fair

Data Scientist | Take Home Challenge

Preferred Language: Python
Preferred Deliverable Format: Jupyter Notebook

The Fair team needs your assistance with analyzing historical loan data. The attached csv data file contains records from a few thousand customer accounts. Refer to the accompanying data dictionary for the definitions of each variable in the csv file. Your assignment is to build a model to predict the probability that someone will default on a loan given the various features in the data set. A default is identified by a loan status listed as in default, over 30 days late, or charged off.

1) Perform any data cleaning and exploratory analysis that you feel is appropriate. Summarize your approach in a few sentences and report some interesting facts about the population in the data.

2) Build a predictive model to help us determine who is likely to default if they are extended a loan. Discuss your overall approach, with a focus on data preparation, classifier selection, and model evaluation.

3) Now that you have trained a model, it is time to put it in production (a web app where a customer can apply and get an immediate loan approval or rejection decision). How would you come up with an appropriate cut off that separates those approved and those rejected? We are not necessarily expecting code for this section, just an analytical explanation for how you would solve this problem in a real business setting.