1. Credit card applications

Commercial banks receive *a lot* of applications for credit cards. Many of them get rejected for many reasons, like high loan balances, low income levels, or too many inquiries on an individual's credit report, for example. Manually analyzing these applications is mundane, error-prone, and time-consuming (and time is money!). Luckily, this task can be automated with the power of machine learning and pretty much every commercial bank does so nowadays. In this notebook, we will build an automatic credit card approval predictor using machine learning techniques, just like the real banks do.

We'll use the Credit Card Approval dataset (http://archive.ics.uci.edu/ml/datasets/credit+approval) from the UCI Machine Learning Repository. The structure of this notebook is as follows:

- First, we will start off by loading and viewing the dataset.
- We will see that the dataset has a mixture of both numerical and non-numerical features, that it contains values from different ranges, plus that it contains a number of missing entries.
- We will have to preprocess the dataset to ensure the machine learning model we choose can make good predictions.
- After our data is in good shape, we will do some exploratory data analysis to build our intuitions.
- Finally, we will build a machine learning model that can predict if an individual's application for a credit card will be accepted.

First, loading and viewing the dataset. We find that since this data is confidential, the contributor of the dataset has anonymized the feature names.

In [27]:
```python
# Import pandas
import pandas as pd

# Load dataset
cc_apps = pd.read_csv("datasets/cc_approvals.data", header=None)

# Inspect data
cc_apps.head()
```

Out[27]:

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	b	30.83	0.000	u	g	w	v	1.25	t	t	1	f	g	00202	0	+
1	a	58.67	4.460	u	g	q	h	3.04	t	t	6	f	g	00043	560	+
2	a	24.50	0.500	u	g	q	h	1.50	t	f	0	f	g	00280	824	+
3	b	27.83	1.540	u	g	w	v	3.75	t	t	5	t	g	00100	3	+
4	b	20.17	5.625	u	g	w	v	1.71	t	f	0	f	s	00120	0	+

In [28]:
```
%%nose
import pandas as pd

def test_cc_apps_exists():
    assert "cc_apps" in globals(), \
        "The variable cc_apps should be defined."

def test_cc_apps_correctly_loaded():
    correct_cc_apps = pd.read_csv("datasets/cc_approvals.data", header=None)
    try:
        pd.testing.assert_frame_equal(cc_apps, correct_cc_apps)
    except AssertionError:
        assert False, "The variable cc_apps should contain the data as present in datasets/cc_approvals.data."
```

Out[28]: 2/2 tests passed

2. Inspecting the applications

The output may appear a bit confusing at its first sight, but let's try to figure out the most important features of a credit card application. The features of this dataset have been anonymized to protect the privacy, but this blog (http://rstudio-pubs-static.s3.amazonaws.com/73039_9946de135c0a49daa7a0a9eda4a67a72.html) gives us a pretty good overview of the probable features. The probable features in a typical credit card application are Gender , Age , Debt , Married , BankCustomer , EducationLevel , Ethnicity , YearsEmployed , PriorDefault , Employed , CreditScore , DriversLicense , Citizen , ZipCode , Income and finally the ApprovalStatus . This gives us a pretty good starting point, and we can map these features with respect to the columns in the output.

