Intro

We would like to develop a model that predicts whether a user 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.

Dataset

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",2018­09­13,1,0.83,0,TRUE,FALSE,FALSE FALSE,"Cm7ofJ1fABISzfx",2018­08­20,6,0.74,2,TRUE,FALSE,FALSE TRUE,"LGygqNnJ79e5Xzi",2018­08­12,3,0.47,0,TRUE,FALSE,FALSE

Conversion: Can be true of false. Means that the user 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 user 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 user

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.

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","rDRFnl6JGc6377l"

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.