LOAN HELPER

Ramakant Khandel
Computer Science Department, Indiana University

Abstract

In this system, I have applied case base reasoning approach to the student loan application domain. The number of applicants for student loans and loan defaulters are increasing rapidly over the past few years. The lenders face huge losses, due to steep increase in the defaulters. Credit histories of young student loan borrowers are not well established and, consequently, less likely to be predictive of future credit behavior. Student's investment in education is made with the expectation that the future financial returns from acquired skills and increased income will outweigh the current cost both direct and indirect. My system helps lenders determine defaulters based on their data.

Introduction

The approach used in this system is case base, to solve the current problem with the help of previously solved problem. It uses distance function to determine similar cases for the new case form to the case base to predict the outcome. Case base reasoning can be seen as a cycle of retrieve, reuse, revise and retain. Case base can be used for classification and regression. The decision making in CBR is dynamic. Hence, I thought it will be quite interesting to test the power of CBR to the financial domain. The main idea behind this system 'If a previous student applicant has not defaulted then there is a high chance that a student with similar profile will also not default'.

Review of Relevant Work

In research for my system, I came across a few systems that are similar to my system.

Predicting Student Loan Default for the University of Texas at Austin

This project comprised two distinct parts: an investigative research portion and a data mining portion. The pure research portion of the project consisted of systematically testing the various groups of academic and demographic data to see which variables were predictive of eventual loan default. Data was entered into the series of logistic regressions incrementally in six different blocks: demographic and background data, high school information, degree and major data, credit hour information, transfer information, and any available financial data. The final model combined the demographic, degree completion, credit hour, and financial variables. Race/ethnicity and gender remained highly significant and accounted for approximately 20% of the variation in default behavior explained by the model. The highest educational degree attained, academic grade level, and school of enrollment variables provided a detailed degree-completion and persistence profile. The data mining model also helps to find variables that would allow either the loan guarantor or the institution to identify at-risk borrowers as early as possible and take intervention measures to help prevent student loan default.

This paper discusses data mining approach to identify defaulters. The results from the case based and data mining models will be interesting to compare.

What Matters in Student Loan Default: A Review of the Research Literature

Student loan default has considered (a) the characteristics of students as they begin college (e.g., family income, race/ethnicity); (b) students' college experiences (e.g., type of institution, field of study, educational outcomes); (c) students' financial aid and the amount of debt they incur; and (d) students' employment and income after college as well as their overall debt.

Nearly all studies that considered the age of the student either while enrolled in school or at the start of the loan repayment period concluded that as age increases so does the likelihood of loan default, even after controlling for other important factors such as income.

Student loan default occurs across the range of students' socioeconomic contexts. The family

structure, the parents' education, the parents' marital status, and the family's eligibility for federal assistance such as Aid to Families with Dependent Children are all proxies for the social and economic capital students can "cash in" to attend college and then later to repay loans. One of the study this paper mentions is that Family structure affects in a number of ways the likelihood of defaulting on loans. First, the greater the number of dependents claimed by a student, the greater the likelihood of loan default.

Students whose parents had higher levels of formal education were less likely to default than first-generation college students.

This paper also mention an important study about Students' academic experiences in postsecondary education—credits attempted, credits completed, credit hours failed, grades, transfer patterns, enrollment patterns, and time to degree/certificate emerge as the strongest predictors of loan default. Students who enroll continuously, enroll in more rather than fewer credit hours, complete their attempted courses (i.e., do not receive incompletes), and graduate within eight semesters are less prone to default on average.

The majority of the research we reviewed suggested that completing a postsecondary program is the strongest single predictor of not defaulting regardless of institution type.

This paper mentions interesting features but due to lack of such dataset, my model cannot be tested for the same.

College on Credit: A Multilevel Analysis of Student Loan Default

This paper discusses about the features such as race, age, gender likely account for a degree of variation in default probability but the nature of these relationships is not entirely clear. It would have been good to consider these feature and test my model based on the same, but it was difficult to create such a dataset manually.