As we can see from our first glance at the data, the dataset has a mixture of numerical and non-numerical features. This can be fixed with some preprocessing, but before we do that, let's learn about the dataset a bit more to see if there are other dataset issues that need to be fixed.

```python
# Print summary statistics
cc_apps_description = cc_apps.describe()
print(cc_apps_description)

print('\n')

# Print DataFrame information
cc_apps_info = cc_apps.info()
print(cc_apps_info)

print('\n')

# Inspect missing values in the dataset
cc_apps.tail(17)
```

```
               2           7         10            14
count  690.000000  690.000000  690.00000    690.000000
mean     4.758725    2.223406    2.40000   1017.385507
std      4.978163    3.346513    4.86294   5210.102598
min      0.000000    0.000000    0.00000      0.000000
25%      1.000000    0.165000    0.00000      0.000000
50%      2.750000    1.000000    0.00000      5.000000
75%      7.207500    2.625000    3.00000    395.500000
max     28.000000   28.500000   67.00000  100000.000000


<class 'pandas.core.frame.DataFrame'>
RangeIndex: 690 entries, 0 to 689
Data columns (total 16 columns):
0    690 non-null object
1    690 non-null object
2    690 non-null float64
3    690 non-null object
4    690 non-null object
5    690 non-null object
6    690 non-null object
7    690 non-null float64
8    690 non-null object
9    690 non-null object
10   690 non-null int64
11   690 non-null object
12   690 non-null object
13   690 non-null object
14   690 non-null int64
15   690 non-null object
dtypes: float64(2), int64(2), object(12)
memory usage: 86.3+ KB
None
```

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
673	?	29.50	2.000	y	p	e	h	2.000	f	f	0	f	g	00256	17	-
674	a	37.33	2.500	u	g	i	h	0.210	f	f	0	f	g	00260	246	-
675	a	41.58	1.040	u	g	aa	v	0.665	f	f	0	f	g	00240	237	-
676	a	30.58	10.665	u	g	q	h	0.085	f	t	12	t	g	00129	3	-
677	b	19.42	7.250	u	g	m	v	0.040	f	t	1	f	g	00100	1	-
678	a	17.92	10.210	u	g	ff	ff	0.000	f	f	0	f	g	00000	50	-
679	a	20.08	1.250	u	g	c	v	0.000	f	f	0	f	g	00000	0	-
680	b	19.50	0.290	u	g	k	v	0.290	f	f	0	f	g	00280	364	-
681	b	27.83	1.000	y	p	d	h	3.000	f	f	0	f	g	00176	537	-
682	b	17.08	3.290	u	g	i	v	0.335	f	f	0	t	g	00140	2	-
683	b	36.42	0.750	y	p	d	v	0.585	f	f	0	f	g	00240	3	-
684	b	40.58	3.290	u	g	m	v	3.500	f	f	0	t	s	00400	0	-
685	b	21.08	10.085	y	p	e	h	1.250	f	f	0	f	g	00260	0	-
686	a	22.67	0.750	u	g	c	v	2.000	f	t	2	t	g	00200	394	-
687	a	25.25	13.500	y	p	ff	ff	2.000	f	t	1	t	g	00200	1	-
688	b	17.92	0.205	u	g	aa	v	0.040	f	f	0	f	g	00280	750	-
689	b	35.00	3.375	u	g	c	h	8.290	f	f	0	t	g	00000	0	-

```
In [30]:    %%nose

            def test_cc_apps_description_exists():
                assert "cc_apps_description" in globals(), \
                    "The variable cc_apps_description should be defined."

            def test_cc_apps_description_correctly_done():
                correct_cc_apps_description = cc_apps.describe()
                assert str(correct_cc_apps_description) == str(cc_apps_description), \
                    "cc_apps_description should contain the output of cc_apps.describe()."

            def test_cc_apps_info_exists():
                assert "cc_apps_info" in globals(), \
                    "The variable cc_apps_info should be defined."

            def test_cc_apps_info_correctly_done():
                correct_cc_apps_info = cc_apps.info()
                assert str(correct_cc_apps_info) == str(cc_apps_info), \
                    "cc_apps_info should contain the output of cc_apps.info()."
```

Out[30]:   4/4 tests passed

3. Splitting the dataset into train and test sets

Now, we will split our data into train set and test set to prepare our data for two different phases of machine learning modeling: training and testing. Ideally, no information from the test data should be used to preprocess the training data or should be used to direct the training process of a machine learning model. Hence, we first split the data and then preprocess it.

