1. Read the data

```
---------------------------------------------------------------------------
FileNotFoundError                         Traceback (most recent call last)
Cell In[10], line 1
----> 1 df = pd.read_csv(r'notebook\data\depression_data.csv')

File c:\ML\mentalhealthapp\venv\lib\site-packages\pandas\io\parsers\readers.py:912, in read_csv(fi
lepath_or_buffer, sep, delimiter, header, names, index_col, usecols, dtype, engine, converters, tr
ue_values, false_values, skipinitialspace, skiprows, skipfooter, nrows, na_values, keep_default_n
a, na_filter, verbose, skip_blank_lines, parse_dates, infer_datetime_format, keep_date_col, date_p
arser, date_format, dayfirst, cache_dates, iterator, chunksize, compression, thousands, decimal, l
ineterminator, quotechar, quoting, doublequote, escapechar, comment, encoding, encoding_errors, di
alect, on_bad_lines, delim_whitespace, low_memory, memory_map, float_precision, storage_options, d
type_backend)
    899 kwds_defaults = _refine_defaults_read(
    900     dialect,
    901     delimiter,
   (...)
    908     dtype_backend=dtype_backend,
    909 )
    910 kwds.update(kwds defaults)
--> 912 return _read(filepath_or_buffer, kwds)

File c:\ML\mentalhealthapp\venv\lib\site-packages\pandas\io\parsers\readers.py:577, in _read(filep
ath_or_buffer, kwds)
    574 _validate_names(kwds.get("names", None))
    576 # Create the parser.
--> 577 parser = TextFileReader(filepath_or_buffer, **kwds)
    579 if chunksize or iterator:
    580     return parser

File c:\ML\mentalhealthapp\venv\lib\site-packages\pandas\io\parsers\readers.py:1407, in TextFileRe
ader.__init__(self, f, engine, **kwds)
   1404     self.options["has_index_names"] = kwds["has_index_names"]
   1406 self.handles: IOHandles | None = None
-> 1407 self._engine = self._make_engine(f, self.engine)

File c:\ML\mentalhealthapp\venv\lib\site-packages\pandas\io\parsers\readers.py:1661, in TextFileRe
ader._make_engine(self, f, engine)
   1659     if "b" not in mode:
   1660         mode += "b"
-> 1661 self.handles = get_handle(
   1662     f,
   1663     mode,
   1664     encoding=self.options.get("encoding", None),
   1665     compression=self.options.get("compression", None),
   1666     memory map=self.options.get("memory_map", False),
   1667     is text=is text,
   1668     errors=self.options.get("encoding errors", "strict"),
   1669     storage_options=self.options.get("storage_options", None),
   1670 )
   1671 assert self.handles is not None
   1672 f = self.handles.handle

File c:\ML\mentalhealthapp\venv\lib\site-packages\pandas\io\common.py:859, in get_handle(path_or_b
uf, mode, encoding, compression, memory_map, is_text, errors, storage_options)
    854 elif isinstance(handle, str):
    855     # Check whether the filename is to be opened in binary mode.
    856     # Binary mode does not support 'encoding' and 'newline'.
    857     if ioargs.encoding and "b" not in ioargs.mode:
    858         # Encoding
--> 859         handle = open(
    860             handle,
    861             ioargs.mode,
    862             encoding=ioargs.encoding,
    863             errors=errors,
    864             newline="",
    865         )
    866     else:
    867         # Binary mode
    868         handle = open(handle, ioargs.mode)

FileNotFoundError: [Errno 2] No such file or directory: 'notebook\\data\\depression_data.csv'
```

2. Get the look & feel of the data

	Name	Age	Marital Status	Education Level	Number of Children	Smoking Status	Physical Activity Level	Employment Status	Income	Alcohol Consumption	Dietary Habits	S Patt
0	Christine Barker	31	Married	Bachelor's Degree	2	Non-smoker	Active	Unemployed	26265.67	Moderate	Moderate	
1	Jacqueline Lewis	55	Married	High School	1	Non-smoker	Sedentary	Employed	42710.36	High	Unhealthy	
2	Shannon Church	78	Widowed	Master's Degree	1	Non-smoker	Sedentary	Employed	125332.79	Low	Unhealthy	(
3	Charles Jordan	58	Divorced	Master's Degree	3	Non-smoker	Moderate	Unemployed	9992.78	Moderate	Moderate	
4	Michael Rich	18	Single	High School	0	Non-smoker	Sedentary	Unemployed	8595.08	Low	Moderate	

	Name	Age	Marital Status	Education Level	Number of Children	Smoking Status	Physical Activity Level	Employment Status	Income	Alcohol Consumption	Dietary Habits
413763	Sean Miller	68	Married	Master's Degree	0	Former	Moderate	Employed	109233.43	Low	Healthy
413764	Christina Brown	26	Single	Bachelor's Degree	0	Current	Active	Employed	96760.97	Low	Healthy
413765	Matthew Jenkins	57	Married	Bachelor's Degree	0	Non-smoker	Sedentary	Employed	77353.26	Moderate	Moderate
413766	Gary Faulkner	71	Married	Associate Degree	2	Non-smoker	Sedentary	Unemployed	24557.08	Moderate	Moderate
413767	Joseph Johnson	62	Widowed	Master's Degree	0	Former	Moderate	Employed	107125.74	Moderate	Healthy

3. Understand the datatypes & basic statistics of the data

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 413768 entries, 0 to 413767
Data columns (total 16 columns):
 #   Column                        Non-Null Count   Dtype  
---  ------                        --------------   -----  
 0   Name                          413768 non-null  object 
 1   Age                           413768 non-null  int64  
 2   Marital Status                413768 non-null  object 
 3   Education Level               413768 non-null  object 
 4   Number of Children            413768 non-null  int64  
 5   Smoking Status                413768 non-null  object 
 6   Physical Activity Level       413768 non-null  object 
 7   Employment Status             413768 non-null  object 
 8   Income                        413768 non-null  float64
 9   Alcohol Consumption           413768 non-null  object 
 10  Dietary Habits                413768 non-null  object 
 11  Sleep Patterns                413768 non-null  object 
 12  History of Mental Illness     413768 non-null  object 
 13  History of Substance Abuse    413768 non-null  object 
 14  Family History of Depression  413768 non-null  object 
 15  Chronic Medical Conditions    413768 non-null  object 
dtypes: float64(1), int64(2), object(13)
memory usage: 50.5+ MB
```

	Age	Number of Children	Income
count	413768.000000	413768.000000	413768.000000
mean	49.000713	1.298972	50661.707971
std	18.158759	1.237054	40624.100565
min	18.000000	0.000000	0.410000
25%	33.000000	0.000000	21001.030000
50%	49.000000	1.000000	37520.135000
75%	65.000000	2.000000	76616.300000
max	80.000000	4.000000	209995.220000

```
No     287943
Yes    125825
Name: History of Mental Illness, dtype: int64
```

```
Marital Status                  False
Education Level                 False
Number of Children              False
Smoking Status                  False
Physical Activity Level         False
Employment Status               False
Income                          False
Alcohol Consumption             False
Dietary Habits                  False
Sleep Patterns                  False
History of Mental Illness       False
History of Substance Abuse      False
Family History of Depression    False
Chronic Medical Conditions      False
Gender                          False
income_groups                   False
age_groups                      False
dtype: bool
```

```
False
```

Feature Engineering.

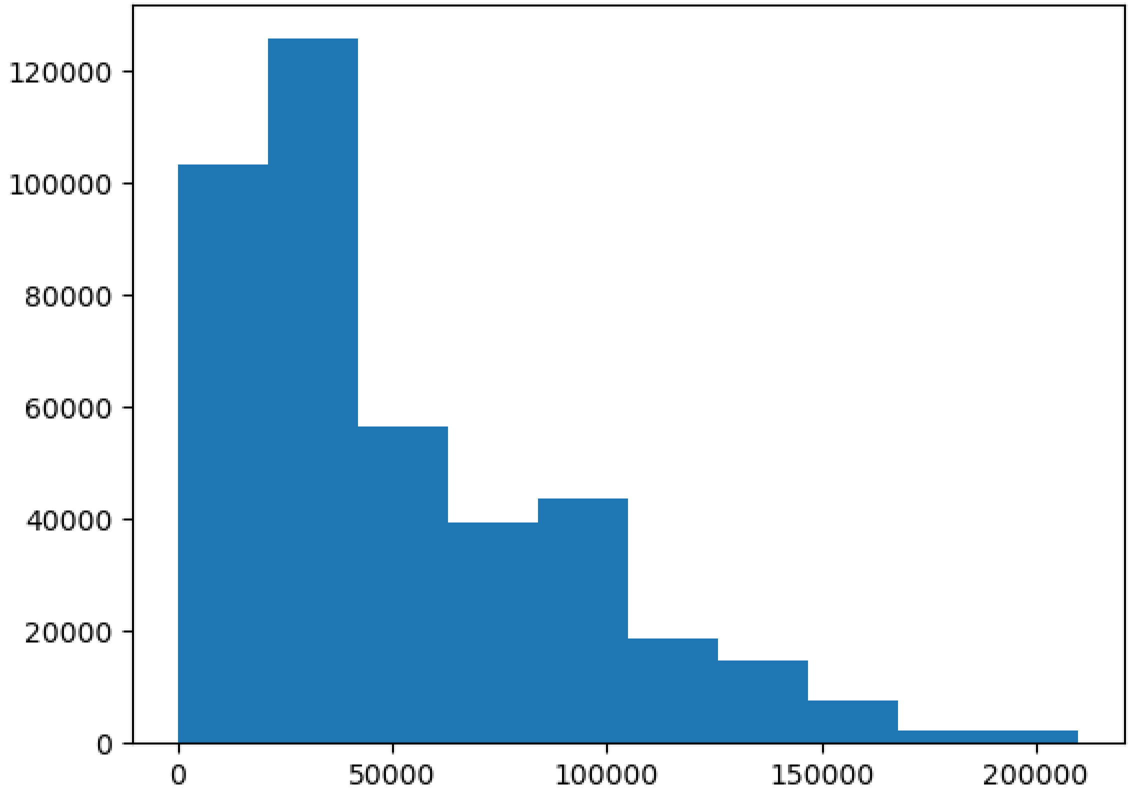
Generate gender from Name

```
male      209232
female    204536
Name: Gender, dtype: int64
```

```
209995.22
```

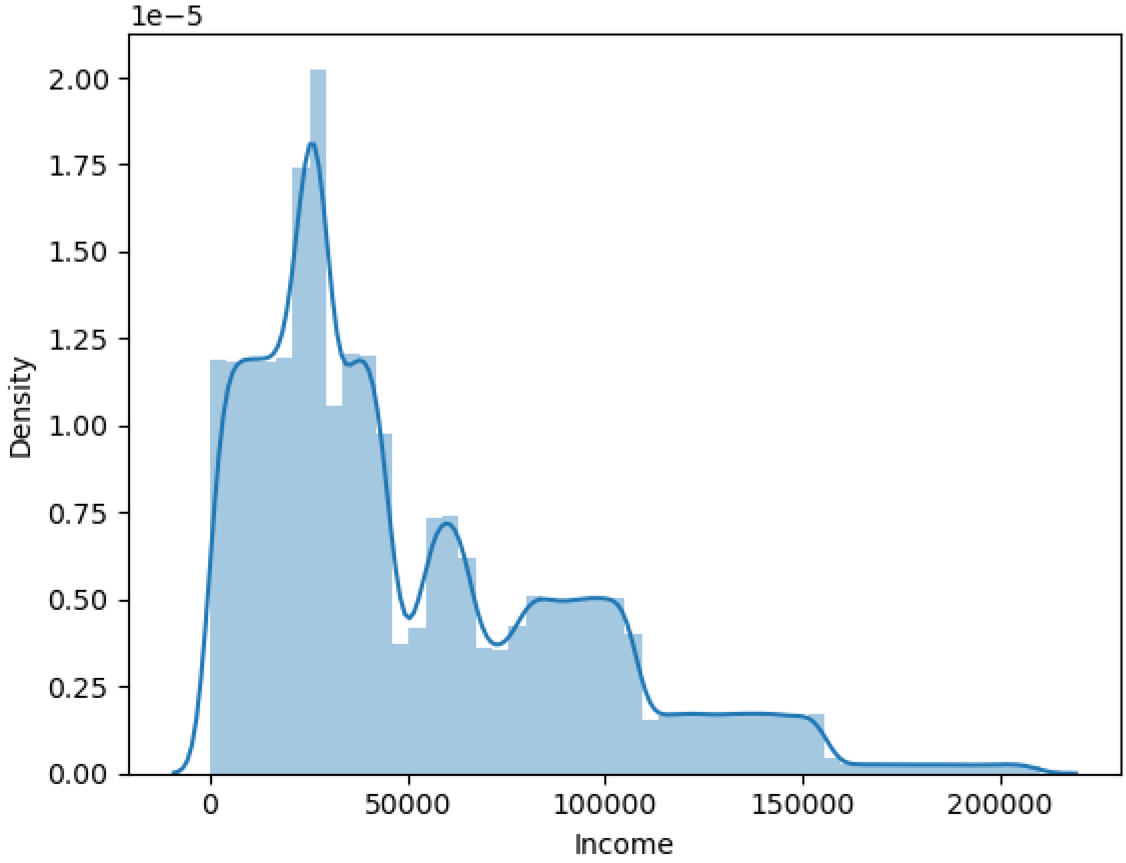
```
History of Mental Illness      No    Yes
Marital Status
Divorced                    23111   9618
Married                    167155  73289
Single                      50690  21420
Widowed                     46987  21498
```

