Background
The data science team from one of our clients have asked us to do an independent model development and buid by making use of the same underlying dataset
which they have used in order for them to do a review and establish whether their inhouse developed model and approach (steps) are sound
The client has not been prescriptive regarding certain criteria, but has given us some guidelines for us to check against as we build out our model.

The main aim of the model is to classify whether a customer will take up the sales offer or not when we cold call them, so the better we discriminate between who will be more likely to
take up the offer and who will not, can help our client to better optimise their internal efforts in terms of cold calling the right customer in order to convert those calls to sales
The secondary aim, is to overlay the revenue / cost impacts based on our classification outcomes, in order to evaluate what the estimated impact will be on the financials by executing campaigns
using this model for selection purposes and what it will yield and cost us

The Dataset
The dataset contains 20 features (10 numerical and 10 categorical) and one response (binary) variable
There are 35,000 observations which can be used for model devlopment and testing purposes
The dataset is pipe (|) delimited

Things to consider
- Check whether all the input variables are relevant to be considered for use in the model
- Can you explain what your data looks like and whether there are some outliers or corrections which have to be done
- How do you check whether there exist any discriminating power amongst the variables
- How do you control for high correlations amongst your variables
- Do you care whether the response variable classes are balanced or not
- Do you mind having variables all at different scales
- How do you encode categorical data and do you group any of these
- Do you bin your numerical data or not?
- Do you split your data for training and testing?
- Do you perform validation checks as you train your models
- How do you decide which model is the best one to use? (Show or use at least 1 non-linear model in the comparison)
- Given the best model you selected, how do you interpret the model predictors?
- How do you ensure the model does not over / underfit the data and that it can be used for prediction purposes?
- How do you know which input variables are the most and least important?
- Do you fit different models and how do you decide which one to optimize?
- How do you optimize your selected model? Why did you choose to optimize it that way?
- How do you settle on which cut off hurdle to use in terms of classifying new data into a 1 or 0?
- How do you interpret the confusion matrix obtained by the final model?

Bonus question:
The Finance, Sales and Risk teams have a vested interest in trying to figure out the potential impact a sales campaign will have to see that we can justify the cost / benefit of running the campaign
The sales team have obtained a list of prospective customers from a data provider which contains 10,000 records.
The expectation is, that we will be utilising the newly created response model you have developed and apply it to the list in order to identify who we should call and who we shouldn't
Assume for now, that the predictions in terms of the confusion matrix of the 10,000 campaign list, follows the same percentages as per your development sample in terms of the confusion matrix

Assuming 100% approval rates (i.e. everyone we call qualifies for a credit product), the risk team has overlaid their predictions on the 10,000 customer list,
and they have classified the customers accordingly into the following 3 risk bands

	Risk distrubution
High	10%
Medium	25%
Low	65%

The Finance team has further broken down the net profit estimates for each of the risk bands based on whether we have a successful offer taken up (sale) or not (not interested)

	Taken up	**Non taken up**
High	285.00	- 300.00
Medium	705.00	- 300.00
Low	1,225.00	- 300.00

By using your confusion matrix, can you calculate the
a) Expected net profit we stand to make by running this campaign on the customers we intent to phone?
b) What is the lost opportunity of net profit due to us misclassifying customers by using your response model?