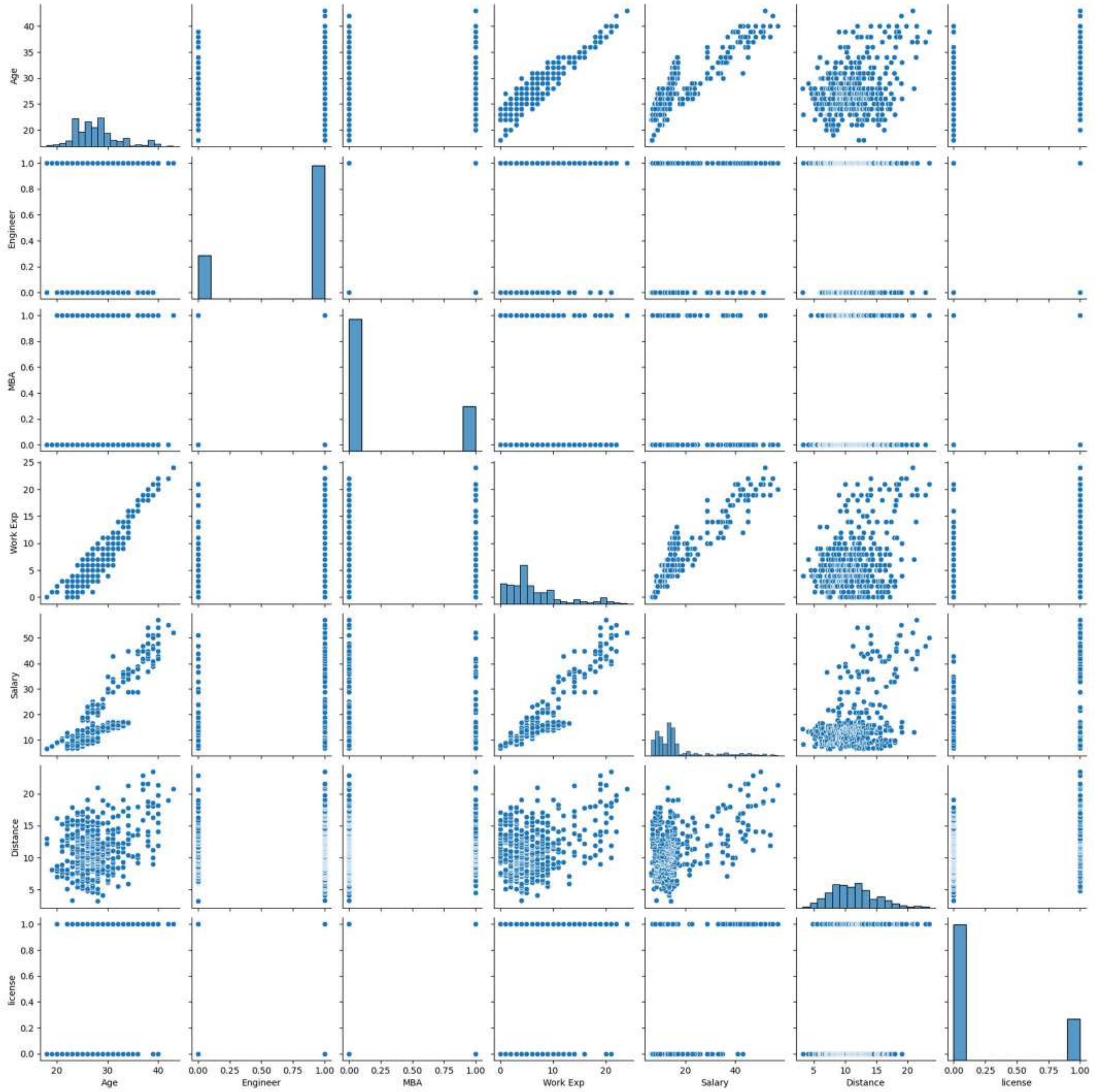


IN LICENCE Column two category are present (0:340, 1:104) Values present in dataset (0 NON LICENCE, 1 LICENCE)



IN TRANSPORT Column two category are present (public :300, private :144) Values present in dataset