```
(array([103438., 125739.,  56380.,  39450.,  43574.,  18596.,  14751.,
          7515.,   2179.,   2146.]),
 array([4.10000000e-01, 2.09998910e+04, 4.19993720e+04, 6.29988530e+04,
        8.39983340e+04, 1.04997815e+05, 1.25997296e+05, 1.46996777e+05,
        1.67996258e+05, 1.88995739e+05, 2.09995220e+05]),
 <BarContainer object of 10 artists>)
```

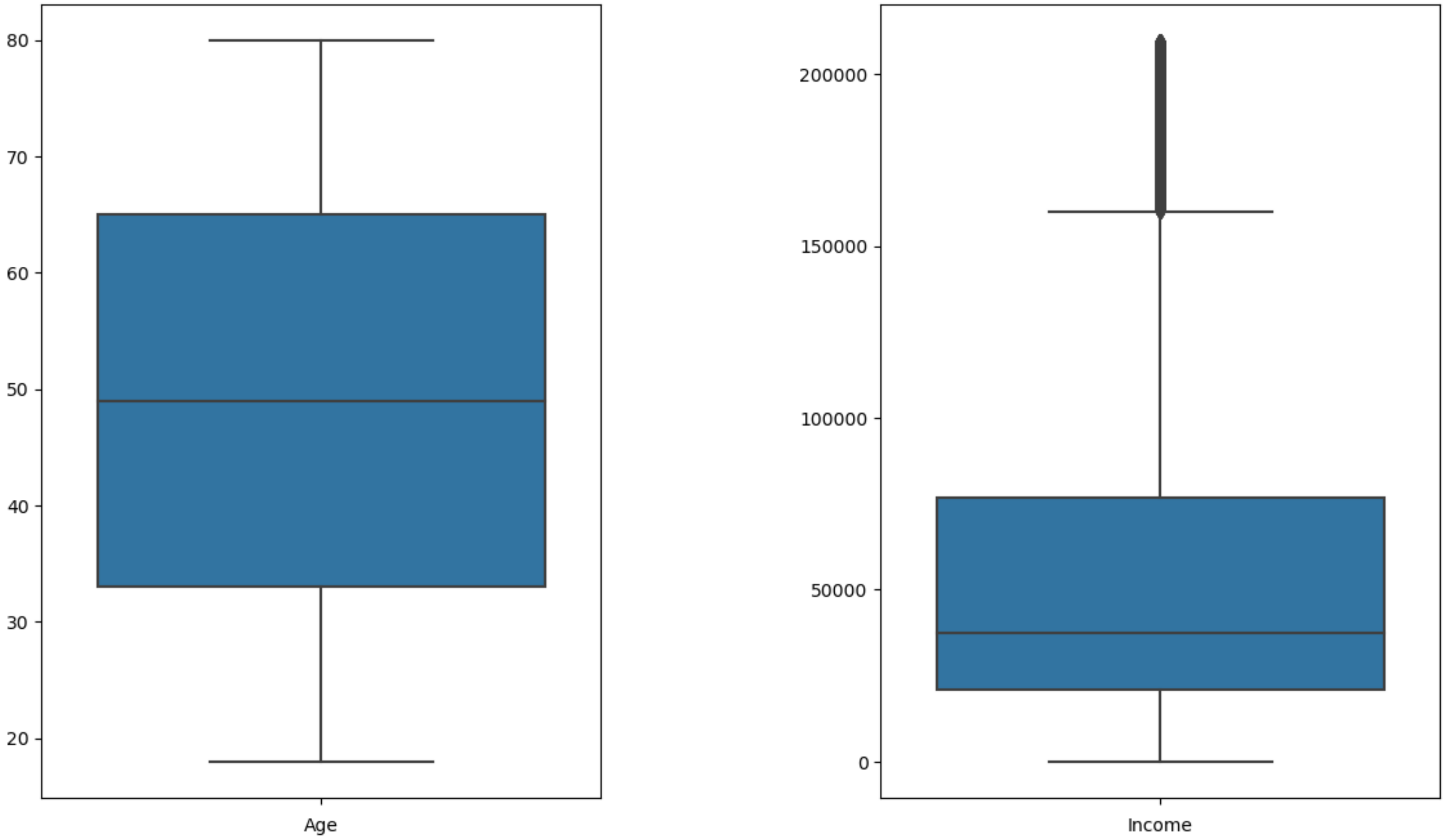


```
C:\Users\admin\anaconda3\lib\site-packages\seaborn\distributions.py:2619: FutureWarning: `distplot
` is a deprecated function and will be removed in a future version. Please adapt your code to use
either `displot` (a figure-level function with similar flexibility) or `histplot` (an axes-level f
unction for histograms).
  warnings.warn(msg, FutureWarning)
```

```
<AxesSubplot:xlabel='Income', ylabel='Density'>
```



```
1.086114667512213
```



The Income feature has the outliers & needs to be treated. I would like to go with capping method.

```
21001.03
```

```
76616.3
```

```
55615.270000000004
```

```
Upper limit 160039.20500000002
Lower limit -62421.875
```

	Age	Marital Status	Education Level	Number of Children	Smoking Status	Physical Activity Level	Employment Status	Income	Alcohol Consumption	Dietary Habits	Sleep Patterns
15	38	Married	PhD	0	Non-smoker	Moderate	Employed	202449.17	High	Healthy	Fair
105	53	Widowed	PhD	3	Former	Sedentary	Employed	169400.38	High	Unhealthy	Good
170	36	Divorced	PhD	3	Non-smoker	Sedentary	Employed	180084.56	High	Unhealthy	Fair
193	64	Married	PhD	1	Non-smoker	Moderate	Employed	193843.44	Low	Healthy	Poor
319	31	Married	PhD	1	Current	Moderate	Employed	177029.40	High	Unhealthy	Poor
...	...	...	...	...	...	...	...	...	...	...	...
413461	77	Widowed	PhD	0	Non-smoker	Moderate	Employed	200828.61	Moderate	Moderate	Fair
413563	76	Married	PhD	1	Former	Sedentary	Employed	164436.49	High	Unhealthy	Poor
413574	41	Divorced	PhD	0	Former	Sedentary	Employed	171921.83	Low	Moderate	Good
413577	30	Married	PhD	1	Non-smoker	Moderate	Employed	169051.46	High	Healthy	Poor
413754	34	Married	PhD	4	Non-smoker	Active	Employed	185657.43	Moderate	Moderate	Fair

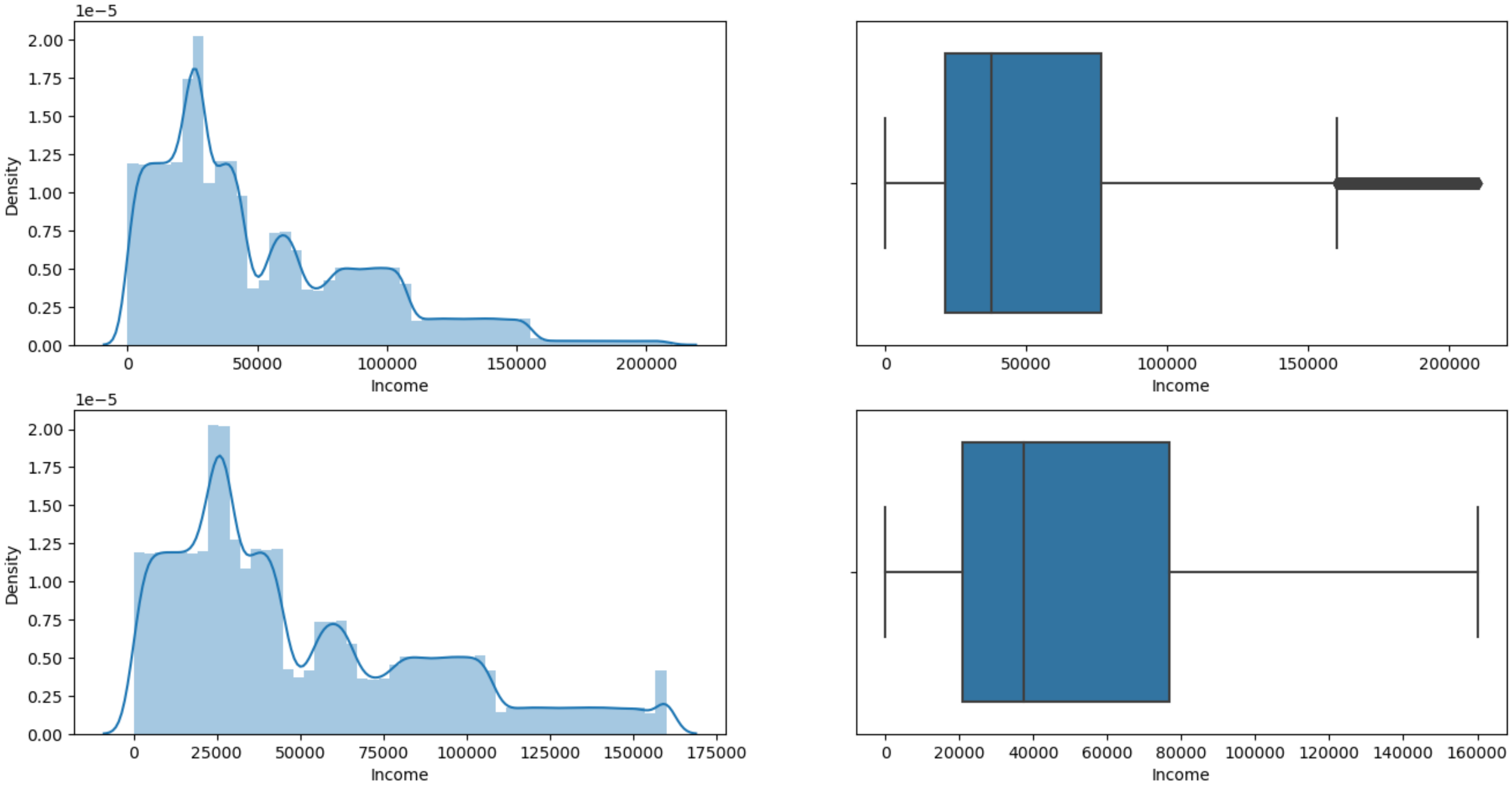
5157 rows × 19 columns

Age	Marital Status	Education Level	Number of Children	Smoking Status	Physical Activity Level	Employment Status	Income	Alcohol Consumption	Dietary Habits	Sleep Patterns	History of Mental Illness	Hist Sub

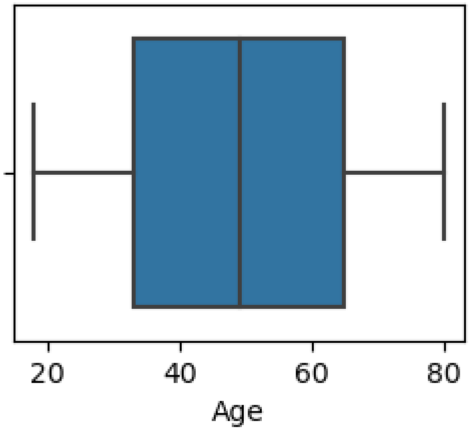
```
array(['PhD'], dtype=object)
```

All the outliers in Income column are PHD holders & rightly so as they have the highest educational degree & are generally paid higher compared to other degree holders.

```
C:\Users\admin\anaconda3\lib\site-packages\seaborn\distributions.py:2619: FutureWarning: `distplot
` is a deprecated function and will be removed in a future version. Please adapt your code to use
either `displot` (a figure-level function with similar flexibility) or `histplot` (an axes-level f
unction for histograms).
  warnings.warn(msg, FutureWarning)
C:\Users\admin\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the foll
owing variable as a keyword arg: x. From version 0.12, the only valid positional argument will be
`data`, and passing other arguments without an explicit keyword will result in an error or misinte
rpretation.
  warnings.warn(
C:\Users\admin\anaconda3\lib\site-packages\seaborn\distributions.py:2619: FutureWarning: `distplot
` is a deprecated function and will be removed in a future version. Please adapt your code to use
either `displot` (a figure-level function with similar flexibility) or `histplot` (an axes-level f
unction for histograms).
  warnings.warn(msg, FutureWarning)
C:\Users\admin\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the foll
owing variable as a keyword arg: x. From version 0.12, the only valid positional argument will be
`data`, and passing other arguments without an explicit keyword will result in an error or misinte
rpretation.
  warnings.warn(
<AxesSubplot:xlabel='Income'>
```

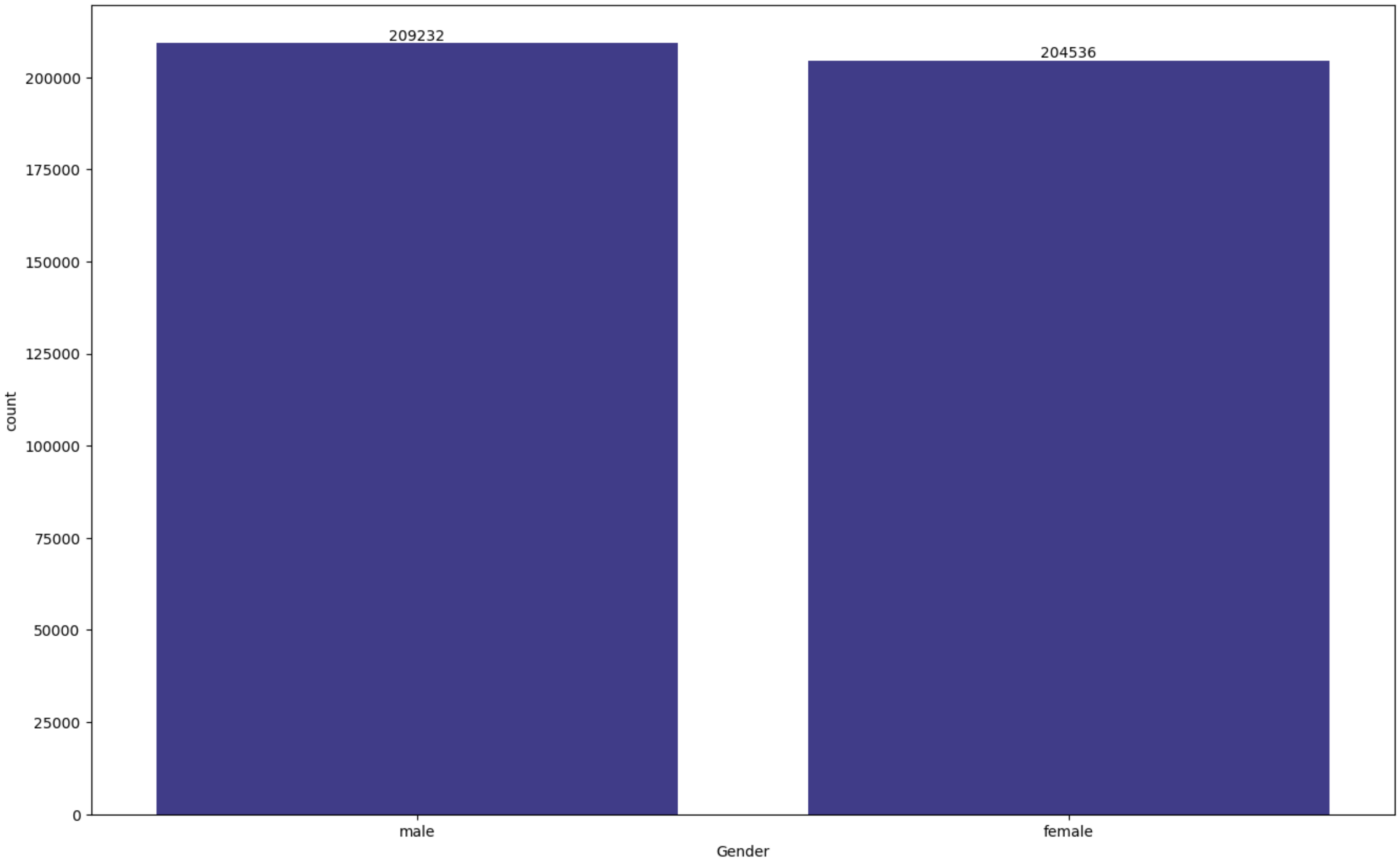


