Rich Helle
July 18, 2018

Capstone Proposal

Domain Background

Lending Club is a peer-to-peer lending platform pairing individual investors with loan seekers. One of the draws of this platform for investors is the potential for higher returns because the traditional intermediary, a costly bank that shares little information, is replaced by a transparent online platform with minimal fees. Lending Club platform allows investors to individually select loans for their investment portfolio, providing certain data collected on each loan to aid the investor.

I invested with Lending Club using the automated investing tool, letting the platform choose the loans for me, this has not proved to be a good strategy. To date my adjusted net annualized return has been a dismal 0.07%. While the weighted average interest rate on my portfolio is above 10%, what's brought my return down to nearly zero is loan defaults which do not get fully repaid. My goal with this project is to use the data provided to determine how I might make a higher return.

Problem Statement

The problem I intend to solve with this project is to identify the feature values that best explain whether a loan will be repaid or default. I plan to use supervised learning because the model will clearly yield the features which influence the repayment or default variable. I'll not be using neural networks because these models do not clearly yield the importance of the features in predicting repayment and default.

Datasets and Inputs

From the link here I will be using the 2007-2011 dataset because each loan has terminated either ful repayment or default. I have removed the loans which were declined, since these were never available to investors. The resultant data consists of 39,786 loans with 145 features. A copy of this data and the data dictionary in the zip file of this capstone proposal.

There is an imbalance in the target class between the number of defaulting and repaid loans. Defaults represent only 14.25%, so there are approximately 6 times as many repaid loans to unrepaid loans.

To guard against overfitting the data, a problem noted in (I-Cheng Yeh, 2009), I will use k-fold cross validation. Additionally there will be a test set, unused in training, used for model evaluation.

The data includes many fields, not all of which are observable at the time of the investment. Since the goal is to select loans based available information, the model input will be restricted to the data available at investment. Screenshots of the information available at investment and the filter options available for sorting loans have been included in the zip file of this capstone proposal.

Solutions Statement

The solution to this problem will be a model which uses the information available at investment to determine which loans are most likely to be repaid. The model will be asked to predict the repayment or default for a test set of the data described above.

Rich Helle
July 18, 2018

Capstone Proposal

Solutions Statement (Continued)

The project will explore at least three models including Support Vectors Machines (SVM), Random Forest (FR) and a naïve Bayes (NB) classifier. Research noted below suggests these will perform well for this project. Given the simplicity of testing additional classification methods I will explore other options and include the findings only if they perform well or offer an unique insight.

Benchmark Model

I will construct a K-nearest Neighbors(KNN) model to compare. The research below suggests this model performs reasonably well and will prove a more difficult mar Additionally I will compare these test set results to the performance of my own Lending Club portfolio, which was chosen via automation on the platform.

Evaluation Metric

Loan default will be the target variable. Model Performance will be determined by the accuracy, recall, precision and F_β of the model.

Accuracy: $$\frac{True\ Positives + True\ Negatives}{Total\ Population}$$

Recall: $$\frac{True\ Positives}{True\ Positives + False\ Negatives}$$

Precision: $$\frac{True\ Positives}{True\ Positives + False\ Positives}$$

F_β: $$(1 + \beta^2) * \frac{Precision * Recall}{\beta^2 * Precision + Recall}$$

To address the imbalance in the target class I will also use the ROC Curve to assess model performance. The ROC cure is true positive rate plotted as a function of the false positive rate.

Project Design

The workflow for this project will begin with importing the data, then pairing down the 145 features down to the observable features at the point on investment.

Next, cleaning the data of any potential errors or any possible default values that would interfere with analysis. For example interest rates for these loans should be positive and less than 100%, so a rate that is negative or greater than 100% would need to be resolved or removed.

This would include exploring the data with summary statistics and visually. Making some observations about the data and the relationships within.

Next will be building a series of supervised models performing grid searches over each models relevant hyperparameters and selecting the models that perform best on the evaluation metrics.

Rich Helle
July 18, 2018

Capstone Proposal

Project Design (Continued)

Finally I'll compare the best model's classification performance against the actual results. I'll also compare the portfolio of loans the model deems likely to repay to the overall loan default statistics noted above. Finally I'll prepare an analysis of the models performance to my own Lending Club portfolio's performance.

To close I'll discuss the features the model identifies as most important in determining which loans will repay and use this to select my next round of investments on Lending Club.

Research exploring the applicability of machine learning models to default predictions have been somewhat successful. (Dinesh Bacham, 2017) found that a RF model out-performed Moody's Analytics RiskCalc, a loan scoring and benchmarking platform offered by the Moody's Analytics.

(I-Cheng Yeh, 2009) examined six different classification methods and discovered KNN and performed best predicting training but failed to generalize to a validation set, while NB, neural networks and classification trees all generalized well to the validation set. To guard against overfitting on the training set, I will use K-fold cross validation.

(Bagherpour, 2017) examined four machine learning algorithms on mortgage default data and noted that random forests and SVM produced excellent results. This research also noted using Synthetic Minority Oversampling Technique (SMOTE) to address the imbalance between defaulting and non-defaulting in the population. I will not attempt this method for this project.

Capstone Proposal

Works Cited

Bagherpour, A. (2017). *Predicting Mortgage Loan Default with Machine.* Riverside: University of California. Retrieved from http://economics.ucr.edu/job_candidates/Bagherpour-Paper.pdf

Dinesh Bacham, D. J. (2017, July). Machine Learning: Challenges, Lessons, and Opportunities in Credit Risk Modeling. *MOODY'S ANALYTICS RISK PERSPECTIVES, IX*(July 2017). Retrieved from https://www.moodysanalytics.com/risk-perspectives-magazine/managing-disruption/spotlight/machine-learning-challenges-lessons-and-opportunities-in-credit-risk-modeling

I-Cheng Yeh, C.-h. L. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications 36*, 2473–2480. Retrieved from https://github.com/wangzongyan/Default-of-credit-card-clients-Data-Project/blob/master/The%20comparisons%20of%20data%20mining%20techniques%20for%20the%20predictive%20accuracy%20of%20probability%20of%20default%20of%20credit%20card%20clients.pdf