Also, features like DriversLicense and ZipCode are not as important as the other features in the dataset for predicting credit card approvals. To get a better sense, we can measure their [statistical correlation (https://realpython.com/numpy-scipy-pandas-correlation-python/)](https://realpython.com/numpy-scipy-pandas-correlation-python/) to the labels of the dataset. But this is out of scope for this project. We should drop them to design our machine learning model with the best set of features. In Data Science literature, this is often referred to as *feature selection*.

```
In [31]:    # Import train_test_split
            from sklearn.model_selection import train_test_split

            # Drop the features 11 and 13
            cc_apps = cc_apps.drop([11, 13], axis=1)

            # Split into train and test sets
            cc_apps_train, cc_apps_test = train_test_split(cc_apps, test_size=0.33, random_state=42)
```

```
In [32]:    %%nose


            def test_columns_dropped_correctly():
                assert cc_apps.shape == (
                    690,
                    14,
                ), "The shape of the DataFrame isn't correct. Did you drop the two columns?"


            def test_data_split_correctly():
                cc_apps_train_correct, cc_apps_test_correct = train_test_split(
                    cc_apps, test_size=0.33, random_state=42
                )
                assert cc_apps_train_correct.equals(cc_apps_train) and cc_apps_test_correct.equals(
                    cc_apps_test
                ), "It doesn't appear that the data splitting was done correctly."
```

Out[32]:   2/2 tests passed

4. Handling the missing values (part i)

Now we've split our data, we can handle some of the issues we identified when inspecting the DataFrame, including:

- Our dataset contains both numeric and non-numeric data (specifically data that are of float64 , int64 and object types). Specifically, the features 2, 7, 10 and 14 contain numeric values (of types float64, float64, int64 and int64 respectively) and all the other features contain non-numeric values.
- The dataset also contains values from several ranges. Some features have a value range of 0 - 28, some have a range of 2 - 67, and some have a range of 1017 - 100000. Apart from these, we can get useful statistical information (like mean , max , and min) about the features that have numerical values.
- Finally, the dataset has missing values, which we'll take care of in this task. The missing values in the dataset are labeled with '?', which can be seen in the last cell's output of the second task.

Now, let's temporarily replace these missing value question marks with NaN.

```
In [33]:   # Import numpy
           import numpy as np

           # Replace the '?'s with NaN in the train and test sets
           cc_apps_train = cc_apps_train.replace('?', np.NaN)
           cc_apps_test = cc_apps_test.replace('?', np.NaN)
```

```
In [34]:   %%nose

           # def test_cc_apps_assigned():
           #     assert "cc_apps" in globals(), \
           #         "After the NaN replacement, it should be assigned to the same variable cc_apps only."


           def test_cc_apps_correctly_replaced():
               cc_apps_fresh = pd.read_csv("datasets/cc_approvals.data", header=None)
               cc_apps_train_correct, cc_apps_test_correct = train_test_split(
                   cc_apps_fresh, test_size=0.33, random_state=42
               )

               correct_cc_apps_replacement_correct_train = cc_apps_train_correct.replace(
                   "?", np.NaN
               )
               correct_cc_apps_replacement_correct_test = cc_apps_test_correct.replace("?", np.NaN)
               string_cc_apps_replacement_train = cc_apps_train_correct.replace("?", "NaN")
               string_cc_apps_replacement_test = cc_apps_test_correct.replace("?", "NaN")
               #     assert cc_apps.to_string() == correct_cc_apps_replacement.to_string(), \
               #         "The code that replaces question marks with NaNs doesn't appear to be correct."
               try:
                   pd.testing.assert_frame_equal(
                       correct_cc_apps_replacement_correct_train, cc_apps_train
                   )

                   pd.testing.assert_frame_equal(
                       correct_cc_apps_replacement_correct_test, cc_apps_test
                   )
               except AssertionError:
                   if string_cc_apps_replacement_train.equals(
                       cc_apps_train
                   ) or string_cc_apps_replacement_test.equals(cc_apps_test):
                       assert (
                           False
                       ), 'It looks like the question marks were replaced by the string "NaN". Missing values should be represented by `np.nan`.'
```

Out[34]:   1/1 tests passed

5. Handling the missing values (part ii)

We replaced all the question marks with NaNs. This is going to help us in the next missing value treatment that we are going to perform.

