Logistic Regression (Example)

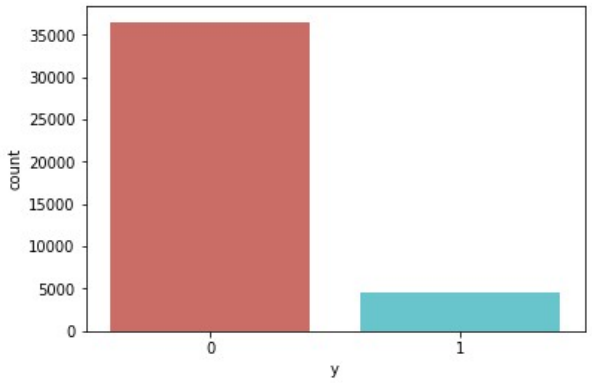
1 - Exploring the data:

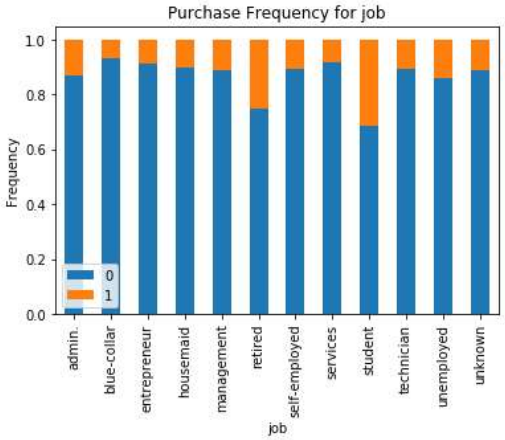
```
RangeIndex: 41188 entries, 0 to 41187
Data columns (total 22 columns):
Unnamed: 0        41188 non-null int64
age               41188 non-null int64
job               41188 non-null object
marital           41188 non-null object
education         41188 non-null object
default           41188 non-null object
housing           41188 non-null object
loan              41188 non-null object
contact           41188 non-null object
month             41188 non-null object
day_of_week       41188 non-null object
duration          41188 non-null int64
campaign          41188 non-null int64
pdays             41188 non-null int64
previous          41188 non-null int64
poutcome          41188 non-null object
emp_var_rate      41188 non-null float64
cons_price_idx    41188 non-null float64
cons_conf_idx     41188 non-null float64
euribor3m         41188 non-null float64
nr_employed       41188 non-null float64
y                 41188 non-null int64
dtypes: float64(5), int64(7), object(10)
memory usage: 6.9+ MB
None
Unnamed: 0        0
age               0
job               0
marital           0
education         0
default           0
housing           0
loan              0
contact           0
month             0
day_of_week       0
duration          0
campaign          0
pdays             0
previous          0
poutcome          0
emp_var_rate      0
cons_price_idx    0
cons_conf_idx     0
euribor3m         0
nr_employed       0
y                 0
dtype: int64
```

