Kellogg School of Management
Northwestern University

Creative Gaming: Propensity-to-Buy Modeling

You run the advanced analytics team for Creative Gaming. You have been tasked to use telemetry data to increase awareness of the Zalon campaign. To get started, you have assembled a dataset of 30,000 *Space Pirates* gamers (i.e., users) named "`data/cg_organic.parquet`". Each row of data contains information for a single current gamer.

Target/Outcome Variable

For each *Space Pirates* gamer, the data contains a variable that recorded whether the gamer had purchased the Zalon campaign since its release two months ago. The variable is named "converted."

Features/Explanatory Variables

The data also contains 19 features that describe the behavior and gameplay of each *Space Pirates* gamer since *Space Pirates*' release.

Your Task

Mi Haruki has asked you to build a model that uses the *Space Pirates* data (described in the main case) to predict which *Space Pirates* gamers are more (less) likely to purchase the Zalon campaign.

Part I: Exploratory Analytics (3 points)

First, to gain an understanding of whether the data is appropriate for predictive analytics, you decide to engage in some exploratory analytics. Please answer the following questions:

1. What is the probability of organically converting to Zalon? (1 point)

2. Generate basic summary statistics for each feature in the data. For numeric variables shown the number of observations, the number of missing values, the number of distinct values, and the mean, min, max, and standard deviation. For non-numeric variables show the number of observations, the number of missing values, the number of distinct values, the most common level, and the least common level (1 point)

3. Generate histograms for all numeric variables and frequency plots for non-numeric variables (1 point)

Part II: Predictive Model (5 points)

Train a logistic regression model with pyrsm or statsmodels. Use all features and use the "training" variable to set a filter ("training == 1").

1. What are the 5 most important features? (1 point)

2. Generate prediction plots for each of these 5 features and summarize your findings (1 point)

3. Create a new variable "pred_logit" with predictions from the logistic regression model for all rows in the data (1 point)

4. Plot gains curves for both the training and test set using the "pred_logit" variable (1 point)

5. Report the AUC of the model in both the training and test set (1 point)

Part III: The Ad-Experiment (12 points)

You have finished building a logistic regression model to predict what kind of *Space Pirates* gamers were more or less likely to purchase the Zalon campaign. After debriefing Mi Haruki on the performance of the predictive model, she lays out the next steps:

> *"We will test the effectiveness of the in-app ad and the predictive model by exposing a random 150,000 customers to a 2-week ad campaign and measuring their conversion to Zalon over the next 2 months. At the same time, we will randomly pick another 30,000 customers and observe their organic upgrade behavior over the same period. This set of customers will not be served any in-app ads.*
>
> *I want you to compare three groups based on this data.*
>
> ***Group 1:** The randomly picked 30,000 Space Pirates gamers who did not receive in-app ads during the experimental period.*
>
> ***Group 2:** A randomly picked 30,000 Space Pirates gamers among the 150,000 who were served in-app ads for Zalon.*
>
> ***Group 3:** A model-selected 30,000 Space Pirates gamers among the 120,000 (after taking out Group 2) who were served in-app ads for Zalon.*
>
> *Please report back to me how well the ads are working in terms of conversion rates and profits and by how much the model improves these metrics, all based on targeting 30,000 customers. To calculate profits, please use*

revenues of $14.99 from selling Zalon. The cost of serving ads to a consumer for 2 weeks is $1.50 in lost coin purchases."

1. Calculate the response rate and profit of group 1. The dataset containing the data for group 1 is named "`data/cg_organic_control.parquet`". (2 points)

The dataset needed to evaluate groups 2 and 3 is "`data/cg_ad_treatment.parquet`". All customers in this dataset were exposed to advertising as part of the experiment.

