FOUNDATIONAL BUSINESS ANALYTICS

COURSEWORK 2019-2020

Release Date: 26th November 2019 (n.b. You will need to download your own specific dataset from Moodle)
Deadline Date: 17th December 2019, 4pm
Submission: via Moodle coursework submission link on the FBA module web page

1. The Problem Definition:

IMBA enterprise is expanding its operations into the banking sector. Their first goal is to build up capital, so IMBA enterprises have come up with a plan: to market a financial product that offers an attractive fixed term interest rate to people who make a sizable deposit (but which can't be withdrawn for a year) - the **"IMBA Platinum Deposit"**.

Cold-calling potential customers to sell this new product is a necessary but difficult and time-consuming task. Only a relatively small proportion of individuals who the company contacts will be receptive to their marketing, making clear how important it is to identify ***which people*** should be targeted. This is the business analytics task you must complete for this coursework, and submit in the style of a professional business report - as though being delivered directly to the company.

Luckily, IMBA enterprises have taken over operations of a banking company that offered a very similar (i.e. almost identical) product. In fact the "IMBA Platinum Deposit" is essentially just a rebranding of that product. And this means they have detailed records of previous marketing calls, all attempting to sell that same product that you can use to inform your model building. This prior data includes information about the characteristics and demographics of people who were called, as well as some features of the calls made themselves. And of course, they also vitally have information as to whether the people contacted ultimately subscribed to the product or not.

This data can shed light on the sort of people who IMBA enterprises should target, and how they should go about it. Your task, as a consultant, is to analyse that historical dataset, and generate a model that can be used to predict if any new individual is a good or bad prospect for the telemarketers to contact (sadly it is just too expensive to blanket contact every prospective customer).

As well as robustly testing, justifying and unpacking your selected model (guided by the CEO's needs, as detailed below), IMBA enterprises also want you to produce some business recommendations - what you think the company should focus on as a result of your investigations. You will submit a formal business report (with a strict 8 page *and 3000 word maximum)* . Additionally you will submit your model implementation, with instructions on how to use it to test new data (written in either Python, Orange3 or some combination - formal specifications are detailed below). Good luck!

2. Important Message from the CEO:

"Of course, to management, an overarching goal is to try and contact every person likely to buy the "IMBA platinum" product. We certainly don't want to be missing potential customers lightly! But we must be realistic - we cannot afford to blanket contact everyone potential target, even though only a relatively small percentage of people will be receptive to our new product! And yes, customers can get annoyed when they're contacted via the phone, especially when they have no interest in our product (indeed some may even not want to deal with us again, but this is rare). But this isn't an issue I want to focus on, as our real business cost here will be the expense of fruitless calls to individuals who aren't interested, wasting costly staff time for us. Avoid this if you can."

3. The Available Dataset:

You will have been provided with a dataset in CSV form **from Moodle (**, containing 6000 prior examples of phone communications to potential customers - and whether they subscribed to the product being offered. Note that customers may have been contacted more than once, although no ID for an individual customer is provided - so each call may be considered independently.

Your training dataset can be downloaded from Moodle. Note that your dataset will be different to other people on the module, so you will expect different results. As you will see from the first line of the datafile (which reflects its header), it follows the schema below:

	Type	Name	Feature Description
1	input	**age**	*The called individual's age in years (numeric)*
2	input	**job**	*The individuals declared job role (categorical: 'management', 'blue-collar', 'technician', 'entrepreneur', 'housemaid', 'services', 'self-employed', 'admin.', 'unemployed', 'student', 'retired', 'unknown')*
3	input	**marital**	*The individual's marital status (categorical: 'divorced', 'married', 'single')*
4	input	**education**	*Declared education level (categorical: 'tertiary', 'primary', 'secondary', 'unknown')*
5	input	**default**	*Does this person have credit they are defaulting on - i.e., unable to pay for. (categorical: 'no','yes')*
6	input	**balance**	*What is the person's current balance at the bank if any? (numeric)*
7	input	**housing**	*Has this person taken out a housing loan? (categorical: 'no', 'yes')*
8	input	**loan**	*Has this person taken out a personal loan? (categorical: 'no', 'yes')*
9	input	**contact**	*Contact communication type (categorical: 'cellular', 'telephone', 'unknown')*
10	input	**day**	*Day of the month the individual was last contacted (numerical)*
11	input	**duration**	*Last contact duration, in seconds (numeric).* ***Important note:*** *this attribute highly affects the output target (e.g., if duration=0 then y='no'), but will not be known for future calls. It may be used within analysis (and please do), but should not be used within a predictive model for new customers.*
12	input	**campaign**	*number of contacts performed during this campaign and for this client (numeric, includes last contact)*
13	input	**pdays**	*The number of days that passed by after the client was last contacted in a previous campaign (numeric; -1 means client was not previously contacted)*
14	input	**previous**	*Prior number of contacts performed* ***before*** *this campaign and for this client (numeric)*
17	input	**poutcome**	*Result of trying to sell the individual something on a previous campaign (categorical: 'unknown', 'other', 'success', 'failure')*
18	output	**y**	*The output feature we must try to understand and predict - whether the call to this individual resulted in a sale (categorical: 'yes', 'no')*

4. Formal Task Specification

- You must provide a classification approach to predict which individuals are more likely to subscribe to "IMBAs Platinum deposit" product. This will require a stage of statistical analysis, a stage of model selection, a stage of final model training, and then an analysis of implications. You may use any software you desire for your analysis, but your model must be produced in either python3 or Orange3 for this coursework (or some combination).
- Your submission will consist of a zip of the files for your model, and a report of a maximum of 8 pages. Your model will be tested on a hidden dataset (with the same schema as the training dataset, but without without the features "duration" and (obviously) "y".

