(https://www.bigdatauniversity.com)

Project: Whether a loan is paid off

Deadline: 2019-08-25 23:59:59

Total marks: 7.0

Your information:

- Fullname:
- Date of birth:
- Place of birth:
- Email:
- Mobile phone:

In this notebook, we practice all the knowledge and skills that we learned in this course.

We apply the **Logistic Algorithm** to predict: "Whether a loan is paid off on in collection" by accuracy evaluation methods.

Lets first load required libraries:

Entrée [1]:
```
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
```

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

Data exploration

*** To predict "Whether a loan is paid off", we need some fields: 'Principal', 'Terms', 'Age', 'Gender', 'Effective_date'

The first things we need to do:

- Identify Variables
- Univariate Analysis
- Bi-variate Analysis
- Handle the Missing Values
- Handle Outlier Values

Tips: Step by step like Chapter2_Ex1_Housing prices

Load Data From CSV File

Entrée [2]:
```
# Read CSV file: loan_train.csv
# code here
```

Entrée [3]:
```
# Understanding to dataset
# shape
# info
# head(), tail()
# describe()
```

Convert 'due_date', 'effective_date' to date time object

Entrée [4]:
```
# code here
```

Data visualization

How many sample of each class is in our data set?

Entrée [5]:
```
# code
```

xxx people have paid off the loan on time while **xxx** have gone into collection

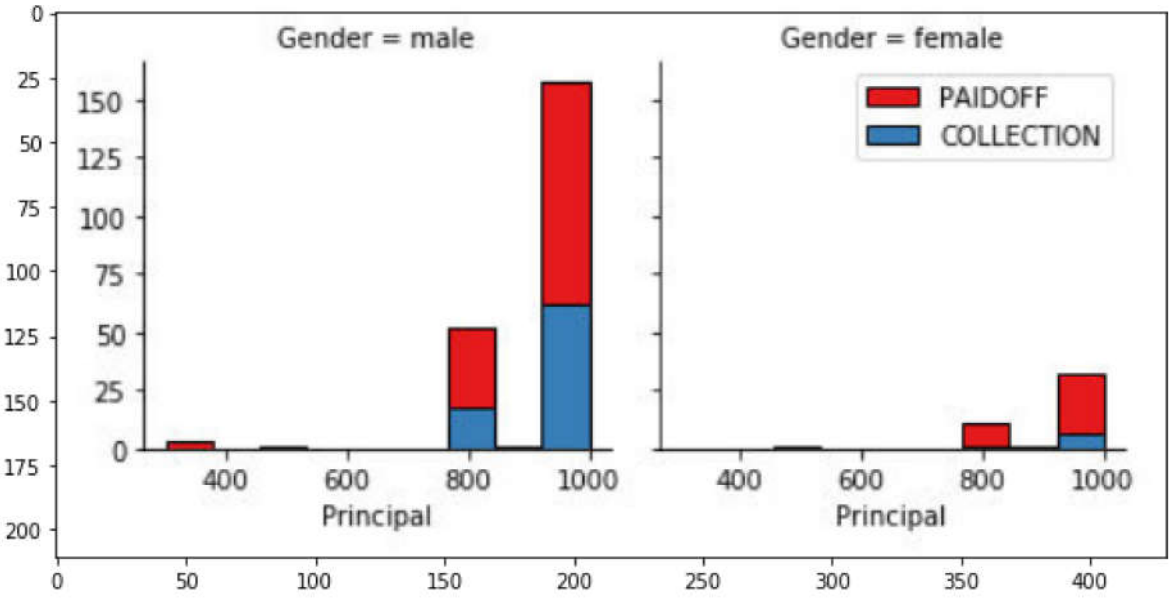
Lets plot some columns to underestand data better:

- Use seaborn or matplotlib to draw some plots like that:

Entrée [6]:
```
import numpy as np
from PIL import Image
import matplotlib.pyplot as plt
```