BIVARIATE ANALYSIS



AGE IS LINEAR RELATION WITH **WORK EXP AND SALARY**

SALARY IS LINEAR RELATION WITH **WORK EXP**

AGE IS HIGHLY COORELATED WITH **WORK EXP AND SALARY**

ENCODE THE DATA

	Age	Gender	Engineer	MBA	Work Exp	Salary	Distance	license	Transport
0	28	Male	0	0	4	14.3	3.2	0	Public Transport
1	23	Female	1	0	4	8.3	3.3	0	Public Transport
2	29	Male	1	0	7	13.4	4.1	0	Public Transport
3	28	Female	1	1	5	13.4	4.5	0	Public Transport
4	27	Male	1	0	4	13.4	4.6	0	Public Transport
...	...	...	...	...	...	...	...	...	...
439	40	Male	1	0	20	57.0	21.4	1	Private Transport
440	38	Male	1	0	19	44.0	21.5	1	Private Transport
441	37	Male	1	0	19	45.0	21.5	1	Private Transport
442	37	Male	0	0	19	47.0	22.8	1	Private Transport
443	39	Male	1	1	21	50.0	23.4	1	Private Transport

First we encoded the object column (GENDER,TRANSPORT) with the help of **mapping.** encode the data is neccessary because in model creation object column does not accept.

	Age	Gender	Engineer	MBA	Work Exp	Salary	Distance	license	Transport
0	28	0	0	0	4	14.3	3.2	0	0
1	23	1	1	0	4	8.3	3.3	0	0
2	29	0	1	0	7	13.4	4.1	0	0
3	28	1	1	1	5	13.4	4.5	0	0
4	27	0	1	0	4	13.4	4.6	0	0
...	...	...	...	...	...	...	...	...	...
439	40	0	1	0	20	57.0	21.4	1	1
440	38	0	1	0	19	44.0	21.5	1	1
441	37	0	1	0	19	45.0	21.5	1	1
442	37	0	0	0	19	47.0	22.8	1	1
443	39	0	1	1	21	50.0	23.4	1	1

then Separate independent column and dependent column and store in (X,y)

	Age	Gender	Engineer	MBA	Work Exp	Salary	Distance	license
0	28	0	0	0	4	14.3	3.2	0
1	23	1	1	0	4	8.3	3.3	0
2	29	0	1	0	7	13.4	4.1	0
3	28	1	1	1	5	13.4	4.5	0
4	27	0	1	0	4	13.4	4.6	0
...	...	...	...	...	...	...	...	...
439	40	0	1	0	20	57.0	21.4	1
440	38	0	1	0	19	44.0	21.5	1
441	37	0	1	0	19	45.0	21.5	1
442	37	0	0	0	19	47.0	22.8	1
443	39	0	1	1	21	50.0	23.4	1

```
y

0      0
1      0
2      0
3      0
4      0
      ..
439    1
440    1
441    1
442    1
443    1
```

SPLIT DATA

Data Split: Split the data into train and test (70:30) for model creation

Help of sklearn model_selection.train_test_split

X_train, X_test, y_train, y_test = train_test_split(
X, y, test_size=0.33, random_state=1234)

then scaling the data help of **sklearn preprocessing .StandardScaler** its neccessary because scaling the standardize the data, mean is o and std is 1

LOGISTIC REGRESSION MODEL

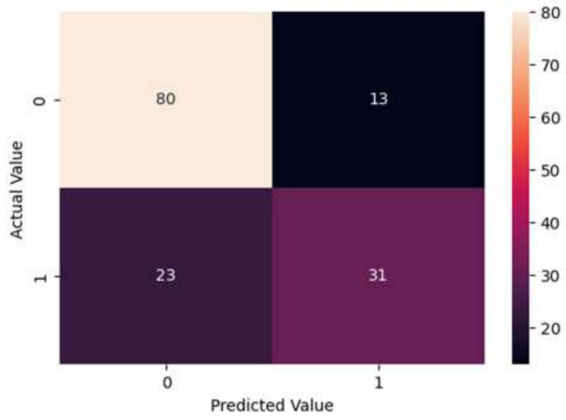
	precision	recall	f1-score	support
0	0.78	0.86	0.82	93
1	0.71	0.59	0.65	54
accuracy			0.76	147
macro avg	0.75	0.73	0.73	147
weighted avg	0.76	0.76	0.76	147

PERFORMACE OF MODEL IS 0.76 it is UNDER FITING Model High bias and High variance

PERFORMACE OF MODEL IS (TP+TN \ TP+TN+FN+FP)

A model with high variance may represent the dataset accurately but could lead to overfitting to noisy or otherwise unrepresentative training data. In comparison, a model with high bias may underfit the training data due to a simpler model that overlooks regularities in the data

Linear Discriminant Analysis



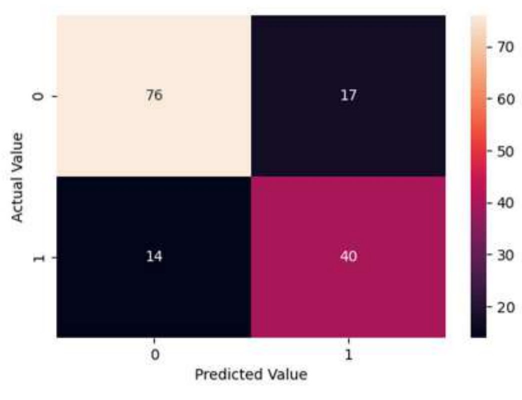
	precision	recall	f1-score	support
0	0.78	0.86	0.82	93
1	0.70	0.57	0.63	54
accuracy			0.76	147
macro avg	0.74	0.72	0.72	147
weighted avg	0.75	0.76	0.75	147

