GOJEK

Analytics Take-home Test

Introduction

Answer the questions and problems below to the best of your ability. You may use any outside resources necessary to help in your answer. For the problems, you may get partial credit if you show/explain your work. You may use Excel, R, Python, Tableau, or any other tool you wish but we highly value answers made using R/Python/Tableau.

Section 1: Exploratory Data Analysis

GOJEK Directors have asked BI analysts to look at the data to understand what has happened during Q1 2016 and what they should do to maximize the revenue for Q2 2016. Given the data in Problem A:
a. What are the main problems that we need to focus on? State your findings clearly.
b. Given the data in Table A & Table B, how will you maximize the revenue in Q2 2016 if we only have a budget of IDR 40,000,000,000 for Q2 2016 while also considering profitability? You are welcome to add any additional assumptions to solve this problem (but do state them clearly).
c. Present your findings from question a and concrete solutions from question b for a management meeting. The goal is to persuade management to make a decision based on your suggestions.

Please refer to the excel sheet for the data, assumptions, and dictionary.

Section 2: Business Case

Our GO-FOOD service in Surabaya performed very well last month - they had 20% more completed orders last month than the month before. The manager of GO-FOOD in Surabaya needs to see what is happening in order to constantly maintain this success for the next month onwards.

Question:
What quantitative methods would you use to evaluate the sudden growth? How would you evaluate the customers' behavior?

GOJEK

Section 3: SQL

Instruction

You will use Google BigQuery for this task. Please follow these steps to access the dataset:

1. Join our Google Group by following this link and click Ask to Join Group
2. After your join request is approved, go to this link to start using BigQuery

The dataset that you will use is **bi-dwh-dev-01.new_york_citibike** which contains the following two tables:

1. citibike_stations
2. citibike_trips

Problem 1

Exclude trips with missing start_station_id from the trip table. From the remaining trips, keep those with start_station_ids that were not present at the station table. What percentage of these trips end up in end_station_ids which are also not present in the station table?

Problem 2

Filter the trip table to include only trips with starttime from 2018-01-01 onwards. Combine usertype, birth_year, and gender into 1. Assume every unique combination represents 1 user. Include users with missing usertype/birth_year/gender.

For every month, classify users into segments based on their trips data that month:

- 0 distinct start_station_name = "inactive"
- 1-10 distinct start_station_name = "casual”
- > 10 distinct start_station_name = "power"

Note that missing month data must be imputed with 0. For example, if user A has info on months 1 and 3 but not 2, then you need to impute month 2 with 0 and therefore classify user A on month 2 as “inactive”.

Questions:

a. For each month in 2018, how many users belong to each segment?
b. For each month in 2018, compute the movements of users between segments for the next month. For example: from January 2018 to February 2018, how many casual users stayed as casual, became power, or became inactive? Do the same for the other groups and the other months in 2018

GOJEK

Section 4: Experiment Design

A product owner from the GO-PULSA product wants to increase the number of active customers and transactions on GO-PULSA for next month. However, the only customers who can use GO-PULSA are GO-JEK customers who use GO-PAY. (You cannot use Cash to buy GO-PULSA.) The PM wants more GO-JEK cash users to convert into a GO-PAY and GO-PULSA user. The Product Owner believes that cash users on GO-JEK would be more likely to use GO-PULSA if only they knew how easy GO-PULSA was to use. The Product Owner wants to give out free GO-PULSA vouchers to see if people will end up using GO-PULSA more often if only they were able to experience it for themselves. However, they only be able to use the vouchers if users have GO-PAY balance.

Your project:
1) Design an experiment to test the Product Owner's hypothesis. Assume you have access to any campaign tools, CRM tools, and a budget of 500 million rupiah.
2) Include details on how you would segment your users in the experiment.
3) What metrics would you look at to evaluate the experiment? How do you know if the Product Owner's hypothesis is correct? Why should the team believe your results?

Section 5: Experiment Post Analysis

You need to determine the optimal voucher discount amount to offer churned GO-PULSA users. You have tested 2 variables in the past 7 days. One experiment tested different voucher amounts while the other experiment was testing the frequency of sending push notifications to our customers.

Experiment Design
Participants were randomly assigned to experimental group and control group. There are 4 types of voucher amounts tested (10K,15K,20K and 25K). The non-frequent reminder group received one push notification per day for 2 days, while the frequent reminder group received one push notification per day for 4 days.

By inspection, it appears that 25K discount draws high attention to the users, but is it statistically significant?

Question
Design the hypothesis and give your suggestion to the manager for the optimal voucher and reminder scheme based on your analysis using proper method

GOJEK

Campaign Performance Table

Reminder Frequency	Voucher Discounts	Target Users	Redeemed Users
Non-Frequent	10K	3043	167
Frequent	10K	3141	204
Non-Frequent	15K	3219	204
Frequent	15K	2928	266
Non-Frequent	20K	2823	299
Frequent	20K	2668	322
Non-Frequent	25K	3076	378
Frequent	25K	2709	478
Control Group	-	3624	41

Section 6: Modelling & R/Python

Instruction

Use this dataset to solve the problems below.

Problem

Using multiple linear regression, predict the total_cbv. Create 1 model for each service.

Forecast period = 2016-03-30, 2016-03-31 and 2016-04-01
Train period = the rest

List of predictors to use:

1. Day of month
2. Month
3. Day of week
4. Weekend/weekday flag (weekend = Saturday & Sunday)

Pre-processing (do it in this order):

GOJEK

1. Remove GO-TIX
2. Keep only `Cancelled` order_status
3. Ensure the complete combinations (cartesian product) of date and service are present
4. Impute missing values with 0
5. Create is_weekend flag predictor (1 if Saturday/Sunday, 0 if other days)
6. One-hot encode month and day of week predictors
7. Standardize all predictors into z-scores using the mean and standard deviation from train-period data only

Evaluation metric: MAPE

Validation: 3-fold scheme. Each validation fold has the same length as the forecast period.

Question 1

After all the pre-processing steps, what is the value of all the predictors for service = GO-FOOD, date = 2016-02-28?

Question 2

Show the first 6 rows of one-hot encoded variables (month and day of the week)

Question 3

Print the first 6 rows of the data after pre-processing for service = GO-KILAT. Sort ascendingly by date

Question 4

Compute the forecast-period MAPE for each service. Display in ascending order based on the MAPE

Question 5

Create graphs to show the performance of each validation fold. One graph one service. x = date, y = total_cbv. Color: black = actual total_cbv, other colors = the fold predictions (there should be 3 other colors). Only show the validation period. For example, if rows 11, 12 and 13 were used for validations, then do not show the other rows in the graphs. Clearly show the month and date on the x-axis