```
C:\Users\admin\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the foll
owing variable as a keyword arg: x. From version 0.12, the only valid positional argument will be
`data`, and passing other arguments without an explicit keyword will result in an error or misinte
rpretation.
  warnings.warn(
<AxesSubplot:xlabel='Age'>
```

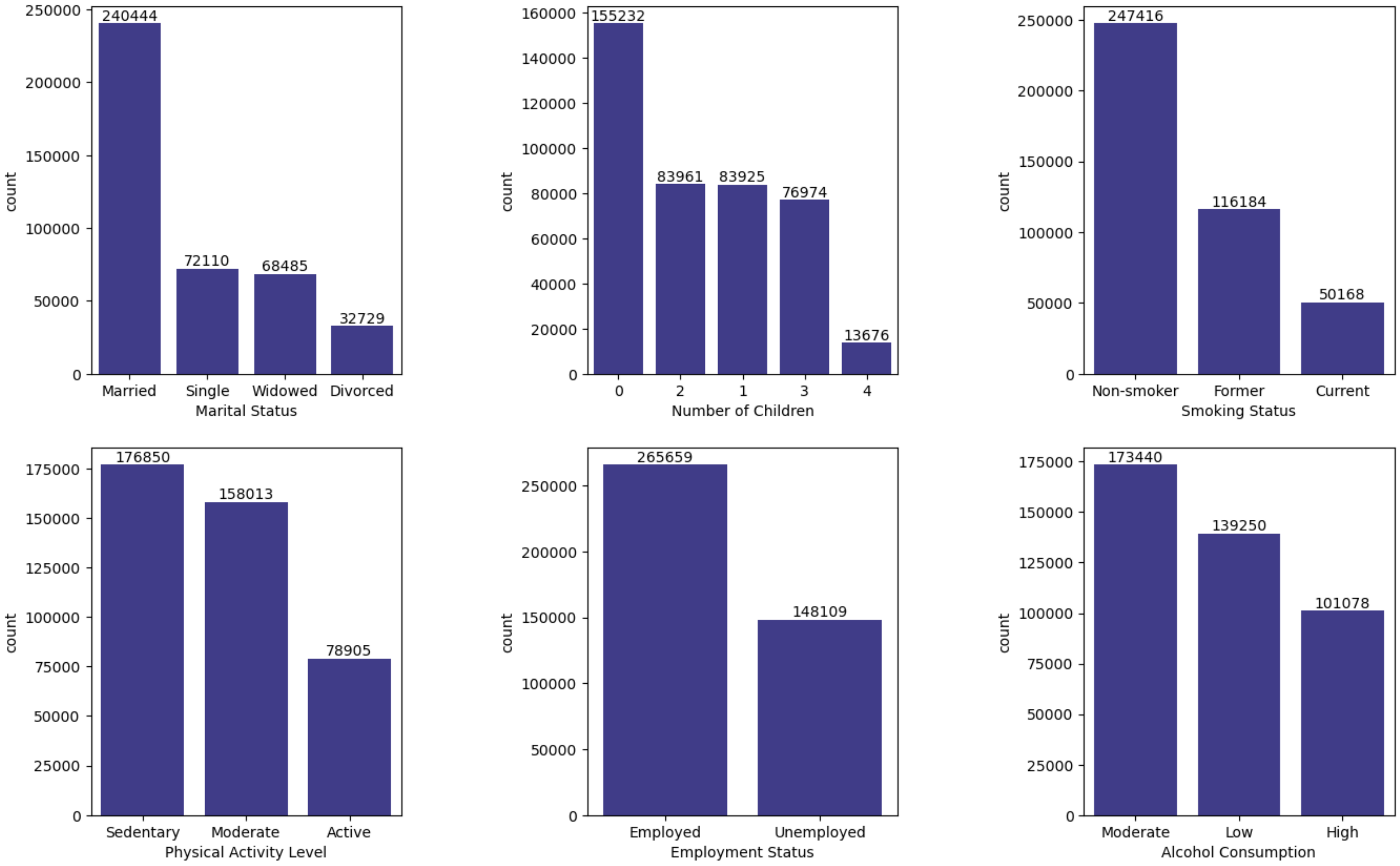


```
array(['Moderate', 'Unhealthy', 'Healthy'], dtype=object)
```

Univariate & Bi-variate Analysis

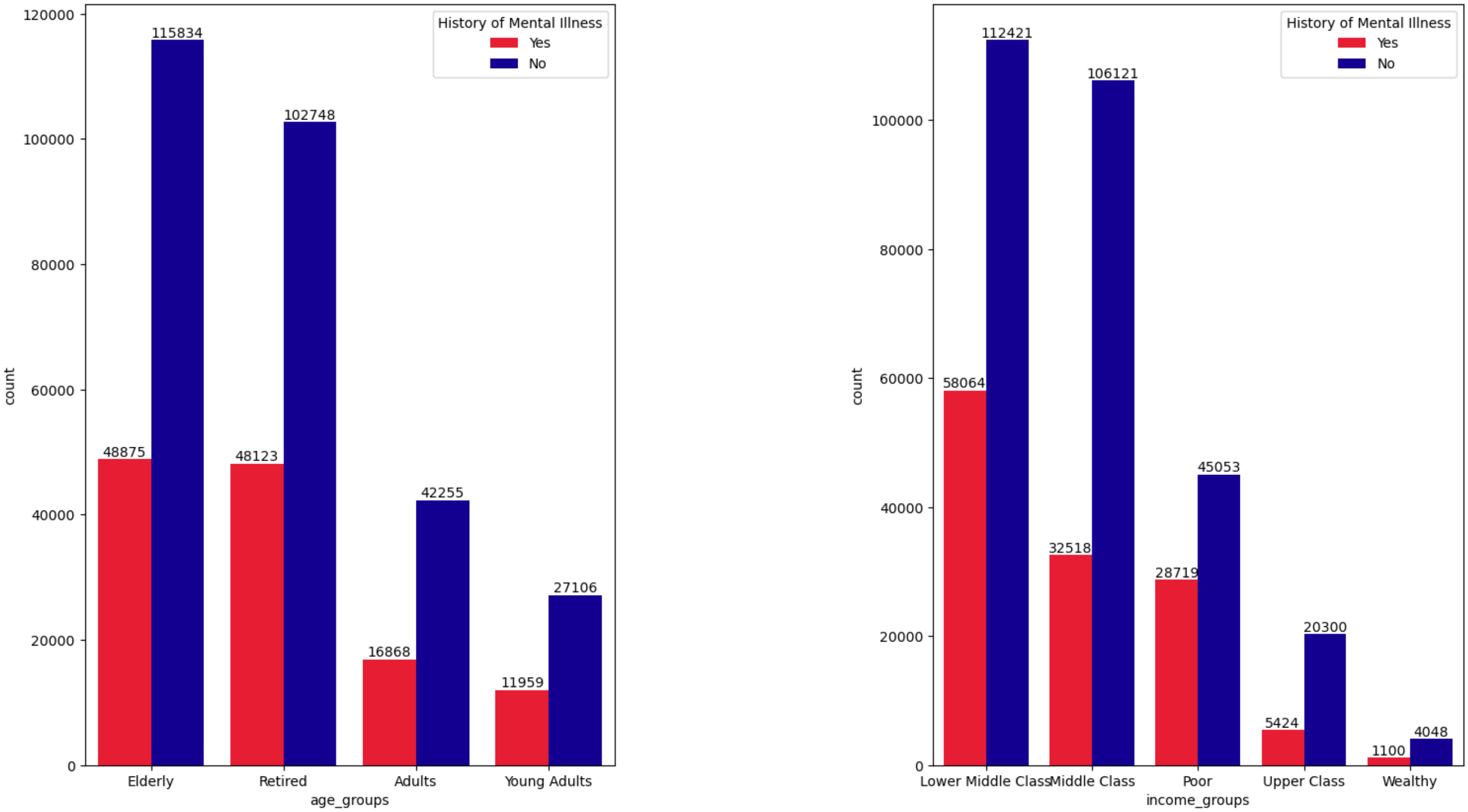


The gender feature seems to be having equal distribution. We can check the correlation between this feature with y variable .



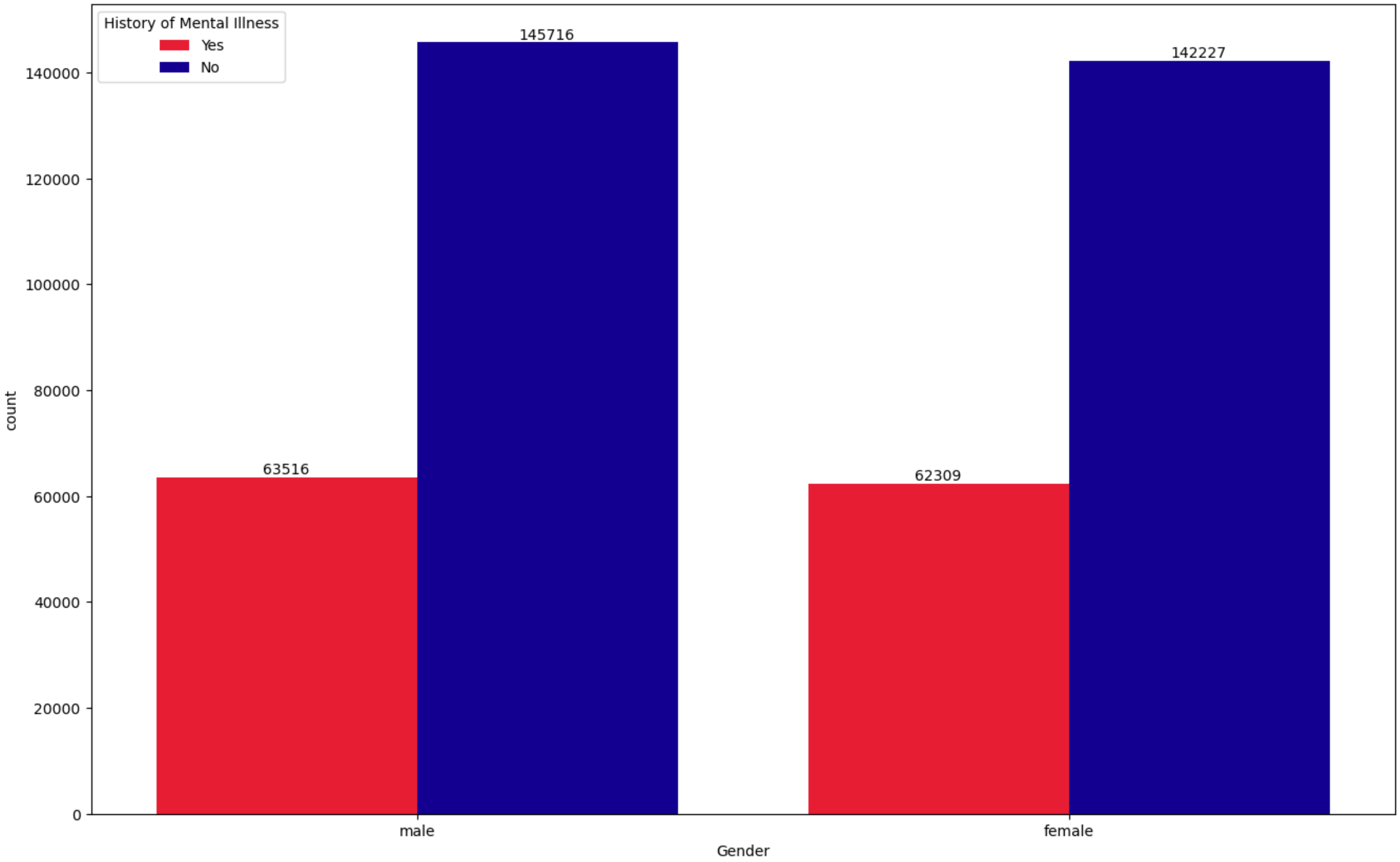
```
80
```

