2. Problem Statement

I decided to treat this as a classification problem by creating a new binary variable affair (did the woman have at least one affair?) and trying to predict the classification for each woman.

Dataset

The dataset I chose is the affairs dataset that comes with Statsmodels. It was derived from a survey of women in 1974 by Redbook magazine, in which married women were asked about their participation in extramarital affairs. More information about the study is available in a 1978 paper from the Journal of Political Economy.

Description of Variables

The dataset contains 6366 observations of 9 variables:

rate_marriage: woman's rating of her marriage (1 = very poor, 5 = very good)

age: woman's age

yrs_married: number of years married

children: number of children

religious: woman's rating of how religious she is (1 = not religious, 4 = strongly religious)

educ: level of education (9 = grade school, 12 = high school, 14 = some college, 16 = college graduate, 17 = some graduate school, 20 = advanced degree)

occupation: woman's occupation (1 = student, 2 = farming/semi skilled/unskilled, 3 = "white collar", 4 =
teacher/nurse/writer/technician/skilled, 5 = managerial/business, 6 = professional with advanced degree)

occupation_husb: husband's occupation (same coding as above)

affairs: time spent in extra-marital affairs

Code to loading data and modules:

In [1]:

```
import numpy as np
import pandas as pd
import statsmodels.api as sm
import matplotlib.pyplot as plt
from patsy import dmatrices
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn import metrics
#from sklearn.cross_validation import cross_val_score
from sklearn.model_selection import cross_val_score
```

In [2]:

```
dta =sm.datasets.fair.load_pandas().data
#add "affair" column: 1 represents having affairs, 0 represents not
dta
```

Out[2]:

	rate_marriage	age	yrs_married	children	religious	educ	occupation	occupation_husb	
0	3.0	32.0	9.0	3.0	3.0	17.0	2.0	5.0	0.
1	3.0	27.0	13.0	3.0	1.0	14.0	3.0	4.0	3.
2	4.0	22.0	2.5	0.0	1.0	16.0	3.0	5.0	1.
3	4.0	37.0	16.5	4.0	3.0	16.0	5.0	5.0	0.
4	5.0	27.0	9.0	1.0	1.0	14.0	3.0	4.0	4.
...	...	...	...	...	...	...	...	...	
6361	5.0	32.0	13.0	2.0	3.0	17.0	4.0	3.0	0.
6362	4.0	32.0	13.0	1.0	1.0	16.0	5.0	5.0	0.
6363	5.0	22.0	2.5	0.0	2.0	14.0	3.0	1.0	0.
6364	5.0	32.0	6.0	1.0	3.0	14.0	3.0	4.0	0.
6365	4.0	22.0	2.5	0.0	2.0	16.0	2.0	4.0	0.

6366 rows × 9 columns

In [3]:

```
dta["affairs"].tail()
```

Out[3]:

```
6361    0.0
6362    0.0
6363    0.0
6364    0.0
6365    0.0
Name: affairs, dtype: float64
```

In [4]:

```
dta['affair'] = (dta.affairs >0).astype(int)
y, X = dmatrices('affair ~ rate_marriage + age + yrs_married + children + religi
y.head()
```

Out[4]:

	affair
0	1.0
1	1.0
2	1.0
3	1.0
4	1.0

In [5]:

```
X = X.rename(columns ={'C(occupation)[T.2.0]':'occ_2', 'C(occupation)[T.3.0]':'oc
 'C(occupation)[T.6.0]':'occ_6', 'C(occupation_husb)[T.2.0]':'occ_husb_2','C(occu
 'C(occupation_husb)[T.4.0]':'occ_husb_4', 'C(occupation_husb)[T.5.0]':'occ_husb_
```

In [6]:

```
y = np.ravel(y)
```

In [7]:

```
model = LogisticRegression()
model = model.fit(X, y)

# check the accuracy on the training set
Accuracy = model.score(X, y)
print("Accuracy of this model is :- {}%   ".format(round(Accuracy*100 , ndigits=
```

```
C:\Users\idofa\Anaconda3\lib\site-packages\sklearn\linear_model\logistic.py:43
2: FutureWarning: Default solver will be changed to 'lbfgs' in 0.22. Specify a
solver to silence this warning.
  FutureWarning)

Accuracy of this model is :- 72.589%
```

In [8]:

```
y.mean()
```

Out[8]:

```
0.3224945020420987
```

In [9]:

```
pd.DataFrame(zip(X.columns, np.transpose(model.coef_)))
```

Out[9]:

	0	1
0	Intercept	[1.489835891324933]
1	occ_2	[0.18806639024440983]
2	occ_3	[0.4989478668156914]
3	occ_4	[0.25066856498524825]
4	occ_5	[0.8390080648117001]
5	occ_6	[0.8339084337443315]
6	occ_husb_2	[0.1906359445867889]
7	occ_husb_3	[0.2978327129263421]
8	occ_husb_4	[0.1614088540760616]
9	occ_husb_5	[0.18777091388972483]
10	occ_husb_6	[0.19401637225511495]
11	rate_marriage	[-0.7031233597323255]
12	age	[-0.05841777448168919]
13	yrs_married	[0.10567653799735635]
14	children	[0.016919266970905608]
15	religious	[-0.3711362653137546]
16	educ	[0.00401650319563816]

In [ ]: