QUESTION 1: Logistic regression and CLV analysis (60 points)

In this first question, you will work with customer data of a large international nonprofit organization, so-called the Arima Foundation (not its real name), which is located in Europe. As for background information, Arima had experienced rapid growth for quite a few years but the donation amounts that the organization was able to raise had been flattened out in the recent months. One major problem was that many donors left the organization shortly after signing up for a recurring donation scheme. As the demand continued to grow, the nonprofit was under intense pressure to address the high customer churn rate if they want to keep financing their activities. In other words, they wanted to answer the question: what could they do to prevent people from churning? You are asked to help the nonprofit by analyzing a monthly dataset simulated based on actual information of 2000 individual customers over a 72 month period between 2012 and 2017. This is a subset of the total customers of the nonprofit who signed up for different recurring donation schemes in the same year 2012. For simplification purposes, let's assume that all donors in the dataset were donating to the nonprofit on a monthly basis until they decided to churn. Note that after people churned, we removed them from the dataset. You can access this simulated dataset through the "arima_data.RDS" file which is attached along with this document. As a backup, I also attached "arima_data.RData" and "arima_data.csv" files, just in case you cannot read the .RDS file.

The arima_data dataset consists of the following variables:

- ID: the id of the individual donor
- Month: the month in the year of the observation (1 = January, 2 = February, ..., 12 = December)
- Year: the year of the observation
- Age: the age of the donor
- Gender: the gender of the donor
- HHIncome: the total annual income of the donor's household (in euros)
- HHSize: the number of people in the donor's household
- RelationshipL: how much time has elapsed since the first recurring donation of the donor to the nonprofit (in months)
- DM: the number of the nonprofit's direct contacts (usually via regular mail, email, or phone) with the donor
- TV_Adv: the nonprofit's advertising expenditure in a given month on the TV channel (in euros)
- Facebook_Adv: the nonprofit's advertising expenditure in a given month on Facebook (in euros)
- Publicity: the number of times the nonprofit had been mentioned on newspapers and similar outlets
- Web_Visits: the number of visits the nonprofit's website got in that month
- Facebook_Page_Views: number of unique weekly page views on Facebook from users logged into Facebook
- churnD: a dummy indicating whether the donor had churned (i.e., terminated his/her recurring donation in that month) (1 = churn, 0 = otherwise)

You are asked to do the following tasks:

1. Explore the data

Check descriptive statistics, outliers, missing/abnormal values, etc., and make plots/graphs if necessary.

2. Churn analysis

2.a. You are asked to use the following variables (IVs) to predict churn: DM, TV_Adv, Facebook_Adv, Publicity. What do you expect regarding the directions of their effects (i.e., positive, negative, or both)? Formulate them in hypotheses. You can find examples of hypothesis development in the journal articles that you were asked to read before our regular lectures. Find literature to support the development of these hypotheses.

2.b. Run the model you created in 2.a. using the whole data set. Report model fit statistics of the model.

2.c. There are some doubts regarding whether it is important to include DM and/or Publicity. Some people even say that you would get a better model without these variables. How would you respond to that? Use what you have learned in this course to support your answers.

2.d. In this part, you are allowed to select any IVs that you like in the given data set to build the best model to predict churn. You are also allowed to create any extra IVs if you want but you must explain clearly how these extra IVs are created (or operationalized). Compare your model here and the model you created in 2.b. and discuss the main findings. Interpret the results of the better model and formulate a conclusion.

2.e. Write a short part about managerial implications based on a better model in 2.d. Furthermore, imagine that you are reporting these findings to the management team of this company, then how would they turn these insights into business actions?

2.f. If you had a chance to come back in time and collect more data, what kind of additional data you would collect to improve the model? Explain why.

3. CLV analysis

Imagine that the nonprofit, in fact, was able to implement some actions to prevent some donors from churning. Particularly, the nonprofit had sent some special gifts (e.g., special newsletters, donor appreciation gifts, etc.) to certain donors who were about to terminate their relationship to persuade them to stay. This program had cost the nonprofit about 5 euros on average per donor. One successful example is a donor with an ID of 10569 who decided to stay with the nonprofit. The nonprofit then simulated the data for this donor in the next months until December 2017. This counterfactual dataset is stored in a file named "CF_data.RDS". Alternatively, you can also access this dataset through the "CF_data.RData" or "CF_data.csv" files. For example, the value of the DM column is projected based on the planned marketing budget for the given time.

