Data Science Technical Task

Please find attached four csv data files:

1. ‘ds-tech-test-credit-file.csv’
2. ‘ds-tech-test-ob-transactions.csv’
3. ‘ds-tech-test-demographics.csv’
4. ‘ds-tech-test-target.csv’

The first contains summarisations of credit file information as of May 2022, the second contains open banking transactions, the third some demographic data on users and the final file contains a flag referring to the delinquency or otherwise of users. The time period is six months (training data: 2021-12 till 2022-05, target: 2022-06 till 2022-12). \*

We would like you to use the data provided to tell us about the relationship between delinquency and the other data we have shared. We recommend this take the form of a model, but it may be another form of analysis if you wish (please explain rationale if so).

The assessment is designed to assess model building capability and ability to apply this critically to a business problem. As part of your analysis, please consider answers to the following questions:

1. Why have you chosen the target variable that you have of the set? What is the impact of this compared to choosing one of the other targets?
2. What are the most important features of your model?
3. What is the relationship between the features and the target? And why does this relationship matter?
4. What do you expect would be done with this model? How might ClearScore or one of the lenders it works with use this model?
5. What are the downsides of your model? Are there any blind spots or model risks?
6. How might you enhance the modelling if given more time or data?

If you are invited to interview, this will consist of:

1. Understanding the procedure you followed (this could be thought processes as well as physical acts)
2. Discussing the decisions you made and the thought processes you followed
3. Explaining the results you have seen and any implications

We also recommend you think about how you would explain your solution to a layperson/non-technical stakeholder. You can prepare a presentation to support this if that helps. If you wish to share code, then it must be in Python.

Finally, remember that the test is supposed to be fun, so please don’t stress about it!

*\* Please refer to the next page for additional information on the datasets provided*

*Some definitions that may be useful:*

* ***Balance Amount*** *– how much is the balance of the account at a point in time*
* ***Credit limit*** *– the amount of credit that is allowed to be taken for that account (as you would get with a credit card)*
* ***Statement Balance Amount*** *– for account types that issue a statement, the last recorded balance amount on the statement*
* ***Statement Payment Amount*** *– for account types that issue a statement, the last recorded payment amount against the statement*
* ***Status One, two, three etc*** *– how many payments the user has missed for that account to date*
* ***Utilisation*** *– the amount of credit used divided by the credit available (as a percentage)*
* ***Category Confidence*** *– for banking transactions, the level of confidence that the category is correct*
* ***One Plus in Window*** *– the user has missed one or more payment in the target period*
* ***Two Plus in Window*** *– the user has missed two or more payments in the target period*
* ***Default in Window*** *– the user has defaulted in the target period*