**Shipt – Interview Challenge – Data Analyst**

**Required Questions**

**Q1. Join these two data sets by “date” and “source\_id”, returning all rows from both regardless of whether there is a match between the two data sets.**

Ans: Since we need all rows from both InterviewData\_Cost and InterviewData\_Rev regardless of whether there is a match between the two data sets, I have joined them using outer join on 'data' and 'source\_id'.

**Python:**

Graphical user interface, text, application

Description automatically generated

**Q2. Join these two data sets by “date” and “source\_id”, returning only the rows from the “Cost” file that have no corresponding date in the “Revenue” file.**

Ans: Here, I have used left join to join the two data sets. As we need all the rows from InterviewData\_Cost, it is placed on the left side of the query.

**Python**:

Graphical user interface, text, application

Description automatically generated

**Q3. Using your result from #1**

1. **What are the Top 4 sources (“source\_id” values) in terms of total revenue generation across this data set?**

Ans: Top 4 sources (“source\_id” values) in terms of total revenue generation across this data set ['PA0527', 'PA0308', 'PA0352', 'PA0552']

**Python**:

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

### b. How would you visualize the monthly revenue for those Top 4 sources?

**Plot:**

**Chart

Description automatically generated**

**Q4. Assuming you’ve read the data into an object called activity\_data, run the following code to build a basic logistic regression model:**

Ans: The accuracy of the model is 58%.

**Python:**

**Graphical user interface, text, application, email

Description automatically generated**

**Q5. Split the data into training and test samples, and build a model over the training data using the following code:**

**Python**:Graphical user interface, text, application

Description automatically generated

**Assess the training data model’s accuracy on the test data. Why does the accuracy change so much?**

The accuracy of the model was 21%, which was less than the accuracy in ques.4. This is because in the previous question, the model was evaluated on the same data on which the model was trained i.e. the activity\_data. So when we evaluate the model that we trained we get high scores, this shows how well our model learnt from our training data.

However, in question 5, the model is evaluated on a new data set and so the accuracy of the model is reduced.

One of the major reason for low accuracy is overfitting. Overfitting models the training data very well.

It takes place when a model learns the detail and noise in the training data well and negatively impacts the performance of the model on new data. This means that the noise or random fluctuations in the training data is picked up and learned as concepts by the model. But, these concepts do not apply to new data.

**Q6. This data comes from a subset of userdata JSON blobs stored in our database. Parse out the values (stored in the “data\_to\_parse” column) into four separate columns. So for example, the four additional columns for the first entry would have values of “N”, “U”, “A7”, and “W”. You can use any R functions/packages you want for this.**

**Python:**

Graphical user interface

Description automatically generated with medium confidence

**Additional Questions – Pick One**

A) Within our web and mobile apps, members can generally find items through search and/or the product category tree (note that you can also search after clicking into a product category, in which case the search is filtered by the chosen category). Let's say that we decide to test a different product category tree. The Product team asks for your help in setting up the test and calling the results. How would you help them: (i) figure out how long we should run this test; (ii) decide what metric to measure; (iii) and then evaluate the test?

**Ans: How would you help them:**

The new product category tree can be tested with the help of A/B testing. A/B testing refers to a randomized experimentation process where two or more versions of a web page, or a page elements are shown to different segments of website visitors at the same time to determine which version leaves the maximum impact and drive business metrics.

This is one of the simplest ways to understand the performance of any website using statistical analysis while spending less time and money.

**i. figure out how long we should run this test:**

The test should run at least for one complete week. This is because for a few websites the conversion rates can be low during weekdays and can increase over the weekends and vice-versa. Considering that the web and mobile app is for Shipt, it can be possible that working people visit the website during weekends, or stay-at-home mothers can visit the website during weekdays. So, to get a valid test data, test should run throughout the week so as to include all possible fluctuations. The duration will also depend on the website traffic. If the traffic is lower, the test will have to run for a longer time.

**ii. decide what metric to measure:**

One of the most important metrics will be click through rate. This will give the percentage of people that clicked on the search product category. This helps to measure the success of marketing efforts. Other metrics would be the bounce rate i.e. the percentage of visitors who clicked on the product category but did not stay there and left, the conversion rate, and number of people who added their products into the cart after searching.

**iii. evaluate the test?:**

In order to evaluate the test, the different metrics that have been measured through the Website Optimizer should be considered. For instance, a higher click through rate shows that the search category is engaging and people are interested in clicking and navigating through the category. A 2% click-through-rate is usually considered good. Similarly, if the bounce rate is high then it shows that the visitors did not find the page or content attractive and so did not stay for a long time.