It also discusses about the academia experience of the students. It mentions an important point wherein a student who has left college without a degree are 10 times more likely than their peers to default. Debt levels should predictably rise in relation to the amount of time a student stays enrolled in college, since longer enrollments are associated with greater levels of debt accumulation. When students accumulate large levels of debt, they have a greater likelihood of defaulting. If students are unable to find employment upon leaving college or become unemployed at some point during repayment, then they may face greater risks of entering into default. Institution Level Predictors of Default can be an important feature as Colleges with high default rates are simply serving a high-risk clientele that is likely to default regardless of which higher education sector they attend. This feature is not considered by my model.

Multivariate Analysis of Student Loan Defaulters at Texas A&M University

The data contains information such as describing high school coursework, SAT scores, college grade point average (GPA), length of attendance at TAMU, graduation status, amounts of financial aid received, financial need assessment, gender, marital status and many other aspects of students' backgrounds and college experiences. It performs statistical analysis proceeds by determining the relationships between borrower characteristics and default behavior within a past population of borrowers. The known outcomes (i.e., default behaviors) of this population serve as the basis for statistical estimation. This analysis will attempt to increase our understanding of the factors that are related to default at Texas A&M University.

One of the goal of variable selection process is to find, among all possible relevant variables, the subset of variables that best explains default behavior.

This paper also contains results after performing multivariate analysis on the dataset. The paper also provides a decreasing order of importance of features based on the analysis performed. The order is GPA< Exit Counselling, Graduation Indicator, College or School Last Attended, Age of Borrower at time of entering repayment, Number of hours failed, Race/Ethnicity of Borrower, Highest Educational Level attained by student's mother, Expected Family Contributor, Number of hours transferred, adjusted gross income and Gender.

I have considered some of the features mentioned in this paper. It would be interesting to compare the results using the features mentioned in the paper and the features considered by my system. Also, the analysis used can be used to determine the weight of feature in order of importance compare to the current which is done manually.

Student Loans and Repayment: Theory, Evidence and Policy

This paper discusses about the current student loan scenario, federal program for the same and an optimal lending arrangement between providing insurance and incentives to borrowers, while ensuring the lender is repaid in expectation. It mentions about the different scenarios wherein student have defaulted on their loan.

This paper discusses about an important feature market uncertainty which can greatly improve the results for retrieval of cases from the case base. It also mentions about other features such as dependents, performance, setting borrowing limit, etc.

The Interplay between Student Loans and Credit Card Debt: Implications for Default Behavior

This paper describes a model to match statistics regarding student loan debit, credit card debt and income of young borrowers with student loans. The borrowers with similar debt levels in the two markets would default on student loans than on credit card loans. It talks about the large risks lenders face from borrowers due to increase in college dropout rates and unemployment rate.

The few parameters discussed in the paper college dropout rate and unemployment rate can improve the results of retrieval of cases form the case base. For my system I have not considered these feature values due to lack of data for the same.

System Workflow

This section describes about the entire flow of the model from the point of case being added to the case base to the point where the decision is made.

Case Base Design

The different features used to represent the student information are as follows: Enroll School, Course, Location, Unemployed, Salary, Credit Score, Course Performance, No Payment Due, Family Income, Max family Credit score.

These features represents student educational performance, family background, school data, on campus job and previous credit.

Location takes value from 1 to 5 depending upon the job opportunities in that city. Course performance takes value from 1 to 4 representing worst, average, good and excellent. No Payment Due represents previous loan which is due. If previous loan is due then the student tis not eligible for new loan. Family Income represents the family income of the student and Max Family Credit Score takes values from 1 to 5 depending upon the maximum credit score of the individual in the family.

The different weights assigned to the features are as follows:

Location	0.25
Unemployed	0.05
Salary	0.05
Course Performance	0.35
Family Income	0.15
Max Family Credit Score	0.15

The weights of the feature change depending upon the input features for an applicant. The weight associated for a feature is distributed based on the proportion of other features in the set. Let's say if the value for family income is not provided then the weight 0.15 is distributed among other features. This distribution is based on the proportion of weights for other features.
So, (0.05/1)*0.15 +0.15=0.1575
The weight of salary after adjusting is 0.1575.

