MercedesBenz

September 4, 2020

[263]:
```
import pandas as pd
import numpy as np
from sklearn.preprocessing import LabelEncoder
import statsmodels.api as sm
from statsmodels.formula.api import ols
from sklearn.decomposition import PCA
from sklearn.model_selection import train_test_split
import xgboost as xgb
from sklearn.metrics import mean_squared_error, r2_score
import matplotlib.pyplot as plt
```

[264]:
```
df_train = pd.read_csv("train.csv")
df_test = pd.read_csv("test.csv")
```

EDA

[265]:
```
df_train.head()
```

[265]:

	ID	y	X0	X1	X2	X3	X4	X5	X6	X8	…	X375	X376	X377	X378	X379	\
0	0	130.81	k	v	at	a	d	u	j	o	…	0	0	1	0	0	
1	6	88.53	k	t	av	e	d	y	l	o	…	1	0	0	0	0	
2	7	76.26	az	w	n	c	d	x	j	x	…	0	0	0	0	0	
3	9	80.62	az	t	n	f	d	x	l	e	…	0	0	0	0	0	
4	13	78.02	az	v	n	f	d	h	d	n	…	0	0	0	0	0	

	X380	X382	X383	X384	X385
0	0	0	0	0	0
1	0	0	0	0	0
2	0	1	0	0	0
3	0	0	0	0	0
4	0	0	0	0	0

[5 rows x 378 columns]

[266]:
```
df_test.head()
```

[266]:

	ID	X0	X1	X2	X3	X4	X5	X6	X8	X10	…	X375	X376	X377	X378	X379	X380	\
0	1	az	v	n	f	d	t	a	w	0	…	0	0	0	1	0	0	

```
1   2   t  b  ai  a  d  b  g  y    0  …    0    0    1    0    0    0
2   3  az  v  as  f  d  a  j  j    0  …    0    0    0    1    0    0
3   4  az  l   n  f  d  z  l  n    0  …    0    0    0    1    0    0
4   5   w  s  as  c  d  y  i  m    0  …    1    0    0    0    0    0

   X382  X383  X384  X385
0     0     0     0     0
1     0     0     0     0
2     0     0     0     0
3     0     0     0     0
4     0     0     0     0

[5 rows x 377 columns]
```

[267]:
```
df_train.shape, df_test.shape
```

[267]:
```
((4209, 378), (4209, 377))
```

[268]:
```
pd.options.display.max_rows = None
```

[269]:
```
df_train.dtypes
```

[269]:
```
ID        int64
y       float64
X0       object
X1       object
X2       object
X3       object
X4       object
X5       object
X6       object
X8       object
X10       int64
X11       int64
X12       int64
X13       int64
X14       int64
X15       int64
X16       int64
X17       int64
X18       int64
X19       int64
X20       int64
X21       int64
X22       int64
X23       int64
X24       int64
```

```
X26       int64
X27       int64
X28       int64
X29       int64
X30       int64
X31       int64
X32       int64
X33       int64
X34       int64
X35       int64
X36       int64
X37       int64
X38       int64
X39       int64
X40       int64
X41       int64
X42       int64
X43       int64
X44       int64
X45       int64
X46       int64
X47       int64
X48       int64
X49       int64
X50       int64
X51       int64
X52       int64
X53       int64
X54       int64
X55       int64
X56       int64
X57       int64
X58       int64
X59       int64
X60       int64
X61       int64
X62       int64
X63       int64
X64       int64
X65       int64
X66       int64
X67       int64
X68       int64
X69       int64
X70       int64
X71       int64
X73       int64
```

```
X74       int64
X75       int64
X76       int64
X77       int64
X78       int64
X79       int64
X80       int64
X81       int64
X82       int64
X83       int64
X84       int64
X85       int64
X86       int64
X87       int64
X88       int64
X89       int64
X90       int64
X91       int64
X92       int64
X93       int64
X94       int64
X95       int64
X96       int64
X97       int64
X98       int64
X99       int64
X100      int64
X101      int64
X102      int64
X103      int64
X104      int64
X105      int64
X106      int64
X107      int64
X108      int64
X109      int64
X110      int64
X111      int64
X112      int64
X113      int64
X114      int64
X115      int64
X116      int64
X117      int64
X118      int64
X119      int64
X120      int64
```

```
X122      int64
X123      int64
X124      int64
X125      int64
X126      int64
X127      int64
X128      int64
X129      int64
X130      int64
X131      int64
X132      int64
X133      int64
X134      int64
X135      int64
X136      int64
X137      int64
X138      int64
X139      int64
X140      int64
X141      int64
X142      int64
X143      int64
X144      int64
X145      int64
X146      int64
X147      int64
X148      int64
X150      int64
X151      int64
X152      int64
X153      int64
X154      int64
X155      int64
X156      int64
X157      int64
X158      int64
X159      int64
X160      int64
X161      int64
X162      int64
X163      int64
X164      int64
X165      int64
X166      int64
X167      int64
X168      int64
X169      int64
```

```
X170      int64
X171      int64
X172      int64
X173      int64
X174      int64
X175      int64
X176      int64
X177      int64
X178      int64
X179      int64
X180      int64
X181      int64
X182      int64
X183      int64
X184      int64
X185      int64
X186      int64
X187      int64
X189      int64
X190      int64
X191      int64
X192      int64
X194      int64
X195      int64
X196      int64
X197      int64
X198      int64
X199      int64
X200      int64
X201      int64
X202      int64
X203      int64
X204      int64
X205      int64
X206      int64
X207      int64
X208      int64
X209      int64
X210      int64
X211      int64
X212      int64
X213      int64
X214      int64
X215      int64
X216      int64
X217      int64
X218      int64
```

```
X219      int64
X220      int64
X221      int64
X222      int64
X223      int64
X224      int64
X225      int64
X226      int64
X227      int64
X228      int64
X229      int64
X230      int64
X231      int64
X232      int64
X233      int64
X234      int64
X235      int64
X236      int64
X237      int64
X238      int64
X239      int64
X240      int64
X241      int64
X242      int64
X243      int64
X244      int64
X245      int64
X246      int64
X247      int64
X248      int64
X249      int64
X250      int64
X251      int64
X252      int64
X253      int64
X254      int64
X255      int64
X256      int64
X257      int64
X258      int64
X259      int64
X260      int64
X261      int64
X262      int64
X263      int64
X264      int64
X265      int64
```

```
X266      int64
X267      int64
X268      int64
X269      int64
X270      int64
X271      int64
X272      int64
X273      int64
X274      int64
X275      int64
X276      int64
X277      int64
X278      int64
X279      int64
X280      int64
X281      int64
X282      int64
X283      int64
X284      int64
X285      int64
X286      int64
X287      int64
X288      int64
X289      int64
X290      int64
X291      int64
X292      int64
X293      int64
X294      int64
X295      int64
X296      int64
X297      int64
X298      int64
X299      int64
X300      int64
X301      int64
X302      int64
X304      int64
X305      int64
X306      int64
X307      int64
X308      int64
X309      int64
X310      int64
X311      int64
X312      int64
X313      int64
```

```
X314      int64
X315      int64
X316      int64
X317      int64
X318      int64
X319      int64
X320      int64
X321      int64
X322      int64
X323      int64
X324      int64
X325      int64
X326      int64
X327      int64
X328      int64
X329      int64
X330      int64
X331      int64
X332      int64
X333      int64
X334      int64
X335      int64
X336      int64
X337      int64
X338      int64
X339      int64
X340      int64
X341      int64
X342      int64
X343      int64
X344      int64
X345      int64
X346      int64
X347      int64
X348      int64
X349      int64
X350      int64
X351      int64
X352      int64
X353      int64
X354      int64
X355      int64
X356      int64
X357      int64
X358      int64
X359      int64
X360      int64
```

```
X361      int64
X362      int64
X363      int64
X364      int64
X365      int64
X366      int64
X367      int64
X368      int64
X369      int64
X370      int64
X371      int64
X372      int64
X373      int64
X374      int64
X375      int64
X376      int64
X377      int64
X378      int64
X379      int64
X380      int64
X382      int64
X383      int64
X384      int64
X385      int64
dtype: object
```