PERFORMACE OF MODEL IS 0.76 it is UNDER FITING Model High bias and High variance

PERFORMACE OF MODEL IS (TP+TN \ TP+TN+FN+FP)

A model with high variance may represent the dataset accurately but could lead to overfitting to noisy or otherwise unrepresentative training data. In comparison, a model with high bias may underfit the training data due to a simpler model that overlooks regularities in the data

Decision Tree Classifier



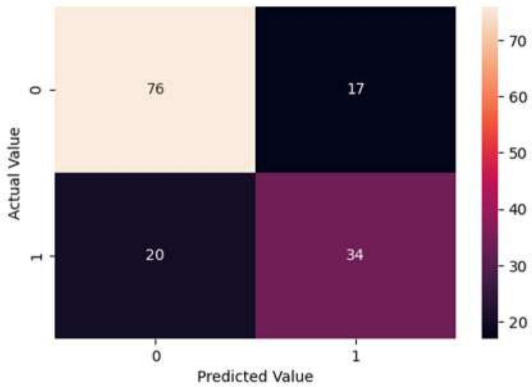
	precision	recall	f1-score	support
0	0.84	0.82	0.83	93
1	0.70	0.74	0.72	54
accuracy			0.79	147
macro avg	0.77	0.78	0.78	147
weighted avg	0.79	0.79	0.79	147

PERFORMACE OF MODEL IS 0.79 it is UNDER FITING Model High bias and High variance

PERFORMACE OF MODEL IS (TP+TN \ TP+TN+FN+FP)

A model with high variance may represent the dataset accurately but could lead to overfitting to noisy or otherwise unrepresentative training data. In comparison, a model with high bias may underfit the training data due to a simpler model that overlooks regularities in the data

Naïve Bayes Model



	precision	recall	f1-score	support
0	0.79	0.82	0.80	93
1	0.67	0.63	0.65	54
accuracy			0.75	147
macro avg	0.73	0.72	0.73	147
weighted avg	0.75	0.75	0.75	147

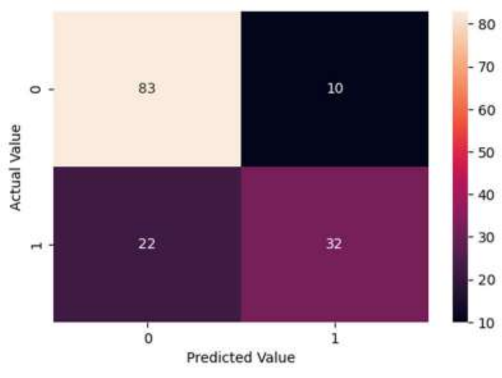
PERFORMACE OF MODEL IS 0.75 it is UNDER FITING Model

High bias and High variance

PERFORMACE OF MODEL IS (TP+TN \ TP+TN+FN+FP)

A model with high variance may represent the dataset accurately but could lead to overfitting to noisy or otherwise unrepresentative training data. In comparison, a model with high bias may underfit the training data due to a simpler model that overlooks regularities in the data

KNN Model



	precision	recall	f1-score	support
0	0.79	0.89	0.84	93
1	0.76	0.59	0.67	54
accuracy			0.78	147
macro avg	0.78	0.74	0.75	147
weighted avg	0.78	0.78	0.78	147

PERFORMACE OF MODEL IS 0.78 it is UNDER FITING Model

High bias and High variance

PERFORMACE OF MODEL IS (TP+TN \ TP+TN+FN+FP)

A model with high variance may represent the dataset accurately but could lead to overfitting to noisy or otherwise unrepresentative training data. In comparison, a model with high bias may underfit the training data due to a simpler model that overlooks regularities in the data

Random Forest Model

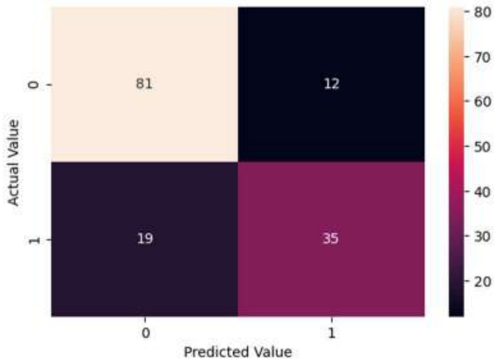
	precision	recall	f1-score	support
0	0.85	0.90	0.88	93
1	0.81	0.72	0.76	54
accuracy			0.84	147
macro avg	0.83	0.81	0.82	147
weighted avg	0.84	0.84	0.83	147