Your report **must strictly** adhere to the following sections, but please take into account the marks available for each in structuring your submission:

Section A: Summarization *[10 marks available]*

- ❑ In this section you must provide a summary statistical analysis of the dataset. Consider how each input feature present is related to the output variable ("y"). Additionally, you may want to examine how they relate to each other. Please feel free to use tables, bar charts, or scatter graphs depending on the feature - it is totally up to you. Note, the point of this section is to be informative rather than overloading your client with information, so also summarize the key analytical points you have observed in the dataset.

Section B: Exploration *[20 marks available]*

- ❑ Apply a decision tree to the dataset to unpack, examine, and identify a first-cut through influencing factors in the data. Which variables appear to be important? Do some combination of variables allow you to identify useful sub-populations in the data? Are all variables useful? Discuss this analysis (linking to your analysis in section A if appropriate). You do not have to visually represent the resulting decision tree, but this is highly recommended and likely to aid your presentation of this initial exploratory analysis.

Section C: Model Evaluation *[25 marks available]*

- ❑ Select at least 3 different classification model classes (selecting only from those we cover in FBA lectures: Logistic Regression, Decision Trees, Random Forests, Naive Bayes Classifier and k-nearest neighbours), and assess their effectiveness in modelling your historical training dataset (which is unique to you) against a point predictor benchmark (i.e. the mode of yes/no's). This should be undertaken in either Python3 or Orange3.
- ❑ In your report, detail the models selected to test and why they were chosen. Detail the parameterizations you chose for each model, explaining why you have chosen the parameters that you have.
- ❑ Describe the evaluation strategy you chose to compare models to each other (including evaluation statistics and performance measures as you see fit) justifying your decisions in full.
- ❑ It is expected that your analysis of the outputs of each of the models will be examined in terms of the **confusion matrices** that they produce. Relate this to your choice of performance measure.
- ❑ Any code/files used in this process may also be submitted, to contribute to your code/file submission mark (see

Section D: Final Assessment *[5 marks available]*

- ❑ Given the analysis in Section C, justify a 'winning' classifier and why you have selected it for your final model, paying close attention to the business case in your consideration of measuring success.

Section E: Model Implementation *[5 marks available for write up]*

- ❑ Having selected the single, best performing model, that model must then be trained against the whole training dataset ready for deployment. This section should specify that choice, and briefly describe the resulting code/project files that are attached with your submission. In particular, this section should be used to supply brief instructions on how the recipient should use your submitted model code/files to process a new test set, and to make new predictions from your model.
- ❑ *N.b., marks awarded here are only for your write up/instructions, with more marks available for the assessment model's implementation code/files- see "further available marks"*

Section F: Business Case Recommendations *[5 marks available]*

- ❑ The final section of your report should summarize the business case to the client (IMBA banking), providing business recommendations for further potential analysis.

Further Available Marks:

- ❑ Overall Presentation of your report, its argument and professionalism
  → *[5 marks available]*
- ❑ The standard of your submitted Evaluation/Final modelling code/workflows. It will be expected in this code/workflow you will also have supplied some means for the user to load in new data (in the same format as your supplied dataset) and make new predictions.
  → *[20 marks available]*
- ❑ The Effectiveness of your model as assessed against our held-back test dataset
  → *[5 marks available]*

Note, that the models you submit will be tested on another external dataset, that I have held out (and which you will not have access to, reflecting the fact that these represent "future" marketing attempts). Thus, as well as receiving marks for your report, your model implementation and how well you have tested, evaluated and justified its construction, there are also additional marks for how well it will predict our hidden test set! Your scores will be relayed to you after submission.

6. Submission

→ In your submission please submit a zip of the following files:

1. Your Final Report (maximum 8 pages).
2. Your Evaluation Code / Workflow files and Final Model Code / Workflow

→ Submissions must be submitted via the moodle submission link

→ Submission must be received by: **17th December 2019, 4pm**

Potential Penalties:

→ Late submissions will lose 5% from their final mark per day.

→ Submitted reports over 8 pages will be received, but only the first 8 pages will be assessed. This is a strict rule.

7. Final Important Note on Plagiarism

→ Each of you have been provided with **a slightly different training dataset**, so expect to have different results to other people. This is obviously to ensure you are working individually, and we will test your resulting model on the dataset you were specifically allocated.

→ All code and workflows will also be examined to ensure there is no repetition between submissions, so while you are able to share ideas and strategies, the implementation and analysis must be 100% your own individual work. Any plagiarised work will immediately receive zero marks, and notified immediately to the School.

8. Some Additional Tips!

- Throughout this coursework, showing thought processes and understanding of how you assess a model *in light of the business case* is more important than the final predictive test result.
- Similarly, and as reflected in the mark scheme, illustrating your understanding of robust model evaluation and comparison is again more important than the final implementation for this coursework.
- You may use any analysis tools to formulate your report, but your submitted model must be implemented in Python or Orange (or both). You can assume the recipient is using python 3 and Orange3 respectively, and has sklearn, scipy, numpy, pandas, matplotlib, seaborn installed. Any further requirements must be clearly specified in your submission with instructions.
- Note the page length available in total, and the available marks for each section to assess how much time and effort to place in each.
- Note that presentation of your work is also being assessed. This is a formal report directed to a business professional, and should be formatted and worded accordingly.
- Using python rather than Orange will not necessarily gain you any extra marks. However, it will likely give opportunity to show off your work with more sophisticated analysis, and increase potential of obtaining higher marks in those respective areas.
- If you choose to illustrate a decision tree - do so for a reason, make it visually useful. No-one wants to see a page of 100's of nodes - so think how best to present the insights it holds!