Samuel Ogbonnaya 2022

Radar – Artificial Intelligence – Machine Learning Project

Predicting the Outcome of Bank Telemarketing

1. Project definition

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).

Make this model as accurate as possible and submit your results!

2. Data

The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

- Detailed information about the dataset can be found here: https://archive.ics.uci.edu/ml/datasets/bank+marketing#
- The dataset can be downloaded from here: bank-additional.zip

Here is a snippet of the data:

	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
0	30	blue-collar	married	basic.9y	no	yes	no	cellular	may	fri	487	2	999	0	nonexistent	-1.8	92.893	-46.2	1.313	5099.1	no
1	39	services	single	high.school	no	no	no	telephone	may	fri	346	4	999	0	nonexistent	1.1	93.994	-36.4	4.855	5191.0	no
2	25	services	married	high.school	no	yes	no	telephone	jun	wed	227	1	999	0	nonexistent	1.4	94.465	-41.8	4.962	5228.1	no
3	38	services	married	basic.9y	no	unknown	unknown	telephone	jun	fri	17	3	999	0	nonexistent	1.4	94.465	-41.8	4.959	5228.1	no
4	47	admin.	married	university.degree	no	yes	no	cellular	nov	mon	58	1	999	0	nonexistent	-0.1	93.200	-42.0	4.191	5195.8	no

As you can see, the some of the data "features" include the following (there are 20 in total):

- Age
- Job
- Marital Status
- Education
- Default
- Loan

You will be using these features to predict the output variable "y":

- "y" - The classification goal is to predict if the client will subscribe (yes/no) a term deposit (variable y).

NOTE:

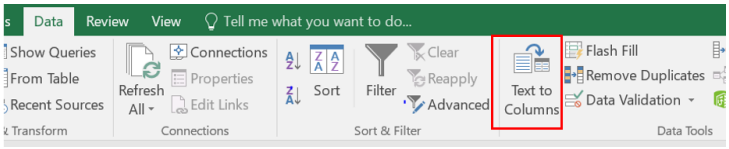
- The **"duration" feature should NOT** be included in the data used to train your classification model – this is to enable us to build a realistic predictive model.
- This duration feature highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known.

3. Pre-requisites

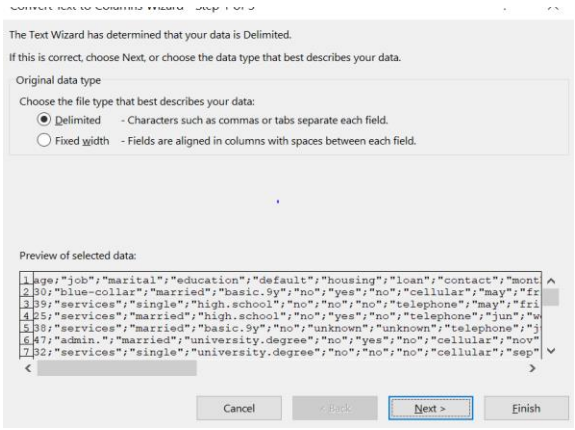
Months 1-3 of “Radar AI Tech Stream - Semester 1” should be completed prior to undertaking this project.

4. How to get started

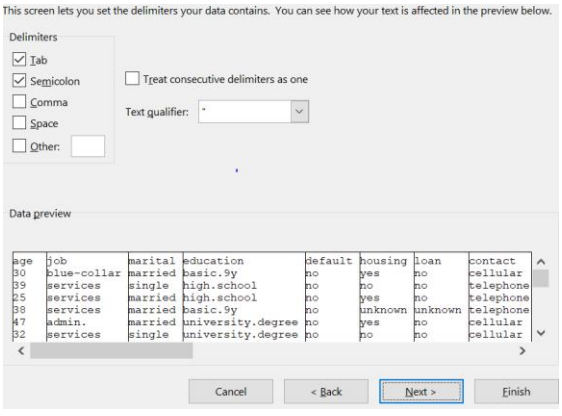
- Download the dataset using the link provided and unzip
- Convert from csv text to columns with excel:
    - Open the bank-additional-full.csv file using excel
    - Select the whole of Column “A”
    - Go to the “Data” Tab and click on the “Text to Columns” button



    - Select the “Delimited” button and click “Next”



    - Select the “Tab” and “Semicolon” delimiters and click Next



    - Click Finish
- Take some time to understand the data
- Open a Google Colab notebook and upload your converted csv file
- Import the “Pandas” Python library and then Import the csv data into Python using Pandas

5. Some things to think about

- Do we have a large dataset? Should we use Cross-Validation?
- This is a binary classification problem, what classifiers are good at solving this type of problem?
- How are we going to get the data into a format that a machine learning model can understand it?
- Do we need to use all of our features? Will the "pdays" help us? If so, how?
- How to handle the categorical variable e.g. one hot encoding
- Consider if the dataset is imbalanced i.e. if there is an unequal distribution of classes within the dataset (e.g. number of "no" >> number of "yes"); how can you handle this problem?
    - See https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/ for different techniques o handle dataset imbalance
    - I would suggest utilising Under-sampling or changing the performance metric for consistency.

6. What to submit

Submit an F1-Scoreaccuracy score from a random test set that makes up 20% (or more) of the data. Also submit your code for feedback! I'll share the best submission(s) with the group so we can all learn from them.

7. Conclusion

This is intentionally a very open ended task – as most real world problems are. There's no right or wrong way of doing this. Having said that, some strategies will result in better accuracies than others.

There are many good solutions to this problem scattered around the internet. I suggest trying to solve it yourself first before using them for guidance.

Enjoy and best of luck!

8. Deadline

Submission deadline Monday 5th April 6am. We will then be moving on to Deep Learning.