```
Elderly         164709
Retired         150871
Adults           59123
Young Adults     39065
Name: age_groups, dtype: int64
```

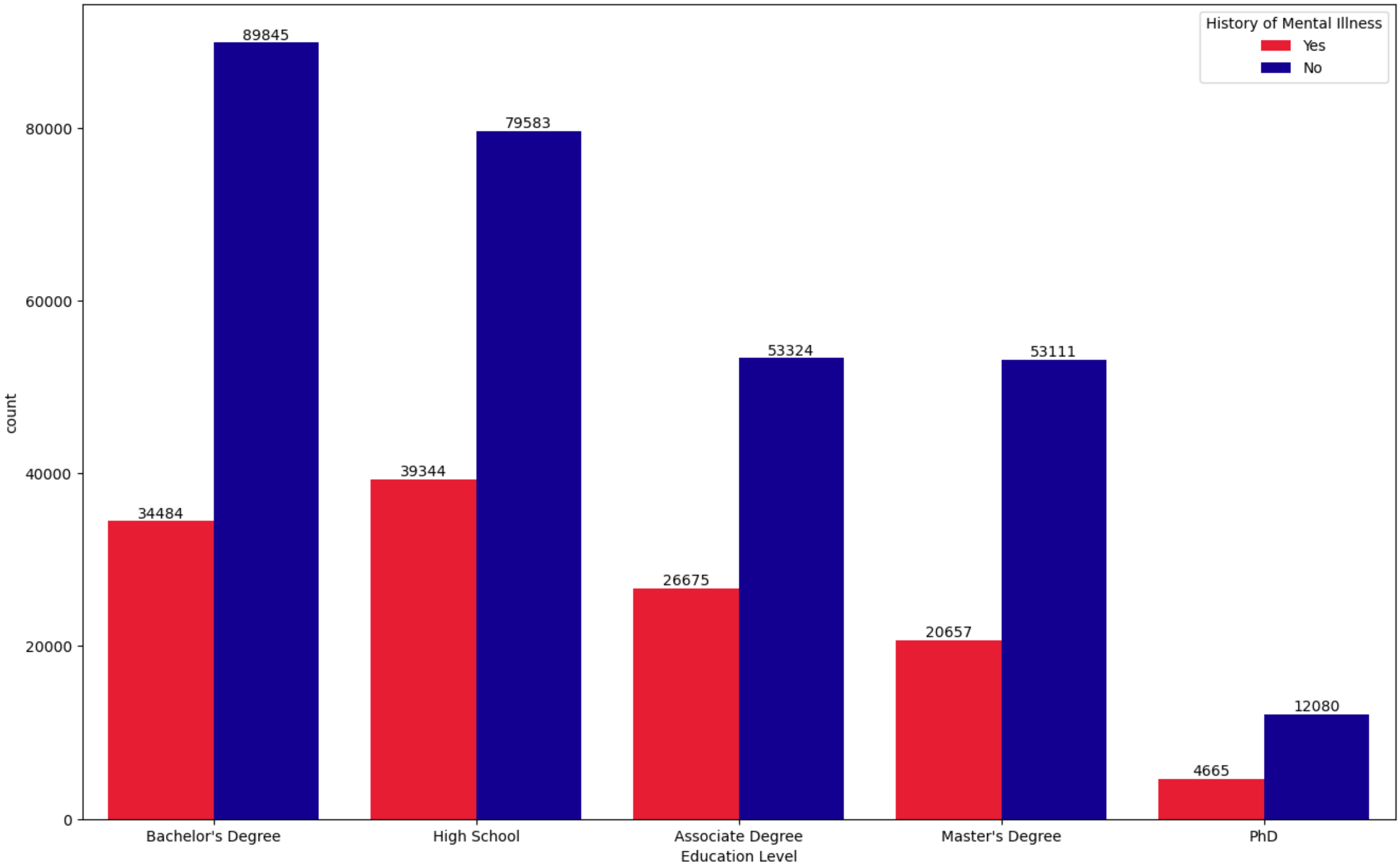


Elderly people have higher risk of mental illness compared to young people.

Financially weaker group of people have higher risk of mental illness compared to rich people.

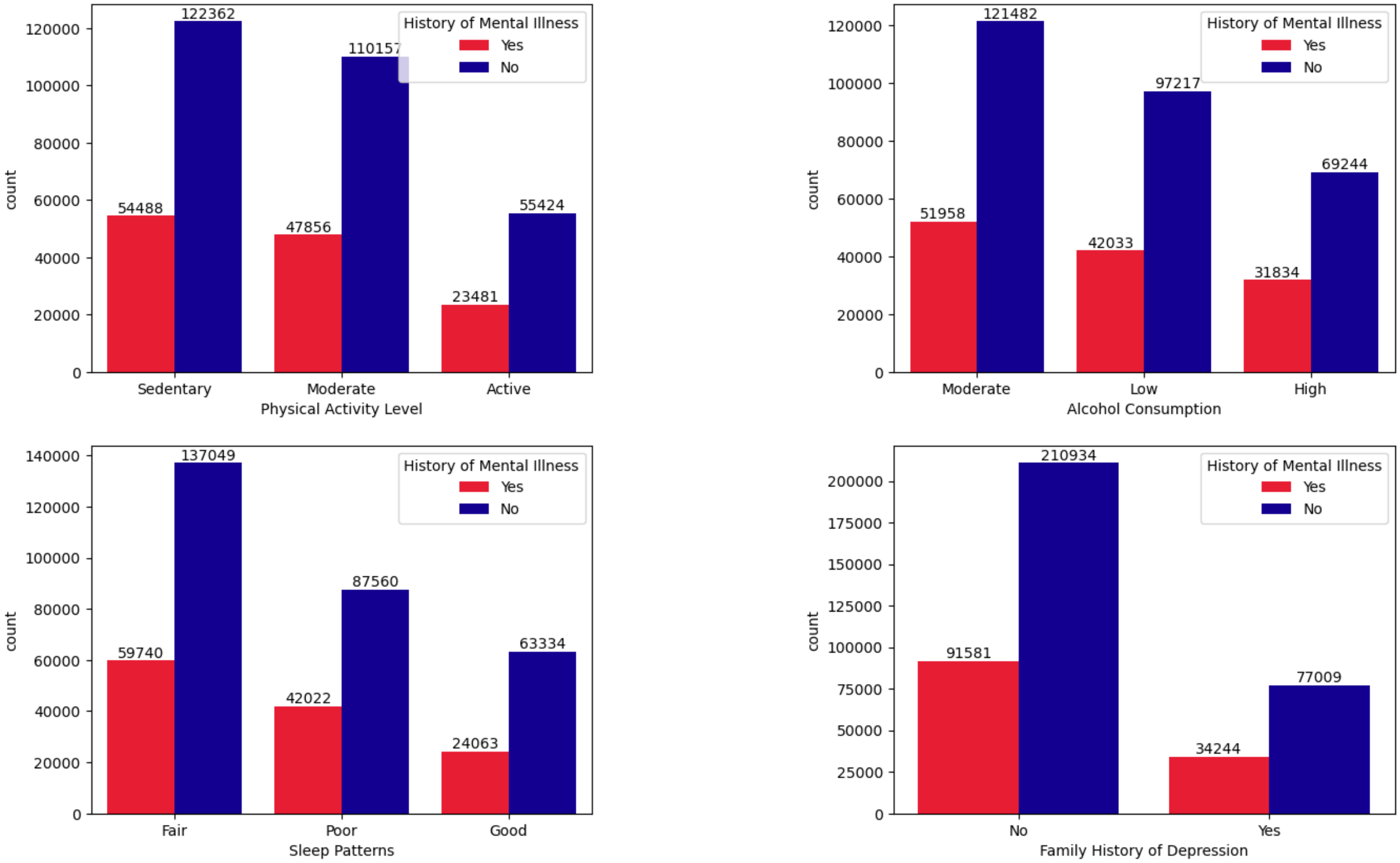


It seems like both males, females have similar risk of health issues.

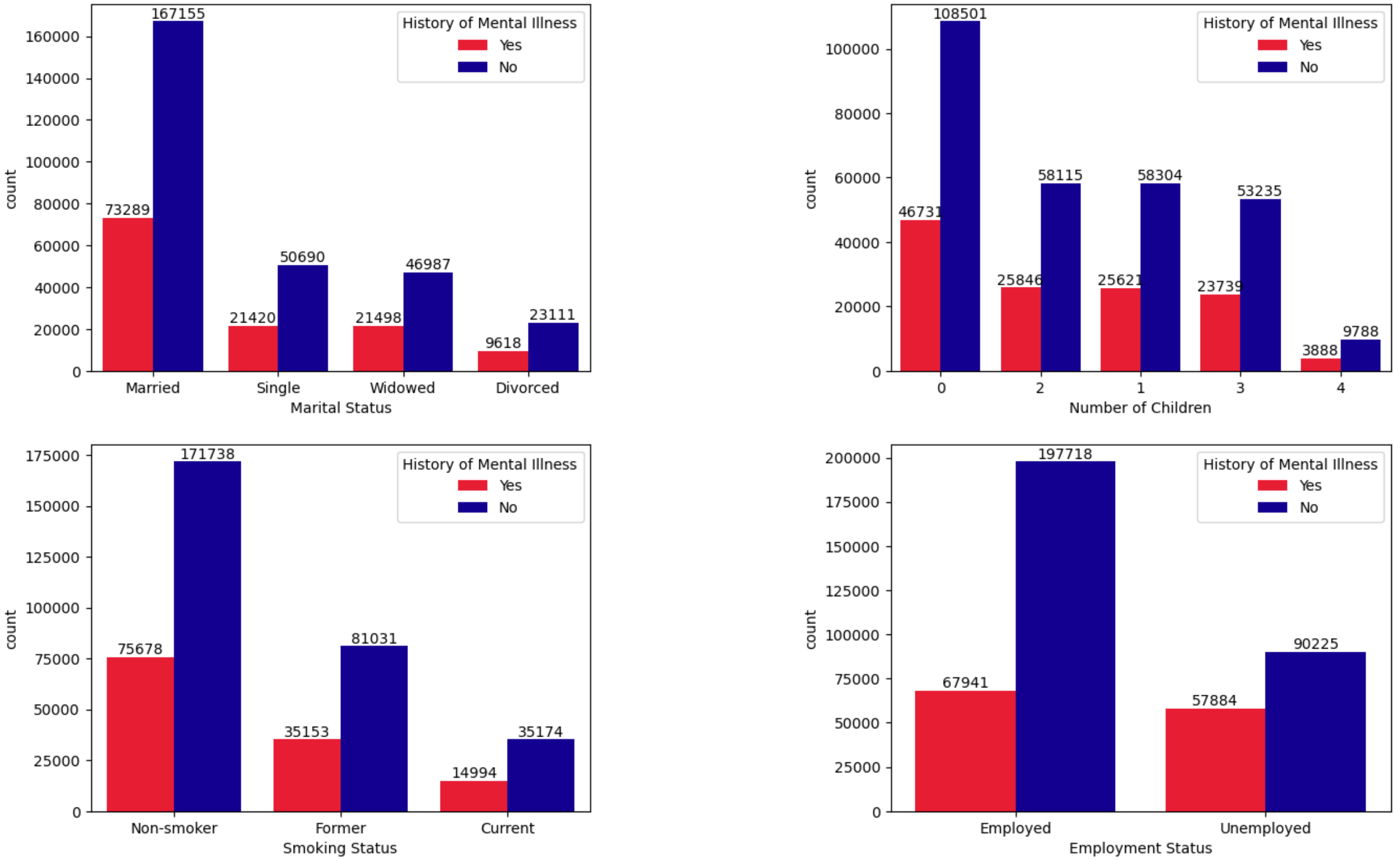


People with higher degree like PhD, Master's have lower risk of mental illness relative to people with High school or

Bachelor's degree.

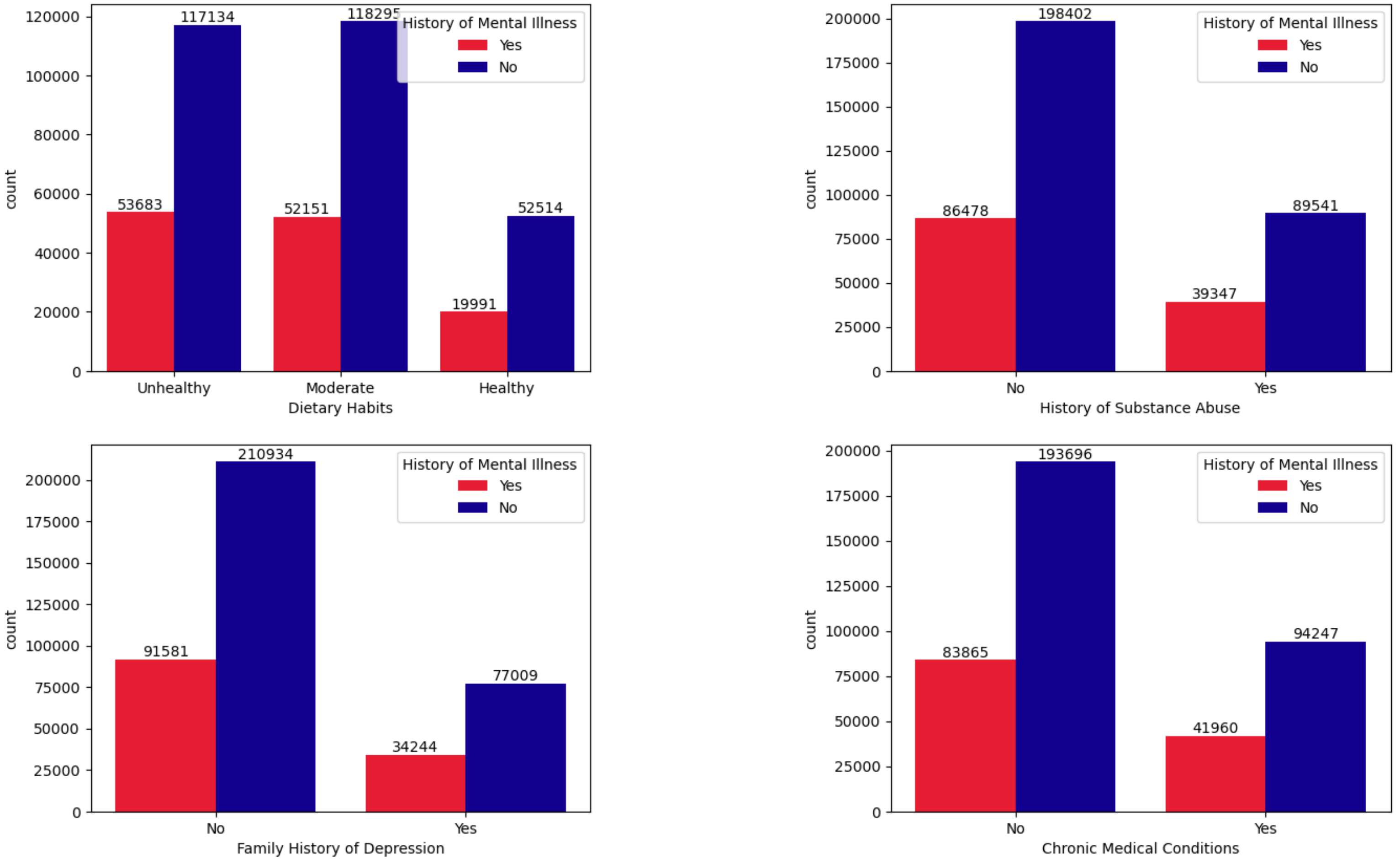


People with Active lifestyle, Good sleeping pattern, High alcohol consumption have lower risk of mental illness.



People with no children , Married people , unemployed have higher chance of mental illness.

Divorced people tend to have less risk as well as the current-moker have less risk.



People who follow healthy lifestyle tend have less risk of mental illness.

Correlation between the categorical variables using chi-square test

