ML0101EN-Proj-Loan-py-v1

January 25, 2023

Classification with Python

In this notebook we try to practice all the classification algorithms that we have learned in this course.

We load a dataset using Pandas library, and apply the following algorithms, and find the best one for this specific dataset by accuracy evaluation methods.

Let's first load required libraries:

[3]:
```
import itertools
import numpy as np
import matplotlib.pyplot as plt
from matplotlib.ticker import NullFormatter
import pandas as pd
import numpy as np
import matplotlib.ticker as ticker
from sklearn import preprocessing
%matplotlib inline
```

[4]:
```
#notice: Disable all warnings
import warnings
warnings.filterwarnings('ignore')
```

0.0.1 About dataset

This dataset is about past loans. The **Loan_train.csv** data set includes details of 346 customers whose loan are already paid off or defaulted. It includes following fields:

Field	Description
Loan_status	Whether a loan is paid off on in collection
Principal	Basic principal loan amount at the
Terms	Origination terms which can be weekly (7 days), biweekly, and monthly payoff schedule
Effective_date	When the loan got originated and took effects
Due_date	Since it's one-time payoff schedule, each loan has one single due date
Age	Age of applicant
Education	Education of applicant
Gender	The gender of applicant

Let's download the dataset

[5]:
```
!wget -O loan_train.csv https://cf-courses-data.s3.us.cloud-object-storage.
↪appdomain.cloud/IBMDeveloperSkillsNetwork-ML0101EN-SkillsNetwork/labs/
↪FinalModule_Coursera/data/loan_train.csv
```

```
--2023-01-25 05:46:57--  https://cf-courses-data.s3.us.cloud-object-
storage.appdomain.cloud/IBMDeveloperSkillsNetwork-ML0101EN-
SkillsNetwork/labs/FinalModule_Coursera/data/loan_train.csv
Resolving cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud (cf-
courses-data.s3.us.cloud-object-storage.appdomain.cloud)… 169.63.118.104
Connecting to cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud (cf-
courses-data.s3.us.cloud-object-storage.appdomain.cloud)|169.63.118.104|:443…
connected.
HTTP request sent, awaiting response… 200 OK
Length: 23101 (23K) [text/csv]
Saving to: 'loan_train.csv'

loan_train.csv      100%[===================>]  22.56K  --.-KB/s    in 0.001s

2023-01-25 05:46:57 (26.4 MB/s) - 'loan_train.csv' saved [23101/23101]
```

0.0.2 Load Data From CSV File

[6]:
```
df = pd.read_csv('loan_train.csv')
df.head()
```

[6]:
```
   Unnamed: 0  Unnamed: 0.1 loan_status  Principal  terms effective_date  \
0           0             0     PAIDOFF       1000     30       9/8/2016
1           2             2     PAIDOFF       1000     30       9/8/2016
2           3             3     PAIDOFF       1000     15       9/8/2016
3           4             4     PAIDOFF       1000     30       9/9/2016
4           6             6     PAIDOFF       1000     30       9/9/2016

    due_date  age             education  Gender
0  10/7/2016   45  High School or Below    male
1  10/7/2016   33              Bechalor  female
2  9/22/2016   27               college    male
3  10/8/2016   28               college  female
4  10/8/2016   29               college    male
```

[7]:
```
df.shape
```

[7]: (346, 10)

0.0.3 Convert to date time object

[8]:
```
df['due_date'] = pd.to_datetime(df['due_date'])
df['effective_date'] = pd.to_datetime(df['effective_date'])
df.head()
```

[8]:
```
   Unnamed: 0  Unnamed: 0.1 loan_status  Principal  terms effective_date  \
0           0             0     PAIDOFF       1000     30     2016-09-08
1           2             2     PAIDOFF       1000     30     2016-09-08
2           3             3     PAIDOFF       1000     15     2016-09-08
3           4             4     PAIDOFF       1000     30     2016-09-09
4           6             6     PAIDOFF       1000     30     2016-09-09

    due_date  age             education  Gender
0 2016-10-07   45  High School or Below    male
1 2016-10-07   33              Bechalor  female
2 2016-09-22   27               college    male
3 2016-10-08   28               college  female
4 2016-10-08   29               college    male
```

