COMP-4448

Christopher Kramer

2021-03-09

University Admittance Probability LogReg

In [57]:
```
import sklearn # PCA, clustering
import pandas as pd # used as ETE data structure, EDA/profiling
import seaborn as sns # used for data import, 2d plots
import matplotlib.pyplot as plt # used for 3d plots
import matplotlib as mpl # used for 3d plots
import numpy as np # used for typing
from dataclasses import dataclass # used for profiling
from pandas_profiling import ProfileReport

import warnings

from typing import NoReturn, Optional, Union # typing
from mpl_toolkits.mplot3d import Axes3D # used for 3d plots
from sklearn.compose import make_column_transformer

from sklearn.preprocessing import OneHotEncoder
from sklearn.preprocessing import MinMaxScaler

sns.set_style('whitegrid')
```

Find another dataset that is suitable for logistic regression. Run a logistic regression on the data using the statsmodel package. Print the results and interpret the parameter coefficients for each input variable: https://www.statsmodels.org/stable/index.html. Evaluate the model as well.

Can the probability of university admission be predicted from test scores and university rating?
https://www.kaggle.com/mohansacharya/graduate-admissions

In [45]:
```
import statsmodels.api as sm
```

In [46]:
```
data = pd.read_csv('Admission_Predict_Ver1.1.csv')
```

In [47]:
```
data.columns
```

Out[47]:
```
Index(['Serial No.', 'GRE Score', 'TOEFL Score', 'University Rating', 'SOP',
       'LOR ', 'CGPA', 'Research', 'Chance of Admit '],
      dtype='object')
```

In [48]:
```
data.columns = [c.strip().lower() for c in data.columns]
```

In [49]:
```
data.columns
```

Out[49]:
```
Index(['serial no.', 'gre score', 'toefl score', 'university rating', 'sop',
       'lor', 'cgpa', 'research', 'chance of admit'],
      dtype='object')
```

In [50]:
```
data = data.drop(['serial no.'], 1)
```

In [51]:
```
X_train, X_test, y_train, y_test = train_test_split(data[data.columns[~data.columns.isin(['chance of admit'])]], data['chance of admit'], test_size=.3)
```

In [52]:
```
model = sm.Logit(endog=y_train, exog=X_train).fit()
```

```
Optimization terminated successfully.
         Current function value: 0.470192
         Iterations 5
```

In [53]:
```
print(model.summary2())
```

```
                          Results: Logit
==================================================================
Model:              Logit            Pseudo R-squared: -1.871
Dependent Variable: chance of admit  AIC:              343.1347
Date:               2021-03-12 13:49 BIC:              370.1402
No. Observations:   350              Log-Likelihood:   -164.57
Df Model:           6                LL-Null:          -57.313
Df Residuals:       343              LLR p-value:      1.0000
Converged:          1.0000           Scale:            1.0000
No. Iterations:     5.0000
------------------------------------------------------------------
                   Coef.  Std.Err.    z    P>|z|   [0.025 0.975]
------------------------------------------------------------------
gre score         -0.0257   0.0145 -1.7666 0.0773 -0.0542 0.0028
toefl score        0.0183   0.0400  0.4578 0.6471 -0.0601 0.0968
university rating  0.1113   0.1660  0.6705 0.5025 -0.2140 0.4366
sop                0.0757   0.2008  0.3768 0.7063 -0.3179 0.4692
lor                0.0774   0.1860  0.4161 0.6773 -0.2872 0.4421
cgpa               0.7133   0.4318  1.6518 0.0986 -0.1331 1.5597
research           0.3844   0.2692  1.4277 0.1534 -0.1433 0.9120
==================================================================
```

In [54]:
```
pd.DataFrame(zip(model.params, np.exp(model.params)), columns=['coef', 'exp_coef'],
             index=data[data.columns[~data.columns.isin(['chance of admit'])]].columns.to_list())
```

Out[54]:

	coef	exp_coef
gre score	-0.025699	0.974629
toefl score	0.018329	1.018498
university rating	0.111278	1.117706
sop	0.075659	1.078595
lor	0.077419	1.080495
cgpa	0.713306	2.040727
research	0.384376	1.468697

All other variables held constant, the odds of being admitted to a school increase by .97 for each increase in GRE test score. Cumulative GPA is the most impactful variable in terms of relative odds of admission increase per variable increase.

Unfortuantely, there are no statistically significant p-values amongst the variables. While cumulative GPA and GRE score come close to $\alpha=.05$ significance and are the best predictors of admission chance, no single variable is promising. To correctly interpret the pseudo R-squared value, another model would need to be fit and compared.

Finally, the LLR p-value shows that the model will almost certainly be outperformed (or at least not perform any worse than) another model which is fit.

In [ ]:
```
```