PERFORMACE OF MODEL IS 0.84 it is UNDER FITING Model
High bias and High variance

PERFORMACE OF MODEL IS (TP+TN \ TP+TN+FN+FP)

A model with high variance may represent the dataset accurately but could lead to overfitting to noisy or otherwise unrepresentative training data. In comparison, a model with high bias may underfit the training data due to a simpler model that overlooks regularities in the data

Boosting Classifier Model using Gradient boost.



	precision	recall	f1-score	support
0	0.81	0.87	0.84	93
1	0.74	0.65	0.69	54
accuracy			0.79	147
macro avg	0.78	0.76	0.77	147
weighted avg	0.79	0.79	0.79	147

PERFORMACE OF MODEL IS 0.79 it is UNDER FITING Model
High bias and High variance

PERFORMACE OF MODEL IS (TP+TN \ TP+TN+FN+FP)

A model with high variance may represent the dataset accurately but could lead to overfitting to noisy or otherwise unrepresentative training data. In comparison, a model with high bias may underfit the training data due to a simpler model that overlooks regularities in the data

TRAINING ERROR	2%	TRAINING ERROR	26%	TRAINING ERROR	<10%
TEST ERROR	25%	TEST ERROR	25%	TEST ERROR	<10%
OVERFITING MODEL		**UNDERFITING MODEL**		**GENALIZED MODEL**	

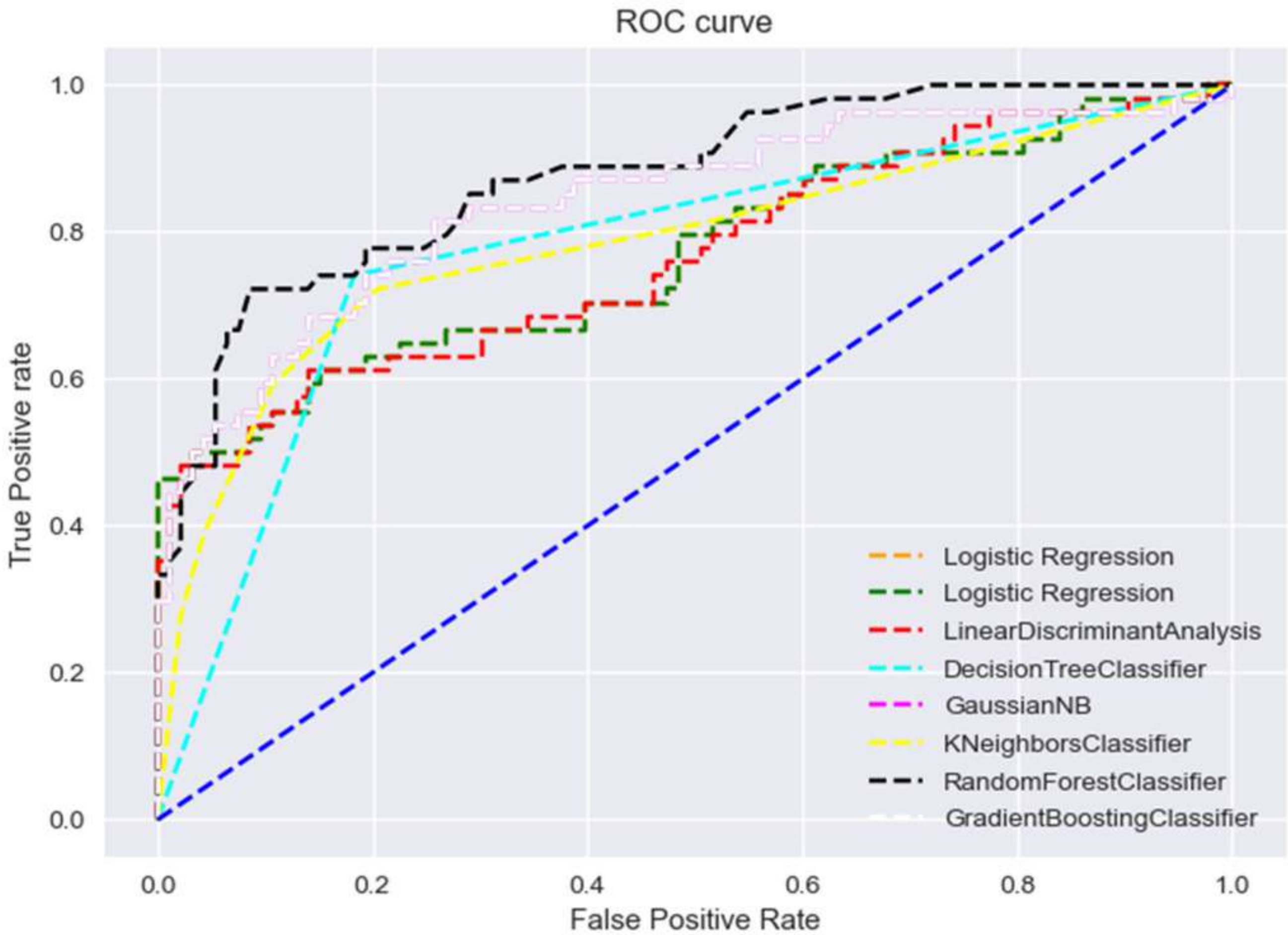
An 75-89% model performance can be considered good, but whether it is sufficient for testing data depends on the specific context and requirements of the task. Generally, a model with an accuracy of 75-89% indicates that it is making correct predictions for 75-89% of the test data points

