**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?**

* First, I would always look to meet with both the business team and the data engineering team, to ensure a thorough understanding of the business needs and the underlying data source.
* Second, always recommend running exploratory analysis on the dataset provided. This helps to find any patterns in customer behavior, and any relationships between demographics or transactional data and the upgrades.
  + Some examples of exploratory analysis to run:
    - Scatter plots of independent variables over time for customers who upgrade and customer who do not upgrade (stay with same product). This may help detect any changes in transactions or behavior to find a “signal” that the customer may upgrade soon. If we can detect any notable differences in behavior patterns between customers who upgrade and those who don’t, this may help us define the variables to use in the modeling process.
    - Histograms of each variable for those who upgrade and those who do not, checking for any differences in distribution. This may help with feature engineering by determining how to slice/dice the data.
    - Analyzing the above for different candidate target variables:
      * Overall Customer Upgrade (Yes, or No), mainly for insights
      * Intended target variable: Customer Upgrades for product ‘x’. i.e. a multiple classification problem, where we look to classify likelihood to upgrade for each product.
    - Run correlations between customer information (demographics, transactions, original product subscription) and the target
* After analyzing the data, I would start the modeling process (see details of modeling approach below)
* Finally, I would provide a recommendation on how the business can use the model. In this case, we would predict which customers are likely to upgrade for each of 4 product types. If a customer is likely to upgrade to 1, 2 or 3 products based on the model’s prediction, then we can target cross-selling to that customer for the appropriate combination of products.
* As a next step, to continue improving on the cross-sell process, we may want to collect additional data points such as the channel used for cross-sell: email, text message, phone call, direct mail, etc. With that additional data, we may also be able to measure which channels of cross-sell are most effective.

**What modeling approaches would you use?**

Since we are looking to identify which customers to pitch to and for which product, I would recommend using a multi-class classification algorithm. A multi-class classification means we are identifying members of more than 2 categories (i.e. we have 4 products). I would recommend a 4-class model so we can identify who is likely to upgrade to add video, to add data, home, and voice. It’s possible that a customer may end up scoring high to upgrade to 1 or more products, in which case we can use the results to see which combination of products to pitch to them.

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

In the model development process, I would train the model with a large portion of the historical customer data (for example, 70% of customers). Then, I’d score the remaining customers (i.e. remaining 30% will be a test dataset) and measure the performance with metrics listed below.

Ideally for all of the below metrics, values close to 1 (100%) are the best:

* Create a Confusion Matrix, and calculate the following:
  + Accuracy: % of customers correctly identified for each product upgrade
  + Recall: Of the customers who upgrade to each product, % of customers we identified
  + Precision: Of the customers who we predict will upgrade to a product, % of customers who actually did upgrade to that product.
* Plot and calculate value for ROC AUC (Receiver Operator Characteristic, Area Under the Curve): Probability that our model predicts better than random selection
* Lift Charts: graph showing the probability a customer upgrades to each product on the y-axis vs. model score on the x-axis (usually, binned into deciles). Good lift charts show low model scores (close to 0) with low probability to upgrade, and high model scores (close to 1) with high probability to upgrade.

Following initial model build, we can also discuss ongoing monitoring of the model’s effectiveness by implementing a validation process. Every month, or every quarter, we could refresh the customer data and re-assess the accuracy of the model to determine if we should re-fit the model. Refitting the model will allow us to account for any significant changes in customer behavior or any changes in upstream data processes.

**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 in this case, as many variables are categorical and others may not be linearly related to the target. Given more time on the project, I would try both modeling techniques to see which performs better to solve our problem.

**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/Visual aids 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 next steps, including: 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

Note: see question 1’s response for definitions of the above metrics. This is similar, with the difference being that we are measuring performance for a binary target (2 classes) vs. a multiclass target in the scenario for question 1.

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

* I’d recommend the following next steps:
  + Perform some additional development for feature engineering and feature selection so that we can find the best model with the given data without overfitting.
  + Request input data from the data engineering team 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. With this, we may be able to get a stronger prediction.
  + If the 300k customer records are only a sample of our customer’s nationwide, then we can consider up-sampling the target customers (customers who upgrade). In the existing dataset, only 2.3% of the customer base upgrades. This leaves us with a relatively small sample size in the target=1 category, so we can try and improve the model’s prediction by sampling more of those who upgrade.
* To further improve the analytics, if the business can provide some measures around cost and revenue-add for correctly and incorrectly identifying a customer for upgrade, then we may be able to better solve an optimization problem for the business. For example, if we know the cost for false positives (attempting cross-sell and failing) and false negatives (missing out on cross-sell), then we can optimize our model accordingly. We can also take into account the estimated revenue for true positives (i.e. attempting cross sell and succeeding). Using the model metrics and these measure of cost & revenue, we can also build an equation for the expected value of our model’s result.

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, or 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”.