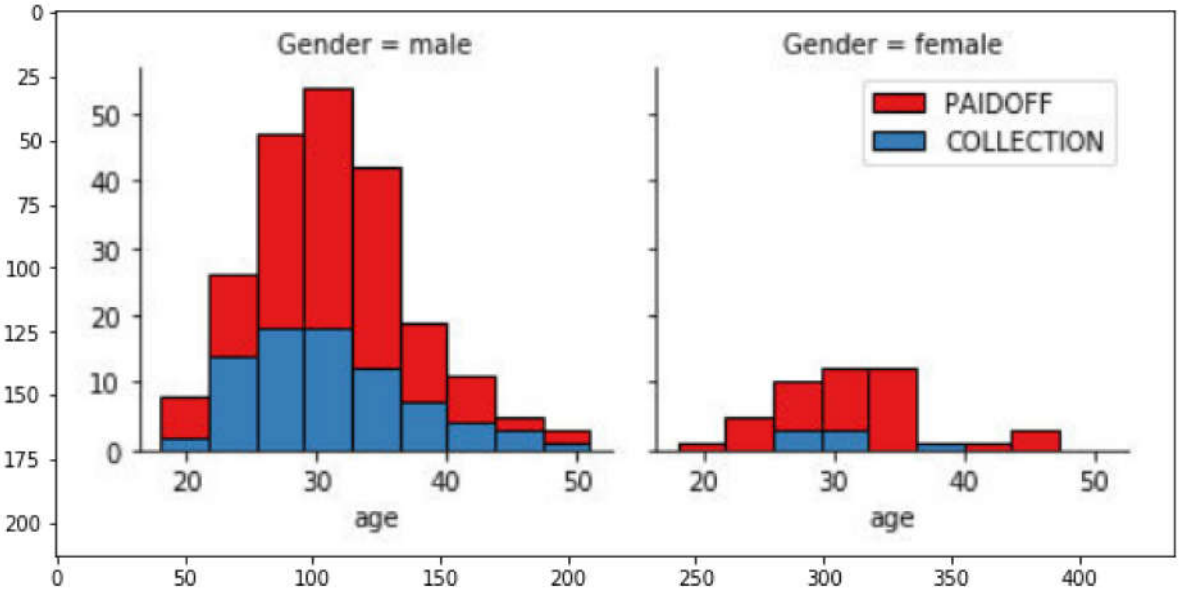
Entrée [7]:
```
img1 = np.array(Image.open('Principal_Male_Female.jpg'))
```

Entrée [8]:
```
plt.figure(figsize=(10,5))
plt.imshow(img1, interpolation='bilinear')
plt.show()
```



Entrée [9]:
```
img2 = np.array(Image.open('Age_Male_Female.jpg'))
```

Entrée [10]:
```
plt.figure(figsize=(10,5))
plt.imshow(img2, interpolation='bilinear')
plt.show()
```



