Machine Learning Nanodegree

Capstone Project Proposal

Predicting Loan Default in the Fannie Mae Single-Family Loan Performance Data

Felipe Ferreira

Udacity
Fall 2017

1 Domain Background

The Federal National Mortgage Association (FNMA), commonly referenced as Fannie Mae, is corporation with the purpose to expand the secondary mortgage market by securitizing mortgages in the form of mortgage-backed securities (MBS). In a nutshell, Fannie Mae acquires mortgage loans from primary lenders such as Wells Fargo, Chase, and Quick Loans, among others. After the mortgages' acquisition, Fannie Mae groups them in mortgage pool, which, by a diversification effect, has lower risk than the risk of each mortgage individually. Until the 2008 financial crisis, these mortgage pools (MBS) sold by Fannie Mae were considered stable investments; however, the crisis showed that these pools were not as safe as people thought. In fact, during the crisis burst, many people defaulted on their mortgages, causing these securities prices to decreases significantly, thereby severely impacting the performance of these investments.

Following the crisis, Fannie Mae with the intention to promote a better understanding of the credit performance of Fannie Mae mortgage loans, started providing loan performance data on a portion of its single-family mortgage loans. The goal of this project is to predict from Fannie Mae single-family performance data, those borrowers who are most at risk of defaulting on their loans

2 Problem Statement

The project's goal is to predict with high degree of accuracy whether a given mortgage will default or not based on the mortgage's characteristics at the time of its acquisition by Fannie Mae. Among the loan features, we will use to make our predictions; we have the debt-to-income ratio, borrower's credit score, and loan amount, and many others.

3 Datasets and Inputs

Fannie Mae has made available on its website[1] a portion of its single-family mortgage loans dataset. The Single Family Fixed Rate Mortgage dataset contains a subset of Fannie Maes 30-year or less, fully amortizing, full documentation, single-family, conventional fixed-rate mortgages.

The loan performance dataset is divided into two files for each acquisition quarter. The **Acquisition file** includes static data at the time of a mortgage loans origination and delivery to Fannie Mae. The **Performance file** contains the monthly performance data of each mortgage loan from the time of Fannie Maes acquisition up until its current status as of the previous quarter. For the project, we will use the data from the 2^o quarter of 2007.

Since we want to predict whether a loan will default given its characteristics at the time of its acquisition, we will merge the **Acquisition file** which contains the features to our model with the **Performance file** which contains our target variable (default or not default) using

[1] `http://www.fanniemae.com/portal/funding-the-market/data/loan-performance-data.html`

the *Loan ID* presented in both files as the key field for the merging. Moreover, since the dataset does not have a field directly informing whether a loan has defaulted or not, we will use the *Foreclosure Date* field as a proxy - if it is populated, it means that the loan has defaulted.

The final dataset has 287,286 samples and 25 features. The features are:

- Sales Channel - Type: Categorical
- Seller Name - Type: Categorical
- Original Interest Rate - Type: Numeric
- Original Unpaid Principal - Type: Numeric
- Original Loan Term - Type: Numeric
- Original Loan-to-Value Ratio - Type: Numeric
- Original Combined Loan-to-Value Ratio - Type: Numeric
- Number of Borrowers - Type: Numeric
- Debt-to-Income Ratio - Type: Numeric
- Borrower Credit Score - Type: Numeric
- First-Time Home Buyer Indicator - Boolean
- Loan Purpose - Type: Categorical
- Property Type - Type: Categorical
- Number of Units - Type: Numeric
- Occupancy Status - Type: Categorical
- Property State - Type: Categorical
- ZIP Code (first 3 digits) - Type: Categorical
- Relocaion Mortgage Indicator - Type: Boolean
- Month of Mortgage Acquisition by Fannie Mae - Type: Numeric
- Year of Mortgage Acquisition by Fannie Mae - Type: Numeric
- Month of the First Payment - Type: Numeric
- Year of the First Payment - Type: Numeric

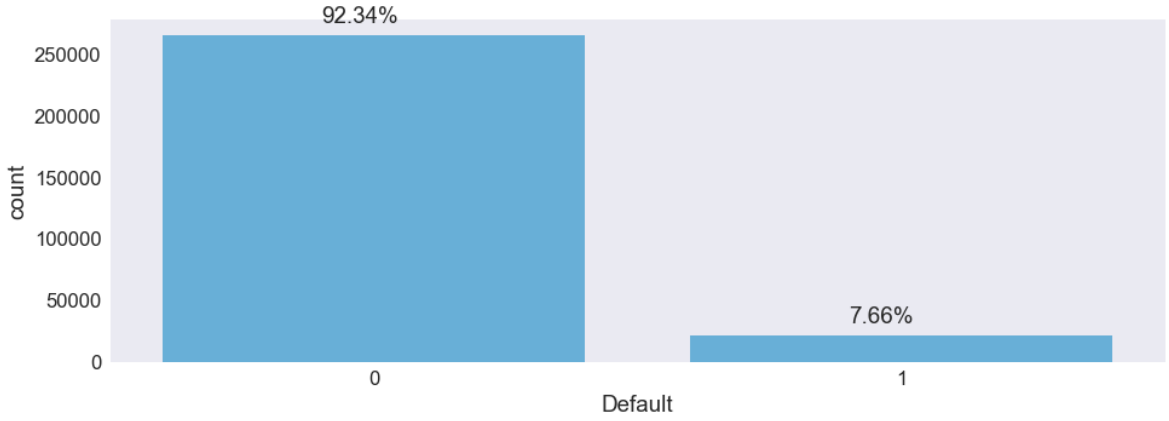

Figure 1: Imbalanced Dataset

Given we have a large number of categorical features, and each has several categories. We plan to use Label Enconding, instead of One-Hot-Encoding. Otherweise, we will deal with a large and sparce input matrix.

From Figure 1, we see that the number of loan that defaulted since their acquisition date are only 7.66% of the samples. Thus, it is clear we are dealing with an imbalanced dataset (when the number of observations in one class is much larger than those belonging to other classes). This behavior was alreasy expected, since loan default prediction belongs a class of problems we can call **Anomaly Detection**. Belonging to this class, we also have identification of diseases ans fraudulent transactions in banks.

4 Solution Statement

The project attempts to predict whether a mortgage loan will default based on its characteristic at the time of the loan acquisition by Fannie Mae.

As ensemble models, particularly Random Forest algorithms, have been extensive and successfully used to solve this type of problem, we will use these type of machine learning algorithms to solve our problem. Besides, Random Forest models are more accessible to interpret, because we can easily inspect which features or variables contribute to the prediction and their relative importance based on their location depthwise in the tree.

5 Benchmark Model

Our benchmark model will be a simple decision tree without any parameter optimisation. Because we can see Random Forests as an extension of Decision Trees, we think the latter might be a good benchmark for the former.

6 Evaluation Metrics

The evaluation metric for the proposed model and the benchmark model is the Receiver Operating Characteristic (ROC) curve. Although sensitivity and specificity could be valid metrics, they are dependent on the right assessment of the appropriate trade-off between sensitivity and specificity. The ROC Curves provide the representation of the range of possible cut points with their associated true positive rate vs. false positive rate. And the area under the curve AUC offers the means to compare two or more models.

7 Project Design

- **Programming Language:** Python 3.6+
- **Libraries:**
  - seaborn==0.8
  - pandas==0.20.3
  - numpy==1.13.3
  - matplotlib==2.1.0
  - scikit-sklearn==0.19.1
  - imbalanced-learn==0.3.1
  - missingno==0.3.7
- **Workflow**
  - Download *Acquisition File* and the *Performance File* from Fannie Mae's website;
  - Preprocess the files - merging, filling in missing values, labeling categorical features, and resampling the data to correct for the imbalanced dataset;
  - Train the baseline model - Decision Tree on the balanced dataset;
  - Train our Random Forest Model;
  - Validate the project model through cross-validation;
  - Depend upon the project model performance compared to the baseline model, we might try to optimize its hyper-parameters;