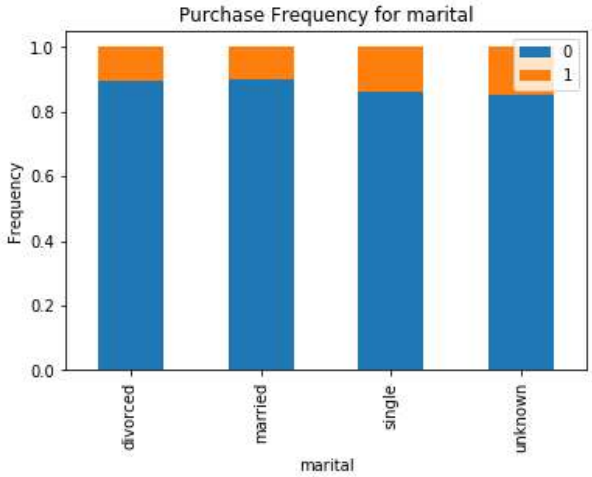
```
(41188, 22)
['Unnamed: 0', 'age', 'job', 'marital', 'education', 'default', 'housing', 'loan',
'contact', 'month', 'day_of_week', 'duration', 'campaign', 'pdays', 'previous', 'poutcome',
'emp_var_rate', 'cons_price_idx', 'cons_conf_idx', 'euribor3m', 'nr_employed', 'y']
admin.           10422
blue-collar       9254
technician        6743
services          3969
management        2924
retired           1720
entrepreneur      1456
self-employed     1421
housemaid         1060
unemployed        1014
student            875
unknown            330
Name: job, dtype: int64
married     24928
single      11568
divorced     4612
unknown        80
Name: marital, dtype: int64
university.degree      12168
high.school             9515
basic.9y                6045
professional.course     5243
basic.4y                4176
basic.6y                2292
unknown                 1731
illiterate                18
Name: education, dtype: int64
no         32588
unknown     8597
yes            3
Name: default, dtype: int64
yes        21576
no         18622
unknown      990
Name: housing, dtype: int64
no         33950
yes         6248
unknown      990
Name: loan, dtype: int64
may    13769
jul     7174
aug     6178
jun     5318
nov     4101
apr     2632
oct      718
sep      570
mar      546
dec      182
Name: month, dtype: int64
thu    8623
mon    8514
wed    8134
tue    8090
fri    7827
```

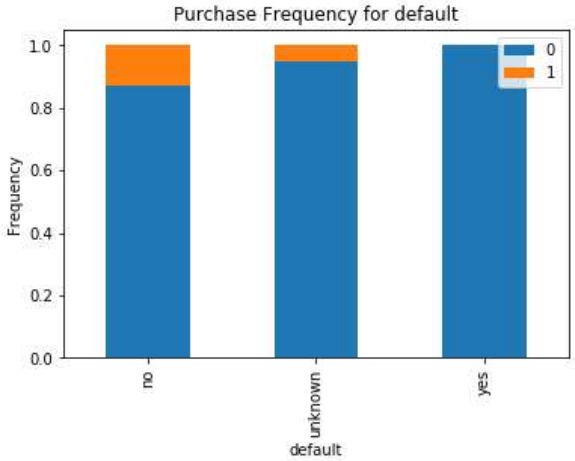
```
Name: day_of_week, dtype: int64
nonexistent    35563
failure         4252
success         1373
Name: poutcome, dtype: int64
0    36548
1     4640
Name: y, dtype: int64
basic                12513
university.degree    12168
high.school           9515
professional.course   5243
unknown               1731
illiterate              18
Name: education, dtype: int64
0    36548
1     4640
Name: y, dtype: int64
```

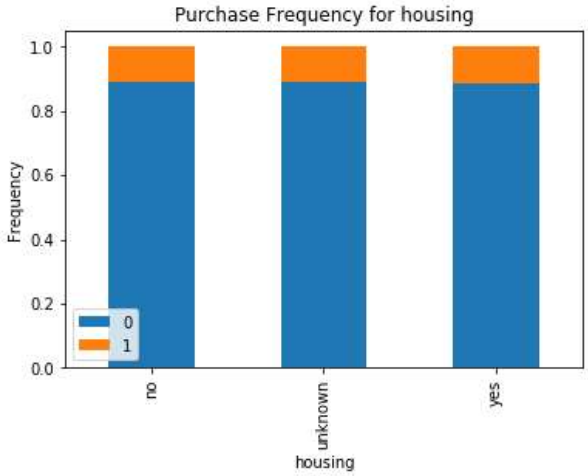
Plot count for our target y [0 or 1]. [No, Yes]

