Machine Learning Engineer Nanodegree

Capstone Proposal

Serge Bouschet
April 25th, 2017

Lending Club Loan Data

Domain Background

The growth of peer to peer lending business over the last decade has helped millions of people achieve financial goals through online marketplaces. The Lending Club connects borrowers and investors by providing qualified loans, leveraging technology to lower the cost from traditional banks.

Loans offered offered through the platform are risk assessed to determine a credit rating and assign appropriate interest rates. However investors are still exposed to loan defaults, so what can data tell us about borrowers attitude to repayment and can we lower exposure to risk? The Lending Club's dataset offers a complete history of loan data through 2005-2017 with many features available to derive insight from this business.

Research Papers and References

Stanford University, Machine Learning Curriculum Project Report:
http://cs229.stanford.edu/proj2015/199_report.pdf
Online Discussion:
https://stats.stackexchange.com/questions/131255/class-imbalance-in-supervised-machine-learning
Online research Paper: https://svds.com/learning-imbalanced-classes/
Book, data mining and knowledge discovery handbook:
http://www3.nd.edu/~dial/publications/chawla2005data.pdf

The Cross Validated post is very instructive and lists different approaches to deal with class imbalanced in supervised learning. The solutions discussed are directly linked to our problem since we'll have to deal with a 90/10 class distribution. I am particularly interested to explore data sampling and cost sensitive techniques in this project.

The stanford research paper is also analysing the lending club data to predict the probability of default, it is an instructive paper that we might want to use to compare our approach and performance. This paper did not pursue any specific method to deal with Class imbalance, however the evaluation method selected takes the nature of the dataset into account.

Problem Statement

Measuring risk is a major concern in peer to peer lending, therefore the ability to predict Loan defaults from the current loan data is the first objective we're going to address in this project. Predicting which loans are likely to default is beneficial in many situations.

For example, to assist investor's decision in selecting a loan. The Lending Club is a marketplace where investors make decision based the information available to them. Making accurate predictions based on this information could lower the risk associated with an investment, this is potentially a very useful tool.

The loan status in the dataset is a categorical variable which offers a very granular set of statuses, including statuses for Late repayments. This status can be either summarized to produce a binary label or can be used as continuous variable to explain the level of risk exposed in the transaction.

Loan Status	**Binary Classifier**
Fully Paid	Current
Current	
Charged Off	Default
Default	
Late (16-30 days)	
Late (31-120 days)	
In Grace Period	

We can find a description of all the loan statuses on the lending club website: https://help.lendingclub.com/hc/en-us/articles/215488038 this reference is really useful to interpret each status.

Default loan prediction on lending club's data presents characteristic of imbalanced dataset classes. Essentially, Current loan statuses account for a very large majority of all classes. A simple benchmark model gives a very high accuracy because most borrowers repay their loan on time and in full. This will need to be addressed when selecting the solution. An appropriate evaluation method is also required to address this.

Datasets and Inputs

The data set offers historical data that covers a period of more than 10 years. Only approved loans are listed in this dataset. All loan applications are reviewed and scored by the Lending Club and then approved in order to be offered on the marketplace to potential investors. During the 10 years covered by the data, the lending process has evolved which is reflected in the data collected, including the credit scoring method or text populated in the *Title* field. From November 2013, regulations have been introduced to offer a new scoring method called VantageScore. This is an important consideration to use during analysis using the credit score feature. We can also observe from the loan statuses that this dataset contains both historical and current data. Some of the loans are effectively still in progress, whether the status shown is current or late, it means these transactions are not final. We will need to

