LOGISTIC REGRESSION

DATA

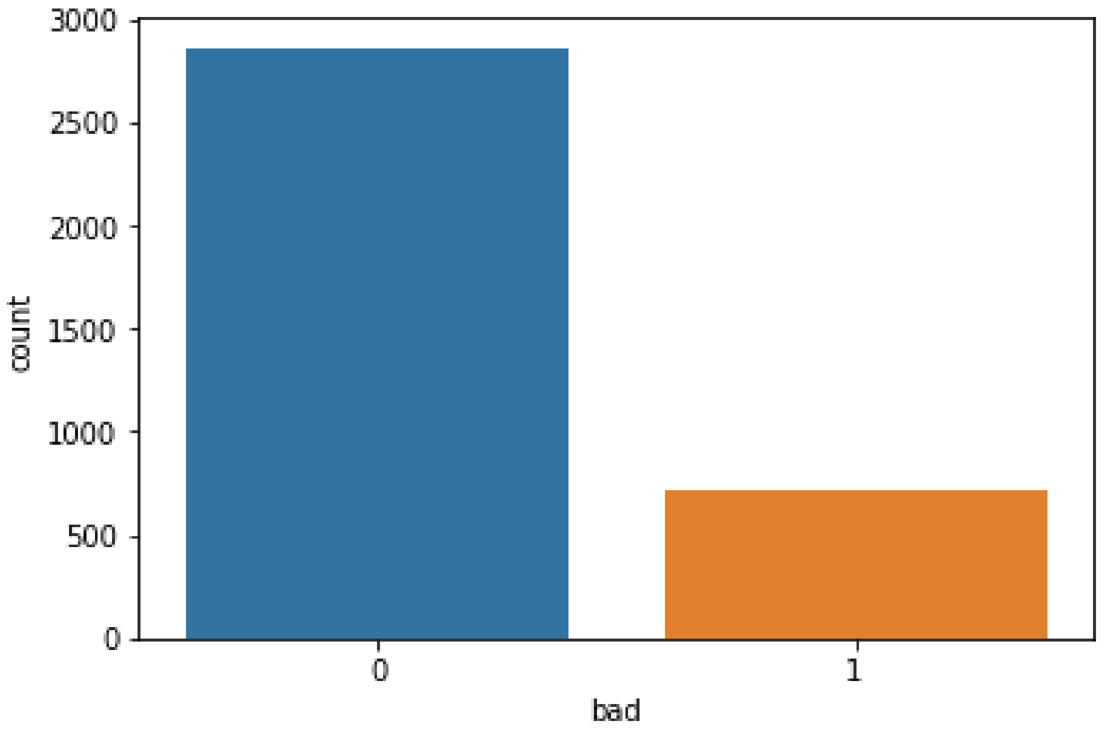
```
information = Classes.Information(data)
information.data_features()
```

```
--------------------------  DATA HEAD --------------------------

   bad   loan   mortdue     value   reason  job   yoj  derog  delinq   clage  \
0    0  81200   18834.0  108355.0  HomeImp  NaN  28.0    0.0     0.0  139.14
1    0  12600  103960.0  127384.0  DebtCon  NaN   2.0    0.0     0.0  129.02
2    0  18000   46865.0   61266.0  DebtCon  NaN   5.0    0.0     0.0  102.59
3    0  10300   57676.0   71027.0  DebtCon  NaN  19.0    0.0     0.0  157.52
4    0   9400   56508.0   78358.0  DebtCon  NaN  17.0    0.0     0.0  141.93

   ninq  clno  debtinc
0   0.0  14.0   34.042
1   0.0  25.0   34.479
2   2.0   9.0   26.354
3   1.0  11.0   33.992
4   0.0  11.0   32.327
```

DATA

```
-------------------------- DATA INFO --------------------------

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3576 entries, 0 to 3575
Data columns (total 13 columns):
bad         3576 non-null int64
loan        3576 non-null int64
mortdue     3262 non-null float64
value       3512 non-null float64
reason      3429 non-null object
job         3409 non-null object
yoj         3264 non-null float64
derog       3149 non-null float64
delinq      3225 non-null float64
clage       3397 non-null float64
ninq        3273 non-null float64
clno        3443 non-null float64
debtinc     2809 non-null float64
dtypes: float64(9), int64(2), object(2)
memory usage: 335.3+ KB
None

-------------------------- DATA SHAPE --------------------------

(3576, 13)
```

DATA



Veriseti içindeki krediyi ödeyip / ödeyememe durumunun yüzdesi :

```
Kredisini Ödeyenler 80.06152125279642
Kredisini Ödemeyenler 19.938478747203582
```

Sınıfların dağılımı dengesizdir.

DATA

```
data.groupby('bad').mean()
```

bad	loan	mortdue	value	yoj	derog	delinq	clage	ninq	clno	debtinc
0	18931.645127	75242.395117	102394.448489	9.031378	0.140732	0.238263	186.338950	1.032692	21.552536	33.179142
1	16915.708275	69029.488140	95308.460184	8.067533	0.716012	1.174888	153.497474	1.780089	21.323572	40.881416

- Kredisini ödeyenlerin(bad =0) , kredi talep miktarı ortalamasının(loan), kredisini ödemeyenlerin(bad =1) kredi talep miktarı ortalamasından yüksektir. Bu durumda kredisini ödeyebilenler yüksek kredi talebinde bulunmuştur diyebiliriz.
- Negatif rapor sayıları fazla olan bireylerin (derog) çoğu kredisini ödememiştir.
- Kredilerini ödeyen bireylerin borç/gelir oranı (debtinc) , kredilerini ödeyemeyen bireylerden daha düşüktür