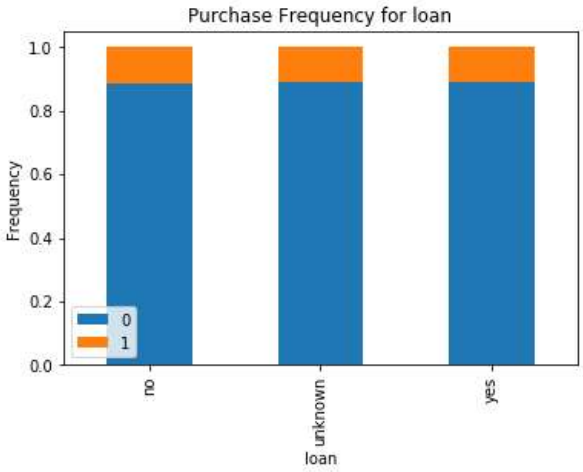


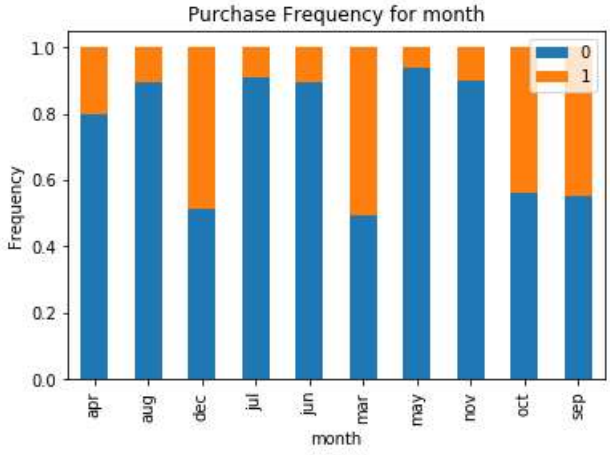


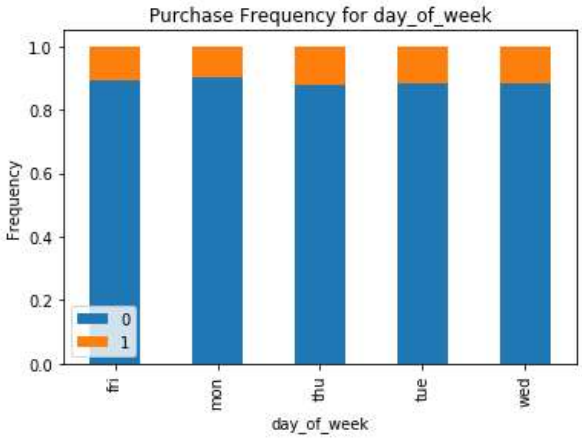


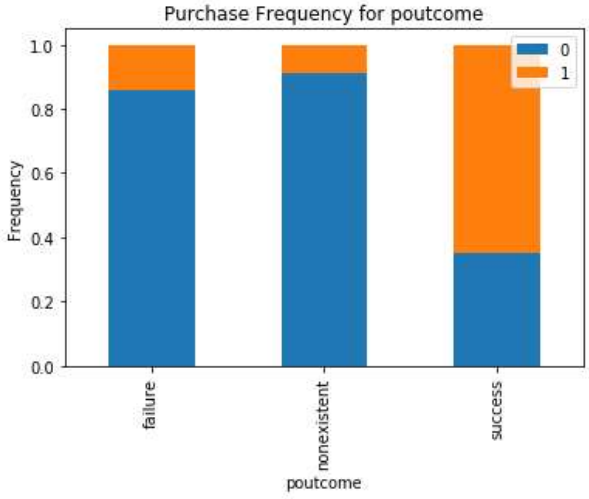


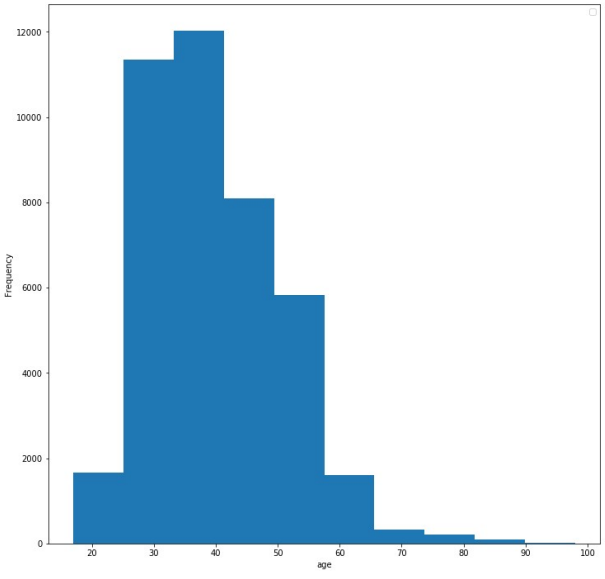

Purchase Frequency for housing
Frequency
1.0
0.8
0.6
0.4
0.2
0.0
0
1
no
unknown
yes
housing


Purchase Frequency for loan
Frequency
1.0
0.8
0.6
0.4
0.2
0.0
0
1
no
unknown
yes
loan

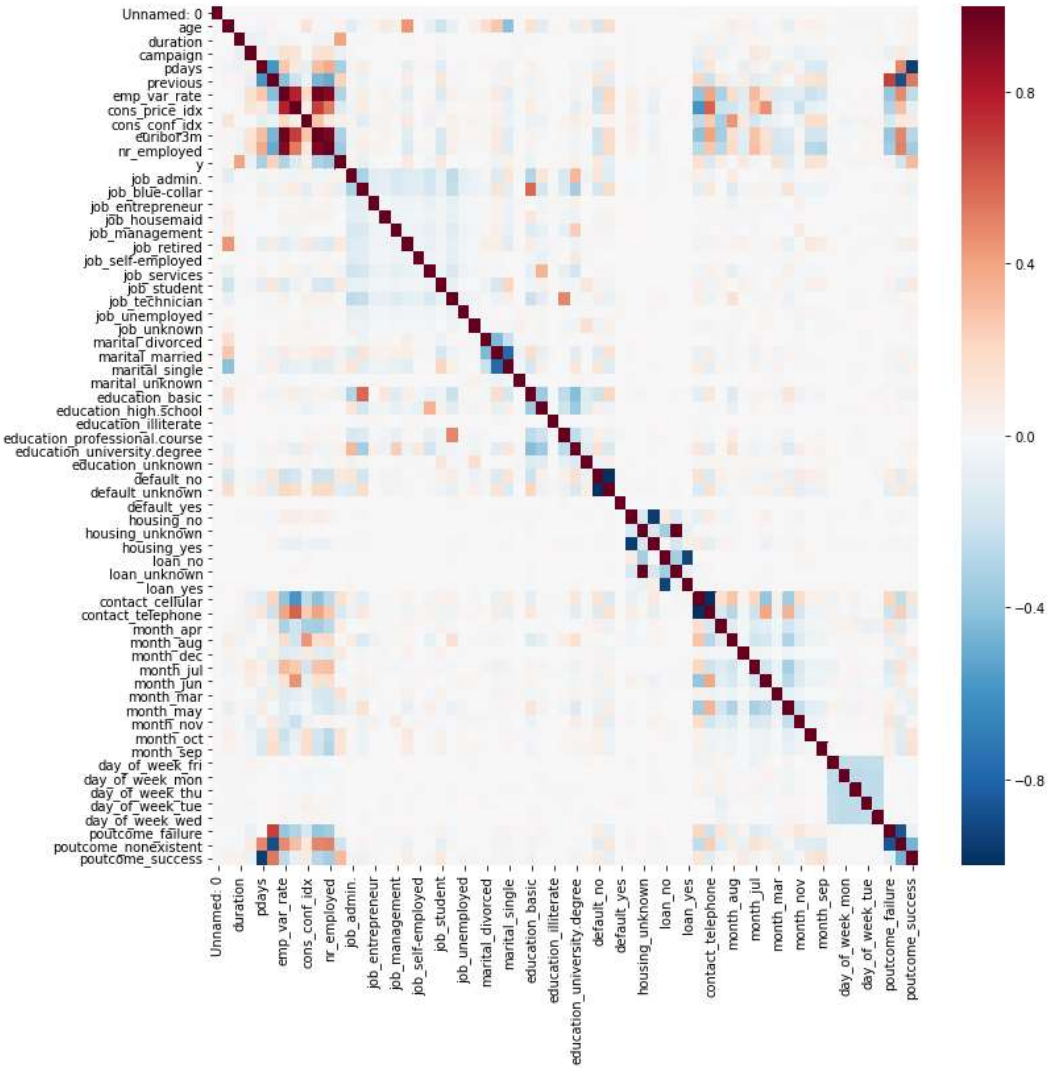

Purchase Frequency for month
Frequency
1.0
0.8
0.6
0.4
0.2
0.0
0
1
apr
aug
dec
jul
jun
mar
may
nov
oct
sep
month


Purchase Frequency for day_of_week
Frequency
1.0
0.8
0.6
0.4
0.2
0.0
0
1
fri
mon
thu
tue
wed
day_of_week


Purchase Frequency for poutcome
Frequency
1.0
0.8
0.6
0.4
0.2
0.0
0
1
failure
nonexistent
success
poutcome

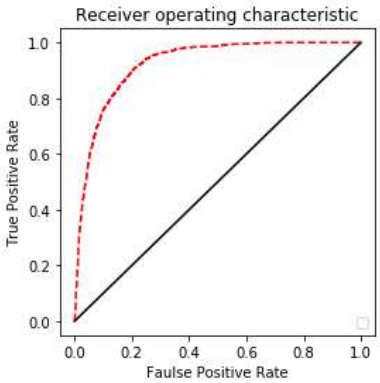

Frequency
12000
10000
8000
6000
4000
2000
0
20
30
40
50
60
70
80
90
100
age

Correlation heatmap for all the features:



The results from the logistic regression:

	precision	recall	f1-score	support
0	0.93	0.97	0.95	10923
1	0.67	0.43	0.52	1434
micro avg	0.91	0.91	0.91	12357
macro avg	0.80	0.70	0.74	12357
weighted avg	0.90	0.91	0.90	12357



```
[[10622   301]
 [  823   611]]
the coefs of logistic regression model are: [[-1.78674812e-06  1.41466144e-03  4.78348634e-03 -2.80446915e-02
  -1.37663442e-03 -3.33748739e-02 -2.64280046e-01  3.34230678e-01
