Uplift data science challenge

I. DATA CHALLENGE : PART I

Uplift is a lending company that finances trips sold through partnered travel providers. Prospective travelers browse travel provider webpages or interact with provider agents, and are presented the opportunity to apply for an Uplift loan at the point of sale. The requested amount to borrow (the "loan amount") is natively the cost of the trip selected, but Uplift may desire to lend more or less money to the applicant, depending on an assessment of their likelihood to repay the loan. Determining optimal loan amounts for applicants is critical to both promoting business with good borrowers as well as mitigating losses from risky borrowers. A data set of historical Uplift borrower characteristics is provided. Using this data, demonstrate your best efforts at the following tasks:

- Identify or derive strong loan default predictors.
- Create a procedure to determine the loan limit for a given applicant.

The rows of the data set represent borrowers, and the columns represent features that you will be using in this exercise. The data set has a total of 22 columns, with the names unmasked for two features:

- default : A binary variable which predicts if a borrower defaulted or not.
- order_amount : The order amount that was requested and issued as a loan to the applicant.

The remaining features are named $\{fn\}_{n=1}^{20}$ and have been masked.

You are free, but not required, to develop any machine learning models that may facilitate your analysis. Use of Python or R is encouraged. Please report your findings in Jupyter Notebooks, R Markdown, or Microsoft Word with embedded figures/tables. Submit your report as a pdf or html file within two weeks of receiving this challenge. Note that there is no *right* answer to this challenge. We are interested in your thought process, analysis approach, and data interpretation. Whether or not you find strong predictors is secondary.

II. CONCEPTUAL QUESTIONS [BONUS] : PART II

Following is a set of conceptual questions to supplement your report from part I. Answering these questions may assist your work in the prior part, but doing so is not necessary. The majority of your evaluation will be based on your performance in part I.

- If our objective is to limit the loan amounts offered, what is the anticipated relationship between 'order_amount' and 'default'?
- Suppose you are tasked with building a logistic regression model predicting 'default'. What kind of relationship between default probability and an ordered feature would you want to observe before including the feature in the model? (Hint: Remember that logistic regression models use features in a linear manner.)
- Would you face similar constraints on input features when training a tree-based model? Why or why not?
- Given a binary classification model predicting 'default', how would you use the model to set the loan limit? Are there limitations to this approach?