An important question that gets raised here is *why are we giving so much importance to missing values*? Can't they be just ignored? Ignoring missing values can affect the performance of a machine learning model heavily. While ignoring the missing values our machine learning model may miss out on information about the dataset that may be useful for its training. Then, there are many models which cannot handle missing values implicitly such as Linear Discriminant Analysis (LDA).

So, to avoid this problem, we are going to impute the missing values with a strategy called mean imputation.

```
In [35]:  # Impute the missing values with mean imputation
          cc_apps_train.fillna(cc_apps_train.mean(), inplace=True)
          cc_apps_test.fillna(cc_apps_train.mean(), inplace=True)

          # Count the number of NaNs in the datasets and print the counts to verify
          print(cc_apps_train.isnull().sum())
          print(cc_apps_test.isnull().sum())
```

```
0    8
1    5
2    0
3    6
4    6
5    7
6    7
7    0
8    0
9    0
10   0
12   0
14   0
15   0
dtype: int64
0    4
1    7
2    0
3    0
4    0
5    2
6    2
7    0
8    0
9    0
10   0
12   0
14   0
15   0
dtype: int64
```

```
In [36]:  %%nose


          def test_cc_apps_correctly_imputed():
              assert (
                  cc_apps_train.isnull().to_numpy().sum() == 39
                  and cc_apps_test.isnull().to_numpy().sum() == 15
              ), "There should be 39 and 15 null values in the training and test sets after your code is run, but there aren't."
```

Out[36]: 1/1 tests passed

6. Handling the missing values (part iii)

We have successfully taken care of the missing values present in the numeric columns. There are still some missing values to be imputed for columns 0, 1, 3, 4, 5, 6 and 13. All of these columns contain non-numeric data and this is why the mean imputation strategy would not work here. This needs a different treatment.

We are going to impute these missing values with the most frequent values as present in the respective columns. This is good practice (https://www.datacamp.com/community/tutorials/categorical-data) when it comes to imputing missing values for categorical data in general.

```
In [37]:   # Iterate over each column of cc_apps_train
           for col in cc_apps_train.columns:
               # Check if the column is of object type
               if cc_apps_train[col].dtypes == 'object':
                   # Impute with the most frequent value
                   cc_apps_train = cc_apps_train.fillna(cc_apps_train[col].value_counts().index[0])
                   cc_apps_test = cc_apps_test.fillna(cc_apps_train[col].value_counts().index[0])

           # Count the number of NaNs in the dataset and print the counts to verify
           print(cc_apps_train.isnull().sum())
           print(cc_apps_test.isnull().sum())
```

```
0     0
1     0
2     0
3     0
4     0
5     0
6     0
7     0
8     0
9     0
10    0
12    0
14    0
15    0
dtype: int64
0     0
1     0
2     0
3     0
4     0
5     0
6     0
7     0
8     0
9     0
10    0
12    0
14    0
15    0
dtype: int64
```

```
In [38]:   %%nose


           def test_cc_apps_correctly_imputed():
               assert (
                   cc_apps_train.isnull().to_numpy().sum() == 0
                   and cc_apps_test.isnull().to_numpy().sum() == 0
               ), "There should be 0 null values after your code is run, but there isn't."
```

Out[38]:   1/1 tests passed

7. Preprocessing the data (part i)

The missing values are now successfully handled.

