Probability and Statistics for AI & DS

Final Exam

Due at 11:59pm on Tuesday, August 30th, 2022

Problem

UlweTel is a telecom services company in the country UlweNagar. Recently, UlweTel has noticed that a good proportion of their customers doesn't recharge on time and thus the company loses out on revenue from these customers. UlweTel is thus trying to get all its customers to recharge on or before their plan expires. UlweTel believe this can help the company retain the customers and maximize the revenue from them.

UlweTel's CEO Vashi Kumari, back from a vacation from the fast growing country Dronagiri, wants to better understand the driving factors of the behaviours of its customer base. Thus, Vashi has called upon the services of its AI&DS team headed by Ghansoli Raja. Ghansoli Raja's very capable AI&DS team has created a dataset comprising of all the features believed to be related to the customer's acquisition state, engagment, payment behaviour, value and experience (data dictionary is attached).

Ghansoli Raja is now looking at you, his finest data scientist, to answer the following questions:

Question 1 (15 points)

Churn rate is a measure of the number of customers who leave a company during a given period. Changes in a business's churn rate can provide valuable insights into an organization. Based on the recharge delay and business priorities, the AI&DS team is interested in marking some of its customers as churned customers. To that purpose:

- a) The team becomes interested in the **number of subjects** that recharge across several days after their plan expires. After an exploratory data analysis which includes drawing the observed histogram of the data, discuss a theoretical distribution that could be used to model such data. (5)
- b) Find the point estimate(s) and 95\% confidence interval(s) for the values of the parameter(s) that provide(s) the best fit to the observed data. (5)
- c) It is customary in the industry to consider the **top x percentile** of the customers with delayed recharges as churners (i.e., the customers with the greatest delay -

after sorting by decreasing recharge delay). Different companies may consider different values of x as they feel appropriate, based on the observed data and related inference. Therefore, typically, the value of x is chosen to be between 10 and 40. Based on the data provided to you by Ghansoli Raja, suggest a value x that the company should use to define the churners. In other words, after how many days since the plan expiration would you suggest that a customer is leaving the company? (5)

Question 2 (35 points)

Vashi Kumari is interested in testing some specific hypotheses on the customers' behavior. Based on the data provided to you, would you reject or fail to reject the following hypotheses?

- a) Customers who are not ready to forego their phone numbers (mnp_flag=1) recharge earlier than the rest (5)
- b) Committed customers (customers with current plan validity more than last cycle plan validity) recharge earlier than the rest (5)
- c) Customers who are not ready to forego their phone numbers (mnp_flag=1) show better engagement (tot_usage_mb and total_sum_duration) (5)
- d) Promotional schemes at the time of activation (activation_offer_code) are associated with the engagement (tot_usage_mb and total_sum_duration) of the customers and thus their timely recharge or recharge delay (5+5)
- e) Customers with better experience (implicit - pcnt_dropped_calls or explicit - cnt_open_qrc) recharge earlier than the rest (5)
- f) Customers who complain (poor explicit experience - cnt_open_qrc) are associated with having poor implicit network experience (pcnt_dropped_calls and any other network quality or coverage feature) (5)

Question 3 (a) (20 points)

One of the main interests for the company is to predict recharge delay (which could be considered a continuous outcome) of each customer, that is on which day before or after the plan expiration the customer will do their next recharge, based on historical data. Ghansoli Raja has asked you do deliver a report based on a model for such predictive effort, given your acclaimed expertise on the field.

Question 3(b) (15 points)

After you deliver your report, Ghansoli Raja wonders about the most important predictors of the recharge behaviour relationship depicted in question 3a. Only some of the features of the **original** dataset should be really meaningful. Then, he suggests to **completely redo your analysis** to see if you can find a (possibly, different) subset of important (significant) predictors.

Question 4 (a) (15 points)

Vashi Kumari has a very busy schedule. She just read the report on the churn rate that you have written in Question 1. She is very happy that you have found a way to identify churned customers, and she suggests you should get a promotion. However, before doing that, she wants to see if you can also build a model to identify the customers who are going to churn based on historical data and your expertise. You go back to your desk, excited about the incoming promotion, and write a report to answer her curiosity.

Bonus: Question 4 (b) (10 points)

Ghansoli Raja reads your report from the previous question. He is a really nice guy, but he is known to have a few pet peeves. So, now he wonders about the most important predictors that predict if a customer might churn. He thinks that only some of the features of the **original** dataset should be really meaningful. Then, he suggests to **completely redo your analysis** to see if you can find a (possibly, different) subset of important (significant) predictors that might help identify churned customers. You go back to your desk, and feel very good about your incoming promotion...

General Instructions

Your assignment is to analyze the data set described above to best address the questions of interest.

Turn in an electronic copy of your final data analysis report as a **well-organized and commented concise** summary of the Python code used in the analysis (in the form of a .ipynb **AND** pdf file).

A well thought out analysis is better than a fully correct analysis with no thought .

Your report will be graded on:

- The appropriateness of your analysis, both scientifically and statistically
- The interpretation of your results
- Your ability to communicate the results (and possible limitations) of your analysis

You need to work on the data analysis project individually

You may not talk to anyone about your analysis other than Michele, Anik, Anindya and Amit.

You can consult online resources to learn more about the cases addressed in the problem, but you may not consult online resources with the goal of finding solutions or help for the data analysis.

Any evidence of communication with other students about the project will be treated as academic dishonesty and will be subject to the consequences

Data Analysis Strategies

- Perform adequate exploratory analysis of the data and provide a complete, yet succinct, presentation of the results including both summary statistics and plots that are relevant to the research questions.
- Clearly state the model building/selection/validation criteria used to address the scientific question(s) of interest
- Perform adequate model diagnostics.
- Provide precise interpretations of the estimated parameters and/or interval estimates in the context of the scientific problem, including assessing statistical and practical significance.
- Your data analysis report should describe the results of your analysis and the conclusions you would reach from those results.
- Start off with descriptive statistics. The goal is to describe the basic characteristics of the sample used to address the question, as well as to present simple descriptive statistics (non-model based) that address the questions. Tables and plots are the key tools. If there are any characteristics of the data that present technical problems that need to be addressed in the modeling, try to present descriptive statistics illustrating those issues. The basic idea is to presage all the issues you will talk about when answering the questions and presenting the models used in statistical inference, insofar as possible with simple descriptive statistics.
- When answering the primary questions, present summaries of the statistical inference obtained from these models (point estimates, confidence intervals, p-values). Highlight any particular issues that materially affected the models used to answer the question (confounding, interactions, nonlinearities, etc.) Tables can often be used to good effect here. Provide interpretations for all parameter estimates of interest. Describe the use of p-values and confidence intervals if they play an important role in your analysis.

DO NOT INCLUDE STRAIGHT OUTPUT FROM STATISTICAL PROGRAMS: they mean little to Vashi Kumari or Ghansoli Raja. They have limited time, they want you to comment on the results for them.

When possible, use words instead of cryptic variable names.

The major theme of the above is to write to the general community rather than to a data scientist. If you cannot explain your findings in a straightforward manner, then the analysis is of little value to a reader (think of a supervisor/collaborator in a work place).

ACADEMIC DISHONESTY POLICY

Students are responsible for adhering to Academic Honesty standards. Any work turned in must be the original and independent work of each student. Academic honesty is a requirement for passing this exam. Any student who compromises the academic integrity of this exam (and of the course) is subject to a failing grade. The work you submit must be your own. Academic dishonesty includes, but is not limited to copying

answers from another student, allowing another student to copy your answers, communicating exam answers to other students during an exam, attempting to use external aids during an exam, or tampering with an exam after it has been corrected and then returning it for more credit. Note that any instance of academic dishonesty will be reported to Jio Institute's Academic Affairs office and may be cause for a failing grade in the course.