

Computer Science Department

CS675 – Introduction to Data Science (CRN: 73405)

Fall 2024

Project #2 / Due 12-Nov-2024

The goal of this assignment is to understand the ‘power’ of various Machine Learning Classification algorithms applied into a dataset. By contrasting these very well-diverse and widely used models within Machine Learning space. The end goal is to find the ‘best’ algorithm to do the job in quest.

Write up **Python/R code** snippets that will device **6 different classification algorithms** on the same dataset. Namely, apply the following ML models:

- 1- **Logistic Regression (LR)**
- 2- **Naive Bayes (NB)**
- 3- **K-Nearest Neighbors (KNN)**
- 4- **Decision Tree (DT)**
- 5- **Random Forest (RF)**
- 6- **XGBoost Algorithm (XGB)**

You should download the following Bank dataset: **Bank Marketing Data Set**

The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if the client will subscribe a term deposit (variable y).

You should get only one (1) dataset, the ‘bank-additonal-full.csv’ with 45 211 records.

<https://archive.ics.uci.edu/dataset/222/bank+marketing>

Download the ‘bank-additional.zip’ file and extract the ‘bank-additonal-full.csv’ file. Read details of what the variable/features mean.

Here is what the file looks like:

bank-additional-full		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q			
1		age	job	marital	education	default	housing	loan	contact	month	day_of_week	duration	campaign	pdays	previous	poutcome	emp.var.rate	cons.price.idx	cons.		
2	56	housemaid	married	basic.4y	no	no	no	telephone	may	mon	261	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191	no
3	57	services	married	high.school	unknown	no	no	telephone	may	mon	149	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191	no
4	37	services	married	high.school	no	yes	no	telephone	may	mon	226	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191	no
5	40	admin.	married	basic.6y	no	no	no	telephone	may	mon	151	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191	no
6	56	services	married	high.school	no	no	yes	telephone	may	mon	307	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191	no
7	45	services	married	basic.9y	unknown	no	no	telephone	may	mon	198	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191	no
8	59	admin.	married	professional.course	no	no	no	telephone	may	mon	139	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191	no
9	41	blue-collar	married	unknown	unknown	no	no	telephone	may	mon	217	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191	no
10	24	technician	single	professional.course	no	yes	no	telephone	may	mon	380	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191	no
11	25	services	single	high.school	no	yes	no	telephone	may	mon	50	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191	no
12	41	blue-collar	married	unknown	unknown	no	no	telephone	may	mon	55	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191	no
13	25	services	single	high.school	no	yes	no	telephone	may	mon	222	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191	no

Perform various Machine Learning activities in order to complete the following tasks along with their output. All work should be done and submitted in a single **Jupyter/Colab Notebook**.

- 1- **Prep the data** in order to be ready to be fed to ML models.
- 2- Split the source dataset into **training** and **test** datasets at a 70%/30% ratio.

3- Run all algorithms with default values and report their **model performance** on the following metrics: -

- Accuracy
- Precision
- Recall
- F1 Harmonic Mean

4- Generate **Classification Report** (for each model) including Confusion Matrices, ROC Curves, and AUCs.

5- Extra points, rerun some of the models by **tuning** some **hyperparameters**.