1 Data visualization and pre-processing

Let's see how many of each class is in our data set

[9]:
```
df['loan_status'].value_counts()
```

[9]:
```
PAIDOFF       260
COLLECTION     86
Name: loan_status, dtype: int64
```

260 people have paid off the loan on time while 86 have gone into collection

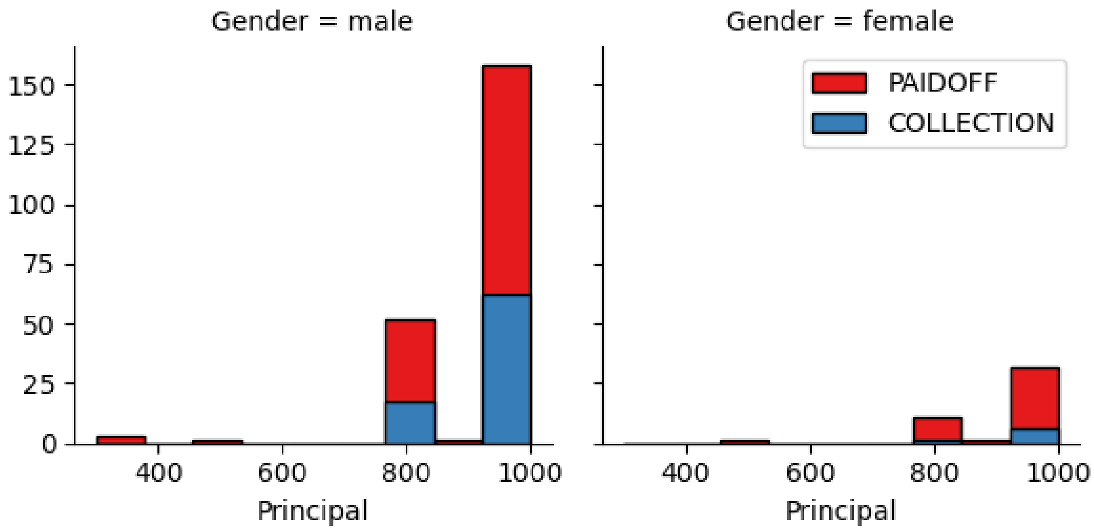
Let's plot some columns to underestand data better:

[10]:
```
# notice: installing seaborn might takes a few minutes
!pip install seaborn
```

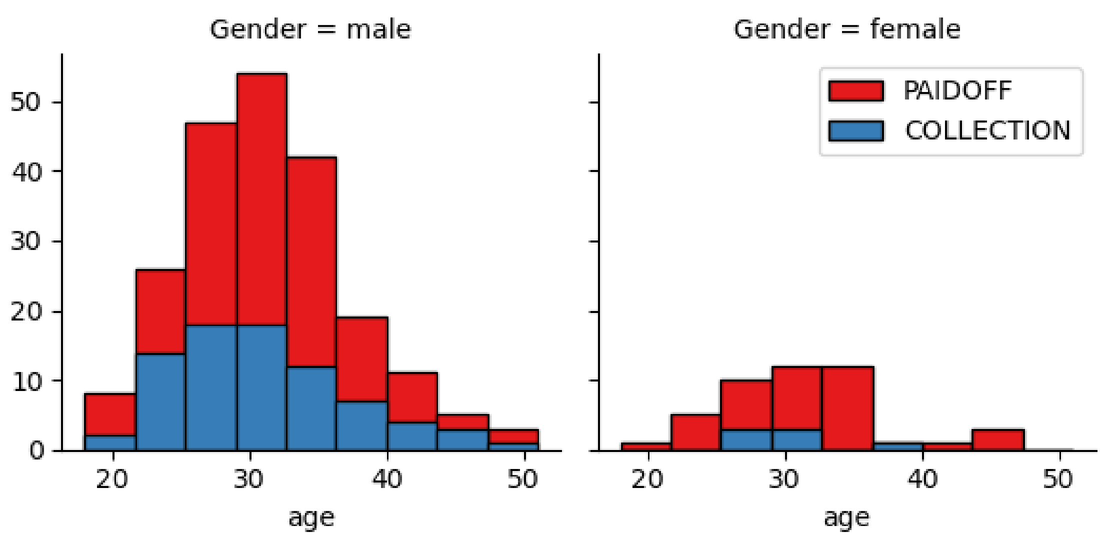
```
Requirement already satisfied: seaborn in
/home/jupyterlab/conda/envs/python/lib/python3.7/site-packages (0.9.0)
Requirement already satisfied: scipy>=0.14.0 in
/home/jupyterlab/conda/envs/python/lib/python3.7/site-packages (from seaborn)
(1.7.3)
Requirement already satisfied: pandas>=0.15.2 in
/home/jupyterlab/conda/envs/python/lib/python3.7/site-packages (from seaborn)
(1.3.5)
Requirement already satisfied: matplotlib>=1.4.3 in
/home/jupyterlab/conda/envs/python/lib/python3.7/site-packages (from seaborn)
(3.5.3)
```

```
Requirement already satisfied: numpy>=1.9.3 in
/home/jupyterlab/conda/envs/python/lib/python3.7/site-packages (from seaborn)
(1.21.6)
Requirement already satisfied: python-dateutil>=2.7 in
/home/jupyterlab/conda/envs/python/lib/python3.7/site-packages (from
matplotlib>=1.4.3->seaborn) (2.8.2)
Requirement already satisfied: packaging>=20.0 in
/home/jupyterlab/conda/envs/python/lib/python3.7/site-packages (from
matplotlib>=1.4.3->seaborn) (21.3)
Requirement already satisfied: cycler>=0.10 in
/home/jupyterlab/conda/envs/python/lib/python3.7/site-packages (from
matplotlib>=1.4.3->seaborn) (0.11.0)
Requirement already satisfied: pyparsing>=2.2.1 in
/home/jupyterlab/conda/envs/python/lib/python3.7/site-packages (from
matplotlib>=1.4.3->seaborn) (3.0.9)
Requirement already satisfied: pillow>=6.2.0 in
/home/jupyterlab/conda/envs/python/lib/python3.7/site-packages (from
matplotlib>=1.4.3->seaborn) (8.1.0)
Requirement already satisfied: kiwisolver>=1.0.1 in
/home/jupyterlab/conda/envs/python/lib/python3.7/site-packages (from
matplotlib>=1.4.3->seaborn) (1.4.4)
Requirement already satisfied: fonttools>=4.22.0 in
/home/jupyterlab/conda/envs/python/lib/python3.7/site-packages (from
matplotlib>=1.4.3->seaborn) (4.38.0)
Requirement already satisfied: pytz>=2017.3 in
/home/jupyterlab/conda/envs/python/lib/python3.7/site-packages (from
pandas>=0.15.2->seaborn) (2022.6)
Requirement already satisfied: typing-extensions in
/home/jupyterlab/conda/envs/python/lib/python3.7/site-packages (from
kiwisolver>=1.0.1->matplotlib>=1.4.3->seaborn) (4.4.0)
Requirement already satisfied: six>=1.5 in
/home/jupyterlab/conda/envs/python/lib/python3.7/site-packages (from python-
dateutil>=2.7->matplotlib>=1.4.3->seaborn) (1.16.0)
```