[270]:
```
df_test.dtypes
```

[270]:
```
ID       int64
X0      object
X1      object
X2      object
X3      object
X4      object
X5      object
X6      object
X8      object
X10      int64
X11      int64
X12      int64
X13      int64
X14      int64
X15      int64
X16      int64
X17      int64
X18      int64
X19      int64
```

```
X20      int64
X21      int64
X22      int64
X23      int64
X24      int64
X26      int64
X27      int64
X28      int64
X29      int64
X30      int64
X31      int64
X32      int64
X33      int64
X34      int64
X35      int64
X36      int64
X37      int64
X38      int64
X39      int64
X40      int64
X41      int64
X42      int64
X43      int64
X44      int64
X45      int64
X46      int64
X47      int64
X48      int64
X49      int64
X50      int64
X51      int64
X52      int64
X53      int64
X54      int64
X55      int64
X56      int64
X57      int64
X58      int64
X59      int64
X60      int64
X61      int64
X62      int64
X63      int64
X64      int64
X65      int64
X66      int64
X67      int64
```

```
X68     int64
X69     int64
X70     int64
X71     int64
X73     int64
X74     int64
X75     int64
X76     int64
X77     int64
X78     int64
X79     int64
X80     int64
X81     int64
X82     int64
X83     int64
X84     int64
X85     int64
X86     int64
X87     int64
X88     int64
X89     int64
X90     int64
X91     int64
X92     int64
X93     int64
X94     int64
X95     int64
X96     int64
X97     int64
X98     int64
X99     int64
X100    int64
X101    int64
X102    int64
X103    int64
X104    int64
X105    int64
X106    int64
X107    int64
X108    int64
X109    int64
X110    int64
X111    int64
X112    int64
X113    int64
X114    int64
X115    int64
```

```
X116    int64
X117    int64
X118    int64
X119    int64
X120    int64
X122    int64
X123    int64
X124    int64
X125    int64
X126    int64
X127    int64
X128    int64
X129    int64
X130    int64
X131    int64
X132    int64
X133    int64
X134    int64
X135    int64
X136    int64
X137    int64
X138    int64
X139    int64
X140    int64
X141    int64
X142    int64
X143    int64
X144    int64
X145    int64
X146    int64
X147    int64
X148    int64
X150    int64
X151    int64
X152    int64
X153    int64
X154    int64
X155    int64
X156    int64
X157    int64
X158    int64
X159    int64
X160    int64
X161    int64
X162    int64
X163    int64
X164    int64
```

```
X165    int64
X166    int64
X167    int64
X168    int64
X169    int64
X170    int64
X171    int64
X172    int64
X173    int64
X174    int64
X175    int64
X176    int64
X177    int64
X178    int64
X179    int64
X180    int64
X181    int64
X182    int64
X183    int64
X184    int64
X185    int64
X186    int64
X187    int64
X189    int64
X190    int64
X191    int64
X192    int64
X194    int64
X195    int64
X196    int64
X197    int64
X198    int64
X199    int64
X200    int64
X201    int64
X202    int64
X203    int64
X204    int64
X205    int64
X206    int64
X207    int64
X208    int64
X209    int64
X210    int64
X211    int64
X212    int64
X213    int64
```

```
X214    int64
X215    int64
X216    int64
X217    int64
X218    int64
X219    int64
X220    int64
X221    int64
X222    int64
X223    int64
X224    int64
X225    int64
X226    int64
X227    int64
X228    int64
X229    int64
X230    int64
X231    int64
X232    int64
X233    int64
X234    int64
X235    int64
X236    int64
X237    int64
X238    int64
X239    int64
X240    int64
X241    int64
X242    int64
X243    int64
X244    int64
X245    int64
X246    int64
X247    int64
X248    int64
X249    int64
X250    int64
X251    int64
X252    int64
X253    int64
X254    int64
X255    int64
X256    int64
X257    int64
X258    int64
X259    int64
X260    int64
```

```
X261    int64
X262    int64
X263    int64
X264    int64
X265    int64
X266    int64
X267    int64
X268    int64
X269    int64
X270    int64
X271    int64
X272    int64
X273    int64
X274    int64
X275    int64
X276    int64
X277    int64
X278    int64
X279    int64
X280    int64
X281    int64
X282    int64
X283    int64
X284    int64
X285    int64
X286    int64
X287    int64
X288    int64
X289    int64
X290    int64
X291    int64
X292    int64
X293    int64
X294    int64
X295    int64
X296    int64
X297    int64
X298    int64
X299    int64
X300    int64
X301    int64
X302    int64
X304    int64
X305    int64
X306    int64
X307    int64
X308    int64
```

```
X309    int64
X310    int64
X311    int64
X312    int64
X313    int64
X314    int64
X315    int64
X316    int64
X317    int64
X318    int64
X319    int64
X320    int64
X321    int64
X322    int64
X323    int64
X324    int64
X325    int64
X326    int64
X327    int64
X328    int64
X329    int64
X330    int64
X331    int64
X332    int64
X333    int64
X334    int64
X335    int64
X336    int64
X337    int64
X338    int64
X339    int64
X340    int64
X341    int64
X342    int64
X343    int64
X344    int64
X345    int64
X346    int64
X347    int64
X348    int64
X349    int64
X350    int64
X351    int64
X352    int64
X353    int64
X354    int64
X355    int64
```

```
X356     int64
X357     int64
X358     int64
X359     int64
X360     int64
X361     int64
X362     int64
X363     int64
X364     int64
X365     int64
X366     int64
X367     int64
X368     int64
X369     int64
X370     int64
X371     int64
X372     int64
X373     int64
X374     int64
X375     int64
X376     int64
X377     int64
X378     int64
X379     int64
X380     int64
X382     int64
X383     int64
X384     int64
X385     int64
dtype: object
```