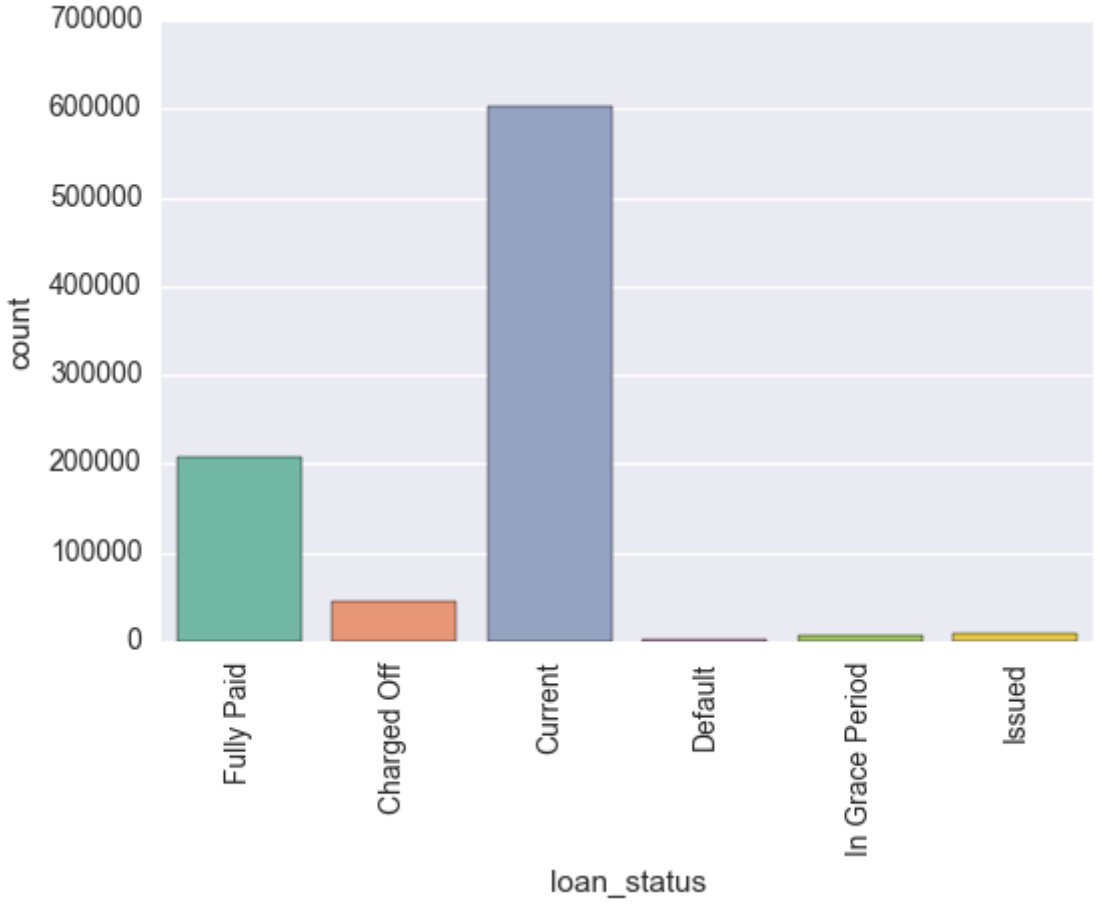
take this into account when building our training dataset as there are risks of misleading our classifier.

In this project we are analysing a dataset published on Kaggle: Lending Club Loan Data. https://www.kaggle.com/wendykan/lending-club-loan-data

The dataset contains over 80 attributes including

- Personal Borrower Data
    - Personal information
    - Financial history and scoring
    - Motivation
- Transaction Information
    - Temporal
    - Financial
    - Statuses

Looking more closely at the loan statuses and in particular into our derived binary classifier (current or default), it is evident that this dataset offers a class imbalanced challenged. The full dataset contains 887, 379 data points covering the entire period of 12 years. Loan status are distributed as follows.



Our aim will be first to run through a feature selection process that eliminates features that have no predictive power, lend to overfitting and leaking data about the outcome of the loan.

Important Features to explore: Distribution, Correlations, Unique Values, Nulls, Outliers

Solution Statement

Our loan default prediction using this dataset is a classic supervised learning, binary classification problem. Our first goal during exploratory data analysis is to determine which slice of the dataset, column projection, status and temporality should be considered for our candidate models.