[11]:
```
import seaborn as sns

bins = np.linspace(df.Principal.min(), df.Principal.max(), 10)
g = sns.FacetGrid(df, col="Gender", hue="loan_status", palette="Set1",␣
 ↪col_wrap=2)
g.map(plt.hist, 'Principal', bins=bins, ec="k")

g.axes[-1].legend()
plt.show()
```



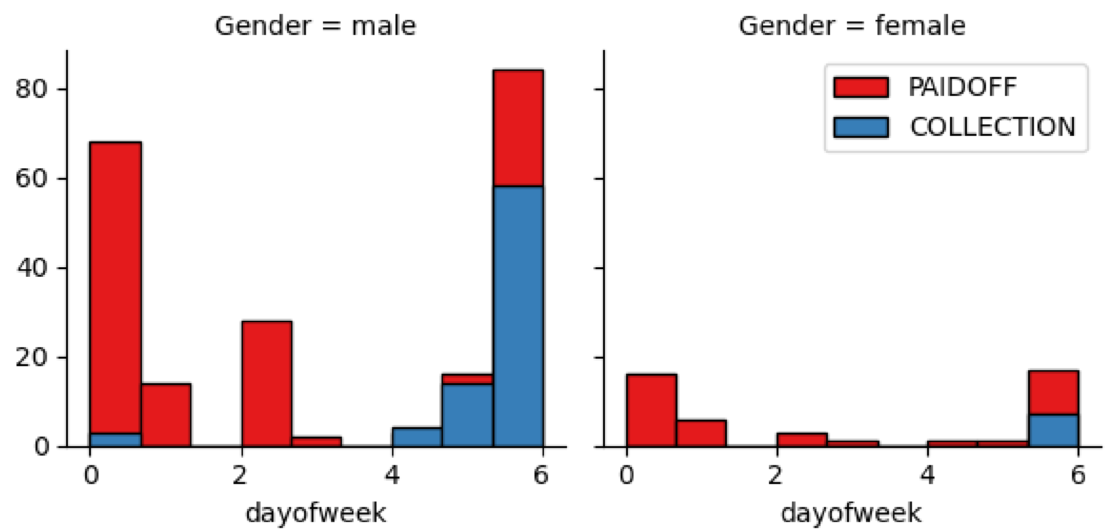

```
[12]: bins = np.linspace(df.age.min(), df.age.max(), 10)
g = sns.FacetGrid(df, col="Gender", hue="loan_status", palette="Set1",␣
 ↪col_wrap=2)
g.map(plt.hist, 'age', bins=bins, ec="k")

g.axes[-1].legend()
plt.show()
```



