Followup Questions for Nicolette Ige – March 7, 2018

* Dummy code outliers to see if the outliers end up being a significant variable in a model

Vivek asked:

1. What are some of the ways used in the industry to implement the models in production or how do

companies put these models to use at scale for future data?

2. Most companies use Agile methodology for project management. How does analytics fit into

Agile framework?

3. Under Oversampling (page7) you mentioned that biased sampling is corrected by adjusting the

predicted probabilities? Does the model adjust the predicted probabilities using the training

dataset and then use it to make the predictions on test dataset or does it make predictions on the

test dataset and then adjust the probabilities? A sequence of events will be helpful in further

understanding.

Chad asked:

**1:** Can you explain more on Tableau. I have always had trouble doing seemly simple things in Tableau and consequently I have done things in R instead, because I know I can get the result I want. My biggest issue is Tableau aggregating everything together

**2:** What are some ways of validating that a sample is representative of the overall population? And if it is not a good representation how do we address this? Gather more data or try and make our data better represent that population? (I know this was gone over to some degree but maybe just some further discussion on this topic)

**3:** Dealing with ambiguous or undefined problems. (Ex. Someone throws data at you with a very poorly defined issue they want you to solve.) How do we handle that?

Amy asked:

1. What number of variables would be too much to complete a casebook?

2. Can you ever say, “This data won’t answer that question” but suggest another question that would

be helpful.

3. Can you ever have data that is too dirty. Meaning, once you clean it, it really doesn’t represent the

population anymore.

Dalton asked:

1. What is the best way to tell if you should use an interaction term or not?

2. Do you lose any interpretability when you oversample for rare events?

3. Are there any situations where you lose predictability by removing outliers?

Bethany asked:

1. Scenario : Lets say we have a data set with roughly 300 variables, 30 of which are categorical

with between 8 and 12 distinct levels. The data set does not have a codebook and thus we don’t

know which variables mean what, which levels mean what, and which categorical variables are

nominal or ordinal.

a. Question: How might we be able to figure out which categorical variables are ordinal or

nominal? (This is an issue I’m currently trying to figure out and haven’t really known

what to do)

2. Question: In data sets that are relatively large, with hundreds of potential predictors, how could

we go about conducting variable reduction? Due to how large the data set/number of predictors

are, there are likely issues with ‘the curse of dimensionality’ and everything appears to be

significant.

3. Question: How does binning a variable affect outliers? Does it have any impact?

Bharat asked:

1: Even though outliers may give insights about the range of values, is it necessary to investigate since we should delete it before modeling?

2: Isn't creating dummy variables actually adding many variables and increasing the complexity of the modeling?

3: What is target population. How do you calculate its univariate characteristics?