

MSiA 421: Data Mining

Professor Malthouse

Homework 1: Due January 21, 11:59pm

You may work in self-selected teams of at most 5 students. Submit one assignment per team, and make sure that all team members have their name on the submitted assignment. [Link to Kaggle](#), [click here](#)

You may use any software you like to complete this exercise, which gives you practice with feature creation and taxonomy data. The data are available on Canvas under Files/Bookdata.

- This will be called the **book case**. You have data from a German retailer that mostly sells books, but also some music and DVDs. I have given you two tables:
 - **booktrain.csv**: This gives the dependent variable for the training set only. You should find 8311 records plus a header.
 - * **id**: unique customer identifier
 - * **logtarg**: dependent variable. It is the **natural** logarithm of the spending in response to an offer mailed on 01AUG2014.
 - **orders.csv**: all order prior to 01AUG2014 for training ($n = 8311$) and test ($n = 25,402$) sets. You should find 627,955 records plus a header.
 - * **id**: unique customer identifier
 - * **orddate**: order date
 - * **ordnum**: order number
 - * **category**: category identifier, 1=fiction; 3=classics; 5=cartoons; 6=legends; 7=philosophy; 8=religion; 9=psychology; 10=linguistics; 12=art; 14=music; 17=art reprints; 19=history; 20=contemporary history; 21=economy; 22=politics; 23=science; 26=computer science; 27=traffic, railroads; 30=maps; 31=travel guides; 35=health; 36=cooking; 37=learning; 38=games and riddles; 39=sports; 40=hobby; 41=nature/animals/plants; 44=encyclopedia; 50=videos, DVDs; 99=non books
 - * **qty**: quantity ordered
 - * **price**: price paid

Create feature variables from **orders**, then merge them with **booktrain**. Fit a regression predicting **logtarg**.

Export a csv file with two variables, the customer id and your prediction. Upload this to Kaggle.