Check for null and unique values for test and train sets

[271]:
```
df_train.isna().sum()
```

[271]:
```
ID      0
y       0
X0      0
X1      0
X2      0
X3      0
X4      0
X5      0
X6      0
X8      0
X10     0
X11     0
```

```
X12     0
X13     0
X14     0
X15     0
X16     0
X17     0
X18     0
X19     0
X20     0
X21     0
X22     0
X23     0
X24     0
X26     0
X27     0
X28     0
X29     0
X30     0
X31     0
X32     0
X33     0
X34     0
X35     0
X36     0
X37     0
X38     0
X39     0
X40     0
X41     0
X42     0
X43     0
X44     0
X45     0
X46     0
X47     0
X48     0
X49     0
X50     0
X51     0
X52     0
X53     0
X54     0
X55     0
X56     0
X57     0
X58     0
X59     0
```

```
X60     0
X61     0
X62     0
X63     0
X64     0
X65     0
X66     0
X67     0
X68     0
X69     0
X70     0
X71     0
X73     0
X74     0
X75     0
X76     0
X77     0
X78     0
X79     0
X80     0
X81     0
X82     0
X83     0
X84     0
X85     0
X86     0
X87     0
X88     0
X89     0
X90     0
X91     0
X92     0
X93     0
X94     0
X95     0
X96     0
X97     0
X98     0
X99     0
X100    0
X101    0
X102    0
X103    0
X104    0
X105    0
X106    0
X107    0
```

X108	0
X109	0
X110	0
X111	0
X112	0
X113	0
X114	0
X115	0
X116	0
X117	0
X118	0
X119	0
X120	0
X122	0
X123	0
X124	0
X125	0
X126	0
X127	0
X128	0
X129	0
X130	0
X131	0
X132	0
X133	0
X134	0
X135	0
X136	0
X137	0
X138	0
X139	0
X140	0
X141	0
X142	0
X143	0
X144	0
X145	0
X146	0
X147	0
X148	0
X150	0
X151	0
X152	0
X153	0
X154	0
X155	0
X156	0

X157	0
X158	0
X159	0
X160	0
X161	0
X162	0
X163	0
X164	0
X165	0
X166	0
X167	0
X168	0
X169	0
X170	0
X171	0
X172	0
X173	0
X174	0
X175	0
X176	0
X177	0
X178	0
X179	0
X180	0
X181	0
X182	0
X183	0
X184	0
X185	0
X186	0
X187	0
X189	0
X190	0
X191	0
X192	0
X194	0
X195	0
X196	0
X197	0
X198	0
X199	0
X200	0
X201	0
X202	0
X203	0
X204	0
X205	0

```
X206    0
X207    0
X208    0
X209    0
X210    0
X211    0
X212    0
X213    0
X214    0
X215    0
X216    0
X217    0
X218    0
X219    0
X220    0
X221    0
X222    0
X223    0
X224    0
X225    0
X226    0
X227    0
X228    0
X229    0
X230    0
X231    0
X232    0
X233    0
X234    0
X235    0
X236    0
X237    0
X238    0
X239    0
X240    0
X241    0
X242    0
X243    0
X244    0
X245    0
X246    0
X247    0
X248    0
X249    0
X250    0
X251    0
X252    0
```

X253	0
X254	0
X255	0
X256	0
X257	0
X258	0
X259	0
X260	0
X261	0
X262	0
X263	0
X264	0
X265	0
X266	0
X267	0
X268	0
X269	0
X270	0
X271	0
X272	0
X273	0
X274	0
X275	0
X276	0
X277	0
X278	0
X279	0
X280	0
X281	0
X282	0
X283	0
X284	0
X285	0
X286	0
X287	0
X288	0
X289	0
X290	0
X291	0
X292	0
X293	0
X294	0
X295	0
X296	0
X297	0
X298	0
X299	0

X300	0
X301	0
X302	0
X304	0
X305	0
X306	0
X307	0
X308	0
X309	0
X310	0
X311	0
X312	0
X313	0
X314	0
X315	0
X316	0
X317	0
X318	0
X319	0
X320	0
X321	0
X322	0
X323	0
X324	0
X325	0
X326	0
X327	0
X328	0
X329	0
X330	0
X331	0
X332	0
X333	0
X334	0
X335	0
X336	0
X337	0
X338	0
X339	0
X340	0
X341	0
X342	0
X343	0
X344	0
X345	0
X346	0
X347	0

```
X348    0
X349    0
X350    0
X351    0
X352    0
X353    0
X354    0
X355    0
X356    0
X357    0
X358    0
X359    0
X360    0
X361    0
X362    0
X363    0
X364    0
X365    0
X366    0
X367    0
X368    0
X369    0
X370    0
X371    0
X372    0
X373    0
X374    0
X375    0
X376    0
X377    0
X378    0
X379    0
X380    0
X382    0
X383    0
X384    0
X385    0
dtype: int64
```