DATA

```
data.groupby('job').mean()
```

job	bad	loan	mortdue	value	yoj	derog	delinq	clage	ninq	clno	debtinc
Mgr	0.232104	19084.598698	83964.704189	108464.106133	8.919318	0.320707	0.594203	174.285822	1.517564	23.097561	35.307687
Office	0.131810	18048.857645	68058.197973	94675.024670	8.103011	0.136905	0.445076	178.784840	0.936803	21.425795	34.158283
Other	0.232006	18006.918239	60064.432343	84251.694202	9.403457	0.313281	0.417183	174.026556	1.333836	19.572139	34.260072
ProfEx	0.166884	18750.717080	92690.971376	128851.319683	8.731349	0.203911	0.376871	196.769973	0.949728	24.503989	32.622049
Sales	0.348485	15251.515152	79856.864407	105960.969231	7.476667	0.450000	0.274194	202.301667	0.772727	24.272727	38.326064
Self	0.295652	27923.478261	102575.392523	147150.513274	7.210185	0.221239	0.551402	176.590526	1.404040	24.271930	36.824762

- Mesleklere göre kredi talep miktarları değişmektedir.(loan)

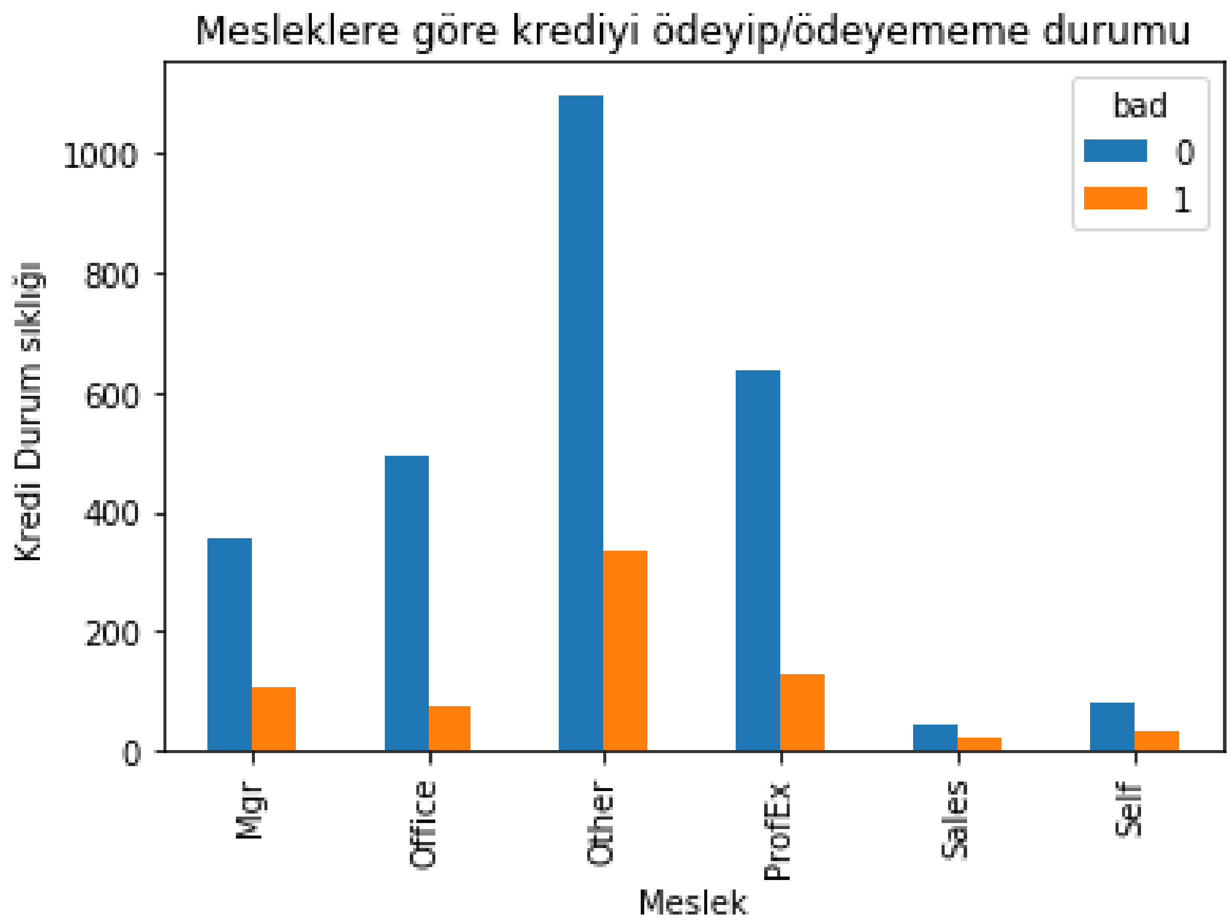
DATA

```
data.groupby('reason').mean()
```

	bad	loan	mortdue	value	yoj	derog	delinq	clage	ninq	clno	debtinc
reason											
DebtCon	0.185576	19868.705188	74483.615277	101611.714495	8.551114	0.261098	0.409427	176.176174	1.343708	22.287742	34.301599
HomeImp	0.230624	15892.911153	73308.909702	100007.497760	9.411429	0.245596	0.455852	185.208453	0.845361	19.905273	33.496014

