# **RVU Data Science Interview: Technical Assessment**

## **Problem**

You are on the RVU data science team and helping a financial services partner. The partner's primary product is a high-value credit card associated with high customer lifetime values (LTVs). The partner is looking to reduce losses from customer defaults by offering a new, lower-value card that supports balance management and helps customers avoid default.

The partner has asked RVU to help increase overall customer LTVs by downgrading risky customers to the new balance management card while keeping good customers in the high-LTV traditional card.

**You are tasked with identifying which customers are likely to default** under the traditional credit card plan and should be downgraded to the balance management card.

## **Timing**

**This task should be time boxed to 4 hours.** *This is a recommendation, we don’t want candidates to feel like they need to spend days on this task!*

You should submit your deliverables as a single, compressed archive, labeled <name>.zip

For example, Jennifer\_Gosset.zip.

This archive should be **emailed to your RVU Talent Team Contact.**

## **Data**

### **Description**

These data consist of attributes about 30,000 traditional credit card customers and their credit card billing and payment history from April to September. Additionally, their default status in October is given. For assessment purposes, these data have been partitioned into 2 sets: - train.csv - test.csv

train.csv includes 80% of the original data with labels.

test.csv contains 20% of the original data without labels.

You will use test.csv to make predictions for submission.

### **Codebook**

* limit\_bal: Amount of the given credit in dollars: it includes both the individual consumer credit and his/her family (supplementary) credit.
* sex: Gender (1 = male; 2 = female).
* education: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).
* marriage: Marital status (1 = married; 2 = single; 3 = others).
* age: Age (year).
* pay\_1:pay\_6: History of past payment. We tracked the past monthly payment records (from April to September) as follows: pay\_1 = the repayment status in September; pay\_2 = the repayment status in August; . . .;pay\_6 = the repayment status in April. The measurement scale for the repayment status is: -1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight months; 9 = payment delay for nine months and above.
* bill\_amt1:bill\_amt6: Amount of bill statement in dollars. bill\_amt1 = amount of bill statement in September; bill\_amt2 = amount of bill statement in August; . . .; bill\_amt6 = amount of bill statement in April.
* pay\_amt1:pay\_amt6: Amount of previous payment in dollars. pay\_amt1 = amount paid in September; pay\_amt2 = amount paid in August; . . .;pay\_amt6 = amount paid in April.
* default\_oct: response in yes,no of default in October

### **Scoring**

You are asked to make predictions on an unlabeled test set test.csv. The predictions should be the probability of defaulting

Pr(default=yes) in [0, 1].

## **Deliverable**

How will this task be assessed?

First, you will be objectively scored on how well your classifier is able to predict a customer's likelihood to default.

Secondly, you will participate in a code review with members of our team to assess your programming skills, methodological choices, and recommendations for solving the business problem.

Your submission archive should contain:

1. <name>\_predictions.csv Your predictions on the test set. Format is described below.

2. code/ a directory containing the code you used to make predictions. This code should be reproducible. That is, given your code and the train and test data sets, we should be able to exactly reproduce your predictions CSV.

***Presentation***

If the candidate is successful in this task submission - the first section of the next stage of the interview process will be to use the work you have done in this task to prepare a presentation.

You will prepare a short deck, and present your findings. Imagine that this presentation is to a non-technical, business audience. Specifically, you should address the following:

* Describe the problem to a general audience
* Describe what you found in your analysis
* Make recommendations to decision makers going forward

Your presentation should range from 5 - 10 minutes, and you should expect 20 minutes of Q&A to follow.

### **Submitting Predictions**

As part of your submission archive, create a CSV with two columns [customer\_id, pr\_y] that contain your predicted probabilities for each customer\_id in the test set.

The file-name should follow the format <name>\_predictions.csv, for example Gosset\_predictions.csv.

Three files are included to help ensure your predictions are in the correct format: - example\_predictions.csv - make\_submission.R - make\_submission.py

example\_predicions.csv is an example of what your submission CSV should look like. Additionally, convenience functions for both python and R have been provided to turn data frames into the appropriately formatted CSV.

### **Code**

You may use anything you would like to make the predictions, however whatever you do should be reproducible. That is, given your code and the train and test data files, we should be able to exactly reproduce your predictions CSV. You may wish to include a README that describes how to reproduce your predictions, given your code. We are not expecting that you include any software with your submission.