Entrée [11]:
```
# code here
```

Pre-processing: Feature selection/extraction

Lets look at the day of the week people get the loan

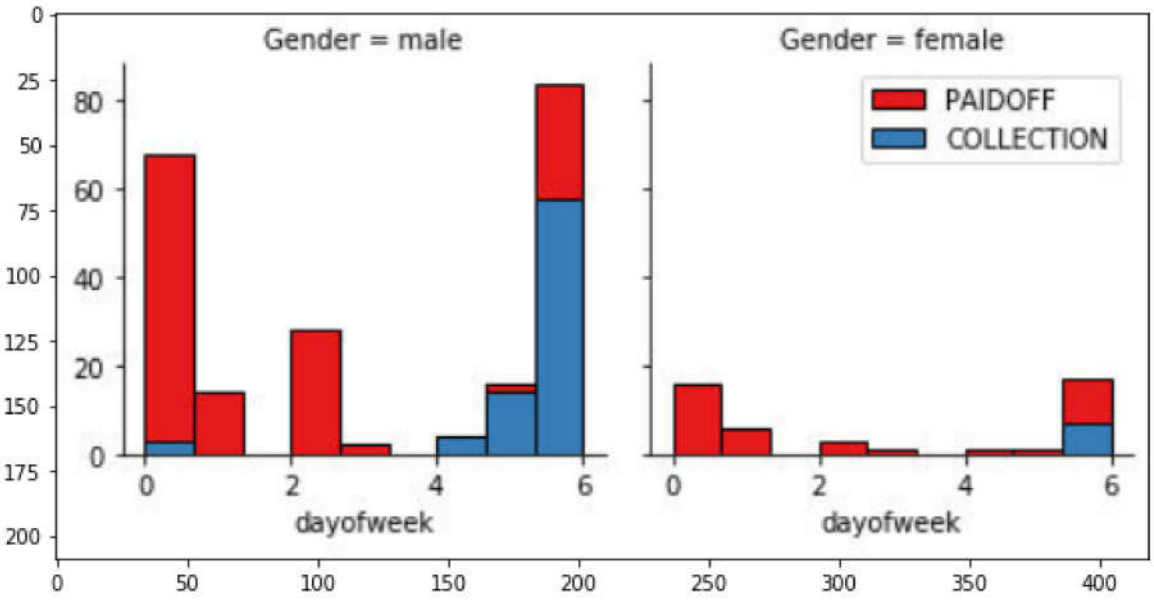
- Make new column 'dayofweek' from 'effective_date'
  - Example: 2016-09-08 => dayofweek is 3 (The day of the week with Monday=0, Sunday=6)
  - Link: https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DatetimeIndex.dayofweek.html (https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DatetimeIndex.dayofweek.html)

Entrée [12]:
```
# code here
```

Lets plot some columns to underestand data better:

Entrée [13]:
```
img3 = np.array(Image.open('day_of_week.jpg'))
```

Entrée [14]:
```
plt.figure(figsize=(10,5))
plt.imshow(img3, interpolation='bilinear')
plt.show()
```



Entrée [15]:
```
# code here
```

We see that people who get the loan at the end of the week dont pay it off, so lets use Feature binarization to set a threshold values less then day 4

- Make new column 'weekend': =1 if 'dayofweek'>3, else =0

Entrée [16]:
```
# code here
```

Convert Categorical features to numerical values

- groupby 'Gender' and count by 'loan_status'

Entrée [17]:
```
# code here
```

xxx % of female pay there loans while only **xxx** % of males pay there loan

Lets convert male to 0 and female to 1:

Entrée [18]:
```
# code here
```

One Hot Encoding

How about education?

- groupby 'education' and count by 'loan_status'

Entrée [19]:
```
# code here
```

Feature befor One Hot Encoding

- Print head() data with 5 columns: 'Principal','terms','age','Gender','education'

Entrée [20]:
```
# code here
```

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

- Make new dataframe **Feature** has: 'Principal','terms','age','Gender','weekend', 'education'
- In **Feature**: Use one hot encoding technique to convert 'education' to binary variable, then drop column 'Master or Above'

Entrée [21]:
```
# code here
```

Feature selection

Lets defind feature sets, X:

- X is input, X = Feature

Entrée [22]:
```
# code here
```

What are our lables?

- y is output, y = 'loan_status' column

Entrée [23]:
```
# code here
```

Normalize Data

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

- Find the suitable Scaler to scale data of X (if we need to do to have a better prediction)

Entrée [24]:
```
# code here
```

Classification

Now, use the training set to build an accurate model. Then use the test set (loan_test.csv) to report the accuracy of the model You should use the following algorithm:

- 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.

Logistic Regression

Entrée [ ]:

Entrée [ ]:

Entrée [ ]:

Model Evaluation using Test set

Entrée [25]:
```
from sklearn.metrics import f1_score
from sklearn.metrics import accuracy_score
# and the others libraries...
```

Load Test set for evaluation

Entrée [26]:
```
# Read CSV file: loan_test.csv
# code here
```

Entrée [27]:
```
# Model Evaluation
```

Entrée [ ]:

Entrée [ ]:

Report

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

Algorithm	Accuracy Score	F1-score
LogisticRegression	?	?

Thanks for completing this project!

Author: Saeed Aghabozorgi (https://ca.linkedin.com/in/saeedaghabozorgi)

Saeed Aghabozorgi (https://ca.linkedin.com/in/saeedaghabozorgi), 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.

Entrée [ ]: