Take Home Assessment

Data Strategist Role
Retina AI

From the time you receive this assessment you will have 48 hours to complete and return this assignment. The goal of this assignment is to test your ability to gather, process/analyze and visualize data. In addition, we are also looking at your ability to apply statistical and machine learning techniques to deduce conclusions from the data and use language understandable by non-technical individuals.

Please provide a Powerpoint, Word or PDF document which answers the questions below. In addition, please attach the source code for your analysis (ideally, this should be SQL, Python and/or R code). We specifically want to see the code you use to read in the files, create plots, perform statistical analysis, etc. Also note that not all of the data is required to do the analysis.

1 Queries [10pt]

Consider the following excerpt from a hypothetical data set in a SQL database. The `page_table` contains all the pages seen by visitors to our website. There are three columns: `visitor_id`, `visit_date`, and `page_name`.

visitor_id	**visit_date**	**page_name**
10512	01/01/2014	home
10512	01/01/2014	product
10512	01/01/2014	checkout
10512	01/01/2014	confirmation
10692	01/02/2014	home
10692	01/02/2014	product
15391	01/03/2014	home
16239	01/04/2014	home

Using pseudo-SQL, write commands and share your imaginary results for the following queries.

i) Sample 10 rows at random,

ii) Get the list of unique `visitor_id`s who land on `home page`,

iii) Get the list of 100 visitors who saw the most unique pages.

2 Analysis [25pt]

For this question you will analyze some anonymous sales data (link), which is stored as a ZIP file of 50 CSV files with time-stamped purchases by customer. Each row has two columns:

- `sale_time`: Time-stamp of sale *e.g.*, 2012-10-01 01:42:22,
- `purchaser_gender`: Gender of the person who purchased (male or female).

Using this data, perform the following tasks:

i) Plot the number of sales by day for all 50 weeks,

ii) It looks like there has been a sudden change in daily sales. On what date was that sudden change?

iii) Is the change in daily sales at the date you selected statistically significant? If so, what is the p-value?

iv) Does the data suggest that the change in daily sales is due to a shift in the proportion of male-vs-female customers?

3 Modeling [50pt]

For this problem, assume you are representing Retina as a Data Strategist. Specifically, you have been summoned by the client to answer their questions and present the results. We would recommend putting together no more than 10-12 slides total. Keep in mind that that the Executive Team and the Data Science Team have different objectives, and to craft your presentation accordingly.

Here you will analyze some customer data (link) for a fictitious e-commerce company that provides animal cookies. Some customers subscribe to receive cookies every month, and others just buy cookies one order at a time. A description of each field can be found in Appendix A.

Each customer is represented by a single row. The fields represent purchase behavior, geographic and marketing data, and the output of a (hidden) model predicting future dollars spent by the customer (`predicted_total_ltv`).

We've provided some points of discussion to include in your presentation:

Executive Team

i) Who are my best customers? What characteristics do they have? Do they exhibit certain behaviors?

ii) Customer acquisition costs are rising quickly. What should I do to acquire higher quality customers?

Data Science Team

iii) What are the greatest predictors of high LTV? (Use `predicted_total_ltv` for LTV values.)

iv) If you built a model here, how do I know how "good" the model is? If another approach was taken, how confident should I be in your findings?

4 Visualization & Narratives [15pt]

Consider the following (hypothetical) employee organizational data set (link); it has three columns:

- `employee_id`: Unique ID for each employee,
- `manager_employee_id`: ID of the manager for the employee (this also shows up in the first column),
- `employee_name`: Full name of employee.

Create a visual plot of the organization chart. Send over the plot, and use it to answer the following questions:

i) Who is the head of the organization?

ii) What is the maximum depth of the organization?

iii) What is the maximum team size?

A Animal Cookie Customer Data

Here we describe how to interpret each field of Animal Cookie's customer dataset. The columns are defined as follows:

- `user_id`: unique ID on a customer level
- `customer_type`: type of customer
- `first_order_type`: type of a customers first order
- `region`: region of customer address
- `subregion`: subregion of customer address
- `state`: state of customer address
- `city`: city of customer address
- `post_code`: postal code of customer address
- `product_preference`: the top product preference of the customer
- `food_preference`: the top food preference of the customer
- `days_to_conversion`: number of days between first website session and first order
- `channel_credit`: the marketing channel credited for acquiring the customer
- `first_order_date`: date of the customers first order
- `first_order_total_revenue`: total revenue of the customers first order
- `90d_total_revenue`: total revenue in the first 90 days after a customers first order
- `first_order_source`: the utm source in which the customer was acquired (for the first order)
- `first_order_medium`: the utm medium in which the customer was acquired (for the first order)
- `predicted_total_ltv`: predicted 5-Year LTV of the customer