2 Pre-processing: Feature selection/extraction

2.0.1 Let's look at the day of the week people get the loan

[13]:
```
df['dayofweek'] = df['effective_date'].dt.dayofweek
bins = np.linspace(df.dayofweek.min(), df.dayofweek.max(), 10)
g = sns.FacetGrid(df, col="Gender", hue="loan_status", palette="Set1",␣
 ↪col_wrap=2)
g.map(plt.hist, 'dayofweek', bins=bins, ec="k")
g.axes[-1].legend()
plt.show()
```



We see that people who get the loan at the end of the week don't pay it off, so let's use Feature binarization to set a threshold value less than day 4

[14]:
```
df['weekend'] = df['dayofweek'].apply(lambda x: 1 if (x>3)  else 0)
df.head()
```

[14]:

	Unnamed: 0	Unnamed: 0.1	loan_status	Principal	terms	effective_date
0	0	0	PAIDOFF	1000	30	2016-09-08
1	2	2	PAIDOFF	1000	30	2016-09-08
2	3	3	PAIDOFF	1000	15	2016-09-08
3	4	4	PAIDOFF	1000	30	2016-09-09
4	6	6	PAIDOFF	1000	30	2016-09-09

	due_date	age	education	Gender	dayofweek	weekend
0	2016-10-07	45	High School or Below	male	3	0
1	2016-10-07	33	Bechalor	female	3	0
2	2016-09-22	27	college	male	3	0

```
3 2016-10-08   28             college  female          4         1
4 2016-10-08   29             college    male          4         1
```

2.1 Convert Categorical features to numerical values

Let's look at gender:

[15]:
```
df.groupby(['Gender'])['loan_status'].value_counts(normalize=True)
```

[15]:
```
Gender  loan_status
female  PAIDOFF        0.865385
        COLLECTION     0.134615
male    PAIDOFF        0.731293
        COLLECTION     0.268707
Name: loan_status, dtype: float64
```

86 % of female pay there loans while only 73 % of males pay there loan

Let's convert male to 0 and female to 1:

[16]:
```
df['Gender'].replace(to_replace=['male','female'], value=[0,1],inplace=True)
df.head()
```

[16]:
```
   Unnamed: 0  Unnamed: 0.1 loan_status  Principal  terms effective_date  \
0           0             0     PAIDOFF       1000     30     2016-09-08
1           2             2     PAIDOFF       1000     30     2016-09-08
2           3             3     PAIDOFF       1000     15     2016-09-08
3           4             4     PAIDOFF       1000     30     2016-09-09
4           6             6     PAIDOFF       1000     30     2016-09-09

    due_date  age             education  Gender  dayofweek  weekend
0 2016-10-07   45  High School or Below       0          3        0
1 2016-10-07   33              Bechalor       1          3        0
2 2016-09-22   27               college       0          3        0
3 2016-10-08   28               college       1          4        1
4 2016-10-08   29               college       0          4        1
```

2.2 One Hot Encoding

How about education?

[17]:
```
df.groupby(['education'])['loan_status'].value_counts(normalize=True)
```

[17]:
```
education             loan_status
Bechalor              PAIDOFF        0.750000
                      COLLECTION     0.250000
High School or Below  PAIDOFF        0.741722
                      COLLECTION     0.258278
Master or Above       COLLECTION     0.500000
```

```
                         PAIDOFF        0.500000
college                  PAIDOFF        0.765101
                         COLLECTION     0.234899
Name: loan_status, dtype: float64
```

Features before One Hot Encoding

[18]:
```
df[['Principal','terms','age','Gender','education']].head()
```

[18]:
```
   Principal  terms  age  Gender             education
0       1000     30   45       0  High School or Below
1       1000     30   33       1              Bechalor
2       1000     15   27       0               college
3       1000     30   28       1               college
4       1000     30   29       0               college
```

Use one hot encoding technique to conver categorical varables to binary variables and append them to the feature Data Frame

[19]:
```
Feature = df[['Principal','terms','age','Gender','weekend']]
Feature = pd.concat([Feature,pd.get_dummies(df['education'])], axis=1)
Feature.drop(['Master or Above'], axis = 1,inplace=True)
Feature.head()
```

[19]:
```
   Principal  terms  age  Gender  weekend  Bechalor  High School or Below  \
0       1000     30   45       0        0         0                     1
1       1000     30   33       1        0         1                     0
2       1000     15   27       0        0         0                     0
3       1000     30   28       1        1         0                     0
4       1000     30   29       0        1         0                     0

   college
0        0
1        0
2        1
3        1
4        1
```

2.2.1 Feature Selection

Let's define feature sets, X:

