ABSTRACT

The Consumer Financial Protection Bureau (CFPB) offers many services and resources for Americans on a number of topics. One of these topics is buying a home, an important financial step for many Americans. The entire process of buying a home and applying for a mortgage can be overwhelming for consumers. The CFPB is here to guide, educate, and help prospective homeowners with this daunting financial transaction.

DESIGN

The CFPB has hired me to come up with a model that will predict how likely someone is to be approved for a loan. The CFBP wants to use my model to provide a resource on their website where a consumer can fill out a brief survey and see how likely they are to get approved for a loan. This can be a useful tool for people considering buying a home but do not know where to start. A tool like this can help inform people on if they might be able to get a mortgage but instead of going through the whole mortgage application process, they simply fill out a survey on their phone or computer.

We are interested in soft predictions in this case. If the model predicted that a prospective homeowner would be denied for a loan, they would certainly want to know if they were close vs if they had no chance. Then depending on their personal situation, they could decide how to proceed.

DATA

The CFPB obtains data on mortgage applications every year for each state. I used the most recent year available (2017) for the state of California due to it's high number of records. The data had approximatley one million rows (each row = one application) and contained information on the loan itself, the applicant, and the location applied for. Features were only considered if they would either be known by the consumer or the CFPB before the actual loan process was started.

ALGORITHMS

Before training and testing any models, I performed data cleaning, feature engineering, and EDA to prepare my data and compile feature combinations to test. I then built an algorithm that could take in training data for a given set of features and would automatically cross validate and score the data. This algorithm is designed to take in an entire training set of unprocessed data it will then standardize numerical features, create dummy columns for categories, and re-balance the training data with down-sampling. This processing and re-sampling happens at each K fold split so that the validation data is completely uncontaminated and separate from the training data. Therefore, once basic high level data cleaning is done on the full dataset, cross validating a new model requires simply calling one function.

While the California dataset was around 1 million records, there were over 14 million records nationwide in the same year. Because of the structure of my algorithm, it can easily be applied to other states, even if those states have different features that best predict loan approval. While I have made a model for California, the I have also provided the CFPB with an algorithm that can make models for other states easily.

TOOLS

I mostly used Pandas and Numpy for data cleaning and manipulation and matplotlib for EDA. Modeling was done with Scikit Learn and XGBoost. Since some models took a long time to train I used a Google Collab instance so that I could run some models in a cloud kernel while simultaneously running other models on my local machine.

OTHER COMMUNICATION – Also see: Slide Deck, Jupyter Notebook