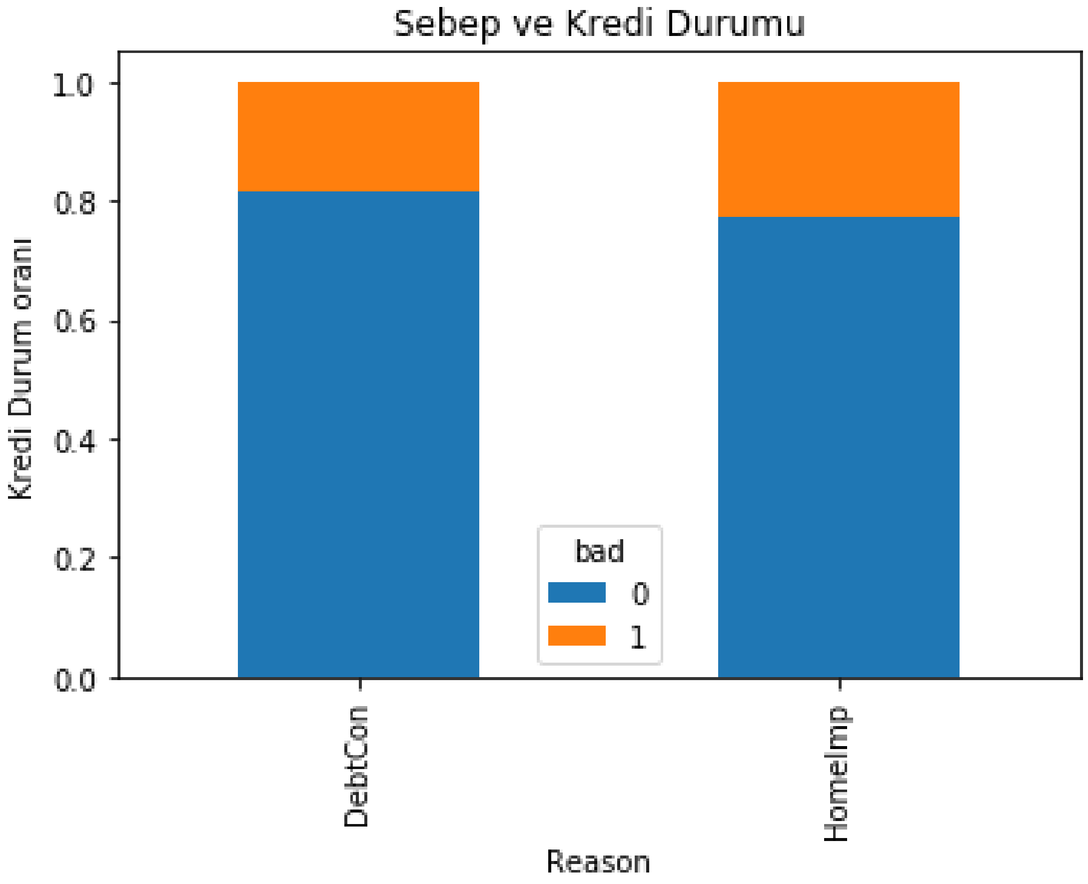
- Ev kredisi alanların kredi talep tutarları , borç kredisi alanların kredi talep tutarlarından düşüktür.(loan)

GÖRSELLEŞTİRME



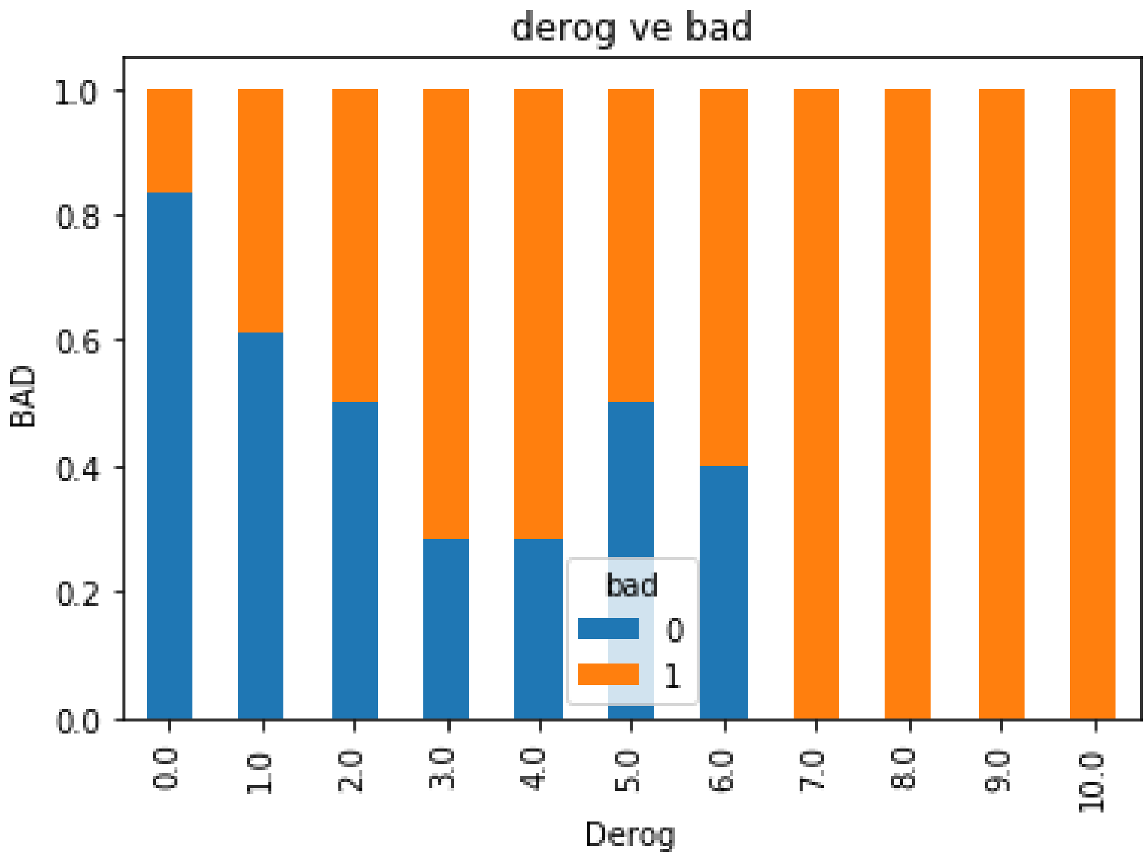
- Krediyi ödeyebilme(bad = 0) , büyük ölçüde mesleklere bağlı .Dolayısıyla iş unvanı, sonuç değişkeninin iyi bir öngörücüsü olabilir.