[20]:
```
X = Feature
X[0:5]
```

[20]:
```
   Principal  terms  age  Gender  weekend  Bechalor  High School or Below  \
0       1000     30   45       0        0         0                     1
1       1000     30   33       1        0         1                     0
```

```
2         1000      15    27       0          0          0                     0
3         1000      30    28       1          1          0                     0
4         1000      30    29       0          1          0                     0

   college
0        0
1        0
2        1
3        1
4        1
```

What are our lables?

```
[21]: y = df['loan_status'].values
      y[0:5]
```

```
[21]: array(['PAIDOFF', 'PAIDOFF', 'PAIDOFF', 'PAIDOFF', 'PAIDOFF'],
            dtype=object)
```

2.3 Normalize Data

Data Standardization give data zero mean and unit variance (technically should be done after train test split)

```
[22]: X= preprocessing.StandardScaler().fit(X).transform(X)
      X[0:5]
```

```
[22]: array([[ 0.51578458,  0.92071769,  2.33152555, -0.42056004, -1.20577805,
              -0.38170062,  1.13639374, -0.86968108],
             [ 0.51578458,  0.92071769,  0.34170148,  2.37778177, -1.20577805,
               2.61985426, -0.87997669, -0.86968108],
             [ 0.51578458, -0.95911111, -0.65321055, -0.42056004, -1.20577805,
              -0.38170062, -0.87997669,  1.14984679],
             [ 0.51578458,  0.92071769, -0.48739188,  2.37778177,  0.82934003,
              -0.38170062, -0.87997669,  1.14984679],
             [ 0.51578458,  0.92071769, -0.3215732 , -0.42056004,  0.82934003,
              -0.38170062, -0.87997669,  1.14984679]])
```

3 Classification

Now, it is your turn, use the training set to build an accurate model. Then use the test set to report the accuracy of the model You should use the following algorithm:

- K Nearest Neighbor(KNN)
- Decision Tree
- Support Vector Machine
- Logistic Regression

___ Notice:___

- You can go above and change the pre-processing, feature selection, feature-extraction, and so on, to make a better model.
- You should use either scikit-learn, Scipy or Numpy libraries for developing the classification algorithms.
- You should include the code of the algorithm in the following cells.

[28]:
```
# We split the X into train and test to find the best k
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=4)
print ('Train set:', X_train.shape,  y_train.shape)
print ('Test set:', X_test.shape,  y_test.shape)
```

```
Train set: (276, 8) (276,)
Test set: (70, 8) (70,)
```

4 K Nearest Neighbor(KNN)

Notice: You should find the best k to build the model with the best accuracy.
warning: You should not use the **loan_test.csv** for finding the best k, however, you can split your train_loan.csv into train and test to find the best **k**.

[29]:
```
from sklearn.neighbors import KNeighborsClassifier
k = 3
#Train Model and Predict
kNN_model = KNeighborsClassifier(n_neighbors=k).fit(X_train,y_train)
kNN_model
```

[29]:
```
KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski',
           metric_params=None, n_jobs=None, n_neighbors=3, p=2,
           weights='uniform')
```

[30]:
```
yhat = kNN_model.predict(X_test)
yhat[0:5]
```

[30]:
```
array(['PAIDOFF', 'PAIDOFF', 'PAIDOFF', 'PAIDOFF', 'PAIDOFF'],
      dtype=object)
```

[31]:
```
# Best k
Ks=15
mean_acc=np.zeros((Ks-1))
std_acc=np.zeros((Ks-1))
ConfustionMx=[];
for n in range(1,Ks):

    #Train Model and Predict
```

```
    kNN_model = KNeighborsClassifier(n_neighbors=n).fit(X_train,y_train)
    yhat = kNN_model.predict(X_test)


    mean_acc[n-1]=np.mean(yhat==y_test);

    std_acc[n-1]=np.std(yhat==y_test)/np.sqrt(yhat.shape[0])
mean_acc
```

[31]:
```
array([0.67142857, 0.65714286, 0.71428571, 0.68571429, 0.75714286,
       0.71428571, 0.78571429, 0.75714286, 0.75714286, 0.67142857,
       0.7       , 0.72857143, 0.7       , 0.7       ])
```

[32]:
```
# Building the model again, using k=7
from sklearn.neighbors import KNeighborsClassifier
k = 7
#Train Model and Predict
kNN_model = KNeighborsClassifier(n_neighbors=k).fit(X_train,y_train)
kNN_model
```

[32]:
```
KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski',
           metric_params=None, n_jobs=None, n_neighbors=7, p=2,
           weights='uniform')
```

5 Decision Tree

[33]:
```
from sklearn.tree import DecisionTreeClassifier
DT_model = DecisionTreeClassifier(criterion="entropy", max_depth = 4)
DT_model.fit(X_train,y_train)
DT_model
```

[33]:
```
DecisionTreeClassifier(class_weight=None, criterion='entropy', max_depth=4,
            max_features=None, max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, presort=False, random_state=None,
            splitter='best')
```

[34]:
```
yhat = DT_model.predict(X_test)
yhat
```

[34]:
```
array(['COLLECTION', 'COLLECTION', 'PAIDOFF', 'PAIDOFF', 'PAIDOFF',
       'PAIDOFF', 'PAIDOFF', 'PAIDOFF', 'PAIDOFF', 'COLLECTION',
       'PAIDOFF', 'COLLECTION', 'PAIDOFF', 'PAIDOFF', 'PAIDOFF',
       'PAIDOFF', 'COLLECTION', 'PAIDOFF', 'COLLECTION', 'PAIDOFF',
       'PAIDOFF', 'COLLECTION', 'COLLECTION', 'COLLECTION', 'PAIDOFF',
```

```
       'COLLECTION', 'COLLECTION', 'PAIDOFF', 'COLLECTION', 'PAIDOFF',
       'COLLECTION', 'COLLECTION', 'COLLECTION', 'PAIDOFF', 'PAIDOFF',
       'PAIDOFF', 'COLLECTION', 'PAIDOFF', 'COLLECTION', 'PAIDOFF',
       'COLLECTION', 'PAIDOFF', 'PAIDOFF', 'COLLECTION', 'PAIDOFF',
       'COLLECTION', 'COLLECTION', 'COLLECTION', 'PAIDOFF', 'PAIDOFF',
       'PAIDOFF', 'PAIDOFF', 'PAIDOFF', 'PAIDOFF', 'PAIDOFF', 'PAIDOFF',
       'PAIDOFF', 'PAIDOFF', 'COLLECTION', 'PAIDOFF', 'PAIDOFF',
       'PAIDOFF', 'PAIDOFF', 'COLLECTION', 'PAIDOFF', 'COLLECTION',
       'PAIDOFF', 'COLLECTION', 'PAIDOFF', 'PAIDOFF'], dtype=object)
```

6 Support Vector Machine

[36]:
```
from sklearn import svm
SVM_model = svm.SVC()
SVM_model.fit(X_train, y_train)
```

[36]:
```
SVC(C=1.0, cache_size=200, class_weight=None, coef0=0.0,
  decision_function_shape='ovr', degree=3, gamma='auto_deprecated',
  kernel='rbf', max_iter=-1, probability=False, random_state=None,
  shrinking=True, tol=0.001, verbose=False)
```

[37]:
```
yhat = SVM_model.predict(X_test)
yhat
```

[37]:
```
array(['COLLECTION', 'PAIDOFF', 'PAIDOFF', 'PAIDOFF', 'PAIDOFF',
       'PAIDOFF', 'COLLECTION', 'COLLECTION', 'PAIDOFF', 'PAIDOFF',
       'PAIDOFF', 'PAIDOFF', 'PAIDOFF', 'PAIDOFF', 'PAIDOFF', 'PAIDOFF',
       'COLLECTION', 'PAIDOFF', 'PAIDOFF', 'PAIDOFF', 'PAIDOFF',
       'PAIDOFF', 'COLLECTION', 'COLLECTION', 'PAIDOFF', 'COLLECTION',
       'COLLECTION', 'PAIDOFF', 'PAIDOFF', 'PAIDOFF', 'PAIDOFF',
       'PAIDOFF', 'PAIDOFF', 'PAIDOFF', 'PAIDOFF', 'PAIDOFF', 'PAIDOFF',
       'PAIDOFF', 'COLLECTION', 'PAIDOFF', 'PAIDOFF', 'PAIDOFF',
       'PAIDOFF', 'COLLECTION', 'PAIDOFF', 'PAIDOFF', 'PAIDOFF',
       'PAIDOFF', 'PAIDOFF', 'PAIDOFF', 'PAIDOFF', 'PAIDOFF', 'PAIDOFF',
       'PAIDOFF', 'PAIDOFF', 'PAIDOFF', 'PAIDOFF', 'PAIDOFF', 'PAIDOFF',
       'PAIDOFF', 'PAIDOFF', 'PAIDOFF', 'PAIDOFF', 'COLLECTION',
       'PAIDOFF', 'PAIDOFF', 'PAIDOFF', 'PAIDOFF', 'PAIDOFF', 'PAIDOFF'],
      dtype=object)
```

[ ]:
```

```

7 Logistic Regression

[38]:
```
from sklearn.linear_model import LogisticRegression
LR_model = LogisticRegression(C=0.01).fit(X_train,y_train)
LR_model
```

