**Technical Interview**

**Background**

Feel free to use any online or other technical resources to support your analysis and submission, but please do not solicit direct advice or help on questions. This technical interview is meant to be completed independently and without assistance on any part of the interview.

**Instructions**

This interview will test the three areas of technical expertise that we often encounter on data science projects –feature engineering, quantitative modeling, and analysis of results. We are looking for candidates that have knowledge and skills across all three areas with distinct expertise in at least one area. Please work in Python.

Build a model to predict the riskiness of a lender (i.e. the probability of default) given historical data. A complete test requires 2 components:

1. All code used to ingest and manipulate data, build models, make predictions, and analyze results
2. A written summary of results and analysis
3. Please complete all main questions before moving to “BONUS” questions.

**Data**

The data provided in the file ‘raw\_data.csv’ is on a customer-level basis and the column titled ‘label’ is the dependent variable (whether the customer defaults within 2 years of opening an account). The other independent variables are attributes of the customer at the time he/she opened the account.

The data dictionary are the definitions of all variables.

**Feature Engineering**

Once you have constructed your initial dataset, perform any data manipulations you think are necessary to develop a reliable predictive model (e.g. imputation, normalization, deduplication, etc.).

**BONUS:**

* Visualize your features and explain why transformations were necessary
* Create additional features that can be used in your model that are not simple manipulations of the raw data. Explain why you think these new variables will be important

**Quantitative Model**

Build a model to predict the probability that a loan application will default. Be sure to make predictions on the test data.

When constructing your model, please explain your approach in detail, including

* Why you selected the ML algorithm
* How you evaluated your model during training
* How you selected your final model to predict the hold-out test data
* Your process for variable selection
* The steps you took to prevent overfitting

**BONUS:**

* Visualize your results and illustrate the strengths and weaknesses of your model

**Results Analysis**

Evaluate the validity of your model and explain the accuracy metrics you used. In addition, to any metrics you use, please include a gains table/lift chart.

Please explain your analysis in detail, including:

* Create a function to build a gains table. The inputs should be predicted probability, the actual label, and the number of bins. Please explain what lift and gains mean in terms of business implication.
* How do you measure the effectiveness of your model? Discuss reasons for choosing one metric over others.
* How would you explain feature importance to a non-technical business stakeholder?
* Describe some of the business impacts of your model; for example, what are the characteristics of high default probability applicants? Is there anything undesirable about how your model assigns default probabilities?

**BONUS:**

* How would you adjust your model to address any issues?
* Create a function to find the AUC