```
Correlated Categorical Features:
Marital Status and History of Mental Illness are correlated with p-value 5.409031887961224e-14
Education Level and History of Mental Illness are correlated with p-value 5.678973600047682e-302
Smoking Status and History of Mental Illness are correlated with p-value 0.0032730141384533966
Physical Activity Level and History of Mental Illness are correlated with p-value 2.55625629390646
35e-07
Employment Status and History of Mental Illness are correlated with p-value 0.0
Alcohol Consumption and History of Mental Illness are correlated with p-value 2.7124703085166094e-
17
Dietary Habits and History of Mental Illness are correlated with p-value 2.1242586440978167e-79
Sleep Patterns and History of Mental Illness are correlated with p-value 3.30469070431477e-129
History of Mental Illness and Family History of Depression are correlated with p-value 0.001689708
2070443303
History of Mental Illness and Chronic Medical Conditions are correlated with p-value 0.00010435052
869828637
History of Mental Illness and income_groups are correlated with p-value 0.0
History of Mental Illness and age_groups are correlated with p-value 1.4262358344020623e-64

Independent Categorical Features:
History of Mental Illness and History of Substance Abuse are independent with p-value 0.2666100524
904718
History of Mental Illness and Fname are independent with p-value 0.11847747615352716
History of Mental Illness and Gender are independent with p-value 0.4571005011796979
```

History of Mental Illness is not correlated with Name , Gender & History of Substance Abuse.

We can filter these features.

```
No     287943
Yes    125825
Name: History of Mental Illness, dtype: int64
```

```
array([ 0.        ,  1.        ,  0.        ,  0.        ,  0.        ,
        1.        ,  0.        ,  0.        ,  0.        ,  0.        ,
        0.        ,  1.        ,  1.        ,  0.        ,  0.        ,
        0.        ,  1.        ,  0.        ,  0.        ,  1.        ,
        0.        ,  1.        ,  0.        ,  1.        ,  0.        ,
        0.        ,  1.        ,  0.        ,  0.        ,  1.        ,
        0.        ,  1.        ,  1.        ,  0.        ,  1.        ,
        0.        ,  0.        ,  0.        ,  0.        ,  1.        ,
        0.        ,  0.        ,  0.        ,  0.5666923 , -0.60745955])
```

```
array([1, 1, 0, ..., 0, 0, 0])
```

```
125825
```

Trying Classifier models on un-balanced raw data.

```
Training Logistic Regression...
```

```
C:\Users\admin\anaconda3\lib\site-packages\sklearn\metrics\_classification.py:1334: UndefinedMetri
cWarning: Precision is ill-defined and being set to 0.0 in labels with no predicted samples. Use `
zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
C:\Users\admin\anaconda3\lib\site-packages\sklearn\metrics\_classification.py:1334: UndefinedMetri
cWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted s
amples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
C:\Users\admin\anaconda3\lib\site-packages\sklearn\metrics\_classification.py:1334: UndefinedMetri
cWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted s
amples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
C:\Users\admin\anaconda3\lib\site-packages\sklearn\metrics\_classification.py:1334: UndefinedMetri
cWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted s
amples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
```

```
Logistic Regression Classification Report:
               precision    recall  f1-score   support

           0       0.70      1.00      0.82     86319
           1       0.00      0.00      0.00     37812

    accuracy                           0.70    124131
   macro avg       0.35      0.50      0.41    124131
weighted avg       0.48      0.70      0.57    124131


Logistic Regression Confusion Matrix:
 [[86319     0]
 [37812     0]]
--------------------------------------------------------------------------------
Training Decision Tree...

Decision Tree Classification Report:
               precision    recall  f1-score   support

           0       0.70      0.69      0.70     86319
           1       0.32      0.34      0.33     37812

    accuracy                           0.58    124131
   macro avg       0.51      0.51      0.51    124131
weighted avg       0.59      0.58      0.58    124131


Decision Tree Confusion Matrix:
 [[59236 27083]
 [24879 12933]]
--------------------------------------------------------------------------------
Training Random Forest...

Random Forest Classification Report:
               precision    recall  f1-score   support

           0       0.70      0.83      0.76     86319
           1       0.34      0.19      0.25     37812

    accuracy                           0.64    124131
   macro avg       0.52      0.51      0.50    124131
weighted avg       0.59      0.64      0.61    124131
```

```
Random Forest Confusion Matrix:
 [[72028 14291]
 [30529  7283]]
--------------------------------------------------------------------------------
Training Gradient Boosting (XGBoost)...

Gradient Boosting (XGBoost) Classification Report:
              precision    recall  f1-score   support

           0       0.70      0.99      0.82     86319
           1       0.39      0.01      0.02     37812

    accuracy                           0.69    124131
   macro avg       0.54      0.50      0.42    124131
weighted avg       0.60      0.69      0.58    124131


Gradient Boosting (XGBoost) Confusion Matrix:
 [[85707   612]
 [37421   391]]
--------------------------------------------------------------------------------
Training Naive Bayes...

Naive Bayes Classification Report:
              precision    recall  f1-score   support

           0       0.75      0.62      0.68     86319
           1       0.37      0.51      0.43     37812

    accuracy                           0.59    124131
   macro avg       0.56      0.57      0.56    124131
weighted avg       0.63      0.59      0.60    124131


Naive Bayes Confusion Matrix:
 [[53746 32573]
 [18341 19471]]
--------------------------------------------------------------------------------
Training LightGBM...
[LightGBM] [Info] Number of positive: 88013, number of negative: 201624
[LightGBM] [Info] Auto-choosing row-wise multi-threading, the overhead of testing was 0.043929 sec
onds.
You can set `force_row_wise=true` to remove the overhead.
And if memory is not enough, you can set `force_col_wise=true`.
[LightGBM] [Info] Total Bins 390
[LightGBM] [Info] Number of data points in the train set: 289637, number of used features: 45
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.303873 -> initscore=-0.828920
[LightGBM] [Info] Start training from score -0.828920

LightGBM Classification Report:
              precision    recall  f1-score   support

           0       0.70      1.00      0.82     86319
           1       0.42      0.00      0.00     37812

    accuracy                           0.70    124131
   macro avg       0.56      0.50      0.41    124131
weighted avg       0.61      0.70      0.57    124131


LightGBM Confusion Matrix:
 [[86308    11]
 [37804     8]]
--------------------------------------------------------------------------------
                         Model  Accuracy  Precision    Recall  F1-Score  \
0          Logistic Regression  0.695386   0.483562  0.695386  0.570445
1                Decision Tree  0.581394   0.588159  0.581394  0.584614
2                Random Forest  0.638930   0.591217  0.638930  0.605087
3  Gradient Boosting (XGBoost)  0.693606   0.602793  0.693606  0.575250
4                  Naive Bayes  0.589837   0.632424  0.589837  0.603893
5                     LightGBM  0.695362   0.611833  0.695362  0.570552

    AUC-ROC
0  0.593136
1  0.513946
2  0.544688
3  0.593340
4  0.591918
5  0.598066
```

Balancing data using SMOTE

```
287943
```

```
287943
```

```
Training Logistic Regression...

Logistic Regression Classification Report:
              precision    recall  f1-score   support

           0       0.57      0.60      0.58     86344
           1       0.58      0.55      0.56     86422

    accuracy                           0.57    172766
   macro avg       0.57      0.57      0.57    172766
weighted avg       0.57      0.57      0.57    172766


Logistic Regression Confusion Matrix:
 [[51439 34905]
 [39058 47364]]
--------------------------------------------------------------------------------
Training Decision Tree...

Decision Tree Classification Report:
              precision    recall  f1-score   support

           0       0.65      0.64      0.64     86344
           1       0.65      0.66      0.65     86422

    accuracy                           0.65    172766
   macro avg       0.65      0.65      0.65    172766
weighted avg       0.65      0.65      0.65    172766


Decision Tree Confusion Matrix:
 [[55223 31121]
 [29768 56654]]
--------------------------------------------------------------------------------
Training Random Forest...

Random Forest Classification Report:
              precision    recall  f1-score   support

           0       0.71      0.78      0.74     86344
           1       0.75      0.68      0.72     86422
    accuracy                           0.73    172766
   macro avg       0.73      0.73      0.73    172766
weighted avg       0.73      0.73      0.73    172766


Random Forest Confusion Matrix:
 [[67032 19312]
 [27461 58961]]
--------------------------------------------------------------------------------
Training Gradient Boosting (XGBoost)...

Gradient Boosting (XGBoost) Classification Report:
              precision    recall  f1-score   support

           0       0.66      0.95      0.78     86344
           1       0.91      0.50      0.65     86422

    accuracy                           0.73    172766
   macro avg       0.78      0.73      0.71    172766
weighted avg       0.78      0.73      0.71    172766


Gradient Boosting (XGBoost) Confusion Matrix:
 [[81970  4374]
 [42907 43515]]
--------------------------------------------------------------------------------
Training Naive Bayes...

Naive Bayes Classification Report:
              precision    recall  f1-score   support

           0       0.59      0.48      0.53     86344
           1       0.56      0.67      0.61     86422

    accuracy                           0.57    172766
   macro avg       0.58      0.57      0.57    172766
weighted avg       0.58      0.57      0.57    172766


Naive Bayes Confusion Matrix:
 [[41453 44891]
 [28861 57561]]
--------------------------------------------------------------------------------
Training LightGBM...
[LightGBM] [Info] Number of positive: 201521, number of negative: 201599
[LightGBM] [Info] Auto-choosing row-wise multi-threading, the overhead of testing was 0.034312 sec
onds.
You can set `force_row_wise=true` to remove the overhead.
And if memory is not enough, you can set `force_col_wise=true`.
[LightGBM] [Info] Total Bins 10597
[LightGBM] [Info] Number of data points in the train set: 403120, number of used features: 45
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.499903 -> initscore=-0.000387
[LightGBM] [Info] Start training from score -0.000387

LightGBM Classification Report:
              precision    recall  f1-score   support

           0       0.66      0.94      0.77     86344
           1       0.89      0.52      0.65     86422

    accuracy                           0.73    172766
   macro avg       0.77      0.73      0.71    172766
weighted avg       0.77      0.73      0.71    172766


LightGBM Confusion Matrix:
 [[80736  5608]
 [41644 44778]]
--------------------------------------------------------------------------------
                       Model  Accuracy  Precision    Recall  F1-Score  \
0        Logistic Regression  0.571889   0.572065  0.571889  0.571646   
1              Decision Tree  0.647564   0.647597  0.647564  0.647541   
2              Random Forest  0.729270   0.731340  0.729270  0.728672   
3  Gradient Boosting (XGBoost)  0.726329   0.782592  0.726329  0.712035   
4                Naive Bayes  0.573110   0.575681  0.573110  0.569385   
5                   LightGBM  0.726497   0.774259  0.726497  0.714086   

    AUC-ROC  
0  0.596138  
1  0.647555  
2  0.793471  
3  0.788423  
4  0.596682  
5  0.789205  
```

Hyper-parameter Tuning

```
Step 1

Step 2

Step 3

Step 4

Step 5.1

Step 5.2

Step 5.3

Step 5.1.1

Fitting 3 folds for each of 5 candidates, totalling 15 fits
[CV] END max_depth=20, min_samples_leaf=4, min_samples_split=5, n_estimators=300; total time= 1.8m
in
[CV] END max_depth=20, min_samples_leaf=4, min_samples_split=5, n_estimators=300; total time= 1.8m
in
[CV] END max_depth=20, min_samples_leaf=4, min_samples_split=5, n_estimators=300; total time= 1.8m
in
[CV] END max_depth=None, min_samples_leaf=2, min_samples_split=2, n_estimators=200; total time= 8.
9min
[CV] END max_depth=None, min_samples_leaf=2, min_samples_split=2, n_estimators=200; total time= 1.
4min
[CV] END max_depth=None, min_samples_leaf=2, min_samples_split=2, n_estimators=200; total time= 1.
4min
[CV] END max_depth=None, min_samples_leaf=1, min_samples_split=5, n_estimators=200; total time= 2.
0min
[CV] END max_depth=None, min_samples_leaf=1, min_samples_split=5, n_estimators=200; total time= 1.
4min
[CV] END max_depth=None, min_samples_leaf=1, min_samples_split=5, n_estimators=200; total time= 1.
4min
[CV] END max_depth=30, min_samples_leaf=1, min_samples_split=2, n_estimators=300; total time=44.1m
in
[CV] END max_depth=30, min_samples_leaf=1, min_samples_split=2, n_estimators=300; total time= 2.6m
in
[CV] END max_depth=30, min_samples_leaf=1, min_samples_split=2, n_estimators=300; total time= 2.5m
in
[CV] END max_depth=20, min_samples_leaf=1, min_samples_split=10, n_estimators=300; total time= 1.9
```