You are asked to do the following tasks:

3.a. Using the model you built in parts 2.a. and 2.b., what is the churn probability of this donor in February 2014? Without using the predict function in R, can you compute it manually based on the model parameters you obtained in 2.b. as well as all the relevant formulas you learned in this course?

3.b. Still using the model you built in parts 2.a. and 2.b., and using the predict function in R, predict the churn probability for this donor in all months available in the new data set.

3.c. Based on the results you got from 3.b., compute the lifetime value of this donor starting from February 2013.

There are some extra information as follows:

- This donor's recurring donation scheme is at 3 euros per month.
- Publicity can be considered as cost-free communication activities
- It costs around 0.65 euros for one DM contact.

- This nonprofit has a total of 100,000 donors and about 65% of them are active on social media. Importantly, donor 10569 is also an active Facebook user.
- The monthly discount rate corresponds to an annual rate of 8.5% and can be computed using the following formula:

$$1 + annual\_rate = (1 + monthly\ discount\_rate)^{number\ of\ months\ per\ year}$$

Please also note that the computation of a donor's expected donation amounts should be conditional upon the donor has not defected (or there is no churn yet at the time). Based on your computation regarding the CLV of this donor, do you think it was a good decision to keep this customer? And what if this donor's recurring donation scheme is at 5 euros per month instead?

QUESTION 2: Text analytics (25 points)

In this question, you will work with actual reviews from Amazon in the Magazine Subscriptions category (collected and shared by Ni, Li, and McAuley (Empirical Methods in Natural Language Processing (EMNLP) 2019)). You can access this dataset through the "magazine_reviews.RDS" file which is attached along with this document. As a backup, I also attached the "magazine_reviews.RData" file, just in case you cannot read the .RDS file.

The data set consists of the following variables:

- overall: product rating (from 1 to 5)
- vote: the number of votes for the helpfulness of a review
- verified: with (TRUE) or without (FALSE) a verified purchase
- reviewTime: time of the review
- reviewerID: ID of the reviewer
- asin: ID of the product
- reviewerName: name of the reviewer
- reviewText: text of the review
- summary: summary of the review
- unixReviewTime: time of the review (in unix format)
- style: a disctionary of the product metadata
- image: link of the product image (if any)

To fasten the modeling process, you are asked to work with only the first 2500 online reviews in the original data set. Using this subset of the data, your tasks are as follows:

1. Explore the data

Check descriptive statistics, outliers, missing/abnormal values, etc., and make plots/graphs if necessary.

2. Data cleaning and pre-processing

Explore the review text and create a word frequency bar chart using lemmatization

3. Sentiment analysis

Perform sentiment analysis using data from part 2. Note that customers of this product category are having a serious negativity bias, meaning that they read, write, and use negative information at least two times more frequently than positive one.

4. Topic modeling

Based on the results from the above parts, run an LDA model with different number of topics and do not trim the DFM (Document-feature matrix). How many topics should we choose? Can you label them?

QUESTION 3: Choice-Based Conjoint Analysis (15 points)

Imagine that you are working in a team of customer analytics specialists and your company is considering expanding their product range (e.g., adding new product features, etc.). The chief marketing officer of your company is asking if you and your team can help conduct a choice-based conjoint experiment to reveal consumer preferences for the existing products in the market. The following issues have been brought up during your most recent meeting with your colleagues:

a) In choice-based conjoint analysis, the importance of an attribute depends on only the number of its levels (e.g., 3 vs. 5 levels) but not the range of its levels (e.g., price ranging from 50 to 150 (NOK) compared to price ranging from 50 to 500 (NOK)).

b) In this type of analysis, hit rate is not a meaningful measure of model fit because different choice sets contain different product alternatives.

c) Because the customer-specific variables (e.g., income, age, gender) are constant across product alternatives, you will not be able to estimate their effects with the same coefficients across all alternatives, unless you interact them with alternative-specific variables as in the example below:

$$V_{i,j} = \beta_1 * Brand_j + \beta_2 * Income_i + \beta_3 * Brand_j * Income_i$$

where $V_{i,j}$ is the systematic part of the utility of product alternative j in J alternatives for respondent i.

d) You can choose any number of choice tasks per each respondent in a choice-based conjoint experiment as you wish.

Do you agree with the above statements? Explain why (or why not).