Kellogg School of Management
Northwestern University

`

Assignment for

Creative Gaming: Propensity Modelling

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

Target/Outcome Variable

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

Features/Explanatory Variables

The data also contained 19 features that describe the behavior and gameplay of *Space Pirates* gamers 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* users are likely to purchase the Zalon campaign.

Part 1: Exploratory Analytics

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 organic probability of converting to Zalon?

2. For each feature, show basic summary statistics.
   Hint: install the "`skimr`" package and use the "`skim_without_charts()`" command (don't forget to add `library(skimr)` in the header of your Rmd file). Don't worry about formatting when you knit to pdf.

Part 2: Predictive Model

1. Create a training and test sample based on the "`cg_organic`" dataframe. Please use the "`sample_train_org`" vector to select observations for both new dataframes.[1] You can use this syntax:

```
cg_organic_train <- cg_organic[sample_train_org,]
cg_organic_test <- cg_organic[-sample_train_org,]
```

   What is the training/test split?

2. Train a logistic regression model using all features
   a. What are the 5 most important features?
   b. For each feature of the top 5, summarize what you learn from marginal effect plots for those variables.
   c. Plot the gains curve in the test sample and report the AUC of the model.

Part 3: The Ad-Experiment

You have finished building a logistic regression model to predict what kind of *Space Pirates* users were most 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, having <u>not</u> served them an in-app ad.*
>
> *I want you to compare three groups based on this data.*
>
> ***Group 1:** The <u>randomly</u> picked 30,000 Space Pirates users who did <u>not</u> receive in-app ads during the experimental period.*
>
> ***Group 2:** A <u>randomly</u> picked 30,000 Space Pirates users among the 150,000 who were served in-app ads for Zalon.*
>
> ***Group 3:** A <u>model-selected</u> 30,000 Space Pirates users 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] If you were sampling outside of this assignment you would create a random sample vector yourself using the command: `sample_train_org <- sample(nrow(organic), x*(nrow(organic))` where x is the fraction you want for the training sample. For the purpose of this assignment I have included `sample_train_org` with the other data to ensure that all teams choose the same train/test split.

1. Calculate the response rate and profit of group 1.
   The dataframe is called "`cg_organic_control`"

2. Calculate the response rate and profit of group 2. To randomly select 30,000 customers, please use the "`sample_random_30000`" vector to select observations from the 150,000 row dataframe "`cg_ad_treatment`" which contains the customers who were exposed to the ad campaign. You can use this syntax to create the sample:

   ```
   cg_ad_random <- cg_ad_treatment[sample_random_30000,]
   ```

3. Calculate the response rate and profit of group 3. To do this please:
   (a) Use the logistic regression model you trained in Part 2 to score the 120,000 customers in "`cg_ad_treatment`" who remained after sampling 30,000 in "`cg_ad_test`". You can use this syntax to create the sample for scoring:

   ```
   cg_ad_scoring <- cg_ad_treatment[-sample_random_30000,]
   ```

   (b) Select the 30,000 customers with best scores and use only these 30,000 (who correspond to group 3) to compute conversion rates and profits of group 3 in the next question.

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."

5. Plot the gains curve in the `cg_ad_scoring` sample you scored in Part 3 Q3 (which used the model trained in Part 2 Q2). Also report the AUC of the model. Compare the gains curve and AUC to the ones you calculated in Part 2 Q2c. Why are they different?

6. What is the purpose of group 1 given that we already had data on organic conversions?

Part 4: Better Data, Better Predictions

Mi Hiruki called for a meeting to plan the 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 users 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 2 (which was trained on "`cg_organic_train`") on the sample of customers who were exposed to the ad campaign ("`cg_ad_random`"). Instead of creating separate training/test datasets, please use the full 30,000 sample "`cg_ad_random`" you created in Part 3, Q2 to train the model.

2. Compare the performance of the original "organic" model from Part 2 and the "ad" model on the "`cg_ad_scoring`" sample you created in Part 3, Q3. Use gains curves and AUC to make the comparison. What do you find?

3. Calculate the profit improvement of using a model trained on ad treatment instead of organic data to target the best 30,000 customers in the "`cg_ad_scoring`" sample.

4. Compare the variable importance plot of the "organic" model and the "ad" model to explain why the performance of the models differ.

Part 5: Better Models, Better Predictions

Mi Haruki had enjoyed by how much the model improved when it was retrained on the ad treatment instead of organic data. However, she knew that the analytics team was not done yet:

"I know we have been trying to keep the modeling simple. Perhaps the logistic regression is what we go with. However, I want to explore using some machine learning models to see whether they improve our predictions."

1. Train a neural network on the sample of customers who were exposed to the ad campaign. (If you time you can also try a random forest!).

   Hints:
   - To use a neural network we use the `nnet` package (please install it first and then load it in your R notebook). Please see the in-class handout "Using Neural Networks in Customer Analytics," (Class8a_NN_demo.pdf) for how to run a neural network.

2. Compare the performance of the neural network "ad" model and the logistic "ad" model from Part 4 on the "cg_ad_scoring" sample. Use gains curves and AUC to make the comparison. What do you find?

3. Calculate the profit improvement of using a neural network instead of a logistic regression (both trained on ad treatment data) to target the best 30,000 customers in the "cg_ad_scoring" sample.