[272]:
```
df_test.isna().any()
```

[272]:
```
ID      False
X0      False
X1      False
X2      False
X3      False
X4      False
```

```
X5      False
X6      False
X8      False
X10     False
X11     False
X12     False
X13     False
X14     False
X15     False
X16     False
X17     False
X18     False
X19     False
X20     False
X21     False
X22     False
X23     False
X24     False
X26     False
X27     False
X28     False
X29     False
X30     False
X31     False
X32     False
X33     False
X34     False
X35     False
X36     False
X37     False
X38     False
X39     False
X40     False
X41     False
X42     False
X43     False
X44     False
X45     False
X46     False
X47     False
X48     False
X49     False
X50     False
X51     False
X52     False
X53     False
X54     False
```

```
X55     False
X56     False
X57     False
X58     False
X59     False
X60     False
X61     False
X62     False
X63     False
X64     False
X65     False
X66     False
X67     False
X68     False
X69     False
X70     False
X71     False
X73     False
X74     False
X75     False
X76     False
X77     False
X78     False
X79     False
X80     False
X81     False
X82     False
X83     False
X84     False
X85     False
X86     False
X87     False
X88     False
X89     False
X90     False
X91     False
X92     False
X93     False
X94     False
X95     False
X96     False
X97     False
X98     False
X99     False
X100    False
X101    False
X102    False
```

```
X103    False
X104    False
X105    False
X106    False
X107    False
X108    False
X109    False
X110    False
X111    False
X112    False
X113    False
X114    False
X115    False
X116    False
X117    False
X118    False
X119    False
X120    False
X122    False
X123    False
X124    False
X125    False
X126    False
X127    False
X128    False
X129    False
X130    False
X131    False
X132    False
X133    False
X134    False
X135    False
X136    False
X137    False
X138    False
X139    False
X140    False
X141    False
X142    False
X143    False
X144    False
X145    False
X146    False
X147    False
X148    False
X150    False
X151    False
```

```
X152    False
X153    False
X154    False
X155    False
X156    False
X157    False
X158    False
X159    False
X160    False
X161    False
X162    False
X163    False
X164    False
X165    False
X166    False
X167    False
X168    False
X169    False
X170    False
X171    False
X172    False
X173    False
X174    False
X175    False
X176    False
X177    False
X178    False
X179    False
X180    False
X181    False
X182    False
X183    False
X184    False
X185    False
X186    False
X187    False
X189    False
X190    False
X191    False
X192    False
X194    False
X195    False
X196    False
X197    False
X198    False
X199    False
X200    False
```

```
X201    False
X202    False
X203    False
X204    False
X205    False
X206    False
X207    False
X208    False
X209    False
X210    False
X211    False
X212    False
X213    False
X214    False
X215    False
X216    False
X217    False
X218    False
X219    False
X220    False
X221    False
X222    False
X223    False
X224    False
X225    False
X226    False
X227    False
X228    False
X229    False
X230    False
X231    False
X232    False
X233    False
X234    False
X235    False
X236    False
X237    False
X238    False
X239    False
X240    False
X241    False
X242    False
X243    False
X244    False
X245    False
X246    False
X247    False
```

```
X248    False
X249    False
X250    False
X251    False
X252    False
X253    False
X254    False
X255    False
X256    False
X257    False
X258    False
X259    False
X260    False
X261    False
X262    False
X263    False
X264    False
X265    False
X266    False
X267    False
X268    False
X269    False
X270    False
X271    False
X272    False
X273    False
X274    False
X275    False
X276    False
X277    False
X278    False
X279    False
X280    False
X281    False
X282    False
X283    False
X284    False
X285    False
X286    False
X287    False
X288    False
X289    False
X290    False
X291    False
X292    False
X293    False
X294    False
```

```
X295    False
X296    False
X297    False
X298    False
X299    False
X300    False
X301    False
X302    False
X304    False
X305    False
X306    False
X307    False
X308    False
X309    False
X310    False
X311    False
X312    False
X313    False
X314    False
X315    False
X316    False
X317    False
X318    False
X319    False
X320    False
X321    False
X322    False
X323    False
X324    False
X325    False
X326    False
X327    False
X328    False
X329    False
X330    False
X331    False
X332    False
X333    False
X334    False
X335    False
X336    False
X337    False
X338    False
X339    False
X340    False
X341    False
X342    False
```

```
X343    False
X344    False
X345    False
X346    False
X347    False
X348    False
X349    False
X350    False
X351    False
X352    False
X353    False
X354    False
X355    False
X356    False
X357    False
X358    False
X359    False
X360    False
X361    False
X362    False
X363    False
X364    False
X365    False
X366    False
X367    False
X368    False
X369    False
X370    False
X371    False
X372    False
X373    False
X374    False
X375    False
X376    False
X377    False
X378    False
X379    False
X380    False
X382    False
X383    False
X384    False
X385    False
dtype: bool
```

[273]:
```
def print_unique_cat_columns(df):
    df_cat = df.select_dtypes(include = 'object')
    df_cat.head()
```

```
    for col in df_cat.columns:
        print(f"column {col}")
        print("------------------------")
        print(df_cat[col].unique())
        print(df_cat[col].value_counts())
```

