National University of Sciences & Technology
School of Electrical Engineering and Computer Science
Department of Software Engineering

CS471: Machine Learning

Class: BESE-6AB

Assignment 01: Logistic Regression

Announcement Date: 28-02-2018
Due Date: 9th March 2018

Instructor: Dr. Muhammad Moazam Fraz

Course Learning Outcomes (CLOs)

	Upon completion of the course, students should demonstrate the ability to:		PLO** Mapping	BT Level*
	CLO 1	Develop an appreciation for what is involved in learning from data.	PLO 1	C1
	CLO 2	Understand a wide variety of learning algorithms.	PLO 2	C2
	CLO 3	Apply a variety of learning algorithms to data for solution development.	PLO 3	C3
	CLO 4	Evaluate various learning algorithms for optimal model selection.	PLO 4	C6
	CLO 5	Develop solutions by using modern machine learning tools / models to solve practical problems.	PLO 5	C5

* BT= Bloom's Taxonomy, C=Cognitive domain, P=Psychomotor domain, A= Affective domain

- *Knowledge(C-1), Comprehension(C-2), Application(C-3), Analysis(C-4), Synthesis(C-5), Evaluation(C-6)*
- *Perception(P-1), Set(P-2), Guided Response(P-3), Mecha*
- *Receiving(A-1), Responding(A-2), Valuing(A nism(P-4), Complete Overt Response(P-5), Adaption(P-6), Organization(P-7) -3), Organization(A-4), Internalizing(A-5)*

*** PLOs are published on departmental website*

Learning Outcome

CLO 2: Understand a wide variety of learning algorithms

Introduction:

Logistic Regression is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.). In other words, the logistic regression model predicts P(Y=1) as a function of X.

Assumptions:

- Binary logistic regression requires the dependent variable to be binary.
- For a binary regression, the factor level 1 of the dependent variable should represent the desired outcome.
- Only the meaningful variables should be included.
- The independent variables should be independent of each other. That is, the model should have little or no multicollinearity.
- The independent variables are linearly related to the log odds.
- Logistic regression requires quite large sample sizes.

Data:

The dataset comes from the UCI Machine Learning repository, and it is related to direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict whether the client will subscribe (1/0) to a term deposit (variable y). The dataset can be downloaded from here.

Tasks

1. Read the dataset and print the following information for analysis.

```
(41188, 21)
['age', 'job', 'marital', 'education', 'default', 'housing', 'loan', 'contact', 'month', 'day_of_week', 'duration', 'campaign',
'pdays', 'previous', 'poutcome', 'emp_var_rate', 'cons_price_idx', 'cons_conf_idx', 'euribor3m', 'nr_employed', 'y']
```

```
In [37]: data.head()
```

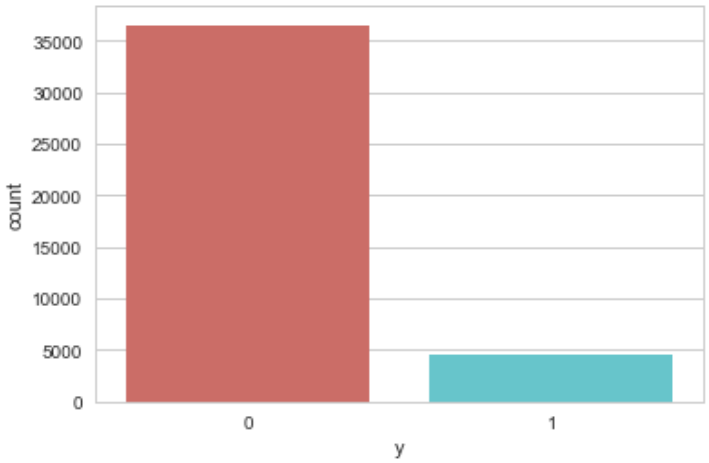
Out[37]:

	age	job	marital	education	default	housing	loan	contact	month	day_of_week	...	campaign	pdays	previous	poutcome	emp_var_rate
0	44	blue-collar	married	basic.4y	unknown	yes	no	cellular	aug	thu	...	1	999	0	nonexistent	1.4
1	53	technician	married	unknown	no	no	no	cellular	nov	fri	...	1	999	0	nonexistent	-0.1
2	28	management	single	university.degree	no	yes	no	cellular	jun	thu	...	3	6	2	success	-1.7
3	39	services	married	high.school	no	no	no	cellular	apr	fri	...	2	999	0	nonexistent	-1.8
4	55	retired	married	basic.4y	no	yes	no	cellular	aug	fri	...	1	3	1	success	-2.9

5 rows × 21 columns

Hint: Use Pandas to read the data and use pandas functions mentioned in slides for cleaning the missing values and attributes

2. Plot the ‘y’ for counts to check the values of 0 and 1 in the prediction also plot the job, martial, load and pooutcome.



Hint: Use the seaborn library function countplot.

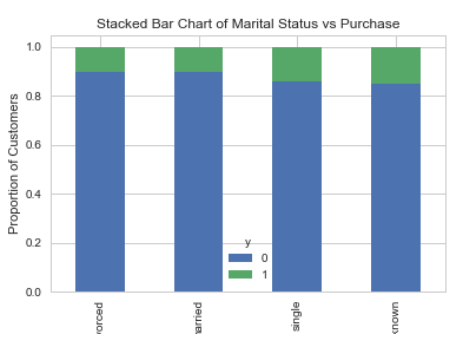
3. Analyze the data using the y,job,martial and education for the insights.