```
min
[CV] END max_depth=20, min_samples_leaf=1, min_samples_split=10, n_estimators=300; total time= 2.3
min
[CV] END max_depth=20, min_samples_leaf=1, min_samples_split=10, n_estimators=300; total time= 2.3
min
Step 5.2.1

Fitting 3 folds for each of 5 candidates, totalling 15 fits
```

```
C:\Users\admin\anaconda3\lib\site-packages\xgboost\core.py:158: UserWarning: [10:10:11] WARNING:
C:\buildkite-agent\builds\buildkite-windows-cpu-autoscaling-group-i-0015a694724fa8361-1\xgboost\xg
boost-ci-windows\src\learner.cc:740:
Parameters: { "use_label_encoder" } are not used.

  warnings.warn(smsg, UserWarning)
```

```
[CV] END learning_rate=0.1, max_depth=3, n_estimators=200, subsample=0.6; total time=   4.4s
```

```
C:\Users\admin\anaconda3\lib\site-packages\xgboost\core.py:158: UserWarning: [10:10:14] WARNING:
C:\buildkite-agent\builds\buildkite-windows-cpu-autoscaling-group-i-0015a694724fa8361-1\xgboost\xg
boost-ci-windows\src\learner.cc:740:
Parameters: { "use_label_encoder" } are not used.

  warnings.warn(smsg, UserWarning)
```

```
[CV] END learning_rate=0.1, max_depth=3, n_estimators=200, subsample=0.6; total time=   2.4s
```

```
C:\Users\admin\anaconda3\lib\site-packages\xgboost\core.py:158: UserWarning: [10:10:16] WARNING:
C:\buildkite-agent\builds\buildkite-windows-cpu-autoscaling-group-i-0015a694724fa8361-1\xgboost\xg
boost-ci-windows\src\learner.cc:740:
Parameters: { "use_label_encoder" } are not used.

  warnings.warn(smsg, UserWarning)
```

```
[CV] END learning_rate=0.1, max_depth=3, n_estimators=200, subsample=0.6; total time=   5.8s
```

```
C:\Users\admin\anaconda3\lib\site-packages\xgboost\core.py:158: UserWarning: [10:10:22] WARNING:
C:\buildkite-agent\builds\buildkite-windows-cpu-autoscaling-group-i-0015a694724fa8361-1\xgboost\xg
boost-ci-windows\src\learner.cc:740:
Parameters: { "use_label_encoder" } are not used.

  warnings.warn(smsg, UserWarning)
```

```
[CV] END learning_rate=0.01, max_depth=3, n_estimators=100, subsample=0.6; total time=   2.7s
```

```
C:\Users\admin\anaconda3\lib\site-packages\xgboost\core.py:158: UserWarning: [10:10:25] WARNING:
C:\buildkite-agent\builds\buildkite-windows-cpu-autoscaling-group-i-0015a694724fa8361-1\xgboost\xg
boost-ci-windows\src\learner.cc:740:
Parameters: { "use_label_encoder" } are not used.

  warnings.warn(smsg, UserWarning)
```

```
[CV] END learning_rate=0.01, max_depth=3, n_estimators=100, subsample=0.6; total time=   2.4s
```

```
C:\Users\admin\anaconda3\lib\site-packages\xgboost\core.py:158: UserWarning: [10:10:28] WARNING:
C:\buildkite-agent\builds\buildkite-windows-cpu-autoscaling-group-i-0015a694724fa8361-1\xgboost\xg
boost-ci-windows\src\learner.cc:740:
Parameters: { "use_label_encoder" } are not used.

  warnings.warn(smsg, UserWarning)
```

```
[CV] END learning_rate=0.01, max_depth=3, n_estimators=100, subsample=0.6; total time=   2.9s
```

```
C:\Users\admin\anaconda3\lib\site-packages\xgboost\core.py:158: UserWarning: [10:10:31] WARNING:
C:\buildkite-agent\builds\buildkite-windows-cpu-autoscaling-group-i-0015a694724fa8361-1\xgboost\xg
boost-ci-windows\src\learner.cc:740:
Parameters: { "use_label_encoder" } are not used.

  warnings.warn(smsg, UserWarning)
```

```
[CV] END learning_rate=0.01, max_depth=10, n_estimators=200, subsample=0.8; total time=  11.5s
```

```
C:\Users\admin\anaconda3\lib\site-packages\xgboost\core.py:158: UserWarning: [10:10:42] WARNING:
C:\buildkite-agent\builds\buildkite-windows-cpu-autoscaling-group-i-0015a694724fa8361-1\xgboost\xg
boost-ci-windows\src\learner.cc:740:
Parameters: { "use_label_encoder" } are not used.

  warnings.warn(smsg, UserWarning)
```

```
[CV] END learning_rate=0.01, max_depth=10, n_estimators=200, subsample=0.8; total time=   9.2s
C:\Users\admin\anaconda3\lib\site-packages\xgboost\core.py:158: UserWarning: [10:10:52] WARNING:
C:\buildkite-agent\builds\buildkite-windows-cpu-autoscaling-group-i-0015a694724fa8361-1\xgboost\xg
boost-ci-windows\src\learner.cc:740:
Parameters: { "use_label_encoder" } are not used.

  warnings.warn(smsg, UserWarning)
[CV] END learning_rate=0.01, max_depth=10, n_estimators=200, subsample=0.8; total time=   7.9s
C:\Users\admin\anaconda3\lib\site-packages\xgboost\core.py:158: UserWarning: [10:11:00] WARNING:
C:\buildkite-agent\builds\buildkite-windows-cpu-autoscaling-group-i-0015a694724fa8361-1\xgboost\xg
boost-ci-windows\src\learner.cc:740:
Parameters: { "use_label_encoder" } are not used.

  warnings.warn(smsg, UserWarning)
[CV] END learning_rate=0.1, max_depth=3, n_estimators=200, subsample=0.8; total time=   3.4s
C:\Users\admin\anaconda3\lib\site-packages\xgboost\core.py:158: UserWarning: [10:11:03] WARNING:
C:\buildkite-agent\builds\buildkite-windows-cpu-autoscaling-group-i-0015a694724fa8361-1\xgboost\xg
boost-ci-windows\src\learner.cc:740:
Parameters: { "use_label_encoder" } are not used.

  warnings.warn(smsg, UserWarning)
[CV] END learning_rate=0.1, max_depth=3, n_estimators=200, subsample=0.8; total time=   3.0s
C:\Users\admin\anaconda3\lib\site-packages\xgboost\core.py:158: UserWarning: [10:11:06] WARNING:
C:\buildkite-agent\builds\buildkite-windows-cpu-autoscaling-group-i-0015a694724fa8361-1\xgboost\xg
boost-ci-windows\src\learner.cc:740:
Parameters: { "use_label_encoder" } are not used.

  warnings.warn(smsg, UserWarning)
[CV] END learning_rate=0.1, max_depth=3, n_estimators=200, subsample=0.8; total time=   2.9s
C:\Users\admin\anaconda3\lib\site-packages\xgboost\core.py:158: UserWarning: [10:11:09] WARNING:
C:\buildkite-agent\builds\buildkite-windows-cpu-autoscaling-group-i-0015a694724fa8361-1\xgboost\xg
boost-ci-windows\src\learner.cc:740:
Parameters: { "use_label_encoder" } are not used.

  warnings.warn(smsg, UserWarning)
[CV] END learning_rate=0.01, max_depth=10, n_estimators=100, subsample=0.6; total time=   4.3s
C:\Users\admin\anaconda3\lib\site-packages\xgboost\core.py:158: UserWarning: [10:11:14] WARNING:
C:\buildkite-agent\builds\buildkite-windows-cpu-autoscaling-group-i-0015a694724fa8361-1\xgboost\xg
boost-ci-windows\src\learner.cc:740:
Parameters: { "use_label_encoder" } are not used.

  warnings.warn(smsg, UserWarning)
[CV] END learning_rate=0.01, max_depth=10, n_estimators=100, subsample=0.6; total time=   4.4s
C:\Users\admin\anaconda3\lib\site-packages\xgboost\core.py:158: UserWarning: [10:11:18] WARNING:
C:\buildkite-agent\builds\buildkite-windows-cpu-autoscaling-group-i-0015a694724fa8361-1\xgboost\xg
boost-ci-windows\src\learner.cc:740:
Parameters: { "use_label_encoder" } are not used.

  warnings.warn(smsg, UserWarning)
[CV] END learning_rate=0.01, max_depth=10, n_estimators=100, subsample=0.6; total time=   4.2s
C:\Users\admin\anaconda3\lib\site-packages\xgboost\core.py:158: UserWarning: [10:11:23] WARNING:
C:\buildkite-agent\builds\buildkite-windows-cpu-autoscaling-group-i-0015a694724fa8361-1\xgboost\xg
boost-ci-windows\src\learner.cc:740:
Parameters: { "use_label_encoder" } are not used.

  warnings.warn(smsg, UserWarning)
Step 5.3.1

Fitting 3 folds for each of 5 candidates, totalling 15 fits
[LightGBM] [Info] Number of positive: 67028, number of negative: 153648
[LightGBM] [Info] Auto-choosing row-wise multi-threading, the overhead of testing was 0.022893 sec
onds.
You can set `force_row_wise=true` to remove the overhead.
```

```
And if memory is not enough, you can set `force_col_wise=true`.
[LightGBM] [Info] Total Bins 347
[LightGBM] [Info] Number of data points in the train set: 220676, number of used features: 45
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.303739 -> initscore=-0.829554
[LightGBM] [Info] Start training from score -0.829554
[CV] END boosting_type=gbdt, learning_rate=0.2, n_estimators=100, num_leaves=50; total time=   1.9
s
[LightGBM] [Info] Number of positive: 67028, number of negative: 153648
[LightGBM] [Info] Auto-choosing row-wise multi-threading, the overhead of testing was 0.010730 sec
onds.
You can set `force_row_wise=true` to remove the overhead.
And if memory is not enough, you can set `force_col_wise=true`.
[LightGBM] [Info] Total Bins 347
[LightGBM] [Info] Number of data points in the train set: 220676, number of used features: 45
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.303739 -> initscore=-0.829554
[LightGBM] [Info] Start training from score -0.829554
[CV] END boosting_type=gbdt, learning_rate=0.2, n_estimators=100, num_leaves=50; total time=   1.9
s
[LightGBM] [Info] Number of positive: 67028, number of negative: 153648
[LightGBM] [Info] Auto-choosing row-wise multi-threading, the overhead of testing was 0.011540 sec
onds.
You can set `force_row_wise=true` to remove the overhead.
And if memory is not enough, you can set `force_col_wise=true`.
[LightGBM] [Info] Total Bins 347
[LightGBM] [Info] Number of data points in the train set: 220676, number of used features: 45
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.303739 -> initscore=-0.829554
[LightGBM] [Info] Start training from score -0.829554
[CV] END boosting_type=gbdt, learning_rate=0.2, n_estimators=100, num_leaves=50; total time=   1.6
s
[LightGBM] [Info] Number of positive: 67028, number of negative: 153648
[LightGBM] [Info] Auto-choosing row-wise multi-threading, the overhead of testing was 0.008373 sec
onds.
You can set `force_row_wise=true` to remove the overhead.
And if memory is not enough, you can set `force_col_wise=true`.
[LightGBM] [Info] Total Bins 347
[LightGBM] [Info] Number of data points in the train set: 220676, number of used features: 45
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.303739 -> initscore=-0.829554
[LightGBM] [Info] Start training from score -0.829554
[CV] END boosting_type=dart, learning_rate=0.2, n_estimators=200, num_leaves=50; total time=  11.8
s
[LightGBM] [Info] Number of positive: 67028, number of negative: 153648
[LightGBM] [Info] Auto-choosing row-wise multi-threading, the overhead of testing was 0.010952 sec
onds.
You can set `force_row_wise=true` to remove the overhead.
And if memory is not enough, you can set `force_col_wise=true`.
[LightGBM] [Info] Total Bins 347
[LightGBM] [Info] Number of data points in the train set: 220676, number of used features: 45
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.303739 -> initscore=-0.829554
[LightGBM] [Info] Start training from score -0.829554
[CV] END boosting_type=dart, learning_rate=0.2, n_estimators=200, num_leaves=50; total time=  11.7
s
[LightGBM] [Info] Number of positive: 67028, number of negative: 153648
[LightGBM] [Info] Auto-choosing row-wise multi-threading, the overhead of testing was 0.012068 sec
onds.
You can set `force_row_wise=true` to remove the overhead.
And if memory is not enough, you can set `force_col_wise=true`.
[LightGBM] [Info] Total Bins 347
[LightGBM] [Info] Number of data points in the train set: 220676, number of used features: 45
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.303739 -> initscore=-0.829554
[LightGBM] [Info] Start training from score -0.829554
[CV] END boosting_type=dart, learning_rate=0.2, n_estimators=200, num_leaves=50; total time=  11.9
s
[LightGBM] [Info] Number of positive: 67028, number of negative: 153648
[LightGBM] [Info] Auto-choosing row-wise multi-threading, the overhead of testing was 0.008941 sec
onds.
You can set `force_row_wise=true` to remove the overhead.
```