GÖRSELLEŞTİRME



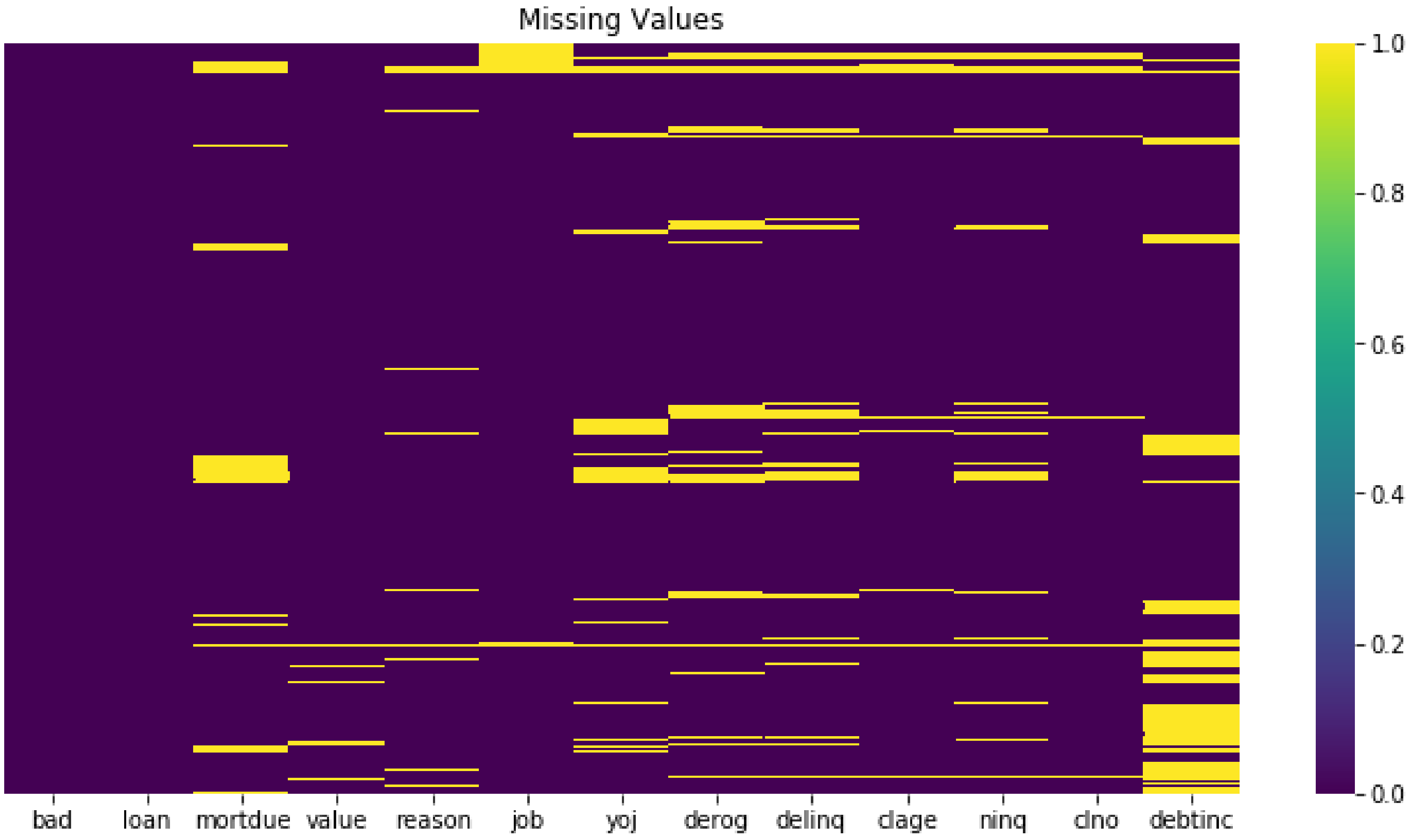
- Kredi talep sebebi , y değişkeni için güçlü bir yorumlayıcı görünmemektedir.

GÖRSELLEŞTİRME



- Negatif raporlara sahip bireylerin kredilerini ödeyip ödeyememelerinde iyi bir öngörücü olabilir

PREPROCESS




```
------------  Missing Values  --

debtinc    767
derog      427
delinq     351
mortdue    314
yoj        312
ninq       303
clage      179
job        167
reason     147
clno       133
value       64
loan         0
bad          0
dtype: int64
```