[274]:
```
print_unique_cat_columns(df_train)
```

```
column X0
------------------------
['k' 'az' 't' 'al' 'o' 'w' 'j' 'h' 's' 'n' 'ay' 'f' 'x' 'y' 'aj' 'ak' 'am'
 'z' 'q' 'at' 'ap' 'v' 'af' 'a' 'e' 'ai' 'd' 'aq' 'c' 'aa' 'ba' 'as' 'i'
 'r' 'b' 'ax' 'bc' 'u' 'ad' 'au' 'm' 'l' 'aw' 'ao' 'ac' 'g' 'ab']
z     360
ak    349
y     324
ay    313
t     306
x     300
o     269
f     227
n     195
w     182
j     181
az    175
aj    151
s     106
ap    103
h      75
d      73
al     67
v      36
af     35
m      34
ai     34
e      32
ba     27
at     25
a      21
ax     19
i      18
aq     18
am     18
u      17
aw     16
l      16
ad     14
```

```
b      11
k      11
au     11
r      10
as     10
bc      6
ao      4
c       3
q       2
aa      2
g       1
ac      1
ab      1
Name: X0, dtype: int64
column X1
------------------------
['v' 't' 'w' 'b' 'r' 'l' 's' 'aa' 'c' 'a' 'e' 'h' 'z' 'j' 'o' 'u' 'p' 'n'
 'i' 'y' 'd' 'f' 'm' 'k' 'g' 'q' 'ab']
aa    833
s     598
b     592
l     590
v     408
r     251
i     203
a     143
c     121
o      82
w      52
z      46
u      37
e      33
m      32
t      31
h      29
f      23
y      23
j      22
n      19
k      17
p       9
g       6
q       3
ab      3
d       3
Name: X1, dtype: int64
column X2
------------------------
```

```
['at' 'av' 'n' 'e' 'as' 'aq' 'r' 'ai' 'ak' 'm' 'a' 'k' 'ae' 's' 'f' 'd'
 'ag' 'ay' 'ac' 'ap' 'g' 'i' 'aw' 'y' 'b' 'ao' 'al' 'h' 'x' 'au' 't' 'an'
 'z' 'ah' 'p' 'am' 'j' 'q' 'af' 'l' 'aa' 'c' 'o' 'ar']
as    1659
ae     496
ai     415
m      367
ak     265
r      153
n      137
s       94
f       87
e       81
aq      63
ay      54
a       47
t       29
k       25
i       25
b       21
ao      20
ag      19
z       19
d       18
ac      13
g       12
ap      11
y       11
x       10
aw       8
at       6
h        6
al       5
q        5
an       5
ah       4
av       4
p        4
au       3
am       1
ar       1
l        1
j        1
af       1
o        1
c        1
aa       1
Name: X2, dtype: int64
```

```
column X3
------------------------
['a' 'e' 'c' 'f' 'd' 'b' 'g']
c    1942
f    1076
a     440
d     290
g     241
e     163
b      57
Name: X3, dtype: int64
column X4
------------------------
['d' 'b' 'c' 'a']
d    4205
a       2
c       1
b       1
Name: X4, dtype: int64
column X5
------------------------
['u' 'y' 'x' 'h' 'g' 'f' 'j' 'i' 'd' 'c' 'af' 'ag' 'ab' 'ac' 'ad' 'ae'
 'ah' 'l' 'k' 'n' 'm' 'p' 'q' 's' 'r' 'v' 'w' 'o' 'aa']
v     231
w     231
q     220
r     215
d     214
s     214
n     212
p     208
m     208
i     207
ae    205
ag    204
ac    200
ab    197
l     195
af    188
ad    185
k     177
c     131
j     125
aa    112
ah     97
o      20
f       7
x       2
```

```
h       1
u       1
g       1
y       1
Name: X5, dtype: int64
column X6
------------------------
['j' 'l' 'd' 'h' 'i' 'a' 'g' 'c' 'k' 'e' 'f' 'b']
g    1042
j    1039
d     625
i     488
l     478
a     206
h     190
k      43
c      38
b      28
f      20
e      12
Name: X6, dtype: int64
column X8
------------------------
['o' 'x' 'e' 'n' 's' 'a' 'h' 'p' 'm' 'k' 'd' 'i' 'v' 'j' 'b' 'q' 'w' 'g'
 'y' 'l' 'f' 'u' 'r' 't' 'c']
j    277
s    255
f    243
n    242
i    237
e    225
r    219
a    210
w    196
v    194
b    190
k    176
o    163
m    155
g    130
t    119
u    119
q    117
h    117
y    116
x    105
d    103
l    101
```

```
c    100
p    100
Name: X8, dtype: int64
```

[275]:
```
print_unique_cat_columns(df_test)
```

```
column X0
-------------------------
['az' 't' 'w' 'y' 'x' 'f' 'ap' 'o' 'ay' 'al' 'h' 'z' 'aj' 'd' 'v' 'ak'
 'ba' 'n' 'j' 's' 'af' 'ax' 'at' 'aq' 'av' 'm' 'k' 'a' 'e' 'ai' 'i' 'ag'
 'b' 'am' 'aw' 'as' 'r' 'ao' 'u' 'l' 'c' 'ad' 'au' 'bc' 'g' 'an' 'ae' 'p'
 'bb']
ak    432
y     348
z     335
x     302
ay    299
t     293
o     246
f     213
w     198
j     171
n     167
aj    162
az    161
s     116
ap    108
al     88
h      64
d      61
e      48
v      40
ai     38
af     34
m      34
am     28
i      25
at     21
u      20
ba     19
a      18
b      13
ad     12
k      12
aq     11
aw     11
r      10
ax      8
```

```
bc      6
l       6
c       6
as      6
au      5
ao      5
g       3
ag      1
an      1
p       1
ae      1
av      1
bb      1
Name: X0, dtype: int64
column X1
------------------------
['v' 'b' 'l' 's' 'aa' 'r' 'a' 'i' 'p' 'c' 'o' 'm' 'z' 'e' 'h' 'w' 'g' 'k'
 'y' 't' 'u' 'd' 'j' 'q' 'n' 'f' 'ab']
aa    826
s     602
l     599
b     596
v     436
r     252
i     189
a     153
c     142
o      81
w      50
u      40
z      31
e      29
m      27
h      27
j      22
y      21
t      18
n      16
f      12
k      12
p      10
g       9
ab      5
q       3
d       1
Name: X1, dtype: int64
column X2
------------------------
```

```
['n' 'ai' 'as' 'ae' 's' 'b' 'e' 'ak' 'm' 'a' 'aq' 'ag' 'r' 'k' 'aj' 'ay'
 'ao' 'an' 'ac' 'af' 'ax' 'h' 'i' 'f' 'ap' 'p' 'au' 't' 'z' 'y' 'aw' 'd'
 'at' 'g' 'am' 'j' 'x' 'ab' 'w' 'q' 'ah' 'ad' 'al' 'av' 'u']
as    1658
ae     478
ai     462
m      348
ak     260
r      155
n      113
s      100
f       85
e       84
ay      78
aq      72
a       44
b       38
t       25
k       25
ag      23
ac      20
ao      19
i       15
z       12
ap      11
p       10
aw       9
d        6
h        6
q        5
g        5
au       5
ad       4
af       4
ab       4
al       4
at       3
ah       3
am       3
w        3
j        2
x        2
an       1
ax       1
y        1
av       1
u        1
aj       1
```

```
Name: X2, dtype: int64
column X3
------------------------
['f' 'a' 'c' 'e' 'd' 'g' 'b']
c    1900
f    1083
a     476
d     274
g     272
e     158
b      46
Name: X3, dtype: int64
column X4
------------------------
['d' 'b' 'a' 'c']
d    4203
b       4
a       1
c       1
Name: X4, dtype: int64
column X5
------------------------
['t' 'b' 'a' 'z' 'y' 'x' 'h' 'g' 'f' 'j' 'i' 'd' 'c' 'af' 'ag' 'ab' 'ac'
 'ad' 'ae' 'ah' 'l' 'k' 'n' 'm' 'p' 'q' 's' 'r' 'v' 'w' 'o' 'aa']
v     246
r     239
p     227
w     218
af    217
ad    213
ac    212
n     209
l     206
s     205
ag    201
q     197
m     197
ae    196
k     193
d     192
i     180
ab    179
j     137
c     121
aa    105
ah     80
o      16
g       8
```

```
f       6
x       2
h       2
y       1
b       1
a       1
t       1
z       1
Name: X5, dtype: int64
column X6
------------------------
['a' 'g' 'j' 'l' 'i' 'd' 'f' 'h' 'c' 'k' 'e' 'b']
g    1073
j    1002
d     589
i     490
l     473
h     218
a     196
k      67
c      40
f      25
b      19
e      17
Name: X6, dtype: int64
column X8
------------------------
['w' 'y' 'j' 'n' 'm' 's' 'a' 'v' 'r' 'o' 't' 'h' 'c' 'k' 'p' 'u' 'd' 'g'
 'b' 'q' 'e' 'l' 'f' 'i' 'x']
e    274
j    256
s    244
f    241
n    236
i    234
r    228
a    202
w    192
v    174
b    172
o    169
m    162
k    154
u    144
g    137
h    128
t    117
q    114
```

```
x    110
d    108
p    106
y    106
c    101
l    100
Name: X8, dtype: int64
```