[38]:
```
LogisticRegression(C=0.01, class_weight=None, dual=False, fit_intercept=True,
          intercept_scaling=1, max_iter=100, multi_class='warn',
          n_jobs=None, penalty='l2', random_state=None, solver='warn',
          tol=0.0001, verbose=0, warm_start=False)
```

[39]:
```
yhat = LR_model.predict(X_test)
yhat
```

[39]:
```
array(['COLLECTION', 'PAIDOFF', 'PAIDOFF', 'PAIDOFF', 'PAIDOFF',
       'PAIDOFF', 'PAIDOFF', 'PAIDOFF', 'PAIDOFF', 'PAIDOFF', 'PAIDOFF',
       'PAIDOFF', 'PAIDOFF', 'PAIDOFF', 'PAIDOFF', 'PAIDOFF',
       'COLLECTION', 'PAIDOFF', 'COLLECTION', 'PAIDOFF', 'PAIDOFF',
       'PAIDOFF', 'COLLECTION', 'PAIDOFF', 'PAIDOFF', 'COLLECTION',
       'COLLECTION', 'PAIDOFF', 'COLLECTION', 'PAIDOFF', 'PAIDOFF',
       'PAIDOFF', 'PAIDOFF', 'PAIDOFF', 'PAIDOFF', 'PAIDOFF',
       'COLLECTION', 'PAIDOFF', 'PAIDOFF', 'PAIDOFF', 'COLLECTION',
       'PAIDOFF', 'PAIDOFF', 'COLLECTION', 'PAIDOFF', 'PAIDOFF',
       'PAIDOFF', 'PAIDOFF', 'PAIDOFF', 'PAIDOFF', 'PAIDOFF', 'PAIDOFF',
       'PAIDOFF', 'PAIDOFF', 'PAIDOFF', 'PAIDOFF', 'PAIDOFF', 'PAIDOFF',
       'PAIDOFF', 'PAIDOFF', 'PAIDOFF', 'PAIDOFF', 'PAIDOFF',
       'COLLECTION', 'PAIDOFF', 'PAIDOFF', 'PAIDOFF', 'PAIDOFF',
       'PAIDOFF', 'PAIDOFF'], dtype=object)
```

8 Model Evaluation using Test set

[41]:
```
from sklearn.metrics import jaccard_similarity_score
from sklearn.metrics import f1_score
from sklearn.metrics import log_loss
```

First, download and load the test set:

[42]:
```
!wget -O loan_test.csv https://s3-api.us-geo.objectstorage.softlayer.net/
 ↪cf-courses-data/CognitiveClass/ML0101ENv3/labs/loan_test.csv
```

```
--2023-01-25 05:57:15--  https://s3-api.us-geo.objectstorage.softlayer.net/cf-
courses-data/CognitiveClass/ML0101ENv3/labs/loan_test.csv
Resolving s3-api.us-geo.objectstorage.softlayer.net (s3-api.us-
geo.objectstorage.softlayer.net)… 67.228.254.196
Connecting to s3-api.us-geo.objectstorage.softlayer.net (s3-api.us-
geo.objectstorage.softlayer.net)|67.228.254.196|:443… connected.
HTTP request sent, awaiting response… 200 OK
```

```
Length: 3642 (3.6K) [text/csv]
Saving to: 'loan_test.csv'

loan_test.csv       100%[===================>]   3.56K  --.-KB/s    in 0s

2023-01-25 05:57:15 (18.2 MB/s) - 'loan_test.csv' saved [3642/3642]
```

8.0.1 Load Test set for evaluation

[43]:
```python
test_df = pd.read_csv('loan_test.csv')
test_df.head()
```

[43]:
```
   Unnamed: 0  Unnamed: 0.1 loan_status  Principal  terms effective_date  \
0           1             1     PAIDOFF       1000     30       9/8/2016
1           5             5     PAIDOFF        300      7       9/9/2016
2          21            21     PAIDOFF       1000     30      9/10/2016
3          24            24     PAIDOFF       1000     30      9/10/2016
4          35            35     PAIDOFF        800     15      9/11/2016

    due_date  age             education  Gender
0  10/7/2016   50              Bechalor  female
1  9/15/2016   35       Master or Above    male
2  10/9/2016   43  High School or Below  female
3  10/9/2016   26               college    male
4  9/25/2016   29              Bechalor    male
```