PREPROCESS

```
p.drop('any')
```

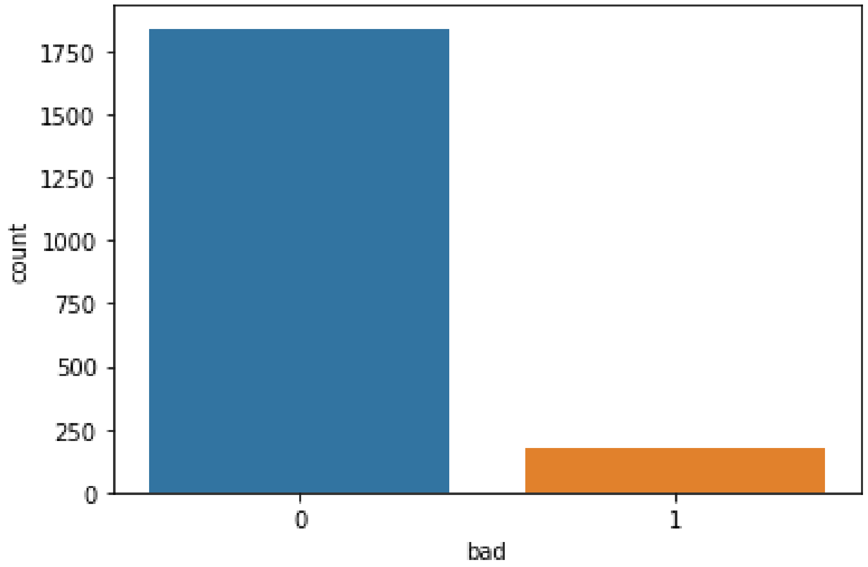
```
Drop Öncesi Data Shape -->    (3576, 13)
Drop Sonrası Data Shape -->    (2018, 13)
------------  Missing Values  -------------

debtinc     0
clno        0
ninq        0
clage       0
delinq      0
derog       0
yoj         0
job         0
reason      0
value       0
mortdue     0
loan        0
bad         0
dtype: int64
```

```
# DUMMIES

HomeImp = pd.get_dummies(data['reason'], drop_first =True)
jobs = pd.get_dummies(data['job'], drop_first =True)
data=pd.concat([data,HomeImp,jobs],axis=1)
data.head()
```

	bad	loan	mortdue	value	reason	job	yoj	derog	delinq	clage	ninq	clno	debtinc	HomeImp	Office	Other	ProfEx	Sales	Self
153	0	18200	94727.0	136877.0	DebtCon	Mgr	15.0	0.0	0.0	168.96	2.0	26.0	36.056	0	0	0	0	0	0
154	0	21700	79240.0	96784.0	DebtCon	Mgr	5.0	0.0	0.0	64.51	6.0	24.0	38.079	0	0	0	0	0	0
155	0	34100	241931.0	36486.0	DebtCon	Mgr	1.0	0.0	2.0	196.01	3.0	50.0	42.459	0	0	0	0	0	0
156	0	8400	62989.0	76718.0	HomeImp	Mgr	3.0	0.0	2.0	131.47	0.0	22.0	29.200	1	0	0	0	0	0
157	0	17400	25859.0	43684.0	DebtCon	Mgr	16.0	1.0	0.0	95.36	1.0	17.0	27.108	0	0	0	0	0	0

```
data.drop(['reason','job'],axis =1,inplace=True)
data.head()
```

PREPROCESS



- Veri seti çok dengesiz dağılmış.
- Bu dengesizliği gidermek için SMOTE Algoritması kullanılarak veri setindeki azınlıkta olan gözlemler çoğaltılır.

```
X = data.loc[:, data.columns != 'bad']
y = data.loc[:, data.columns == 'bad']
```

```
os_data_X, os_data_y = p.SMOTE(X,y)
```

```
length of oversampled data is  2574
Y değişkeni  0  : 1287
Y değişkeni  1  : 1287
```

LOGISTIC REGRESSION MODEL

```
model = Classes.GridSearchHelper()
model.LogReg(os_data_X, os_data_y)
```

```
------------------        STATS   MODELS   --------------------

Optimization terminated successfully.
         Current function value: 0.538330
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:                    bad   No. Observations:                 2574
Model:                          Logit   Df Residuals:                     2558
Method:                           MLE   Df Model:                           15
Date:                Wed, 05 Aug 2020   Pseudo R-squ.:                  0.2234
Time:                        22:45:08   Log-Likelihood:                -1385.7
converged:                       True   LL-Null:                       -1784.2
Covariance Type:            nonrobust   LLR p-value:                3.732e-160
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
loan       -1.291e-05   5.59e-06     -2.309      0.021   -2.39e-05   -1.95e-06
mortdue    -7.179e-07   2.71e-06     -0.264      0.791   -6.04e-06     4.6e-06
value       1.086e-06   2.28e-06      0.476      0.634   -3.39e-06    5.56e-06
yoj           -0.0410      0.007     -5.622      0.000      -0.055      -0.027
derog          0.6062      0.090      6.742      0.000       0.430       0.782
delinq         1.0166      0.077     13.151      0.000       0.865       1.168
clage         -0.0047      0.001     -7.387      0.000      -0.006      -0.003
ninq           0.2003      0.033      5.991      0.000       0.135       0.266
clno          -0.0363      0.006     -6.349      0.000      -0.047      -0.025
debtinc        0.0479      0.004     11.122      0.000       0.039       0.056
HomeImp       -0.2302      0.115     -2.002      0.045      -0.456      -0.005
Office        -1.0009      0.179     -5.600      0.000      -1.351      -0.651
Other         -0.3617      0.137     -2.634      0.008      -0.631      -0.093
ProfEx        -0.3326      0.160     -2.078      0.038      -0.646      -0.019
Sales          0.9157      0.376      2.434      0.015       0.178       1.653
Self           0.4353      0.335      1.298      0.194      -0.222       1.093
==============================================================================
```