Check out of sample values in each category of columns of test data

[276]:
```
df_train_cat = df_train.select_dtypes(include='object')
df_test_cat = df_test.select_dtypes(include='object')

for col in df_train_cat.columns:
    out_of_sample_values = [v for v in df_test_cat[col].unique() if v not in␣
 ↪df_train_cat[col].unique()]
    if len(out_of_sample_values) > 0:
        print(df_test_cat[col].value_counts()[out_of_sample_values])
```

```
av    1
ag    1
an    1
ae    1
p     1
bb    1
Name: X0, dtype: int64
aj    1
ax    1
ab    4
w     3
ad    4
u     1
Name: X2, dtype: int64
t    1
b    1
a    1
z    1
Name: X5, dtype: int64
```

Filling out of sample values by the mode value of the column

[277]:
```
for col in df_train_cat.columns:
    out_of_sample_values = [v for v in df_test_cat[col].unique() if v not in␣
 ↪df_train_cat[col].unique()]
    if len(out_of_sample_values) > 0:
        df_test_cat.loc[df_test_cat[col].isin(out_of_sample_values), col] =␣
 ↪df_test_cat[col].mode()[0]
```

```
/usr/local/lib/python3.7/site-packages/pandas/core/indexing.py:671:
```

```
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  self._setitem_with_indexer(indexer, value)
/usr/local/lib/python3.7/site-packages/ipykernel_launcher.py:4:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  after removing the cwd from sys.path.
```

1 Apply label encoder

[278]:
```
columns = df_train_cat.columns

for col in columns:
    le = LabelEncoder()
    df_train_cat[col] = le.fit_transform(df_train_cat[col])
    df_test_cat[col] = le.transform(df_test_cat[col])
```

```
/usr/local/lib/python3.7/site-packages/ipykernel_launcher.py:5:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  """
/usr/local/lib/python3.7/site-packages/ipykernel_launcher.py:6:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
```

[279]:
```
df_train_cat.head()
```

[279]:

	X0	X1	X2	X3	X4	X5	X6	X8
0	32	23	17	0	3	24	9	14
1	32	21	19	4	3	28	11	14
2	20	24	34	2	3	27	9	23

```
3  20  21  34   5   3  27  11   4
4  20  23  34   5   3  12   3  13
```

[280]:
```
df_test_cat.head()
```

[280]:
```
   X0  X1  X2  X3  X4  X5  X6  X8
0  20  23  34   5   3  25   0  22
1  40   3   7   0   3  25   6  24
2  20  23  16   5   3  25   9   9
3  20  13  34   5   3  25  11  13
4  43  20  16   2   3  28   8  12
```

2 Verify variance of target variable with the categorical features using ANOVA

[285]:
```
temp_df = pd.concat((df_train['y'],df_train_cat), axis = 1)
```

[286]:
```
model = ols("y~C(X0)+C(X1)+C(X2)+C(X3)+C(X4)+C(X5)+C(X6)+C(X8)",data=temp_df).
 ↪fit()
anova_table = sm.stats.anova_lm(model, typ=2)
anova_table
#model.summary()
```

[286]:
```
                sum_sq      df          F        PR(>F)
C(X0)     206535.179273    46.0  66.318128  0.000000e+00
C(X1)       1651.036730    26.0   0.937949  5.539314e-01
C(X2)       4868.705236    43.0   1.672404  3.962290e-03
C(X3)        359.486951     6.0   0.884969  5.048417e-01
C(X4)        590.113729     3.0   2.905431  3.343685e-02
C(X5)       6498.750565    28.0   3.428214  3.029852e-09
C(X6)        778.815154    11.0   1.045774  4.023786e-01
C(X8)       1478.635578    24.0   0.910010  5.887769e-01
Residual  272298.979688  4022.0        NaN           NaN
```

From above analysis we consider p-values < 0.05, the independent categorical variable affects significantly the output are X0, X2, X4, X5

[287]:
```
model = ols("y~C(X0)+C(X2)+C(X4)+C(X5)",data=temp_df).fit()
anova_table = sm.stats.anova_lm(model, typ=2)
anova_table
```

[287]:
```
                sum_sq      df          F        PR(>F)
C(X0)     238351.657972    46.0  76.484348  0.000000e+00
C(X2)       5249.014700    43.0   1.801862  1.071471e-03
C(X4)        554.148092     3.0   2.726571  4.257719e-02
```

```
C(X5)       6688.684638    28.0   3.526102  1.119647e-09
Residual  277016.021071  4089.0        NaN           NaN
```