   4.70881028e-02 -2.93207525e-01 -6.07012122e-03  3.40650271e-02
  -7.41903764e-02 -8.95548632e-03  7.47022851e-04  3.01965818e-03
   3.09878560e-02 -1.84334321e-03 -1.78796029e-02  2.20743430e-02
   6.98106149e-03  7.32634512e-03  1.13812796e-03 -4.02487752e-03
  -4.78455471e-02  5.63130046e-02 -9.71947063e-04 -6.61066945e-02
  -2.10036795e-02  7.28958076e-04  1.19306856e-02  7.15400573e-02
   6.38130591e-03  6.74681533e-02 -6.39766602e-02 -2.08602664e-05
  -1.46069603e-03 -5.41458002e-03  1.03459089e-02  1.35040361e-02
  -5.41458002e-03 -4.61882322e-03  1.00688494e-01 -9.72178606e-02
   1.97389714e-02  3.05875023e-02  2.49388778e-03  4.36834400e-02
   3.73942387e-02  5.77341256e-02 -1.97911271e-01 -5.33980298e-03
   1.05016325e-02  4.58790831e-03 -5.40020705e-03 -1.71540817e-02
   8.46875970e-03  1.00659291e-02  7.49023280e-03 -4.76749533e-02
   3.86317911e-02  1.25137951e-02]]
The accuracy of prediction is  0.9090394108602412
The area under the ROC curve is: 0.6992621802599176
```