```
And if memory is not enough, you can set `force_col_wise=true`.
[LightGBM] [Info] Total Bins 347
[LightGBM] [Info] Number of data points in the train set: 220676, number of used features: 45
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.303739 -> initscore=-0.829554
[LightGBM] [Info] Start training from score -0.829554
[CV] END boosting_type=dart, learning_rate=0.2, n_estimators=200, num_leaves=31; total time=   9.3
s
[LightGBM] [Info] Number of positive: 67028, number of negative: 153648
[LightGBM] [Info] Auto-choosing row-wise multi-threading, the overhead of testing was 0.010446 sec
onds.
You can set `force_row_wise=true` to remove the overhead.
And if memory is not enough, you can set `force_col_wise=true`.
[LightGBM] [Info] Total Bins 347
[LightGBM] [Info] Number of data points in the train set: 220676, number of used features: 45
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.303739 -> initscore=-0.829554
[LightGBM] [Info] Start training from score -0.829554
[CV] END boosting_type=dart, learning_rate=0.2, n_estimators=200, num_leaves=31; total time=  10.3
s
[LightGBM] [Info] Number of positive: 67028, number of negative: 153648
[LightGBM] [Info] Auto-choosing row-wise multi-threading, the overhead of testing was 0.011120 sec
onds.
You can set `force_row_wise=true` to remove the overhead.
And if memory is not enough, you can set `force_col_wise=true`.
[LightGBM] [Info] Total Bins 347
[LightGBM] [Info] Number of data points in the train set: 220676, number of used features: 45
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.303739 -> initscore=-0.829554
[LightGBM] [Info] Start training from score -0.829554
[CV] END boosting_type=dart, learning_rate=0.2, n_estimators=200, num_leaves=31; total time=   9.9
s
[LightGBM] [Info] Number of positive: 67028, number of negative: 153648
[LightGBM] [Info] Auto-choosing row-wise multi-threading, the overhead of testing was 0.007598 sec
onds.
You can set `force_row_wise=true` to remove the overhead.
And if memory is not enough, you can set `force_col_wise=true`.
[LightGBM] [Info] Total Bins 347
[LightGBM] [Info] Number of data points in the train set: 220676, number of used features: 45
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.303739 -> initscore=-0.829554
[LightGBM] [Info] Start training from score -0.829554
[CV] END boosting_type=gbdt, learning_rate=0.1, n_estimators=200, num_leaves=31; total time=   2.6
s
[LightGBM] [Info] Number of positive: 67028, number of negative: 153648
[LightGBM] [Info] Auto-choosing row-wise multi-threading, the overhead of testing was 0.026696 sec
onds.
You can set `force_row_wise=true` to remove the overhead.
And if memory is not enough, you can set `force_col_wise=true`.
[LightGBM] [Info] Total Bins 347
[LightGBM] [Info] Number of data points in the train set: 220676, number of used features: 45
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.303739 -> initscore=-0.829554
[LightGBM] [Info] Start training from score -0.829554
[CV] END boosting_type=gbdt, learning_rate=0.1, n_estimators=200, num_leaves=31; total time=   3.6
s
[LightGBM] [Info] Number of positive: 67028, number of negative: 153648
[LightGBM] [Info] Auto-choosing row-wise multi-threading, the overhead of testing was 0.011196 sec
onds.
You can set `force_row_wise=true` to remove the overhead.
And if memory is not enough, you can set `force_col_wise=true`.
[LightGBM] [Info] Total Bins 347
[LightGBM] [Info] Number of data points in the train set: 220676, number of used features: 45
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.303739 -> initscore=-0.829554
[LightGBM] [Info] Start training from score -0.829554
[CV] END boosting_type=gbdt, learning_rate=0.1, n_estimators=200, num_leaves=31; total time=   2.6
s
[LightGBM] [Info] Number of positive: 67028, number of negative: 153648
[LightGBM] [Info] Auto-choosing row-wise multi-threading, the overhead of testing was 0.010616 sec
onds.
You can set `force_row_wise=true` to remove the overhead.
```

```
And if memory is not enough, you can set `force_col_wise=true`.
[LightGBM] [Info] Total Bins 347
[LightGBM] [Info] Number of data points in the train set: 220676, number of used features: 45
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.303739 -> initscore=-0.829554
[LightGBM] [Info] Start training from score -0.829554
[CV] END boosting_type=dart, learning_rate=0.1, n_estimators=300, num_leaves=100; total time=  26.
6s
[LightGBM] [Info] Number of positive: 67028, number of negative: 153648
[LightGBM] [Info] Auto-choosing row-wise multi-threading, the overhead of testing was 0.013189 sec
onds.
You can set `force_row_wise=true` to remove the overhead.
And if memory is not enough, you can set `force_col_wise=true`.
[LightGBM] [Info] Total Bins 347
[LightGBM] [Info] Number of data points in the train set: 220676, number of used features: 45
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.303739 -> initscore=-0.829554
[LightGBM] [Info] Start training from score -0.829554
[CV] END boosting_type=dart, learning_rate=0.1, n_estimators=300, num_leaves=100; total time=  23.
4s
[LightGBM] [Info] Number of positive: 67028, number of negative: 153648
[LightGBM] [Info] Auto-choosing row-wise multi-threading, the overhead of testing was 0.010147 sec
onds.
You can set `force_row_wise=true` to remove the overhead.
And if memory is not enough, you can set `force_col_wise=true`.
[LightGBM] [Info] Total Bins 347
[LightGBM] [Info] Number of data points in the train set: 220676, number of used features: 45
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.303739 -> initscore=-0.829554
[LightGBM] [Info] Start training from score -0.829554
[CV] END boosting_type=dart, learning_rate=0.1, n_estimators=300, num_leaves=100; total time=  27.
0s
[LightGBM] [Info] Number of positive: 100542, number of negative: 230472
[LightGBM] [Info] Auto-choosing col-wise multi-threading, the overhead of testing was 0.024448 sec
onds.
You can set `force_col_wise=true` to remove the overhead.
[LightGBM] [Info] Total Bins 347
[LightGBM] [Info] Number of data points in the train set: 331014, number of used features: 45
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.303739 -> initscore=-0.829554
[LightGBM] [Info] Start training from score -0.829554
Best model: Random Forest with F1 score: 0.5046067373878573
RandomForestClassifier(max_depth=30, n_estimators=300, random_state=42)
```

Best Model

```
RandomForestClassifier(max_depth=30, n_estimators=300, random_state=42)

Step 1

Step 2

Step 3

Step 4

Step 5.1

Step 5.2

Step 5.3

Step 5.1.1

Fitting 3 folds for each of 5 candidates, totalling 15 fits
[CV] END max_depth=20, min_samples_leaf=4, min_samples_split=5, n_estimators=300; total time= 3.9m
in
[CV] END max_depth=20, min_samples_leaf=4, min_samples_split=5, n_estimators=300; total time= 4.2m
in
```

```
[CV] END max_depth=20, min_samples_leaf=4, min_samples_split=5, n_estimators=300; total time= 4.3m
in
[CV] END max_depth=None, min_samples_leaf=2, min_samples_split=2, n_estimators=200; total time= 3.
2min
[CV] END max_depth=None, min_samples_leaf=2, min_samples_split=2, n_estimators=200; total time= 3.
4min
[CV] END max_depth=None, min_samples_leaf=2, min_samples_split=2, n_estimators=200; total time= 3.
4min
[CV] END max_depth=None, min_samples_leaf=1, min_samples_split=5, n_estimators=200; total time= 3.
9min
[CV] END max_depth=None, min_samples_leaf=1, min_samples_split=5, n_estimators=200; total time= 3.
3min
[CV] END max_depth=None, min_samples_leaf=1, min_samples_split=5, n_estimators=200; total time= 3.
0min
[CV] END max_depth=30, min_samples_leaf=1, min_samples_split=2, n_estimators=300; total time= 5.2m
in
[CV] END max_depth=30, min_samples_leaf=1, min_samples_split=2, n_estimators=300; total time= 4.4m
in
[CV] END max_depth=30, min_samples_leaf=1, min_samples_split=2, n_estimators=300; total time= 4.1m
in
[CV] END max_depth=20, min_samples_leaf=1, min_samples_split=10, n_estimators=300; total time= 3.3
min
[CV] END max_depth=20, min_samples_leaf=1, min_samples_split=10, n_estimators=300; total time= 3.3
min
[CV] END max_depth=20, min_samples_leaf=1, min_samples_split=10, n_estimators=300; total time= 3.4
min
Step 5.2.1

Fitting 3 folds for each of 5 candidates, totalling 15 fits
```

```
C:\Users\admin\anaconda3\lib\site-packages\xgboost\core.py:158: UserWarning: [11:40:26] WARNING:
C:\buildkite-agent\builds\buildkite-windows-cpu-autoscaling-group-i-0015a694724fa8361-1\xgboost\xg
boost-ci-windows\src\learner.cc:740:
Parameters: { "use_label_encoder" } are not used.

  warnings.warn(smsg, UserWarning)
```

```
[CV] END learning_rate=0.1, max_depth=3, n_estimators=200, subsample=0.6; total time=   5.1s
```

```
C:\Users\admin\anaconda3\lib\site-packages\xgboost\core.py:158: UserWarning: [11:40:31] WARNING:
C:\buildkite-agent\builds\buildkite-windows-cpu-autoscaling-group-i-0015a694724fa8361-1\xgboost\xg
boost-ci-windows\src\learner.cc:740:
Parameters: { "use_label_encoder" } are not used.

  warnings.warn(smsg, UserWarning)
```

```
[CV] END learning_rate=0.1, max_depth=3, n_estimators=200, subsample=0.6; total time=   4.8s
```

```
C:\Users\admin\anaconda3\lib\site-packages\xgboost\core.py:158: UserWarning: [11:40:36] WARNING:
C:\buildkite-agent\builds\buildkite-windows-cpu-autoscaling-group-i-0015a694724fa8361-1\xgboost\xg
boost-ci-windows\src\learner.cc:740:
Parameters: { "use_label_encoder" } are not used.

  warnings.warn(smsg, UserWarning)
```

```
[CV] END learning_rate=0.1, max_depth=3, n_estimators=200, subsample=0.6; total time=   6.3s
```

```
C:\Users\admin\anaconda3\lib\site-packages\xgboost\core.py:158: UserWarning: [11:40:42] WARNING:
C:\buildkite-agent\builds\buildkite-windows-cpu-autoscaling-group-i-0015a694724fa8361-1\xgboost\xg
boost-ci-windows\src\learner.cc:740:
Parameters: { "use_label_encoder" } are not used.

  warnings.warn(smsg, UserWarning)
```

```
[CV] END learning_rate=0.01, max_depth=3, n_estimators=100, subsample=0.6; total time=   4.2s
```

```
C:\Users\admin\anaconda3\lib\site-packages\xgboost\core.py:158: UserWarning: [11:40:46] WARNING:
C:\buildkite-agent\builds\buildkite-windows-cpu-autoscaling-group-i-0015a694724fa8361-1\xgboost\xg
boost-ci-windows\src\learner.cc:740:
Parameters: { "use_label_encoder" } are not used.

  warnings.warn(smsg, UserWarning)
```

```
[CV] END learning_rate=0.01, max_depth=3, n_estimators=100, subsample=0.6; total time=   3.6s
```

```
C:\Users\admin\anaconda3\lib\site-packages\xgboost\core.py:158: UserWarning: [11:40:50] WARNING:
C:\buildkite-agent\builds\buildkite-windows-cpu-autoscaling-group-i-0015a694724fa8361-1\xgboost\xg
boost-ci-windows\src\learner.cc:740:
Parameters: { "use_label_encoder" } are not used.

  warnings.warn(smsg, UserWarning)
 [CV] END learning_rate=0.01, max_depth=3, n_estimators=100, subsample=0.6; total time=   4.1s
C:\Users\admin\anaconda3\lib\site-packages\xgboost\core.py:158: UserWarning: [11:40:55] WARNING:
C:\buildkite-agent\builds\buildkite-windows-cpu-autoscaling-group-i-0015a694724fa8361-1\xgboost\xg
boost-ci-windows\src\learner.cc:740:
Parameters: { "use_label_encoder" } are not used.

  warnings.warn(smsg, UserWarning)
 [CV] END learning_rate=0.01, max_depth=10, n_estimators=200, subsample=0.8; total time=  27.5s
C:\Users\admin\anaconda3\lib\site-packages\xgboost\core.py:158: UserWarning: [11:41:22] WARNING:
C:\buildkite-agent\builds\buildkite-windows-cpu-autoscaling-group-i-0015a694724fa8361-1\xgboost\xg
boost-ci-windows\src\learner.cc:740:
Parameters: { "use_label_encoder" } are not used.

  warnings.warn(smsg, UserWarning)
 [CV] END learning_rate=0.01, max_depth=10, n_estimators=200, subsample=0.8; total time=  21.1s
C:\Users\admin\anaconda3\lib\site-packages\xgboost\core.py:158: UserWarning: [11:41:43] WARNING:
C:\buildkite-agent\builds\buildkite-windows-cpu-autoscaling-group-i-0015a694724fa8361-1\xgboost\xg
boost-ci-windows\src\learner.cc:740:
Parameters: { "use_label_encoder" } are not used.

  warnings.warn(smsg, UserWarning)
 [CV] END learning_rate=0.01, max_depth=10, n_estimators=200, subsample=0.8; total time=  18.5s
C:\Users\admin\anaconda3\lib\site-packages\xgboost\core.py:158: UserWarning: [11:42:02] WARNING:
C:\buildkite-agent\builds\buildkite-windows-cpu-autoscaling-group-i-0015a694724fa8361-1\xgboost\xg
boost-ci-windows\src\learner.cc:740:
Parameters: { "use_label_encoder" } are not used.

