Springboard Data Science Career Track | Nikki Seegars

Capstone Project I Proposal: Predicting Mortgage Loan Defaults

Speedy Mortgage is a startup internet lending company focused on supplying mortgages to customers through a completely online process. With the mortgage lending business being very mature and competitive, Speedy is focused on providing an efficient process for their customers. Speedy would also like to minimize default rates of buyers so that their mortgages can be sold in the secondary mortgage market so Speedy will be able to generate more income.

Problem Statement: Speedy needs to classify potential customers into two groups, those likely to default and those unlikely to default.

Data Source: Fannie Mae single-family loan performance data (2000-2018 Q2)
https://www.fanniemae.com/portal/funding-the-market/data/loan-performance-data.html
Fannie Mae is a government sponsored enterprise that was created in 1938 to assist in the housing market recovery from the effects of the Great Depression. Fannie Mae along with a similar entity, Freddie Mac, purchases loans originated and secured through other lending entities and sells them in the bond market.

Fannie Mae has made available a large dataset of over 35 million mortgage loans however there a quite a few exclusions from this dataset including a lot of risky loans such as interest only and balloon amortization. A complete list of exclusions can be found on the websites FAQ for this dataset.

The data files are split between two types of text files, acquisition and performance data with a text file for each quarter of the year. A glossary with a detailed description of the columns is also available on Fannie Mae's website. The acquisition files have 25 columns and the performance files have 31 columns.

Key columns in acquisition files: loan identifier, original interest rate, original loan term, origination date, original loan-to-value (LTV), number of borrowers, original debt to income ratio, borrower credit score at origination, first time home buyer indicator, loan purpose, property type, zip code short, primary mortgage insurance percent, mortgage insurance type

Key columns in performance files: loan identifier, current interest rate, current actual UPB, loan age, maturity date, metropolitan statistical area (MSA), current loan

delinquency status, modification flag, foreclosure date, net sale proceeds, principal forgiveness amount, foreclosure principle write-off amount

Evaluation: Several values will be considered in order to group potential borrowers into either a likely to default or not likely to default category. A model will be developed that factors in borrower's credit score, original interest rate, original debt to income ratio, first time home buyer indicator and zip code. With this information, Speedy Mortgage will be able to determine which geographic areas and types of borrowers to market in an effort to reduce the quantity of potential mortgage loan defaults.