The dataset I’m proposing to use for Capstone Project 2 is Kaggle’s Click-Through Rate Prediction dataset:

<https://www.kaggle.com/competitions/avazu-ctr-prediction>

I agreed with the earlier assessment that the previous 3 datasets I chose were too easy and I decided to go for a competition dataset.

In online advertising, click-through rate (CTR) is a very important metric for evaluating ad performance. As a result, click prediction systems are essential and widely used for sponsored search and real-time bidding.

About the dataset: The dataset we’re using is by Criteo from the Kaggle competition, where Criteo is an internet advertising company. Criteo’s goal is to increase its CTR among consumers who have previously an advertiser’s website. The dataset consists of Criteo’s traffic over 7 days with the first column indicating whether an ad has been clicked or not. There are 13 numerical features and 26 categorical features in this dataset.

My proposal is to use the CTR dataset for prediction and feature selection. The main goal would be to create a classification model that accurately predicts clickthroughs for Criteo. Some potential algorithms I’d try are logistic, Naïve Bayes, KNN, random forest and support vector machines.

Another goal would be figuring out which features are most useful for classification. My proposal is to use EDA, stepwise regression as well as various regularization algorithms to understand which predictors are actually the most useful. This is important both for inference as well as preventing overfitting.