Credit Default Analysis – Christian de Chenu

Dataset from:

https://www.kaggle.com/c/GiveMeSomeCredit

Overview:

The goal of this project is to predict the probability that a borrower (represented by each line of the data set) will experience financial distress in the next two years.

Description of data set

Dataset contains historical information on 250,000 borrowers with the dependent variable being **SeriousDlqin2yrs** which is a binary variable and refers to whether or not a default occurred for the particular borrower. The strict meaning of a default is whether the person experienced 90 days or more of "delinquency". There are also 10 independent variables with values for each borrower also defined in the Variables section below.

Hypothesis

I do not have a strict hypothesis yet, but while there are only a small number of variables in this data set, I view each one as being rich with potential impact on the dependent variable. For example intuitively independent variables such

as the age of the borrower, their income, whether they have been late with payments before, their amount of debt, all will likely have an effect on whether the borrower defaults.
Number of dependents (spouse + children) could affect the dependent variable in either direction (may be more stable in their finances if the borrower has a family or potentially could work the other way and may be in a better position to pay debts if they have no dependents).

Statistical methods you plan to use and why

It is not obvious to me yet which methods are possible for me to use but I have mentioned a few below:
-Regression
-Logistic Regression
Regression seems and obvious choice and data looks like it would work with this methodology (multiple independent variables and one dependent variable). It may be possible to use Logistic Regression to predict a binary response of default/not-default rather than producing a probability.

I also have my eye on the following methods, though I currently struggle to know whether they would be suitable:
-K Nearest Neighbors Classification
-Naïve Bayes Classification
-K-means clustering
-Decision trees and random forests

What business applications do you think your findings will have?

Financial institutions need to decide who can have a credit line and on what terms and this can make or break financial transactions. Credit scoring algorithms, which make a guess at the probability of default, are the method banks use to determine whether or not a loan or a credit line should be granted. This project is concerned with this very issue in predicting the probability that somebody will experience financial distress in the next two years. In my current role I am often arguing with the bank over the credit worthiness of a client (and thus the credit charge the bank decides to impose).

Variables

Variable Name	Description	Type
SeriousDlqin2yrs	**Person experienced 90 days past due delinquency or worse**	**Y/N**
RevolvingUtilizationOfUnsecuredLines	Total balance on credit cards and personal lines of credit except real estate and no installment debt like car loans divided by the sum of credit limits	percentage
age	Age of borrower in years	integer
NumberOfTime30-59DaysPastDueNotWorse	Number of times borrower has been 30-59 days past due but no worse in the last 2 years.	integer
DebtRatio	Monthly debt payments, alimony,living costs divided by monthy gross income	percentage
MonthlyIncome	Monthly income	real
NumberOfOpenCreditLinesAndLoans	Number of Open loans (installment like car loan or mortgage) and Lines of credit (e.g. credit cards)	integer
NumberOfTimes90DaysLate	Number of times borrower has been 90 days or more past due.	integer
NumberRealEstateLoansOrLines	Number of mortgage and real estate loans including home equity lines of credit	integer
NumberOfTime60-89DaysPastDueNotWorse	Number of times borrower has been 60-89 days past due but no worse in the last 2 years.	integer
NumberOfDependents	Number of dependents in family excluding themselves (spouse, children etc.)	integer