Machine Learning Assignment Spring 2025

Telecom Customer Targeting

Project Objective: Develop predictive models that will identify whether a customer will take out a new contract in response to marketing.

Context:

Note: **this is a fictional scenario. The numbers will not reflect actual customer behaviour.**

Wallace Communications is a telecoms company that has spent the past few years digging up the streets of UK cities to connect people to their network. They would now like to enter the mobile telecoms market so are planning a marketing campaign encouraging existing landline customers to take out mobile contracts. However, call centre costs for the campaign are high, so they want to target customers most likely to take out contracts to avoid wasted calls. This is where you come in. Wallace Communications have data for over 50000 customers with their responses to previous marketing. The data tells us if each customer responded to the marketing by setting up a new contract "new_contract_this_campaign".

Your Task:

You must develop logistic regression, decision tree and neural network models that will identify the customers that will probably take out a new contract. You can use Orange, Python, R, or any machine learning package of your choice. The data for the assignment is in a file **wallacecommunications.csv**, which has been provided for you. The data dictionary is given at the end of this document. You must follow the correct methodology to use the data to build and test your models.

What to Submit:

You must submit a **single page** infographic poster showing the results of your analysis. Create the poster using a word processor (like Word) or a presentation package (like PowerPoint). Set the page size to A3 and use a 10pt font. You can choose the layout, but you must include:

1. Put your student number (not your name) at the top of the poster
2. A list of the steps you took to carry out the project, including details of the train / validate / test split that you chose;
3. A table showing which variables you used and whether your model treats them as numeric (continuous or discrete) or categorical. Explain one consequence of your choices;

4. A single example of how you used a histogram or bar plot to detect an error in the data and what you did to fix that error. State the data cleaning operation you carried out;
5. A table showing the different models and hyper parameters you trained, along with the correct metric for each; Add a sentence on how you chose the hyper parameter values.
6. A justification of the choice of the final model and a confusion matrix showing its results on the correct data split. Add a comment on what the confusion matrix tells us about how useful the model will be.
7. One additional insight about the data or models you gained from the process.
8. One or more references that are used to justify decisions that you made and cite any materials that you used (such as illustrations).

Important: Make sure your poster has the sections listed above. Missing sections will cost you a lot of marks.

Save your poster as a PDF and submit it on canvas. Check the canvas page for the deadline and other submission rules. You do not need to submit your code or Orange workflow.

Marks: Marks will be allocated in line with the university postgraduate common marking scheme, which for the poster can be summarised as follows:

Mark Range	**Requirements / Achievement**
0 – 39 (clear fail)	Many sections missing, no models built, little work completed
40 – 49 (marginal fail)	Most sections completed, but with methodological errors and poor or no interpretation
50 – 59 (pass)	All sections completed with correct methodology, but with poor or no interpretation of results
60 – 69 (merit)	All sections completed with correct methodology, good interpretation of results, good justification of decisions
70 – 100 (distinction)	All sections completed with correct methodology, excellent interpretation of results, good justification of decisions with references, excellent presentation, excellent insight.

Any Questions?

If you have any questions about this assignment, please post them on Teams under “Assessment Questions”. If circumstances outside your control mean you are unable to meet the deadline, please apply for an extension under “Extension Request” on Canvas.

MOST IMPORTANT OF ALL!!

Read this part very carefully. You must work on this assignment completely on your own. Do not ask classmates for advice and do not share your answers with anybody. Do not copy any work from the internet. If you do look up something online, make sure you include a reference to it. Marks will not be given to work that is clearly not your own.

More details on academic integrity can be found here:
https://canvas.stir.ac.uk/courses/14736/pages/academic-integrity-and-academic-writing-support

Statement on the use of AI
For this coursework, the ethical and intentional use of Generative Artificial Intelligence Tools (AI) is permitted (with the exception of the use of AI for the specific purpose of writing or amending code and generate figures, which is not permitted as this assignment tests your ability to write code or visual tools to implement machine learning pipelines).

Whenever AI tools are used you should:

- Cite as a source, any AI tool used in completing your assignment.
- Acknowledge how you have used AI in your work.
- Using AI without citation or against assessment guidelines falls within the definition of plagiarism or cheating, depending on the circumstances, under the current Academic Integrity Policy, and will be treated accordingly. Making false or misleading statements as to the extent, and how AI was used, is also an example of "dishonest practice" under the policy.
- AI tools that might be relevant for this assignment include: ChatGPT and Copilot.
- Please contact Dr Brownlee, Dr Adeel, or Dr for specific guidance.

Data Dictionary:

Each row of this set relates to a customer's interactions during one marketing campaign "this campaign". This is historical data, so for each one we know whether the customer ended up taking out a new contract (that is, responded positively to the campaign). Each customer may also have been reached by a previous marketing campaign, and if so then the details of that contact are recorded under the features with "previous_campaign" in their name.

Variable	Description
ID	unique identifier for this record
town	home town for customer
country	country for customer's home address
age	customer's age
job	customer's job
married	customer's marital status
education	customer's highest educational qualification level obtained
arrears	has the customer failed to pay a recent bill?
current_balance	current amount in customer's landline account in pounds (could be negative)
housing	is the customer a homeowner?
has_tv_package	has the customer got an additional TV and data package on their landline
last_contact	type of communication used for previous call to customer
conn_tr	connection type grouping ID (related to the connection for their landline)
last_contact_this_campaign_month	last contact month of year
last_contact_this_campaign_day	last contact day of the month
this_campaign	number of contacts performed during this campaign and for this client
days_since_last_contact_previous_campaign	number of days that passed by after the client was last contacted from a previous campaign (-1 means the client has never been contacted before)
contacted_during_previous_campaign	number of contacts performed before this campaign and for this client
outcome_previous_campaign	outcome of the previous marketing campaign
new_contract_this_campaign	has the client taken out a new contract? [target variable]