Lucas Beane - lhb7tz

Karan Gadiya - khg8mh

Murugesan Ramakrishnan - mr6rx

**Kaggle Competition 4 Reflection**

**Who might care about this problem and why?**

Most companies with an online store would love to have a functional solution to the problem of predicting how much revenue customers will net them. It would be a great way to have some indication of whether their money in maintaining the online store is well-spent or if it should potentially be used elsewhere for a better profit.

Economists could be interested in this sort of problem as well, as a sort of analysis for trends in online spending, especially over time. This could potentially give them another form of input for any investment strategies, too.

Psychologists would definitely be interested in analyzing the shopping habits of customers, particularly in the relatively young world of online shopping. This could lead to an improved understanding of human response to various digital stimuli and possibly even a better grasp on the problem of addiction (to online shopping, but could possibly be generalized).

**Why was this problem challenging?**

The main challenge of working with this data set was shockingly the size of the data set. It was an unforeseen obstacle at the beginning of the competition because the original training and testing sets were efficiently read by most accessible machines. But due to the data leak, the supplied training and testing sets grew by factors of 16 and 5, respectively, leading to time-out issues in our go-to R IDE RStudio.

A minor obstacle was that most of our data was in a form that needed additional cleaning, but that was easily addressed with some quick parsing and division of the variables.

Another minor difficulty was our selection of the threshold values for our models, especially in our logistic regression model. We chose a value by adjusting the value and viewing the effects it had on our predictions.

**What other problems resemble this problem?**

A similar approach with similar data could be applied to revenue for physical stores. For example, a superstore that’s able to identify individual customers could gather, keep, and analyze similar data for a prediction on how much a customer will spend under given conditions. This is a little less plausible for obvious reasons, but it seems like an application that many companies would be interested in if not already actively invested in.

**What might account for the differing performance levels of the mandatory models?**

The most apparent explanation for differing performance in our various models is that they perform differently by design. We cannot expect them to perform similarly, nor should we expect to know going in which model will perform the best on our data. Beside that, the condition boundary for the prediction of whether or not revenue will be generated from a given visit is non-linear, which means linear regression will perform very poorly compared to our other models. This is a contributing factor to our random forest model’s exceptional performance compared to our other models. Our spline’s poor performance can be explained by the lack of knowledge of where to place the knots for each variable.