Laon Approval Classifier Project.
Matthew Litschewski
Bellevue University
DSC 680_ Spring 2021

- Any surprises from your domain from these data?

  I have experience in the domain of this data. I worked in mortgage sales both from a business to consumer and business to business stand point, with a majority of that tenure representing a mortgage lender to various brokers. I have closely worked in pre-approving candidates before submitting them to underwriting. Once in underwriting I worked closely to clear stipulation to get loans to final approval. The information in the data set

- The dataset is what you thought it was?

  The data used in this project was from an Analytics Vidhya challenge so it was fairly straight forward. The nice part about using a dataset like this is it was already in a form that was easier to use. In previous projects it has been a struggle to just get data together. Then once assembled you spend time just getting it into a structure that will make it easier to manipulate and feed into the algorithm and this is all before doing any EDA or feature engineering. Some data being categorical has had to be changed to either binary or some other numerical equivalent. I am trying to avoid things like one hot encoding etc.

- Have you had to adjust your approach or research questions?

  So far this project has been fairly straightforward as it is a binary classifier problem and lends itself to simple Logistic Regression and other standard classifier modeling techniques. I have also had to be aware that the target variable is unbalanced and approach the train, test split in a way that will make sure the model does not bias towards the heavier waited class. This can be accomplished in the train test split command line by adding stratify = 'yes' this will make sure to balance the classes when setting up the train test split. There are other techniques like over sampling the minor class or oversampling the major class.
- Is your method working?

  I feel like my methods are showing results. The results I am getting are better than a 50/50 random guess, so that is good. I am averaging about a 0.74 accuracy in most of the models. Although when reviewing my code I did notice I didn't adjust for the imbalance in the classes so I am going to expect and accuracy increase once I correct that issue. Also I have yet to put together the ensemble technique so I am not sure what if any impact that will have on model performance.

- What challenges are you having?

One of the greatest challenges happened when I was training to change the categorical binary features into numeric. There was a type error with one the variables that giving a trace back error when the function I was using to do this was applied. This took several trail and error attempts using pandas and numpy to get the variable figured out and moved into a form that the function would except.