There is still some minor but essential data preprocessing needed before we proceed towards building our machine learning model. We are going to divide these remaining preprocessing steps into two main tasks:

1. Convert the non-numeric data into numeric.
2. Scale the feature values to a uniform range.

First, we will be converting all the non-numeric values into numeric ones. We do this because not only it results in a faster computation but also many machine learning models (like XGBoost) (and especially the ones developed using scikit-learn) require the data to be in a strictly numeric format. We will do this by using the `get_dummies()` method from pandas.

```
In [39]:   # Convert the categorical features in the train and test sets independently
           cc_apps_train = pd.get_dummies(cc_apps_train)
           cc_apps_test = pd.get_dummies(cc_apps_test)

           # Reindex the columns of the test set aligning with the train set
           cc_apps_test = cc_apps_test.reindex(columns=cc_apps_train.columns, fill_value=0)
```

```
In [40]:   %%nose


           def test_label_encoding_done_correctly():
               for col in cc_apps_train.columns:
                   if (
                       np.issubdtype(cc_apps_train[col].dtype, np.number) != True
                       and np.issubdtype(cc_apps_test[col].dtype, np.number) != True
                   ):
                       assert "It doesn't appear that all of the non-numeric columns were converted to numeric using fit_transform and transform."
```

Out[40]:   1/1 tests passed

8. Preprocessing the data (part ii)

Now, we are only left with one final preprocessing step of scaling before we can fit a machine learning model to the data.

Now, let's try to understand what these scaled values mean in the real world. Let's use $CreditScore$ as an example. The credit score of a person is their creditworthiness based on their credit history. The higher this number, the more financially trustworthy a person is considered to be. So, a $CreditScore$ of 1 is the highest since we're rescaling all the values to the range of 0-1.

```
In [41]:   # Import MinMaxScaler
           from sklearn.preprocessing import MinMaxScaler

           # Segregate features and labels into separate variables
           X_train, y_train = cc_apps_train.iloc[:, :-1].values, cc_apps_train.iloc[:, [-1]].values
           X_test, y_test = cc_apps_test.iloc[:, :-1].values, cc_apps_test.iloc[:, [-1]].values

           # Instantiate MinMaxScaler and use it to rescale X_train and X_test
           scaler = MinMaxScaler(feature_range=(0, 1))
           rescaledX_train = scaler.fit_transform(X_train)
           rescaledX_test = scaler.transform(X_test)
```

```
In [42]:   %%nose


           def test_training_range_set_correctly():
               min_value_in_rescaledX_train = np.amin(rescaledX_train)
               max_value_in_rescaledX_train = np.amax(rescaledX_train)
               assert (
                   min_value_in_rescaledX_train == 0.0 and max_value_in_rescaledX_train == 1.0
               ), "Did you correctly fit and transform the `X_train` data?"


           def test_xtest_created():
               assert (
                   "rescaledX_test" in globals()
               ), "Did you correctly use the fitted `scaler` to transform the `X_test` data?"
```

Out[42]:   2/2 tests passed

9. Fitting a logistic regression model to the train set

Essentially, predicting if a credit card application will be approved or not is a classification (https://en.wikipedia.org/wiki/Statistical_classification) task. According to UCI, our dataset contains more instances that correspond to "Denied" status than instances corresponding to "Approved" status. Specifically, out of 690 instances, there are 383 (55.5%) applications that got denied and 307 (44.5%) applications that got approved.

This gives us a benchmark. A good machine learning model should be able to accurately predict the status of the applications with respect to these statistics.

Which model should we pick? A question to ask is: *are the features that affect the credit card approval decision process correlated with each other?* Although we can measure correlation, that is outside the scope of this notebook, so we'll rely on our intuition that they indeed are correlated for now. Because of this correlation, we'll take advantage of the fact that generalized linear models perform well in these cases. Let's start our machine learning modeling with a Logistic Regression model (a generalized linear model).

```
In [43]:   # Import LogisticRegression
           from sklearn.linear_model import LogisticRegression

           # Instantiate a LogisticRegression classifier with default parameter values
           logreg = LogisticRegression()

           # Fit logreg to the train set
           logreg.fit(rescaledX_train,y_train)
```

Out[43]:   LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
                   intercept_scaling=1, max_iter=100, multi_class='ovr', n_jobs=1,
                   penalty='l2', random_state=None, solver='liblinear', tol=0.0001,
                   verbose=0, warm_start=False)
```

```
In [44]:    %%nose


            def test_logreg_defined():
                assert (
                    "logreg" in globals()
                ), "Did you instantiate LogisticRegression in the logreg variable?"


            def test_logreg_defined_correctly():
                logreg_correct = LogisticRegression()
                assert str(logreg_correct) == str(
                    logreg
                ), "The logreg variable should be defined with LogisticRegression() only."
```

Out[44]:   2/2 tests passed

10. Making predictions and evaluating performance

But how well does our model perform?

We will now evaluate our model on the test set with respect to classification accuracy (https://developers.google.com/machine-learning/crash-course/classification/accuracy). But we will also take a look the model's confusion matrix (http://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/). In the case of predicting credit card applications, it is important to see if our machine learning model is equally capable of predicting approved and denied status, in line with the frequency of these labels in our original dataset. If our model is not performing well in this aspect, then it might end up approving the application that should have been approved. The confusion matrix helps us to view our model's performance from these aspects.

```
In [45]:    # Import confusion_matrix
            from sklearn.metrics import confusion_matrix

            # Use logreg to predict instances from the test set and store it
            y_pred = logreg.predict(rescaledX_test)

            # Get the accuracy score of logreg model and print it
            print("Accuracy of logistic regression classifier: ", logreg.score(rescaledX_test,y_test))

            # Print the confusion matrix of the logreg model
            confusion_matrix(y_test,y_pred)
```

Accuracy of logistic regression classifier: 1.0

Out[45]:   array([[103,   0],
                  [0, 125]])

```
In [46]:    %%nose

            def test_ypred_defined():
                assert "y_pred" in globals(),\
                    "The variable y_pred should be defined."

            def test_ypred_defined_correctly():
                correct_y_pred = logreg.predict(rescaledX_test)
                assert str(correct_y_pred) == str(y_pred),\
                    "The y_pred variable should contain the predictions as made by LogisticRegression on rescaledX_test."
```

Out[46]:   2/2 tests passed

11. Grid searching and making the model perform better

Our model was pretty good! In fact it was able to yield an accuracy score of 100%.

For the confusion matrix, the first element of the of the first row of the confusion matrix denotes the true negatives meaning the number of negative instances (denied applications) predicted by the model correctly. And the last element of the second row of the confusion matrix denotes the true positives meaning the number of positive instances (approved applications) predicted by the model correctly.

But if we hadn't got a perfect score what's to be done?. We can perform a grid search (https://machinelearningmastery.com/how-to-tune-algorithm-parameters-with-scikit-learn/) of the model parameters to improve the model's ability to predict credit card approvals.

scikit-learn's implementation of logistic regression (http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html) consists of different hyperparameters but we will grid search over the following two:

- tol
- max_iter

```
In [47]:   # Import GridSearchCV
           from sklearn.model_selection import GridSearchCV

           # Define the grid of values for tol and max_iter
           tol = [0.01, 0.001 ,0.0001]
           max_iter = [100, 150, 200]

           # Create a dictionary where tol and max_iter are keys and the lists of their values are the corresponding values
           param_grid = dict(tol=tol, max_iter=max_iter)
```

```
In [48]:   %%nose


           def test_tol_defined():
               assert "tol" in globals(), "The variable tol should be defined."


           def test_max_iter_defined():
               assert "max_iter" in globals(), "The variable max_iter should be defined."


           def test_tol_defined_correctly():
               correct_tol = [0.01, 0.001, 0.0001]
               assert (
                   correct_tol == tol
               ), "It looks like the tol variable is not defined with the list of correct values."


           def test_max_iter_defined_correctly():
               correct_max_iter = [100, 150, 200]
               assert (
                   correct_max_iter == max_iter
               ), "It looks like the max_iter variable is not defined with a list of correct values."


           def test_param_grid_defined():
               assert "param_grid" in globals(), "The variable param_grid should be defined."


           def test_param_grid_defined_correctly():
               correct_param_grid = dict(tol=tol, max_iter=max_iter)
               assert str(correct_param_grid) == str(
                   param_grid
               ), "It looks like the param_grid variable is not defined properly."
```

Out[48]:   6/6 tests passed

12. Finding the best performing model

We have defined the grid of hyperparameter values and converted them into a single dictionary format which `GridSearchCV()` expects as one of its parameters. Now, we will begin the grid search to see which values perform best.

We will instantiate `GridSearchCV()` with our earlier `logreg` model with all the data we have. We will also instruct `GridSearchCV()` to perform a cross-validation (https://www.dataschool.io/machine-learning-with-scikit-learn/) of five folds.

We'll end the notebook by storing the best-achieved score and the respective best parameters.

While building this credit card predictor, we tackled some of the most widely-known preprocessing steps such as **scaling**, **label encoding**, and **missing value imputation**. We finished with some **machine learning** to predict if a person's application for a credit card would get approved or not given some information about that person.

```
In [49]:   # Instantiate GridSearchCV with the required parameters
           grid_model = GridSearchCV(estimator=logreg, param_grid=param_grid, cv=5)

           # Fit grid_model to the data
           grid_model_result = grid_model.fit(rescaledX_train, y_train)

           # Summarize results
           best_score, best_params = grid_model_result.best_score_, grid_model_result.best_params_
           print("Best: %f using %s" % (best_score, best_params))

           # Extract the best model and evaluate it on the test set
           best_model = grid_model_result.best_estimator_
           print("Accuracy of logistic regression classifier: ", best_model.score(rescaledX_test,y_test))
```

Best: 1.000000 using {'max_iter': 100, 'tol': 0.01}
Accuracy of logistic regression classifier: 1.0

```
In [50]:  %%nose

          correct_grid_model = GridSearchCV(estimator=logreg, param_grid=param_grid, cv=5)
          correct_grid_model_result = correct_grid_model.fit(rescaledX_train, y_train)


          def test_grid_model_defined():
              assert "grid_model" in globals(), "The variable grid_model should be defined."


          def test_grid_model_defined_correctly():
              # correct_grid_model = GridSearchCV(estimator=logreg, param_grid=param_grid, cv=5)
              assert str(correct_grid_model) == str(
                  grid_model
              ), "It doesn't appear that `grid_model` was defined correctly."


          def test_grid_model_results_defined():
              assert (
                  "grid_model_result" in globals()
              ), "The variable `grid_model_result` should be defined."


          def test_grid_model_result_defined_correctly():
              #    correct_grid_model = GridSearchCV(estimator=logreg, param_grid=param_grid, cv=5)
              #    correct_grid_model_result = correct_grid_model.fit(rescaledX, y)
              assert str(correct_grid_model_result) == str(
                  grid_model_result
              ), "It doesn't appear that `grid_model_result` was defined correctly."


          def test_best_score_defined_correctly():
              #    correct_grid_model = GridSearchCV(estimator=logreg, param_grid=param_grid, cv=5)
              #    correct_grid_model_result = correct_grid_model.fit(rescaledX, y)
              correct_best_score = correct_grid_model_result.best_score_
              assert (
                  correct_best_score == best_score
              ), "It looks like the variable `best_score` is not defined correctly."


          def test_best_params_defined_correctly():
              #    correct_grid_model = GridSearchCV(estimator=logreg, param_grid=param_grid, cv=5)
              #    correct_grid_model_result = correct_grid_model.fit(rescaledX, y)
              correct_best_params = correct_grid_model_result.best_params_
              assert (
                  correct_best_params == best_params
              ), "It looks like the variable `best_params` is not defined correctly."


          def test_best_model_defined_correctly():
              correct_best_model = correct_grid_model_result.best_estimator_
              assert (
                  str(correct_best_model) == str(best_model)
              ), "It looks like the variable `best_model` is not defined correctly."
```

Out[50]: 7/7 tests passed