CHE384 Data to Decisions
Chris Mack, University of Texas at Austin

Model Building Contest

A retail company wants to understand the customer purchase behavior (specifically, purchase amount) against various products of different categories. They have provided purchase summary data for various customers from last month. The data set also contains customer demographics, product details, and the total purchase amount for one month. Now, they want to build a model to predict the purchase amount of a customer against various products which will help them to create personalized offers.

The dataset “BlackFridayTrain.csv” contains the training data for this contest. There are 400,000 rows of data. Here are the variables in the data set:

Variable	**Description**
User_ID	User ID
Gender	Sex of User
Age	Age in years (middle of bin)
Salary	Household Salary in dollars
City_Category	Category of the City (A,B,C)
Stay_In_Current_City_Years	Number of years living in current city (max = 4)
Marital_Status	Marital Status (1 = married)
Product_Category	Product Category Code
Purchase	Purchase Amount (response) in cents

Your goal is to build a model with the best ability to predict purchase amount. To test how well you did, I have data for an additional 150,000 customers. We’ll compare predicted to actual for this data set and compute the MSPR (mean square predicted residuals). The model with the lowest MSPR wins!