In order to tackle our imbalanced class problem, we'll also need explore a method do sample our data, then run a supervised learning classification method to predict a binary loan status (1: Default, 0: Current).

The following classifiers will be evaluated against our metrics. These have been considered as they have the ability to produce our target metrics (sensitivity and precision):

- Logistic regression
- Random Forest
- SVM

In addition to these we may also evaluate XGBoost and Multi Layer Perceptron Models.

Benchmark Model

Our benchmark for this problem is to evaluate the accuracy of predicting that all loans do not default using the loan status feature available. Non default loans are either assigned "Current" or "Fully Paid" statuses. If we turn this status into a binary class, we can achieve 91.2% correct prediction by assuming that all loans never default. However while this accuracy score is very high, it is also less than satisfactory when it comes to predicting defaults. The true positive rate from this benchmark is 0% (sensitivity), making those predictions pretty much useless.

Benchmark: All loans are current (score: 91.2%*)
*The actual benchmark score will be refined during data analysis and feature selection. We anticipate records to be dropped from the training set which will probably modify this figure

Evaluation Metrics

Evaluating imbalanced datasets (91.2% accuracy with benchmark model).

With such an imbalanced asset classes, it is important to address the accuracy paradox from the outset and consider other metrics for evaluation such as precision, recall and F1 score.

$$Precision = TP \div (TP + FP)$$
$$Recall = TP \div (TP + FN)$$

F1 score is the weighted average of the precision and recall, where best value is 1 and worst 0. It is mathematically defined as follows:

$$F1 = 2 \times (Precision \times Recall) \div (Precision + Recall)$$

ROC curve analysis should also a useful tool to measure and understand performance of an algorithm with respect of precision and recall.

Our goal for this project is to achieve 50% True Positive Rate (or Recall), predicting loan defaults correctly. That means predicting at least one out of two loan default correctly. It's an arbitrary goal and this project will determine how challenging this target actually is.

Project Design

Our analysis will step through four key analysis domains

1. Data Exploration and Preparation
    - Descriptive data analysis: statistical & aggregation and correlations

In this section, our aim will be to get an in depth understanding of the data. Data distributions and correlations will be at the center of this analysis.

    - Outlier analysis and data cleansing

This section will attempt to determine what data sample will be considered to train our model. Is all historical data from 2005 to 2012 is worth considering? Outlier analysis also plays a role and will be covered in this section.

    - Feature selection and Feature engineering

Our objective will be to determine which features should be kept and which features can be engineered to add information and improve our model training. We will eliminate features that do not provide any predictive power, we'll use correlation and distribution information, analysing skewed and missing data.

    - Data preparation

This consists transforming data using one-hot encoding, data normalisation, data scaling and splitting data into training and test sets.

2. Supervised Learning Classifier
    - Benchmark evaluation

Run a simple logistic regression classifier against selected features and training and test sample, use this as a baseline to measure our evaluation metrics against the following algorithms:

        - Decision tree
        - SVM
        - Random Forest
    - Optionally we may evaluate more algorithms

    - Multi Layer Perceptron
    - XGBoost

3. Model evaluation and Tuning
    - Assess performance

As well as accuracy, precision recall and F1 score will be produced and analysed. We'll also plot ROC curves. In this step our goal will be to measure performance of loan default prediction for each classifier in particular true positive rate.

    - Investigate Data Sampling techniques

In an attempt to rebalance the dataset we will look at whether random sampling makes any difference to the performance. Does clustering help uncover better predictions using selected clusters (under sampling method)?

    - Investigate Cost sensitive learning

This section will look at opportunities to run algorithm that can "penalise" bad predictions (false negatives) and ensure that the classification is more sensitive to prediction loan defaults correctly

    - Model Tuning

Using tools such as grid search can help tune our selected model parameters

4. Conclusion
    - Learnings
    - Future improvements