When evaluating model performance, it's important to consider the nature of the problem you're trying to solve and the baseline or industry standards for that specific task. Some tasks may require higher accuracy rates, while others may have lower expectations. Additionally, you should also consider other metrics such as precision, recall, F1-score, or area under the ROC curve, depending on the nature of the problem.

Furthermore, it's important to ensure that the test data is representative of the real-world data the model is likely to encounter. If the test data differs significantly from the training data, the model's performance may not generalize well to new, unseen data.

ROC AND AUC SCORE

Logistic Regression Model	.76
Linear Discriminant Analysis	.76
Decision Tree Classifier	.77
Naïve Bayes Model	.84
KNN Model	.78
Random Forest ModeL	.87
Boosting Classifier Model Gradient boosting	.84



ROC (Receiver Operating Characteristic) and AUC (Area Under the Curve) are metrics commonly used to evaluate the performance of binary classification models.

The ROC curve is a graphical representation of the true positive rate (sensitivity) against the false positive rate (1-specificity) at different classification thresholds. It shows how well the model can distinguish between the positive and negative classes as the threshold for classifying instances is varied.

The AUC score is the area under the ROC curve. It provides a single numerical value that represents the overall performance of the model

BEST PERFORM MODEL

RANDOM FOREST MODEL IS PERFORM BEST ALL OF THE MODEL

ACCURACY IS 0.84

	precision	recall	f1-score	support
0	0.85	0.90	0.88	93
1	0.81	0.72	0.76	54
accuracy			0.84	147
macro avg	0.83	0.81	0.82	147
weighted avg	0.84	0.84	0.83	147

BUSINESS INSIGHTS

Most of the people use public transport, compared to private transport, it is almost approx 1.7-2 times . If 5 people go from home to office by private transport, then 8 to 10 people go to office by public transport.

The distance between home and office for most of the employers is between 5 to 18-20 kms. and the mostly male emolyes are there Approox Ratio of MALE : FEMALE(5:2)

As the number of people is increasing, work experience and salary are also increasing along with it.

EmploysApprox salary between 6-18 LPA

TEXT MINING

Deal (Dependent Variable) and Description columns into a separate data frame

	deal	description
0	False	Bluetooth device implant for your ear.
1	True	Retail and wholesale pie factory with two reta...
2	True	Ava the Elephant is a godsend for frazzled par...
3	False	Organizing, packing, and moving services deliv...
4	False	Interactive media centers for healthcare waiti...
...	...	...
490	True	Zoom Interiors is a virtual service for interi...
491	True	Spikeball started out as a casual outdoors gam...
492	True	Shark Wheel is out to literally reinvent the w...
493	False	Adriana Montano wants to open the first Cat Ca...
494	True	Sway Motorsports makes a three-wheeled, all-el...

total 495 rows × 2 columns

two corpora, one for those who secured a Deal, the other for those who did not secure a deal.

	deal	description
0	False	Bluetooth device implant for your ear.
3	False	Organizing, packing, and moving services deliv...
4	False	Interactive media centers for healthcare waiti...
6	False	A mixed martial arts clothing line looking to ...
7	False	Attach Noted is a detachable "arm" that holds ...
...	...	...
482	False	Buck Mason makes high-quality men's clothing i...
484	False	Frameri answers the question, "Why aren't your...
485	False	The Paleo Diet Bar is a nutrition bar that is ...
488	False	Sunscreen Mist adds another point of access fo...
493	False	Adriana Montano wants to open the first Cat Ca...