  warnings.warn(smsg, UserWarning)
 [CV] END learning_rate=0.1, max_depth=3, n_estimators=200, subsample=0.8; total time=   5.7s
C:\Users\admin\anaconda3\lib\site-packages\xgboost\core.py:158: UserWarning: [11:42:08] WARNING:
C:\buildkite-agent\builds\buildkite-windows-cpu-autoscaling-group-i-0015a694724fa8361-1\xgboost\xg
boost-ci-windows\src\learner.cc:740:
Parameters: { "use_label_encoder" } are not used.

  warnings.warn(smsg, UserWarning)
 [CV] END learning_rate=0.1, max_depth=3, n_estimators=200, subsample=0.8; total time=   7.3s
C:\Users\admin\anaconda3\lib\site-packages\xgboost\core.py:158: UserWarning: [11:42:15] WARNING:
C:\buildkite-agent\builds\buildkite-windows-cpu-autoscaling-group-i-0015a694724fa8361-1\xgboost\xg
boost-ci-windows\src\learner.cc:740:
Parameters: { "use_label_encoder" } are not used.

  warnings.warn(smsg, UserWarning)
 [CV] END learning_rate=0.1, max_depth=3, n_estimators=200, subsample=0.8; total time=   6.0s
C:\Users\admin\anaconda3\lib\site-packages\xgboost\core.py:158: UserWarning: [11:42:21] WARNING:
C:\buildkite-agent\builds\buildkite-windows-cpu-autoscaling-group-i-0015a694724fa8361-1\xgboost\xg
boost-ci-windows\src\learner.cc:740:
Parameters: { "use_label_encoder" } are not used.

  warnings.warn(smsg, UserWarning)
 [CV] END learning_rate=0.01, max_depth=10, n_estimators=100, subsample=0.6; total time=   9.8s
C:\Users\admin\anaconda3\lib\site-packages\xgboost\core.py:158: UserWarning: [11:42:31] WARNING:
C:\buildkite-agent\builds\buildkite-windows-cpu-autoscaling-group-i-0015a694724fa8361-1\xgboost\xg
boost-ci-windows\src\learner.cc:740:
Parameters: { "use_label_encoder" } are not used.

  warnings.warn(smsg, UserWarning)
 [CV] END learning_rate=0.01, max_depth=10, n_estimators=100, subsample=0.6; total time=   9.9s
```

```
C:\Users\admin\anaconda3\lib\site-packages\xgboost\core.py:158: UserWarning: [11:42:41] WARNING:
C:\buildkite-agent\builds\buildkite-windows-cpu-autoscaling-group-i-0015a694724fa8361-1\xgboost\xg
boost-ci-windows\src\learner.cc:740:
Parameters: { "use_label_encoder" } are not used.

  warnings.warn(smsg, UserWarning)
[CV] END learning_rate=0.01, max_depth=10, n_estimators=100, subsample=0.6; total time=  10.0s
C:\Users\admin\anaconda3\lib\site-packages\xgboost\core.py:158: UserWarning: [11:42:51] WARNING:
C:\buildkite-agent\builds\buildkite-windows-cpu-autoscaling-group-i-0015a694724fa8361-1\xgboost\xg
boost-ci-windows\src\learner.cc:740:
Parameters: { "use_label_encoder" } are not used.

  warnings.warn(smsg, UserWarning)
Step 5.3.1

Fitting 3 folds for each of 5 candidates, totalling 15 fits
[LightGBM] [Info] Number of positive: 153463, number of negative: 153675
[LightGBM] [Info] Auto-choosing row-wise multi-threading, the overhead of testing was 0.047518 sec
onds.
You can set `force_row_wise=true` to remove the overhead.
And if memory is not enough, you can set `force_col_wise=true`.
[LightGBM] [Info] Total Bins 10871
[LightGBM] [Info] Number of data points in the train set: 307138, number of used features: 45
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.499655 -> initscore=-0.001380
[LightGBM] [Info] Start training from score -0.001380
[CV] END boosting_type=gbdt, learning_rate=0.2, n_estimators=100, num_leaves=50; total time=   3.1
s
[LightGBM] [Info] Number of positive: 153463, number of negative: 153676
[LightGBM] [Info] Auto-choosing row-wise multi-threading, the overhead of testing was 0.049018 sec
onds.
You can set `force_row_wise=true` to remove the overhead.
And if memory is not enough, you can set `force_col_wise=true`.
[LightGBM] [Info] Total Bins 10867
[LightGBM] [Info] Number of data points in the train set: 307139, number of used features: 45
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.499653 -> initscore=-0.001387
[LightGBM] [Info] Start training from score -0.001387
[CV] END boosting_type=gbdt, learning_rate=0.2, n_estimators=100, num_leaves=50; total time=   3.1
s
[LightGBM] [Info] Number of positive: 153464, number of negative: 153675
[LightGBM] [Info] Auto-choosing row-wise multi-threading, the overhead of testing was 0.040361 sec
onds.
You can set `force_row_wise=true` to remove the overhead.
And if memory is not enough, you can set `force_col_wise=true`.
[LightGBM] [Info] Total Bins 10863
[LightGBM] [Info] Number of data points in the train set: 307139, number of used features: 45
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.499657 -> initscore=-0.001374
[LightGBM] [Info] Start training from score -0.001374
[CV] END boosting_type=gbdt, learning_rate=0.2, n_estimators=100, num_leaves=50; total time=   3.0
s
[LightGBM] [Info] Number of positive: 153463, number of negative: 153675
[LightGBM] [Info] Auto-choosing row-wise multi-threading, the overhead of testing was 0.048754 sec
onds.
You can set `force_row_wise=true` to remove the overhead.
And if memory is not enough, you can set `force_col_wise=true`.
[LightGBM] [Info] Total Bins 10871
[LightGBM] [Info] Number of data points in the train set: 307138, number of used features: 45
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.499655 -> initscore=-0.001380
[LightGBM] [Info] Start training from score -0.001380
[CV] END boosting_type=dart, learning_rate=0.2, n_estimators=200, num_leaves=50; total time=  27.8
s
[LightGBM] [Info] Number of positive: 153463, number of negative: 153676
[LightGBM] [Info] Auto-choosing row-wise multi-threading, the overhead of testing was 0.045380 sec
onds.
You can set `force_row_wise=true` to remove the overhead.
And if memory is not enough, you can set `force_col_wise=true`.
```

```
[LightGBM] [Info] Total Bins 10867
[LightGBM] [Info] Number of data points in the train set: 307139, number of used features: 45
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.499653 -> initscore=-0.001387
[LightGBM] [Info] Start training from score -0.001387
[CV] END boosting_type=dart, learning_rate=0.2, n_estimators=200, num_leaves=50; total time=  27.6
s
[LightGBM] [Info] Number of positive: 153464, number of negative: 153675
[LightGBM] [Info] Auto-choosing row-wise multi-threading, the overhead of testing was 0.059656 sec
onds.
You can set `force_row_wise=true` to remove the overhead.
And if memory is not enough, you can set `force_col_wise=true`.
[LightGBM] [Info] Total Bins 10863
[LightGBM] [Info] Number of data points in the train set: 307139, number of used features: 45
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.499657 -> initscore=-0.001374
[LightGBM] [Info] Start training from score -0.001374
[CV] END boosting_type=dart, learning_rate=0.2, n_estimators=200, num_leaves=50; total time=  28.1
s
[LightGBM] [Info] Number of positive: 153463, number of negative: 153675
[LightGBM] [Info] Auto-choosing row-wise multi-threading, the overhead of testing was 0.047733 sec
onds.
You can set `force_row_wise=true` to remove the overhead.
And if memory is not enough, you can set `force_col_wise=true`.
[LightGBM] [Info] Total Bins 10871
[LightGBM] [Info] Number of data points in the train set: 307138, number of used features: 45
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.499655 -> initscore=-0.001380
[LightGBM] [Info] Start training from score -0.001380
[CV] END boosting_type=dart, learning_rate=0.2, n_estimators=200, num_leaves=31; total time=  26.9
s
[LightGBM] [Info] Number of positive: 153463, number of negative: 153676
[LightGBM] [Info] Auto-choosing row-wise multi-threading, the overhead of testing was 0.052242 sec
onds.
You can set `force_row_wise=true` to remove the overhead.
And if memory is not enough, you can set `force_col_wise=true`.
[LightGBM] [Info] Total Bins 10867
[LightGBM] [Info] Number of data points in the train set: 307139, number of used features: 45
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.499653 -> initscore=-0.001387
[LightGBM] [Info] Start training from score -0.001387
[CV] END boosting_type=dart, learning_rate=0.2, n_estimators=200, num_leaves=31; total time=  25.1
s
[LightGBM] [Info] Number of positive: 153464, number of negative: 153675
[LightGBM] [Info] Auto-choosing row-wise multi-threading, the overhead of testing was 0.052547 sec
onds.
You can set `force_row_wise=true` to remove the overhead.
And if memory is not enough, you can set `force_col_wise=true`.
[LightGBM] [Info] Total Bins 10863
[LightGBM] [Info] Number of data points in the train set: 307139, number of used features: 45
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.499657 -> initscore=-0.001374
[LightGBM] [Info] Start training from score -0.001374
[CV] END boosting_type=dart, learning_rate=0.2, n_estimators=200, num_leaves=31; total time=  38.0
s
[LightGBM] [Info] Number of positive: 153463, number of negative: 153675
[LightGBM] [Info] Auto-choosing row-wise multi-threading, the overhead of testing was 0.057067 sec
onds.
You can set `force_row_wise=true` to remove the overhead.
And if memory is not enough, you can set `force_col_wise=true`.
[LightGBM] [Info] Total Bins 10871
[LightGBM] [Info] Number of data points in the train set: 307138, number of used features: 45
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.499655 -> initscore=-0.001380
[LightGBM] [Info] Start training from score -0.001380
[CV] END boosting_type=gbdt, learning_rate=0.1, n_estimators=200, num_leaves=31; total time=   7.8
s
[LightGBM] [Info] Number of positive: 153463, number of negative: 153676
[LightGBM] [Info] Auto-choosing row-wise multi-threading, the overhead of testing was 0.057011 sec
onds.
You can set `force_row_wise=true` to remove the overhead.
And if memory is not enough, you can set `force_col_wise=true`.
```

```
[LightGBM] [Info] Total Bins 10867
[LightGBM] [Info] Number of data points in the train set: 307139, number of used features: 45
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.499653 -> initscore=-0.001387
[LightGBM] [Info] Start training from score -0.001387
[CV] END boosting_type=gbdt, learning_rate=0.1, n_estimators=200, num_leaves=31; total time=   7.7
s
[LightGBM] [Info] Number of positive: 153464, number of negative: 153675
[LightGBM] [Info] Auto-choosing row-wise multi-threading, the overhead of testing was 0.054723 sec
onds.
You can set `force_row_wise=true` to remove the overhead.
And if memory is not enough, you can set `force_col_wise=true`.
[LightGBM] [Info] Total Bins 10863
[LightGBM] [Info] Number of data points in the train set: 307139, number of used features: 45
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.499657 -> initscore=-0.001374
[LightGBM] [Info] Start training from score -0.001374
[CV] END boosting_type=gbdt, learning_rate=0.1, n_estimators=200, num_leaves=31; total time=   7.4
s
[LightGBM] [Info] Number of positive: 153463, number of negative: 153675
[LightGBM] [Info] Auto-choosing row-wise multi-threading, the overhead of testing was 0.062695 sec
onds.
You can set `force_row_wise=true` to remove the overhead.
And if memory is not enough, you can set `force_col_wise=true`.
[LightGBM] [Info] Total Bins 10871
[LightGBM] [Info] Number of data points in the train set: 307138, number of used features: 45
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.499655 -> initscore=-0.001380
[LightGBM] [Info] Start training from score -0.001380
[CV] END boosting_type=dart, learning_rate=0.1, n_estimators=300, num_leaves=100; total time= 1.5m
in
[LightGBM] [Info] Number of positive: 153463, number of negative: 153676
[LightGBM] [Info] Auto-choosing row-wise multi-threading, the overhead of testing was 0.058857 sec
onds.
You can set `force_row_wise=true` to remove the overhead.
And if memory is not enough, you can set `force_col_wise=true`.
[LightGBM] [Info] Total Bins 10867
[LightGBM] [Info] Number of data points in the train set: 307139, number of used features: 45
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.499653 -> initscore=-0.001387
[LightGBM] [Info] Start training from score -0.001387
[CV] END boosting_type=dart, learning_rate=0.1, n_estimators=300, num_leaves=100; total time= 1.3m
in
[LightGBM] [Info] Number of positive: 153464, number of negative: 153675
[LightGBM] [Info] Auto-choosing row-wise multi-threading, the overhead of testing was 0.047332 sec
onds.
You can set `force_row_wise=true` to remove the overhead.
And if memory is not enough, you can set `force_col_wise=true`.
[LightGBM] [Info] Total Bins 10863
[LightGBM] [Info] Number of data points in the train set: 307139, number of used features: 45
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.499657 -> initscore=-0.001374
[LightGBM] [Info] Start training from score -0.001374
[CV] END boosting_type=dart, learning_rate=0.1, n_estimators=300, num_leaves=100; total time= 1.1m
in
[LightGBM] [Info] Number of positive: 230195, number of negative: 230513
[LightGBM] [Info] Auto-choosing row-wise multi-threading, the overhead of testing was 0.063700 sec
onds.
You can set `force_row_wise=true` to remove the overhead.
And if memory is not enough, you can set `force_col_wise=true`.
[LightGBM] [Info] Total Bins 10867
[LightGBM] [Info] Number of data points in the train set: 460708, number of used features: 45
[LightGBM] [Info] [binary:BoostFromScore]: pavg=0.499655 -> initscore=-0.001380
[LightGBM] [Info] Start training from score -0.001380
Best model: Random Forest with F1 score: 0.7489743037747075
RandomForestClassifier(min_samples_split=5, n_estimators=200, random_state=42)
```

Random Forest Classifier gives best performance with F1

score: 0.74

We can export this model in .pkl file to be used for prediction