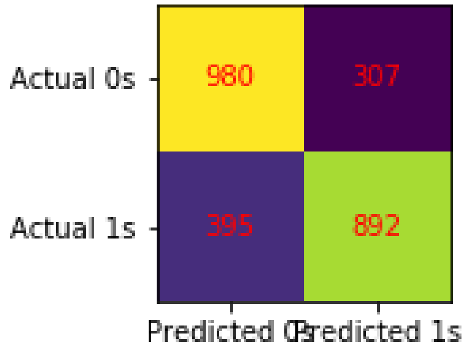
LOGISTIC REGRESSION MODEL

```
------------------     SCIKIT LEARN MODEL  --------------------

Intercept :    [-0.36524632]

Coefficient  :  [[-1.11297114e-05  3.84759933e-07  5.71598274e-07 -4.25266309e-02
   6.58148234e-01  1.02746457e+00 -4.56320793e-03  1.94418147e-01
  -3.38400204e-02  4.73588658e-02 -1.72856182e-01 -4.49158818e-01
  -7.45860071e-02 -7.06819214e-02  1.55828665e-01  9.08299959e-02]]
```

Classification Report :

	precision	recall	f1-score	support
0	0.71	0.76	0.74	1287
1	0.74	0.69	0.72	1287
accuracy			0.73	2574
macro avg	0.73	0.73	0.73	2574
weighted avg	0.73	0.73	0.73	2574

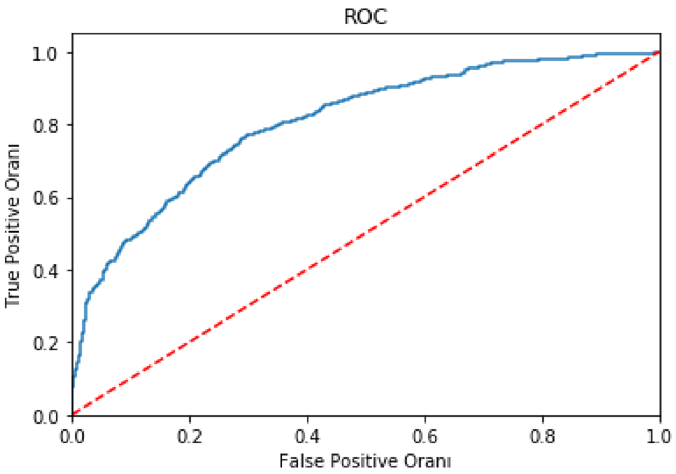
--------------- CONFUSION MATRIS ---------



Accuracy Score :   0.7272727272727273

LOGISTIC REGRESSION MODEL

-------------- ROC CURVE --------------------



-------------- TRAIN - TEST SPLIT --------------------

Accuracy Score : 0.7533980582524272
Classification Report :

	precision	recall	f1-score	support
0	0.75	0.78	0.76	263
1	0.76	0.72	0.74	252
accuracy			0.75	515
macro avg	0.75	0.75	0.75	515
weighted avg	0.75	0.75	0.75	515

Cross Validation Score : 0.7087445573294631

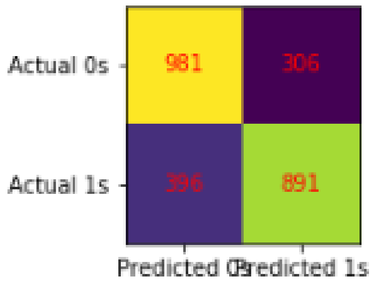
LOGISTIC REGRESSION MODEL

```
os_data_X.drop(['mortdue'],axis =1,inplace=True)
```

P değeri 0.05 den yüksek olan ‘mortdue’ sütunu veri setinden silinip tekrar model kuruldu

```
------------------     STATS  MODELS  --------------------

Optimization terminated successfully.
         Current function value: 0.538344
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:                    bad   No. Observations:                 2574
Model:                          Logit   Df Residuals:                     2559
Method:                           MLE   Df Model:                           14
Date:                Wed, 05 Aug 2020   Pseudo R-squ.:                  0.2233
Time:                        22:47:04   Log-Likelihood:                -1385.7
converged:                       True   LL-Null:                       -1784.2
Covariance Type:            nonrobust   LLR p-value:                 5.021e-161
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
loan       -1.269e-05   5.53e-06     -2.295      0.022   -2.35e-05   -1.85e-06
value       5.379e-07   9.63e-07      0.559      0.576   -1.35e-06    2.43e-06
yoj           -0.0407      0.007     -5.645      0.000      -0.055      -0.027
derog          0.6085      0.090      6.794      0.000       0.433       0.784
delinq         1.0161      0.077     13.163      0.000       0.865       1.167
clage         -0.0047      0.001     -7.416      0.000      -0.006      -0.003
ninq           0.2000      0.033      5.987      0.000       0.135       0.266
clno          -0.0367      0.005     -6.715      0.000      -0.047      -0.026
debtinc        0.0479      0.004     11.137      0.000       0.039       0.056
HomeImp       -0.2302      0.115     -2.002      0.045      -0.456      -0.005
Office        -0.9953      0.177     -5.609      0.000      -1.343      -0.648
Other         -0.3557      0.135     -2.626      0.009      -0.621      -0.090
ProfEx        -0.3262      0.158     -2.062      0.039      -0.636      -0.016
Sales          0.9218      0.376      2.454      0.014       0.186       1.658
Self           0.4397      0.335      1.311      0.190      -0.218       1.097
==============================================================================
```

```
--------------- CONFUSION MATRIS  ---------------------
```




```
Accuracy  Score :   0.7272727272727273
Classification Report  :
              precision    recall  f1-score   support

           0       0.71      0.76      0.74      1287
           1       0.74      0.69      0.72      1287

    accuracy                           0.73      2574
   macro avg       0.73      0.73      0.73      2574
weighted avg       0.73      0.73      0.73      2574
```