2. Calculate the response rate and profit of group 2. Set a filter "rnd_30k == 1" to select the 30,000 customers from the 150,000 rows in "`cg_ad_treatment`". (2 points)

3. Calculate the response rate and profit of group 3. To do this please:
   a. Use the logistic regression model you trained in Part II to predict the probability of purchasing the Zalon campaign (i.e., score) all 150,000 gamers in the "`cg_ad_treatment`". (1 point)

   b. Select the 30,000 customers with best predictions (scores) that are not part of group 2. Use only these 30,000 to compute conversion rates and profits of group 3. (1 point)

4. Answer Mi Haruki's question: "Please report back to me how well the ads are working in terms of *conversion rates and profits* and by how much the model improves these metrics, all based on targeting 30,000 customers." (2 points)

5. Plot the gains curve for all customers that are not in group 2 (i.e., rnd_30k == 0). Also report the AUC of the for this set of customers. Compare the gains curve and AUC to the ones you calculated in Part II.4 and II.5. Why are they different? (2 points)

6. Why would CG have collecting data for group 1 (cg_organic_control) given that they already had data on organic conversions from the cg_organic data? (2 points)

Part IV: Better Data, Better Predictions (8 points)

Mi Hiruki called for a meeting to discuss next steps. She explained:

"Before we roll out the campaign globally, we want to see whether we can use the experimental data to retrain the model. The idea is to model trial in response to the in-app ad—not just organic conversion as we did initially. We know that the in-app ad, on average, increases Zalon conversions. However, if the in-app ad works for people who would not have purchased the Zalon campaign organically, updating the model based on the in-app ad data should improve predictive performance.

Let's retrain the model based on the randomly chosen Space Pirates gamers we messaged in the experiment and see how well the updated model compares to the original model in a test sample."

1. Retrain the logistic regression model from Part II on the random sample of 30K customers in "`cg_ad_treatment`" and generate predictions for all 150,000 customers. Label this retrained model "clf_ad" and assign the predictions to a new variable "pred_logit_ad". Use pyrsm or statsmodels for estimation. (2 points)

2. Compare the performance of the original "organic" model from Part II and the new "ad" model across the 120,000K customers that are not in group 2. Use gains curves and AUC to make the comparison. What do you find? (2 points)

3. Calculate the profit improvement of using the "ad" model instead of the "organic" model to target the best 30,000 customers in the "`cg_ad_treatment`" data that are not in "rnd_30k == 1". (2 points)

4. Compare the permutation importance plot of the "organic" and the "ad" model. Explain why you think the plots differ. (2 points)

Part V: Better Models, Better Predictions (12 points)

Mi Haruki was impressed by how much performance improved when the logistic regression model was retrained on the ad treatment data rather than organic data. However, she knew that the analytics team was not done yet:

"I know we have been trying to keep the models simple and perhaps the logistic regression is what we go with. However, I want to you apply machine learning models to see if they can help improve our predictions and performance further."

1. Train and tune a neural network on the random sample of customers who were exposed to the ad campaign (i.e., rnd_30k == 1). Select two hyper parameters for your grid to tune on. Use both pyrsm and sklearn for estimation and check the similarity of your results (2 points).

2. Compare the performance of the neural network "ad" model and the logistic regression "ad" model from Part IV using data from the 120,000K customers that are not in group 2. Use gains curves and AUC for the comparison. What do you find? (2 points)

3. Calculate the profit improvement of using the neural network "ad" model and the logistic regression "ad" model to target the best 30,000 customers out of the 120,000K customers that are not in group 2. (2 points)

4. Train and tune a random forest on the random sample of customers who were exposed to the ad campaign (i.e., rnd_30k == 1). Select two hyper parameters for your grid to tune on, one of which should be "max_features". Choose at least 5 different values for "max_features" for your grid. (2 points)

5. Compare the performance of the random forest “ad” model and the logistic regression “ad” model from Part IV using data from the 120,000K customers that are not in group 2. Use gains curves and AUC for the comparison. What do you find? (2 points)

6. Calculate the profit improvement of using the random forest “ad” model and the logistic regression “ad” model to target the best 30,000 customers out of the 120,000K customers that are not in group 2. (2 points)

Part VI: Generative AI (5 points)

Please describe how you used Generative AI-tools like ChatGPT to support your work on this assignment and enhance your learning. Create a pdf where you organize your interactions with AI and comment on what things did and did not go well. Bring any questions you may have about the assignment and the support you received from GenAI to class so we can discuss.

Make sure to include:
- Specific examples of prompts you used
- How the AI responses helped or hindered your understanding
- Any limitations or challenges you encountered
- Key insights gained from using GenAI tools
- Questions that arose during your interactions with AI
- How GenAI complemented your learning process

Note: No matter how you used Generative AI-tools, you will be expected to understand and talk meaningfully about the work you submitted for this assignment. You may be called on in class to walk us through your thought process and calculations.