Prithvi Kannan

UID: 405110096

Collaborators:

Problem 1:

* Highly skewed responses. People are more likely to post about their experience on twitter if they had a really good or bad experience versus neutral.
* Twitter audience is not necessarily representative of LA Dept of Health patrons. People on Twitter are likely younger, so this method may ignore older patients.

Problem 2:

* Features
  + past GPA – decimal
  + percent A in professor’s class – decimal
  + student interest level – scale of 1-5
  + past experience in subject – scale of 1-5
  + past history of attendance – decimal
* We can create a survey which we can randomly distribute to students to learn about personal information. To ensure coverage of all student segments we distribute the survey over email, since all students have access and must use email to get class information. We can find out past grade distributions via Bruinwalk records.

Problem 3:

* Replace with median – this is one of the easiest ways to handle null values. This can be a great method if there are very few null values, and the data follows a normal distribution. However, on sparse datasets, our median may not be representative of the underlying distribution.
* Replace with a constant value – here we replace all null fields of the
* Hot deck – in this method we replace null fields with a randomly chosen value from similar individuals in sample. This ensures that the imputed value will be bounded by the true values in the sample while still providing variance in the imputations. However,
* Regression – instead of picking an arbitrary value to impute with, we can train a model to predict the missing value given the remaining features. For example, if we are missing weight, but have height, age, and waist measurement, we can train a model to predict weight using the complete fields. This works well when we have enough additional features related to the feature we are trying to impute. This may not work so well if the data is sparse or if the columns are not related. In regression, we lose variability around predicted values.

Problem 4:

Problem 5:

1. Idk
2. Idk
3. False. If the question being asked is novel enough, then we may also need to create the dataset using standard data collection methods.
4. Idk
5. True. If we are building systems based on biased data, then our new outputs will carry over those biases in predictions.

Problem 6: