Prediciton challenge notes. App - ML 2020

Juan Ignacio Lucas Mation

Expected MSE: 7.5

Questions

- What is the correct protocol for dealing with missing values in some X? In reg-world to solve the problem ofX having missings for some observations, one creates a new dummy, I_{Xmi}, to indicate missing values and then replaces the missing values in X by any arbitrary number (eg: zero, or the average of X) and then include both X and I_{Xmi} in the regression. We did something similar for the missings in our variables, replacing them by the average of X. However, we are woried about the ML algorithm not picking up the missing indicator dummy (I_{Xmi}), making the procedure sensitive to the imputed value.
- How to incorporate the fact that Y is a count data into the ML estimation? In regression we would use a poisson link function. But we did not incorporate that into our procedure.

Lessons learnt

- Juan: Missings are also data. A non response is acatually a usefull information. Adding a dummy gives the model the possibility to pick this information.
- We spent 95% of the time to understand and prepare the data
- Lucas : ML stack in R
- Beucause we had relativelly small number of observations, we needed to be more thoughtful and parcimonious in choosing variables to include and creating features to reduced the number of variables fed to the model.
- Prediction is hard!