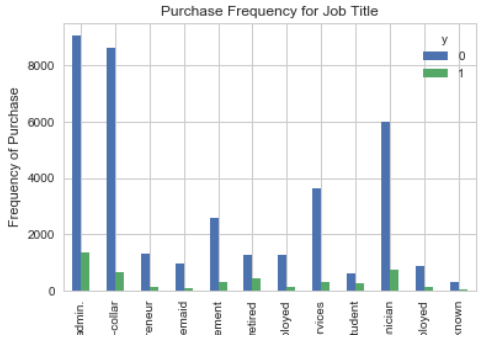
y	age	duration	campaign	pdays	previous	emp_var_rate	cons_price_idx	cons_conf_idx	euribor3m	nr_employed
0	39.911185	220.844807	2.633085	984.113878	0.132374	0.248875	93.603757	-40.593097	3.811491	5176.166600
1	40.913147	553.191164	2.051724	792.035560	0.492672	-1.233448	93.354386	-39.789784	2.123135	5095.115991

job	age	duration	campaign	pdays	previous	emp_var_rate	cons_price_idx	cons_conf_idx	euribor3m	nr_employed	y
admin.	38.187296	254.312128	2.623489	954.319229	0.189023	0.015563	93.534054	-40.245433	3.550274	5164.125350	0.129726
blue-collar	39.555760	264.542360	2.558461	985.160363	0.122542	0.248995	93.656656	-41.375816	3.771996	5175.615150	0.068943
entrepreneur	41.723214	263.267857	2.535714	981.267170	0.138736	0.158723	93.605372	-41.283654	3.791120	5176.313530	0.085165
housemaid	45.500000	250.454717	2.639623	960.579245	0.137736	0.433396	93.676576	-39.495283	4.009645	5179.529623	0.100000
management	42.362859	257.058140	2.476060	962.647059	0.185021	-0.012688	93.522755	-40.489466	3.611316	5166.650513	0.112175
retired	62.027326	273.712209	2.476744	897.936047	0.327326	-0.698314	93.430786	-38.573081	2.770066	5122.262151	0.252326
self-employed	39.949331	264.142153	2.660802	976.621393	0.143561	0.094159	93.559982	-40.488107	3.689376	5170.674384	0.104856
services	37.926430	258.398085	2.587805	979.974049	0.154951	0.175359	93.634659	-41.290048	3.699187	5171.600126	0.081381
student	25.894857	283.683429	2.104000	840.217143	0.524571	-1.408000	93.331613	-40.187543	1.884224	5085.939086	0.314286
technician	38.507638	250.232241	2.577339	964.408127	0.153789	0.274566	93.561471	-39.927569	3.820401	5175.648391	0.108260
unemployed	39.733728	249.451677	2.564103	935.316568	0.199211	-0.111736	93.563781	-40.007594	3.466583	5157.156509	0.142012
unknown	45.563636	239.675758	2.648485	938.727273	0.154545	0.357879	93.718942	-38.797879	3.949033	5172.931818	0.112121

Hint: Use Pandas groupby function for this.

4. Visualize the joint data e.g. job and y, martial and y, education and y for the insights.



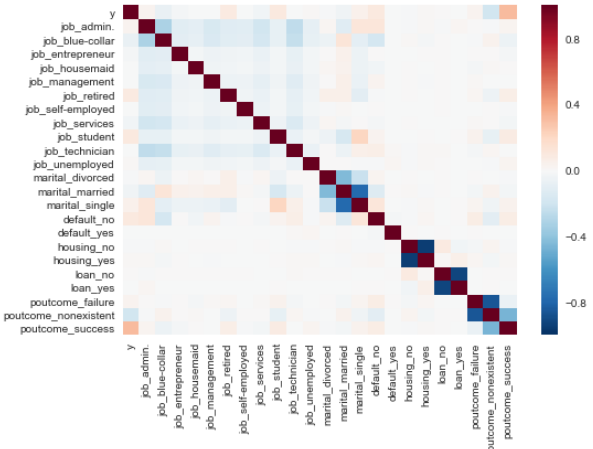


Hint: Use pandas crosstab function to get the desired data and plot using matplotlib bar charts

5. Create dummy variables, that is variables with only two values, zero and one. Use the following columns 'job', 'marital', 'default', 'housing', 'loan', 'poutcome'.
Hint: Use the pandas function get_dummies()

6. Drop the unknown columns [12, 16, 18, 21, 24].
Hint: Use the pandas drop function for this and drop the above mentioned columns

7. Check the independence between the independent variables by drawing the heat map of the data



Hint: Use the seaborn heatmap function for this

8. Split the data into training and test sets.
Hint: X = data.iloc[:,1:] and y = data.iloc[:,0] then use the sklearn function train_test_split()

9. Fit logistic regression to the training set.
Hint: Use sklearn LogisticRegression() class for this and then use fit() method to train the classifier.

10. Predicting the test set results and creating confusion matrix.
    Hint: Use sklearn confusion_matrix() function for confusion matrix and classifier predict() method for the predictions.

11. Print the Accuracy of the classifier using the score() method of the classifier.

12. Compute precision, recall, F-measure and support.

    Hint: Use sklearn classification_report() function for this.

Submission Instructions

1. Please create a Jyputer Notebook for the tasks, include proper comments and submit it.
2. The Name of Notebook should be YOUR NAME_YOURCMSID.ipynb and upload it on LMS.
3. Your code in Notebook shoud run seamlessly. Failure in running code will earn a zero credit.
4. The code should be fully documented explaining each step you had implemented to earn the full credit.

Please note that Failing to follow naming and coding conventions and failing to run the code seamlessly will result in ZERO credit.