[44]:
```python
## Preprocessing
test_df['due_date'] = pd.to_datetime(test_df['due_date'])
test_df['effective_date'] = pd.to_datetime(test_df['effective_date'])
test_df['dayofweek'] = test_df['effective_date'].dt.dayofweek
test_df['weekend'] = test_df['dayofweek'].apply(lambda x: 1 if (x>3)  else 0)
test_df['Gender'].replace(to_replace=['male','female'],␣
 ↪value=[0,1],inplace=True)
test_Feature = test_df[['Principal','terms','age','Gender','weekend']]
test_Feature = pd.concat([test_Feature,pd.get_dummies(test_df['education'])],␣
 ↪axis=1)
test_Feature.drop(['Master or Above'], axis = 1,inplace=True)
test_X = preprocessing.StandardScaler().fit(test_Feature).
 ↪transform(test_Feature)
test_X[0:5]
```

[44]:
```
array([[ 0.49362588,  0.92844966,  3.05981865,  1.97714211, -1.30384048,
         2.39791576, -0.79772404, -0.86135677],
       [-3.56269116, -1.70427745,  0.53336288, -0.50578054,  0.76696499,
        -0.41702883, -0.79772404, -0.86135677],
       [ 0.49362588,  0.92844966,  1.88080596,  1.97714211,  0.76696499,
```

```
        -0.41702883,  1.25356634, -0.86135677],
       [ 0.49362588,  0.92844966, -0.98251057, -0.50578054,  0.76696499,
        -0.41702883, -0.79772404,  1.16095912],
       [-0.66532184, -0.78854628, -0.47721942, -0.50578054,  0.76696499,
         2.39791576, -0.79772404, -0.86135677]])
```

[45]:
```
test_y = test_df['loan_status'].values
test_y[0:5]
```

[45]:
```
array(['PAIDOFF', 'PAIDOFF', 'PAIDOFF', 'PAIDOFF', 'PAIDOFF'],
      dtype=object)
```

[46]:
```
knn_yhat = kNN_model.predict(test_X)
print("KNN Jaccard index: %.2f" % jaccard_similarity_score(test_y, knn_yhat))
print("KNN F1-score: %.2f" % f1_score(test_y, knn_yhat, average='weighted') )
```

```
KNN Jaccard index: 0.67
KNN F1-score: 0.63
```

[47]:
```
DT_yhat = DT_model.predict(test_X)
print("DT Jaccard index: %.2f" % jaccard_similarity_score(test_y, DT_yhat))
print("DT F1-score: %.2f" % f1_score(test_y, DT_yhat, average='weighted') )
```

```
DT Jaccard index: 0.72
DT F1-score: 0.74
```

[48]:
```
SVM_yhat = SVM_model.predict(test_X)
print("SVM Jaccard index: %.2f" % jaccard_similarity_score(test_y, SVM_yhat))
print("SVM F1-score: %.2f" % f1_score(test_y, SVM_yhat, average='weighted') )
```

```
SVM Jaccard index: 0.80
SVM F1-score: 0.76
```

[49]:
```
LR_yhat = LR_model.predict(test_X)
LR_yhat_prob = LR_model.predict_proba(test_X)
print("LR Jaccard index: %.2f" % jaccard_similarity_score(test_y, LR_yhat))
print("LR F1-score: %.2f" % f1_score(test_y, LR_yhat, average='weighted') )
print("LR LogLoss: %.2f" % log_loss(test_y, LR_yhat_prob))
```

```
LR Jaccard index: 0.74
LR F1-score: 0.66
LR LogLoss: 0.57
```

9 Report

You should be able to report the accuracy of the built model using different evaluation metrics:

Algorithm	Jaccard	F1-score	LogLoss
KNN	?	?	NA
Decision Tree	?	?	NA
SVM	?	?	NA
LogisticRegression	?	?	?

Want to learn more?

IBM SPSS Modeler is a comprehensive analytics platform that has many machine learning algorithms. It has been designed to bring predictive intelligence to decisions made by individuals, by groups, by systems – by your enterprise as a whole. A free trial is available through this course, available here: SPSS Modeler

Also, you can use Watson Studio to run these notebooks faster with bigger datasets. Watson Studio is IBM's leading cloud solution for data scientists, built by data scientists. With Jupyter notebooks, RStudio, Apache Spark and popular libraries pre-packaged in the cloud, Watson Studio enables data scientists to collaborate on their projects without having to install anything. Join the fast-growing community of Watson Studio users today with a free account at Watson Studio

Thanks for completing this lesson!

Author: Saeed Aghabozorgi

Saeed Aghabozorgi, PhD is a Data Scientist in IBM with a track record of developing enterprise level applications that substantially increases clients' ability to turn data into actionable knowledge. He is a researcher in data mining field and expert in developing advanced analytic methods like machine learning and statistical modelling on large datasets.

9.1 Change Log

Date (YYYY-MM-DD)	Version	Changed By	Change Description
2020-10-27	2.1	Lakshmi Holla	Made changes in import statement due to updates in version of sklearn library
2020-08-27	2.0	Malika Singla	Added lab to GitLab

##