[288]:
```
columns = ["X0","X2","X4","X5"] #cat columns to be considered
```

[289]:
```
df_train = df_train.drop(df_train.select_dtypes(include='object').columns, axis␣
↪= 1)
df_test = df_test.drop(df_test.select_dtypes(include='object').columns, axis =␣
↪1)
```

[290]:
```
df_train.head()
```

[290]:

	ID	y	X10	X11	X12	X13	X14	X15	X16	X17	…	X375	X376	X377	\
0	0	130.81	0	0	0	1	0	0	0	0	…	0	0	1	
1	6	88.53	0	0	0	0	0	0	0	0	…	1	0	0	
2	7	76.26	0	0	0	0	0	0	0	1	…	0	0	0	
3	9	80.62	0	0	0	0	0	0	0	0	…	0	0	0	
4	13	78.02	0	0	0	0	0	0	0	0	…	0	0	0	

	X378	X379	X380	X382	X383	X384	X385
0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0
2	0	0	0	1	0	0	0
3	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0

[5 rows x 370 columns]

[291]:
```
df_train = pd.concat((df_train, df_train_cat[columns]), axis = 1)
df_test = pd.concat((df_test, df_test_cat[columns]), axis = 1)
```

[292]:
```
df_train.shape, df_test.shape
```

[292]:
```
((4209, 374), (4209, 373))
```

[293]:
```
df_train.head()
```

[293]:

	ID	y	X10	X11	X12	X13	X14	X15	X16	X17	…	X379	X380	X382	\
0	0	130.81	0	0	0	1	0	0	0	0	…	0	0	0	
1	6	88.53	0	0	0	0	0	0	0	0	…	0	0	0	
2	7	76.26	0	0	0	0	0	0	0	1	…	0	0	1	
3	9	80.62	0	0	0	0	0	0	0	0	…	0	0	0	
4	13	78.02	0	0	0	0	0	0	0	0	…	0	0	0	

	X383	X384	X385	X0	X2	X4	X5
0	0	0	0	32	17	3	24
1	0	0	0	32	19	3	28

```
2     0     0     0  20  34   3  27
3     0     0     0  20  34   3  27
4     0     0     0  20  34   3  12

[5 rows x 374 columns]
```

[294]:
```
X_train = df_train.drop(['ID','y'], axis = 1)
X_test = df_test.drop('ID', axis = 1)
```

[295]:
```
y_train = df_train["y"]
```

Before implementing the PCA it is required to feature scale the data, we have data in categorical variables which are having large values compared to discrete columns having 0 or 1. We will Feature scale only categorical variables after LabelEncoder is applied

[296]:
```
from sklearn.preprocessing import StandardScaler
```

[297]:
```
sc = StandardScaler()
X_train[columns] = sc.fit_transform(X_train[columns])
X_test[columns] = sc.transform(X_test[columns])
```

[298]:
```
X_train[columns].head()
```

[298]:
```
          X0        X2        X4        X5
0  0.163012 -0.028122  0.028938  1.292117
1  0.163012  0.155388  0.028938  1.776974
2 -0.710560  1.531709  0.028938  1.655760
3 -0.710560  1.531709  0.028938  1.655760
4 -0.710560  1.531709  0.028938 -0.162454
```

3 Perform Dimensionality Reduction using PCA

[299]:
```
pca = PCA()
X_pca = pca.fit(X_train)
```

[300]:
```
X_pca.explained_variance_ratio_[:25].sum()
```

[300]:
```
0.8084641770868322
```

[301]:
```
pca = PCA(n_components = 25)
X_train_pca = pca.fit_transform(X_train)
X_test_pca = pca.transform(X_test)
```

[302]:
```
X_train_pca.shape, X_test_pca.shape
```

[302]:
```
((4209, 25), (4209, 25))
```

[309]:
```
Xtrain, Xtest, ytrain, ytest =  train_test_split(X_train_pca, y_train,␣
↪test_size = 0.1, random_state = 42)
```

[310]:
```
train_dmatrix = xgb.DMatrix(data = Xtrain, label = ytrain)
test_dmatrix = xgb.DMatrix(data = Xtest, label = ytest)
```

params = {'eta': np.arange(0.1, 0.5,0.1), 'max_depth': np.arange(3,12,1), 'objective': ['reg:squarederror'], #'n_estimators': [10,20,50,80,100], 'alpha': np.arange(10,150,10), 'lambda': np.arange(10,150,10) }

[311]:
```
params = {'eta': 0.1,
          'max_depth': 3,
          'objective': 'reg:squarederror',
          'alpha': 50,
          'eval_metric': 'rmse',
          'booster': 'dart'
         }
```

[306]:
```
def xgb_r2_score(preds, dtrain):
    labels = dtrain.get_label()
    return 'r2', r2_score(labels, preds)
```

[312]:
```
watchlist = [(train_dmatrix, 'train'), (test_dmatrix, 'test')]
```

[318]:
```
xgbmodel = xgb.train(params, train_dmatrix, 120, watchlist,␣
↪early_stopping_rounds=50, feval=xgb_r2_score, maximize = True,␣
↪verbose_eval=10)
```

