**Candidate technical review questionnaire**

1. **Business ask**

You are in a meeting with a product owner, they just launched an additional customer care center of excellence (CoE), and have opportunity to cross-sell/upsell customers with additional products. They need your help in identifying which customers to pitch to and what product.

At your disposal is historical data for the last 12 months of customer upgrades, their demographic and transaction data. The business has 4 products - video, data, home security and Voice. A customer might be subscribed to a single product, a combination of two or three products or have all 4 products.

a). Describe how you would go about providing a solution to the business? What modeling approaches would you use? How would you measure the effectiveness of your solutions?

* Find correlations between customer info (demographics, transactions, original product subscription) and the upgrade
* See if there is a “signal” that the customer may upgrade soon using the historical data. Plot transactions over time and see if there is any specific pattern in transactions that preceed an upgrade. If we can detect a pattern, this might help us define some input variables for a model to predict when a customer is likely to upgrade.

Classification to predict likelihood of upgrade to ‘x’ product

Train the model with a large portion of the historical customer data (for example, 70% of customers). Score the remaining customers (i.e. remaining 30%) and measure the performance with metrics such as accuracy score (add definition), AUC, Lift chart.

**b). Now assume, the business partner has prepared a pipe delimited dataset for you. The dataset contains 128 variables and 300,000 records. The dataset contains records of customer that were serviced by the CoE in the last 9 months, and the analyst created a binary indicator (1 = YES, 0 = NO) of which customers were cross-sold as the target variable. The remaining independent attributes are based on the snapshot view of the customer’s data as of the time of service at the CoE.**

**What modeling approaches would you use?**

The aim is to predict a binary target variable (1 = yes, 0 = no). Therefore, I would recommend trying a random forest classification or a logistic regression. With a tree-based method like random forest classification, we can fit an initial model with minimal variable transformation. This will allow us to see if we can get any predictive power and also look at variable importance from the tree based method. An alternative option may be to use logistic regression, but this may require quite a bit more variable transformation for any variables that are not linearly related to the target.

**What would your output to the business look like?**

Output to the business would be in the form of a write-up or presentation with visualizations to show the important insights in the data. Generally, I would include the following information:

* A high level overview of the business objective and the modeling approach.
* Graphs to clearly to relay any relevant patterns we saw in the data (i.e. where the target=1 group differs from the target =0 group).
* The final model definition, with variables listed in order of importance.
* Model performance (i.e. how accurate is the model prediction)
* Recommendations for model’s business usage going forward, the frequency of the model scoring process, and plans for ongoing model validation.

**How would you measure the effectiveness of your solutions?**

Initially, I would measure the effectiveness of the model by training the model on 70% of the customers, and predicting values for the remaining 30% (test data set). We can measure the effectiveness of the classification in several ways:

* Accuracy
* Recall
* Precision
* ROC AUC
* Lift Charts

**Based on your analysis, what recommendations and or insights do you have for the business?**

I’d recommend that we collect the input data at a monthly level for all customers, instead of using a snapshot at a point it time. This would provide the option to look at the customer’s behavior over time and possibly detect signals in the changes in behavior that correspond to the event of cross-sell & upgrade.

1. **Logical / process questions**

**If you randomly type a 6 digit number on a note, what is the probability that you can see the same number if you flip your note upside down? How would you explain your answer to a 6 year old?**

* There are 10 total digits: 0123456789
* 3 of the digits look the same when you turn them upside down: 0, 1, and 8
* If we want to type a 6 digit number and make sure it’s the same when turned upside down:
  + The first 3 digits of the number can be any combination of 0’s, 1’s, and 8’s.
  + The second 3 digits (digits 4,5,6) must be the reverse of the first 3 digits. (A palindrome)
  + For example: 018810
* The probability is as follows:
  + Probability that a random digit is a 0,1, 8 = 3/10
  + Whenever you want to get the probability of more than 1 event occurring, you multiply the probability of each event together. I.e. prob(event1)\*prob(event2)\*prob(event3)
  + So, the probability that you get a 3 digit number made up of 0’s, 1’s, and 8’s is (3/10)\*(3/10)\*(3/10) = (3/10)^3
  + The next 3 digits (digits 4,5,6) have to be in the exact reverse order of the first 3 digits, so there is only 1 possibly choice for each digit.
  + This means the probability is (1/10)\*(1/10)\*(1/10) = (1/10)^3
  + Now that you have the probability of the 1st 3 digits and 2nd 3 digits, you can multiply them together. The final answer is (3/10)^3 \* (1/10)^3 = 0.000027 or 27/1000000
* To explain it to a 6 year old, I would draw the example on a piece of paper, so they can see what numbers look the same when turned upside down. I’d also simply the explanation quite a bit by describing probability as “the chance that this could happen” and instead of talking about “multiplying”, I could explain it as 3/10 “3 times” and 1/10 “3 times”.