LOGISTIC REGRESSION MODEL

```
--------------- TRAIN - TEST SPLIT  ---------------------

Accuracy  Score  :  0.7572815533980582
Classification Report  :
              precision    recall  f1-score   support

           0       0.75      0.79      0.77       263
           1       0.77      0.73      0.75       252

    accuracy                           0.76       515
   macro avg       0.76      0.76      0.76       515
weighted avg       0.76      0.76      0.76       515

Cross Validation Score  :   0.7186224707589858
```

LOGISTIC REGRESSION MODEL

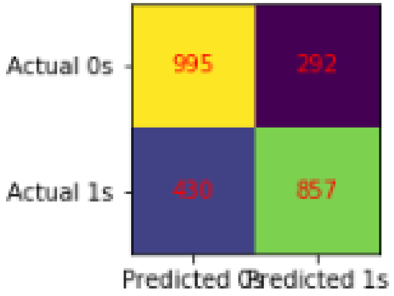
```
os_data_X.drop(['value'],axis =1,inplace=True)
model.LogReg(os_data_X, os_data_y)
```

```
------------------      STATS  MODELS  --------------------

Optimization terminated successfully.
         Current function value: 0.538404
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:                    bad   No. Observations:                 2574
Model:                          Logit   Df Residuals:                     2560
Method:                           MLE   Df Model:                           13
Date:                Wed, 05 Aug 2020   Pseudo R-squ.:                  0.2232
Time:                        22:48:23   Log-Likelihood:                -1385.9
converged:                       True   LL-Null:                       -1784.2
Covariance Type:            nonrobust   LLR p-value:                 7.327e-162
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
loan       -1.118e-05   4.79e-06     -2.333      0.020   -2.06e-05   -1.79e-06
yoj           -0.0407      0.007     -5.638      0.000      -0.055      -0.027
derog          0.6067      0.090      6.769      0.000       0.431       0.782
delinq         1.0142      0.077     13.163      0.000       0.863       1.165
clage         -0.0046      0.001     -7.402      0.000      -0.006      -0.003
ninq           0.1992      0.033      5.973      0.000       0.134       0.265
clno          -0.0365      0.005     -6.686      0.000      -0.047      -0.026
debtinc        0.0484      0.004     11.567      0.000       0.040       0.057
HomeImp       -0.2247      0.114     -1.963      0.050      -0.449      -0.000
Office        -1.0001      0.177     -5.641      0.000      -1.347      -0.653
Other         -0.3662      0.134     -2.728      0.006      -0.629      -0.103
ProfEx        -0.3158      0.157     -2.010      0.044      -0.624      -0.008
Sales          0.9184      0.376      2.442      0.015       0.181       1.656
Self           0.4528      0.334      1.355      0.176      -0.202       1.108
==============================================================================
```

P değeri 0.05 den yüksek olan ‘value’ sütunu veri setinden silinip tekrar model kuruldu

LOGISTIC REGRESSION MODEL

--------------- CONFUSION MATRIS ----------------------



Accuracy Score : 0.7195027195027195

Classification Report :

	precision	recall	f1-score	support
0	0.70	0.77	0.73	1287
1	0.75	0.67	0.70	1287
accuracy			0.72	2574
macro avg	0.72	0.72	0.72	2574
weighted avg	0.72	0.72	0.72	2574

-------------- TRAIN - TEST SPLIT ---------------------

Accuracy Score : 0.7339805825242719

Classification Report :

	precision	recall	f1-score	support
0	0.71	0.81	0.76	263
1	0.76	0.66	0.71	252
accuracy			0.73	515
macro avg	0.74	0.73	0.73	515
weighted avg	0.74	0.73	0.73	515

Cross Validation Score : 0.7144775036284471

PCA(PRINCIPLE COMPONENT ANALYSIS)

DATA

```
data = pd.read_excel("HW_Data_Set.xlsx")
```

```
information = Classes.Information(data)
information.data_features()
```

```
-------------------------- DATA HEAD --------------------------

   ind_5  ind_6  ind_8      ind_9     ind_10     ind_12     ind_13  ind_14  \
0     19     17  100.0  85.714286  14.285714  72.363515  60.808814   23.80
1     24     19  100.0  78.571429  21.428571  74.275883  64.366798   11.45
2     30     24  100.0  71.428571  28.571429  75.140402  65.915803    8.75
3     37     30  100.0  64.285714  35.714286  76.677846  68.584234    7.80
4     41     37  100.0  57.142857  42.857143  81.603007  76.455495   14.90

   ind_15  ind_16  ...  ind_416  ind_418  ind_420  ind_422  ind_424   ind_426  \
0   17.62   11.73  ...    -49.6      -54     -152     -353      1.0  0.498547
1   18.16   12.22  ...    -55.6      -60     -158     -359      1.0  0.537088
2   17.86   12.28  ...    -58.4      -60     -160     -362      1.0  0.615169
3   14.76   12.61  ...    -61.8      -65     -166     -367      1.0  0.661517
4   11.92   14.25  ...    -79.8      -86     -186     -388      1.0  0.747204

    ind_428  20_target  50_target  90_target
0  0.701906  15.135802  35.625252  36.997753
1  0.690833  15.143348  35.643013  37.016198
2  0.693040  15.146870  35.651301 -37.024805
3  0.673418  15.153283   0.000000 -37.040483
4  0.700522 -15.179065 -35.727079 -37.103503

[5 rows x 136 columns]
```

DATA

```
data = data[data['ind_420'] != '?']
data = data[data['ind_422'] != '?']

#dummy

RED = pd.get_dummies(data['ind_109'], drop_first =True)
data=pd.concat([data,RED],axis=1)
data.drop(['ind_109'],axis =1,inplace=True)
```

```
X = data.iloc[:, 0:132]
y = data.loc[:, data.columns == '20_target']
```