The features salary, family income is neutralize before calculating the similarity score. The salary and family income below their respective averages is 0 and above their respective average is 1.

In order to match the new applicant with the cases in the case base, the model calculates Euclidean distance for all the features based on the weight of the features for all the cases in case base. The lesser the distance more similar the new case is with the new applicant. The class label from top k cases which are at a minimum distance from the new case is then considered to predict the class label for the new applicant.

Learning in the system:

The learning of the model is dependent on the feedback received. For e.g.: If the system suggest a loan depending upon the values of the features for an

applicant. In the future let's say the student defaults. A study of this failure can result in determining new features, adjusting the weight of the features, handling special cases etc. Thus in the future if similar case is received, it will learn from its previous mistake and make decisions based on the new learning.

Issues Faced

The increase case base for different scenarios will help the model to determine result with increased accuracy for cases. As the case base increases the time complexity will also increases. To improve the time complexity, the case base can be distributed. The case base for my model is not that large. Hence I have not considered the distributed case base.

One of the important issues with CBR is increasing the scale of one dimension increases the importance of that feature. To overcome this, I normalize the value for salary and family income. The normalize function for salary is the difference between the salary is divided by 1000 and family income is divided by 10000. The results were biased even after normalizing the family income. This is because let's say for e.g.: Family income for household A is 150000 and B is 15000. After normalizing it the difference is 13.5. The scales for all the other features is from 1 to 5. This results in biasness. To overcome this, Average family income is calculated for all the cases and salary below the average is considered 0 and above average is considered 1. The results were better using the second neutralizing technique compared to the first one. The first neutralizing technique introduced a biased for the respective feature while calculating the similarity score. To remove the biasedness in the score second technique was chosen.

Evaluation

The model was tested on 9 different scenarios wherein the values for different features were changed. The system produced correct results on 5/9 cases and wrong results on 4/9 cases. The details of the same can be found in output_b552.txt file.

Analysis of Strengths and Weakness

While working on the model for this system and studying different systems with similar objectives, I have realized several strengths and weakness of the system.

Strengths of the system

It's easy to maintain.

Systems learn by acquiring new cases and also from the feedback received for previous prediction.

New features can be added to system as it is not time consuming. The weights of the feature needs to be adjusted to reflect their relative importance.

Weakness of the system

The major weakness of this system is that all the parameters values considered while retrieving similar cases from case base are numerical. This limits the system in a way that all the feature values needs ot be converted to numeric form. This results in power of expressiveness for features.

Also, there can be case wherein below average feature values might deny applicant loan, but the applicant performs exceptionally in the course in future. Dealing or handling outliers is a major issue.

Also, the model is tested on dataset created using both the UCI student loan dataset and manual process. The size of the dataset is also not huge. Therefore, it will be interesting to test the model on huge dataset with actual feature values for students.

The average of salary needs to be calculated each time a new case is added to the case base.

Conclusion

I have tried to predict defaulters based on the information of the student using Case Base Reasoning. Given its strength and weakness, I believe this system can be a good example to determine the applicability of CBR domain in the prediction of loan defaulters.

Future Work

I think it would be interesting to compare the result from the data mining models with the results from my system. Also, it would be interesting to see the behavior of the system after the addition of new features such as market uncertainty, college dropout rate, college default rate etc. The system should also be scalable i.e. the time does not increase as linearly as the size of the case base increases.

References

http://economics.virginia.edu/sites/economics.virginia.edu/files/macro/Ionescu.pdf

http://www.nber.org/papers/w20849.pdf

http://muse.jhu.edu/journals/review_of_higher_education/v037/37.2.hillman.html

http://files.eric.ed.gov/fulltext/EJ905712.pdf

http://publications.nasfaa.org/cgi/viewcontent.cgi?article=1072&context=jsfa