Capstone Project

Sriharsha M S

Machine Learning Engineer Nanodegree

Jan 18th, 2016

Lending Club *Investor ML*

Predict whether individual loan application will be charged off?

Domain Background

LendingClub is an online financial community that brings together creditworthy borrowers and savvy investors so that both can benefit financially. They replace the high cost and complexity of bank lending with a faster, smarter way to borrow and invest. Since 2007, LendingClub has been bringing borrowers and investors together, transforming the way people access credit. Over the last 10 years, LendingClub helped millions of people take control of their debt, grow their small businesses, and invest for the future.

LendingClub balances different investors on its platform. The mechanics of the platform allow LendingClub to meet the objectives of many different types of investors, including retail investors. Here's how it works. Once loans are approved to the LendingClub platform, they are randomly allocated at a grade and term level either to a program designed for retail investors purchasing interests in fractions of loans (e.g. LendingClub Notes) or to a program intended for institutional investors. This helps ensure that investors have access to comparable quality loans no matter which type of investor they are. LendingClub goal is to meet incoming investor demand for interests in fractional loans as much as possible.

The design of LendingClub platform emphasizes how important retail investors are to LendingClub. For LendingClub retail investors are key component of our diverse marketplace strategy. Retail investors are—and will always be—the heart of the LendingClub marketplace.

Problem Statement

This project "*Investor ML*" tool specifically built keeping retail investors in mind. *Investor ML* aims to predict probability risk of borrower being charged off and not fully pay loan amount, called "**Risk Rate %**". LendingClub can provide "**Risk Rate %**" predictions from Investor ML for each loan of borrower as an additional indicator for investors to make investment decisions. Retail investors use loan information, borrower information like fico score, loan grade etc. and decide to invest in fractional loans. Additional statistic "**Risk Rate %**" learned from historical loans may also act as additional information and help retail investors to diversify their investment.
The datasets are provided by LendingClub, we will use lending data from 2007-2011 and be trying to classify and predict whether or not the borrower paid back their loan in full. Data is available to download https://www.lendingclub.com/info/download-data.action.

Datasets and Inputs

The datasets are provided by LendingClub, we will use lending data from 2007-2011 and be trying to classify and predict whether or not the borrower paid back their loan in full. Data is available to download https://www.lendingclub.com/info/download-data.action.
Data fields: Some of the data fields listed below

LoanStatNew	Description
annual_inc	The self-reported annual income provided by the borrower during registration.
delinq_2yrs	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years
dti	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.
emp_length	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
fico_range_low	The lower boundary range the borrower's FICO at loan origination belongs to.
home_ownership	The home ownership status provided by the borrower during registration or obtained from the credit report. Our values are: RENT, OWN, MORTGAGE, OTHER
inq_last_6mths	The number of inquiries in past 6 months (excluding auto and mortgage inquiries)
int_rate	Interest Rate on the loan
last_fico_range_low	The lower boundary range the borrower's last FICO pulled belongs to.
loan_amnt	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
loan_status	Current status of the loan
open_acc	The number of open credit lines in the borrower's credit file.
purpose	A category provided by the borrower for the loan request.
revol_bal	Total credit revolving balance
revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
term	The number of payments on the loan. Values are in months and can be either 36 or 60.
total_acc	The total number of credit lines currently in the borrower's credit file
verification_status	Indicates if income was verified by LC, not verified, or if the income source was verified

Solution Statement

The *Investor ML* solution will be predictions of either loan status will be fully paid or not. First, I will analyze the dataset to choose the features from all available 150 features. Identify the features that will be available during loan application process. Analyze, understand dataset and gain good understanding of dataset. Then I will prepare features and targets label required for classification. Perform various data processing steps like missing value imputations, scaling features, one-hot encoding of categorical features.

For training models, I will compare logistic regression, ensemble models since this is a classification problem. Finally, I will select the best model for this problem and fine tune parameter to get best accuracy.

Benchmark Model

For this problem, the benchmark model will be naïve predictor, naïve predictor is designed using Gaussian NB model. I will try to beat its performance of naïve predictor with other algorithms.

Evaluation Metrics

Investor ML, is particularly interested in predicting who will fully pay back the loan. It would seem that using accuracy as a metric for evaluating a particular model's performance would be appropriate. Additionally, identifying someone that does not fully pay back loan would be detrimental to *Investor ML*, since they are looking to invest on individual who will pay back loan. Therefore, a model's ability to precisely predict those who fully back is more important than the model's ability to recall those individuals. We can use F-beta score as a metric that considers both precision and recall:

$$\mathbf{F\beta} = (1 + \mathbf{\beta}2) \cdot \frac{\mathbf{precision} \cdot \mathbf{recall}}{(\mathbf{\beta}2 \cdot \mathbf{precision}) + \mathbf{recall}}$$

In particular, when β=0.5, more emphasis is placed on precision. This is called the **F**0.5 score (or F-score for simplicity). Looking at the distribution of classes (those who fully pay loan, and those who don’t), it's clear most individuals fully pay back loan. This can greatly affect accuracy, since we could simply say "this person will pay back loan" and generally be right, without ever looking at the data! Making such a statement would be called naive, since we have not considered any information to substantiate the claim. It is always important to consider the naive prediction for your data, to help establish a benchmark for whether a model is performing well. That been said, using that prediction would be pointless: If we predicted all people pay back loan, *Investor ML* would identify all loan application as fully payable loans.

Project Design

Load the dataset downloaded from LendingClub, I will first take glimpse of the data see what the shape is and what are the different features available. Identify how many missing values(NaN's) are present. If data is not populated at least 40%, I won't be using those columns. Then I will profile the data, learn each feature of the dataset, develop intuitions on features and how it is relevant to the target loan application status. As data contains some features that will not available during loan application stage, I will remove these features. Based on the developed intuitions, I will select the features and prepare the dataset that can be used train the model. I will perform some graph visualization for better understanding of the data distribution. I will scale the numerical features, I will one-hot encode the categorical features.

To train models, I plan to choose different models to compare. Because this is a classification problem, a few approaches in my head would be Logistic Regression, boosted decision trees and other ensemble methods. Using grid search cross-validation I can find which model performs best, and then use that one to tweak relative parameters.

I expect to spend 60% of the time on understanding the data, cleaning, processing part and 40% of the time on training models and tweaking parameters.

Reference

1. https://www.lendingclub.com/info/download-data.action
2. https://www.lendingclub.com
3. http://scikit-learn.org/stable/index.html
4. https://github.com/pandas-profiling/pandas-profiling