244 rows × 2 columns Secured a Deal (TRUE)

	deal	description
1	True	Retail and wholesale pie factory with two reta...
2	True	Ava the Elephant is a godsend for frazzled par...
5	True	One of the first entrepreneurs to pitch on Sha...
9	True	An educational record label and publishing hou...
10	True	A battery-operated cooking device that siphons...
...	...	...
489	True	SynDaver Labs makes synthetic body parts for u...
490	True	Zoom Interiors is a virtual service for interi...
491	True	Spikeball started out as a casual outdoors gam...
492	True	Shark Wheel is out to literally reinvent the w...
494	True	Sway Motorsports makes a three-wheeled, all-el...

251 rows × 2 columns did not secure deal. (FALSE)

Secured a Deal (TRUE)

```
fd.most_common(40)

[('make', 28),
 ('free', 23),
 ('children', 21),
 ('designed', 21),
 ('easy', 20),
 ('natural', 20),
 ('line', 19),
 ('products', 19),
 ('use', 19),
 ('water', 19),
 ('one', 18),
 ('way', 18),
 ('without', 18),
 ('kids', 17),
 ('new', 15),
 ('product', 15),
 ('online', 15),
 ('built', 14),
 ('system', 14),
 ('offers', 13),
 ('look', 13),
 ('service', 13),
 ('keep', 13),
 ('home', 13),
 ('box', 13),
 ('easier', 12),
 ('ingredients', 12),
 ('need', 12),
 ('using', 12),
 ('fun', 12),
 ('design', 12),
 ('light', 12),
 ('skin', 12),
 ('get', 11),
 ('play', 11),
 ('time', 11),
 ('well', 11),
 ('coffee', 11),
 ('safe', 11),
 ('quality', 11)]
```

Characters for both the corpuses.

did not secure deal. (FALSE)

```
[('designed', 19),
 ('make', 19),
 ('use', 17),
 ('water', 17),
 ('system', 16),
 ('one', 15),
 ('product', 15),
 ('online', 15),
 ('bottle', 14),
 ('without', 14),
 ('fun', 13),
 ('products', 13),
 ('device', 12),
 ('home', 12),
 ('line', 11),
 ('kids', 11),
 ('way', 11),
 ('premium', 11),
 ('natural', 11),
 ('allows', 11),
 ('people', 11),
 ('balm', 11),
 ('women', 10),
 ('cards', 10),
 ('featuring', 10),
 ('help', 10),
 ('helps', 10),
 ('every', 10),
 ('free', 10),
 ('hair', 10),
 ('accessories', 9),
 ('service', 9),
 ('instead', 9),
 ('unique', 9),
 ('food', 9),
 ('full', 9),
 ('eco', 9),
 ('traditional', 9),
 ('comes', 9),
 ('available', 9)]
```

```
from nltk.corpus import stopwords
# Make a list of english stopwords
stopwords = nltk.corpus.stopwords.words("english")

# Extend the list with your own custom stopwords
my_stopwords = ['also','made','makes','like','this','even','company']
stopwords.extend(my_stopwords)
```

text_token: after stoping stop word

	deal	description	text_token
1	True	retail and wholesale pie factory with two reta...	[retail, wholesale, pie, factory, two, retail,...
2	True	ava the elephant is a godsend for frazzled par...	[ava, elephant, godsend, frazzled, parents, yo...
5	True	one of the first entrepreneurs to pitch on sha...	[one, first, entrepreneurs, pitch, shark, tank...
9	True	an educational record label and publishing hou...	[educational, record, label, publishing, house...
10	True	a battery-operated cooking device that siphons...	[battery, operated, cooking, device, siphons, ...

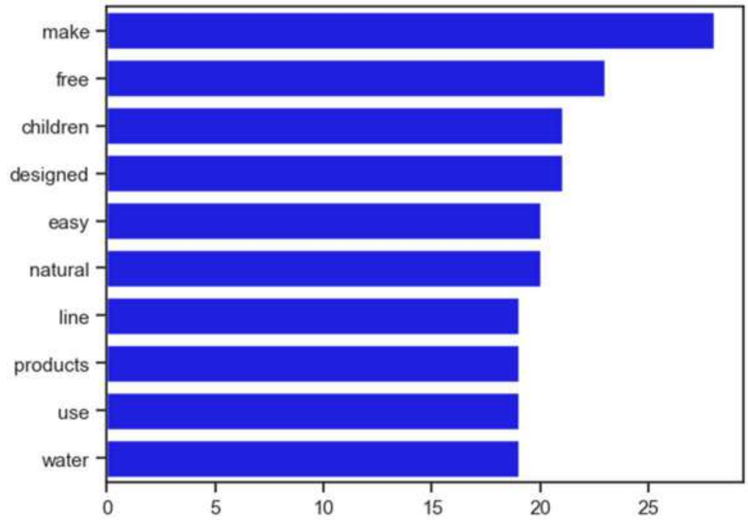
did not secure deal. (FALSE)

	deal	description	text_token
0	False	bluetooth device implant for your ear.	[bluetooth, device, implant, ear]
3	False	organizing, packing, and moving services deliv...	[organizing, packing, moving, services, delive...
4	False	interactive media centers for healthcare waiti...	[interactive, media, centers, healthcare, wait...

Secured a Deal (TRUE) top 10 word use with count after stopword

TOP 3 IS MAKE, FREE, CHILDREN

```
[('make', 28),
 ('free', 23),
 ('children', 21),
 ('designed', 21),
 ('easy', 20),
 ('natural', 20),
 ('line', 19),
 ('products', 19),
 ('use', 19),
 ('water', 19)]
```

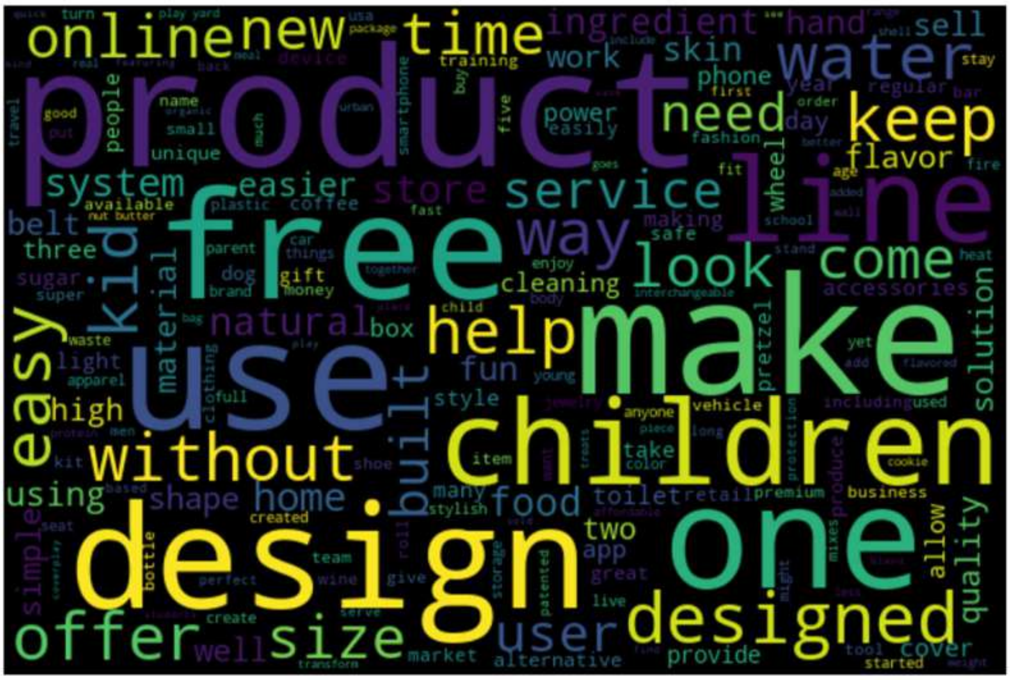


did not secure deal. (FALSE) top 10 word use with count after stopword

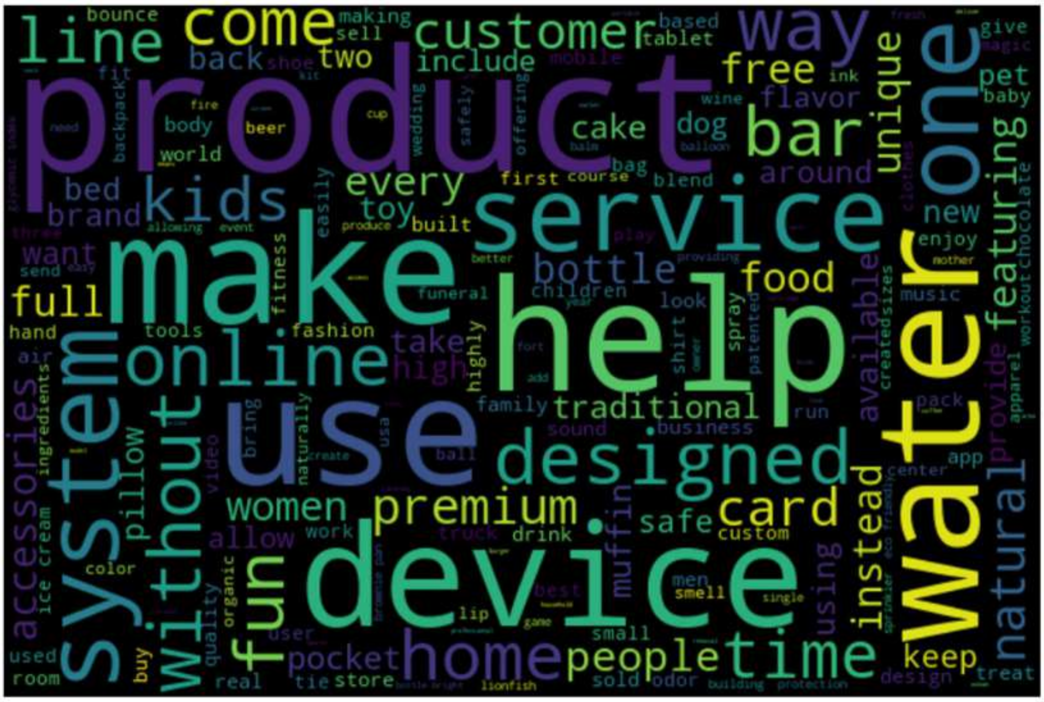
TOP 3 IS DESIGNED, MAKE, USE

```
[('designed', 19),
 ('make', 19),
 ('use', 17),
 ('water', 17),
 ('system', 16),
 ('one', 15),
 ('product', 15),
 ('online', 15),
 ('bottle', 14),
 ('without', 14)]
```

