Module4_Tutorial_GAMClassifcation

October 29, 2020

[1]:
```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
from sklearn.model_selection import train_test_split
import statsmodels.formula.api as smf
import statsmodels.api as sm

from sklearn.metrics import confusion_matrix, roc_curve, auc
from pygam import LogisticGAM, LinearGAM, GAM, s, f, l
```

[2]:
```
#Credit Card Defaults, first prep data
```

[3]:
```
mydata = pd.read_csv('UCI_Credit_Card_prepped.csv', index_col=0)
```

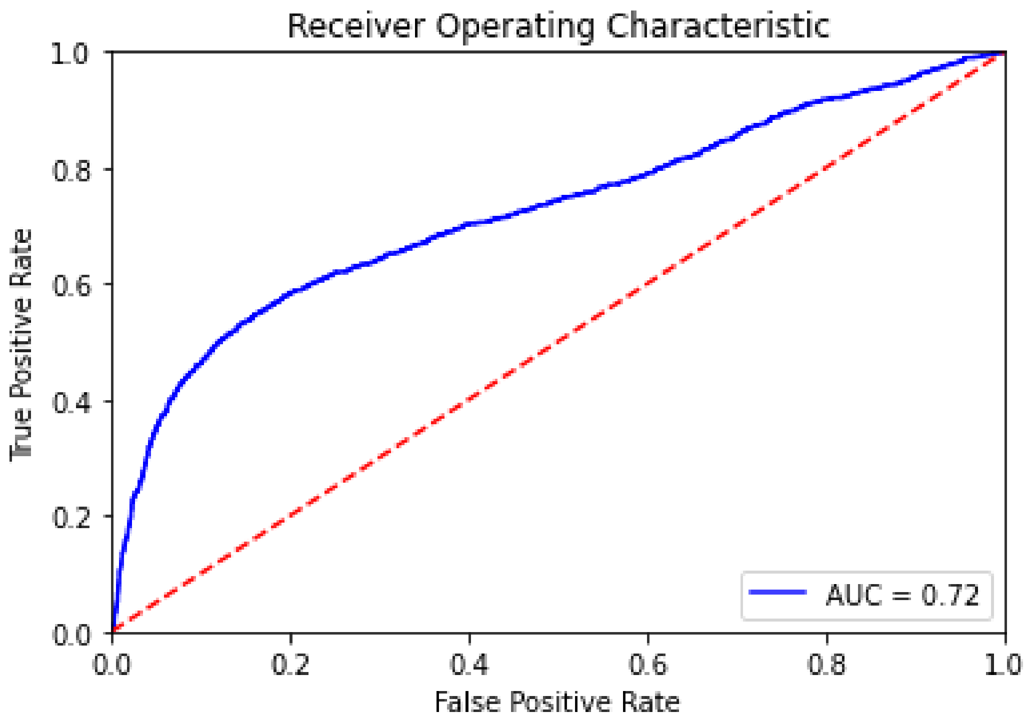
[4]:
```
train, test = train_test_split(mydata, test_size=0.25)
```

[5]:
```
#LINEAR MODEL AGAIN
g1 = smf.glm(formula='default ~ LIMIT_BAL + SEX + EDUCATION + MARRIAGE + AGE +␣
 ↪PAY_0 + PAY_2 + PAY_3 + PAY_4 + PAY_5 + PAY_6 + BILL_REC + BILL_AVG +␣
 ↪PAY_AMT1 + PAY_AMT2 + PAY_AMT3 + PAY_AMT4 + PAY_AMT5 + PAY_AMT6',␣
 ↪data=train, family=sm.families.Binomial())
g1_res = g1.fit()
g1_pred = g1_res.predict(test)
g1_pred_labels = np.zeros(len(test))
g1_pred_labels[g1_pred >0.5] = 1
confusion_matrix(test['default'], g1_pred_labels)
```

[5]:
```
array([[5680,  185],
       [1223,  412]])
```

[6]:
```
fpr, tpr, threshold = roc_curve(test['default'], g1_pred)
roc_auc = auc(fpr, tpr)
plt.title('Receiver Operating Characteristic')
plt.plot(fpr, tpr, 'b', label = 'AUC = %0.2f' % roc_auc)
plt.legend(loc = 'lower right')
plt.plot([0, 1], [0, 1],'r--')
plt.xlim([0, 1])
```

```
plt.ylim([0, 1])
plt.ylabel('True Positive Rate')
plt.xlabel('False Positive Rate')
plt.show()
```