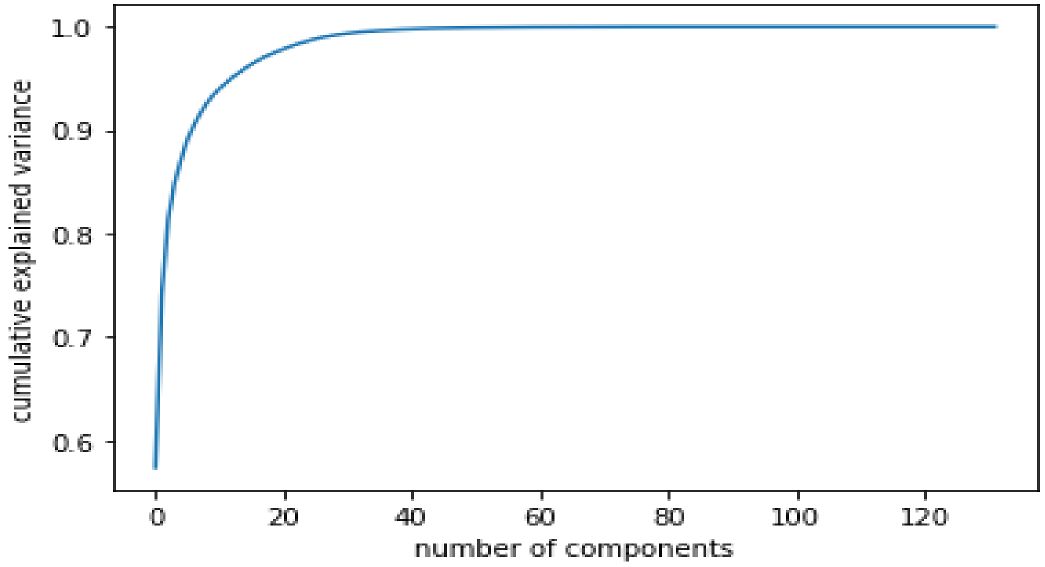
Veri seti içindeki eksik gözlem bulunan satırlar silinip veriseti X ve y değişkenlerine bölündü

PCA MODEL

```python
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X,y,test_size = 0.2, random_state = 42)
```

```python
from sklearn.decomposition import PCA
pca = PCA(whiten = True)
pca.fit(X_train)
```

```python
plt.plot(np.cumsum(pca.explained_variance_ratio_))
plt.xlabel('number of components')
plt.ylabel('cumulative explained variance');
```



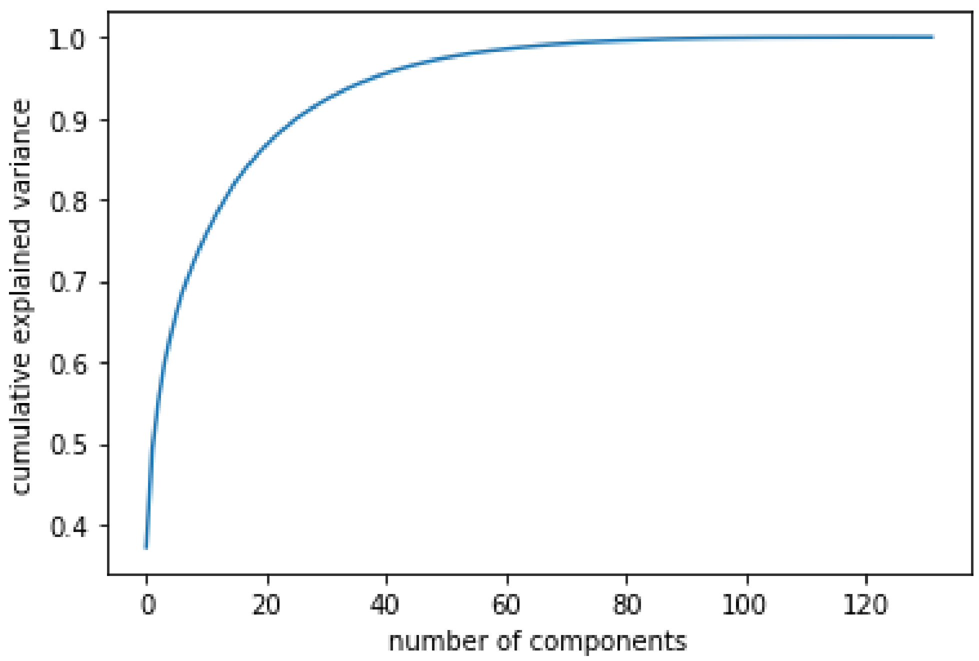
N_components = 30 için yaklaşık %1 lik veri kaybı olduğu görülüyor

```python
print("Sum : ",sum(pca2.explained_variance_ratio_))
```

```
Sum :   0.9927338972077313
```

PCA MODEL

```python
from sklearn.preprocessing import StandardScaler

sc = StandardScaler()
X_std = sc.fit_transform(X_train)
pca3 = PCA()
X_pca = pca3.fit(X_std)
```



X değişkeni StandardScaler() metodu ile normalize edildiğinde n_component = 30 için daha fazla veri kaybı oldu. Yaklaşık %8 .

```python
pca3 = PCA(n_components = 30)
X_pca = pca3.fit_transform(X_std)
print("Sum : ",sum(pca3.explained_variance_ratio_))
```

```
Sum :  0.9188112389729786
```

LINEAR MODEL

```
from sklearn.linear_model import LinearRegression

lm = LinearRegression()
pcr_model = lm.fit(X_pca, y_train)
```

```
y_pred = pcr_model.predict(X_pca)
```

```
from sklearn.metrics import mean_squared_error, r2_score
np.sqrt(mean_squared_error(y_train, y_pred))
```

```
14.526665790088147
```