Introduction

We would like to develop a model that predicts whether a Pinner will convert (buy) or not. We have decided to start with a simple model that uses a transactions table we readily have available. The objective is to spend less than a day to develop a **logistic regression** model and create a baseline measurement on how well we can predict a conversion.

Data Sets

We have dumped our transactions table into the **transactions.csv** file. The first few lines of this file look like this:

"conversion","session_id","session_dt","num_impressions","avg_relevance","num_search","train","test","score"
FALSE,"JtAYcYYL7cU7rnY",2018-09-13,1,0.83,0,TRUE,FALSE,FALSE
FALSE,"Cm7ofJ1fABISzfx",2018-08-20,6,0.74,2,TRUE,FALSE,FALSE
TRUE,"LGygqNnJ79e5Xzi",2018-08-12,3,0.47,0,TRUE,FALSE,FALSE

In table format, the data looks like this:

conversion	session_id	session_dt	num_impressions	avg_relevance	num_search	train	test	score
FALSE	JtAYcYYL7cU7rnY	2018-09-13	1	0.83	0	TRUE	FALSE	FALSE
	Cm7ofJ1fABISzfx	2018-08-20	6	0.74	2	TRUE	FALSE	FALSE
	LGygqNnJ79e5Xzi	2018-08-12	3	0.47	0	TRUE	FALSE	FALSE

Here is the data dictionary:

Conversion: Can be true of false. Means that the Pinner bought something or not
Session_id: A unique identifier that is assigned to the session. A user will typically have multiple sessions. The user_id for a session_id can be found on another table, which is also provided.
Session_dt: The calendar date of the session
Num_impressions: Number of ads shown to the Pinner that are related to the purchase we are trying to predict
Avg_relevance: A score that ranges from 0 to 1. The higher the score, the more relevant were the ads that we showed the Pinner

Num_search: Number of searches the user ran in this session
Train: If TRUE, it means data in this row can be used when training your model. Only use the data with train = TRUE while doing any model development work
Test: If TRUE, it means you can use this row to do any testing related work. Only one of **train** or **test** columns will be TRUE for any given row so this column is actually redundant.
Score: If TRUE, it means you should use this row to calculate the goodness of fit metrics for your model. Please see the “Assumptions You Can Make” section below for further details.

The look up table that can be used to match session_ids with user_ids is dumped into the **sessions.csv** file. First few rows look like this:

"user_id","session_id"
"LGATzZ5aXh","SyQgH9ETmu66sO1"
"LGATzZ5aXh","LT3MW5Jzswi84Ix"
"5Y3rXvHLVP","5QHd36ZquRIkeNt"
"5Y3rXvHLVP","HcUq8m79AUH7QQb"
"5Y3rXvHLVP","ILTfjKTvxsQ8mWI"
"vGymFuQa3q","rDRFnl6JGc6377I"

In table format, the data looks like this:

"user_id"	"session_id"
"LGATzZ5aXh"	"SyQgH9ETmu66sO1"
"LGATzZ5aXh"	"LT3MW5Jzswi84Ix"
"5Y3rXvHLVP"	"5QHd36ZquRIkeNt"
"5Y3rXvHLVP"	"HcUq8m79AUH7QQb"
"5Y3rXvHLVP"	"ILTfjKTvxsQ8mWI"
"vGymFuQa3q"	"rDRFnl6JGc6377I"

Here is the data dictionary:

User_id: The unique id of the user who has the session
Session_id: The id of the session that was referenced in the transactions.csv file.

Assumptions You Can Make:

- The data you received is not perfect

- More than 1/3 of the time you dedicate to this exercise will be spent on feature engineering
- Well over 90% **accuracy** is possible with a simple logistic regression model if you have the right model form and features. Assuming that you have a data frame named analysis_df, a column named conversion with the labels that are provided and another column named pred_conv that has your predicted conversions, the R code to calculate **accuracy** is:
  ```
  > mean(analysis_df$conversion[analysis_df$score==T] == analysis_df$pred_conv[analysis_df$score==T])
  ```
- Your recruiting team will look at your code and all the things that you have analyzed / tried to ensure that your model is valid.

Deliverables

- A deck with 8 slides in it with the following:
  - Slide 1: Cover page
  - Slide 2: Executive Summary
  - Slide 3: Findings from your exploratory data analysis
  - Slide 4: Feature engineering efforts
  - Slide 5: Model forms you have tried and your final model, including the full regression results
  - Slide 6: Justification of the model: How did you validate it? Do coefficients make sense? Typical problems you checked for.
  - Slide 7: Performance metrics for your model
  - Slide 8: Ideas for future development
- Your Python or R code

Please e-mail these 2 files to your recruiter.