HGDS 200 - Foundations of Data-Driven Analysis 2018/2019 - Handout 1

12/17/2018 - 1/7/2019

Name: Connor Voglewede

The following should be completed after reading Chapter 1 (Introduction) of Applied Predictive Modeling.

1. What are some prediction tasks of interest to Red Ventures?
   1. There are a ton of prediction tasks of interest to Red Ventures, across vertical and throughout the sales funnel. One example is a credit model that attempt to predict high, medium or low risk credit for customers on ATV. Another example is a call prority model that predicts customer conversion rate, so we can optimize our sales queue ordering.
2. Can you think of an RV prediction tasks where interpretability should take precedence over prediction accuracy?
   1. Yes, I built a customer churn prediction model for HughesNet that required interpretability. I scrapped logistic models in favor of a more basic tree model because implementation of a dynamic sales script required buy-in from our HughesNet sales agents. If I presented them with a black-box for how we determined who got “harder-to-sell-with” scripts, we’d have a compliance issue with our agents. I was able to find a more interpretable method to make convincing our sales team a little easier.
3. Can you think of any heuristics for deciding when one might favor interpretability over prediction accuracy or vice versa?
   1. I would think that it greatly depends on your audience and implementation strategy. If the decision to implement a model requires sign-off from a group, that group’s comfort-level with interpretability will be a driving force for how to construct the model. If the group is hesitant to approve black-box methodologies, it’s probably not wise to provide them with one. Additionally, it’s not just management that must be comfortable with the methodology. Some users of the model may need to be convinced so they aren’t temped to go against the predictions.
4. The authors give four common reasons why models fail - stating that the most important reason (in their opinion) is ”over-fitting.” Do some research to come up with your own definition of over-fitting and some ways you might diagnose and overcome this problem.
   1. Overfitting is the process of building a model that works narrowly for the collected sample but fails to represent a population accurately. One way to both diagnose and fix the problem would be to use k-fold cross validation in the model validation phase. Acquiring additional test data is likely costly or impossible, but k-fold cross validation allows us to repeat the process of test-train splitting and evaluate a model’s performance on test data k times. A model is much less likely to overfit if it demonstrates a low RMSE on k validations than if it is evaluated on a single train-test comparison.