Secured a Deal (TRUE)



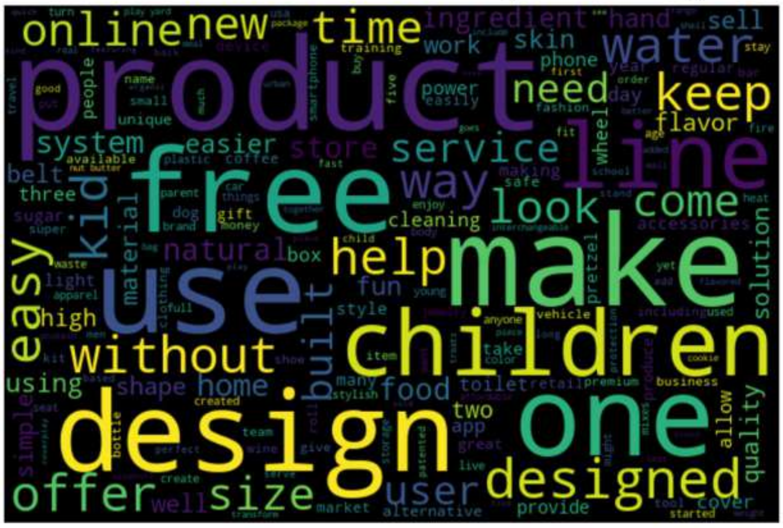
did not secure deal. (FALSE)



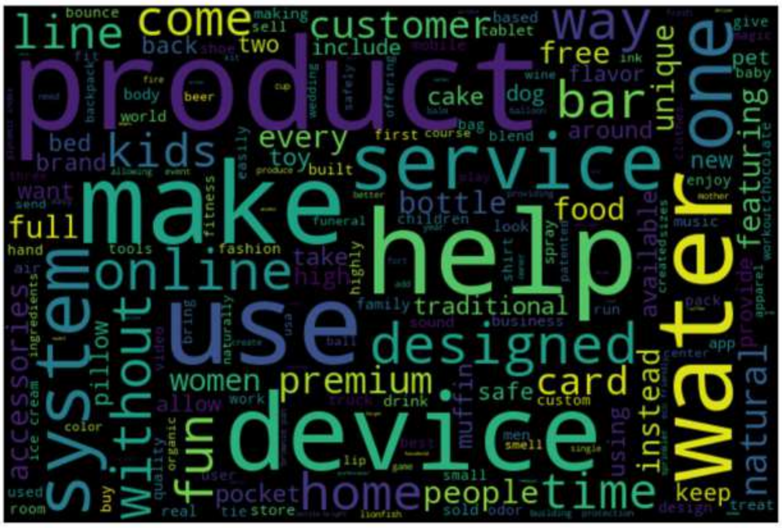
A word cloud is a visual representation of text data, where words are displayed in varying sizes based on their frequency or importance. It provides a quick overview of the most common or significant words in a given text or collection of texts.

Looking at the word clouds, is it true that the entrepreneurs who introduced devices are less likely to secure a deal based on your analysis

Secured a Deal (TRUE)



did not secure deal. (FALSE)



it is true that the entrepreneurs who introduced devices are less likely to secure a deal based on our analysis because, enetrepeneurs introduced devices, names these are looking at the worldcloud in a very small size We can easily see that is highlighted and in big size letters, these are not perfect coorelated to device and device specification, so it is true less likely to secure a deal