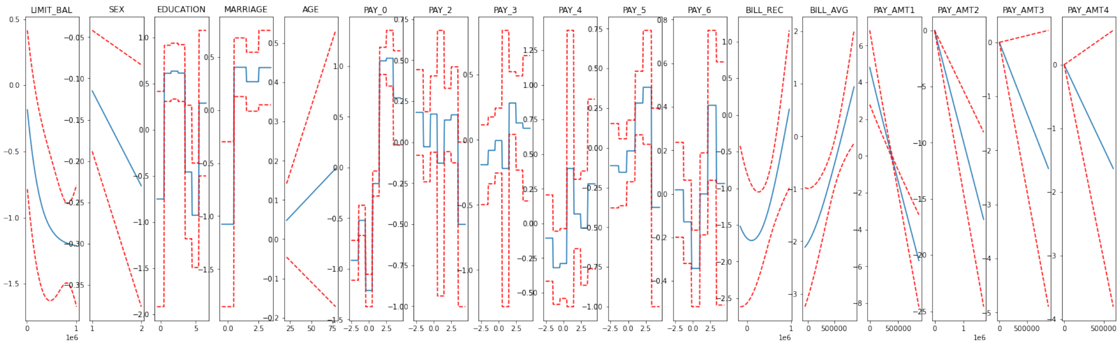


[7]:
```
X_train = train.drop(columns=['default'])
y_train = train['default']
```

[8]:
```
#GAM

gam = LogisticGAM(s(0,n_splines=5) + l(1) + f(2) + f(3) + l(4) + f(5) + f(6) +
↪f(7) + f(8) + f(9) + f(10) + s(11, n_splines=5) + s(12, n_splines=5) + s(13,
↪n_splines=5) + l(14) + l(15) + l(16)).fit(X_train, y_train)
XX = gam.generate_X_grid
plt.rcParams['figure.figsize'] = (28, 8)
fig, axs = plt.subplots(1, 17)
titles = list(X_train.columns)
for i, ax in enumerate(axs):
    XX = gam.generate_X_grid(term=i)
    ax.plot(XX[:, i], gam.partial_dependence(term=i, X=XX))
    ax.plot(XX[:, i], gam.partial_dependence(term=i, X=XX, width=.95)[1],
↪c='r', ls='--')
    ax.set_title(titles[i]);
```

```
plt.show()
```

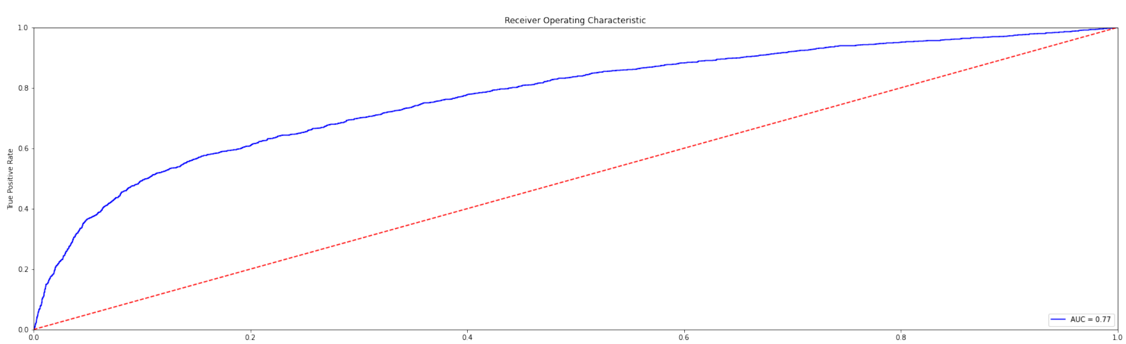


[9]:
```
X_test = test.drop(columns=['default'])
y_test = test['default']
gam_preds = gam.predict_proba(X_test)
```

[10]:
```
gam_pred_labels = np.zeros(len(test))
gam_pred_labels[gam_preds >0.5] = 1
confusion_matrix(test['default'], gam_pred_labels)
```

[10]:
```
array([[5588,  277],
       [1051,  584]])
```

[11]:
```
fpr, tpr, threshold = roc_curve(y_test, gam_preds)
roc_auc = auc(fpr, tpr)
plt.title('Receiver Operating Characteristic')
plt.plot(fpr, tpr, 'b', label = 'AUC = %0.2f' % roc_auc)
plt.legend(loc = 'lower right')
plt.plot([0, 1], [0, 1],'r--')
plt.xlim([0, 1])
plt.ylim([0, 1])
plt.ylabel('True Positive Rate')
plt.xlabel('False Positive Rate')
plt.show()
```



[ ]:

```

```