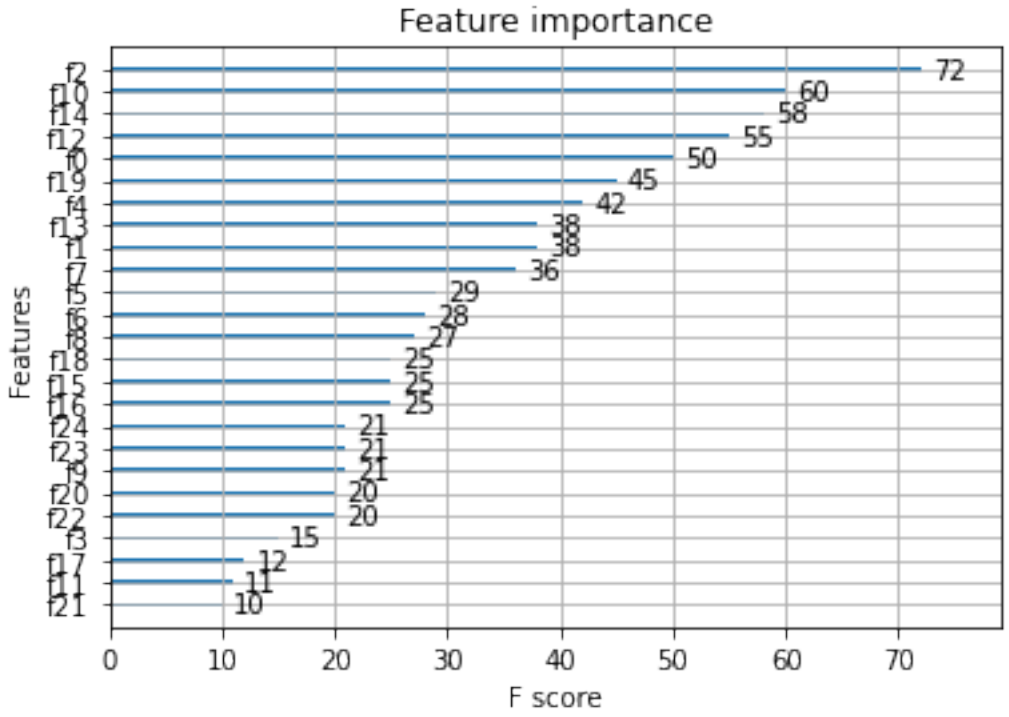
```
[0]     train-rmse:90.98556     test-rmse:91.28316      train-r2:-49.90643
test-r2:-57.03867
Multiple eval metrics have been passed: 'test-r2' will be used for early
stopping.

Will train until test-r2 hasn't improved in 50 rounds.
[10]    train-rmse:33.20081     test-rmse:33.22361      train-r2:-5.77837
test-r2:-6.68829
[20]    train-rmse:14.56063     test-rmse:14.22725      train-r2:-0.30373
test-r2:-0.40987
[30]    train-rmse:9.78567      test-rmse:9.23938       train-r2:0.41114
test-r2:0.40540
[40]    train-rmse:8.72537      test-rmse:8.22134       train-r2:0.53184
test-r2:0.52922
[50]    train-rmse:8.41406      test-rmse:7.99230       train-r2:0.56465
test-r2:0.55508
[60]    train-rmse:8.23741      test-rmse:7.89608       train-r2:0.58274
test-r2:0.56573
```

```
[70]    train-rmse:8.10653      test-rmse:7.90025       train-r2:0.59589
test-r2:0.56527
[80]    train-rmse:7.99334      test-rmse:7.88795       train-r2:0.60710
test-r2:0.56662
[90]    train-rmse:7.91286      test-rmse:7.88173       train-r2:0.61497
test-r2:0.56731
[100]   train-rmse:7.83589      test-rmse:7.86478       train-r2:0.62242
test-r2:0.56917
[110]   train-rmse:7.75303      test-rmse:7.84991       train-r2:0.63037
test-r2:0.57080
[119]   train-rmse:7.68057      test-rmse:7.85479       train-r2:0.63724
test-r2:0.57026
```

[320]:
```
xgb.plot_importance(xgbmodel)
plt.show
```

[320]:
```
<function matplotlib.pyplot.show(close=None, block=None)>
```



[321]:
```
xgbmodel.get_score(importance_type='weight')
```

[321]:
```
{'f2': 72,
 'f7': 36,
 'f0': 50,
 'f12': 55,
```

```
 'f10': 60,
 'f1': 38,
 'f4': 42,
 'f9': 21,
 'f14': 58,
 'f16': 25,
 'f23': 21,
 'f3': 15,
 'f5': 29,
 'f19': 45,
 'f15': 25,
 'f13': 38,
 'f18': 25,
 'f24': 21,
 'f22': 20,
 'f20': 20,
 'f21': 10,
 'f17': 12,
 'f8': 27,
 'f6': 28,
 'f11': 11}
```

[322]:
```
predictions = xgbmodel.predict(test_dmatrix)
np.column_stack((ytest,predictions))[:20]
```

[322]:
```
array([[ 97.94      ,  96.88946533],
       [ 96.41      ,  98.37599945],
       [105.83      , 112.74359131],
       [ 79.09      ,  80.77397919],
       [108.69      , 107.45851898],
       [ 94.6       ,  98.32183075],
       [ 84.48      ,  90.27283478],
       [110.24      , 100.25112915],
       [120.8       , 103.79337311],
       [122.66      , 113.08356476],
       [ 85.94      ,  76.84662628],
       [ 88.05      ,  92.71746063],
       [ 90.01      ,  93.67673492],
       [140.25      , 103.20804596],
       [ 98.25      ,  93.07572174],
       [101.59      ,  94.21078491],
       [105.43      , 112.19467163],
       [ 91.94      ,  95.30301666],
       [ 93.02      ,  95.28708649],
       [110.2       , 115.00834656]])
```

[323]:
```
np.sqrt(mean_squared_error(ytest, predictions))
```

[323]:
```
7.854789023991354
```

[324]:
```
r2_score(ytest, predictions)
```

[324]:
```
0.5702605198675629
```

4 Prediction for test dataset

[333]:
```
valid_dmatrix = xgb.DMatrix(data = X_test_pca)

test_df_pred = pd.DataFrame(xgbmodel.predict(valid_dmatrix).reshape(-1,1),␣
↪columns = ['y'])
```

[334]:
```
df_test = pd.concat((df_test, test_df_pred), axis=1)
df_test.head()
```

[334]:

	ID	X10	X11	X12	X13	X14	X15	X16	X17	X18	…	X380	X382	X383	\
0	1	0	0	0	0	0	0	0	0	0	…	0	0	0	
1	2	0	0	0	0	0	0	0	0	0	…	0	0	0	
2	3	0	0	0	0	1	0	0	0	0	…	0	0	0	
3	4	0	0	0	0	0	0	0	0	0	…	0	0	0	
4	5	0	0	0	0	1	0	0	0	0	…	0	0	0	

	X384	X385	X0	X2	X4	X5	y
0	0	0	20	34	3	25	76.181961
1	0	0	40	7	3	25	93.918526
2	0	0	20	16	3	25	85.020744
3	0	0	20	34	3	25	76.405983
4	0	0	43	16	3	28	109.892303

[5 rows x 374 columns]

[336]:
```
df_test.to_csv("Predicted_test.csv", index=False)
```

[ ]:
```

```