Data were summarized using means and standard deviations or frequency and percentages as appropriate. For discrete count data, data were summarized using medians and interquartile range. Additional exploratory data analyses were conducted to examine normality of continuous variables and cross-tabulations of categorical variables between outcome levels. Bivariate relationships were examined using correlations (point-biserial between categorical and continuous, and Pearson correlation between continuous variables) and plotted in a correlation matrix (darker colors represent stronger correlations).

Our main outcome of interest is whether or not an individual defaults when asking for a loan. Traditionally, associations between predictors and outcome are tested individually through univariate regression models, then a subset of those covariates are selected for a multivariate regression analysis if univariate analyses yielded significant results. Given the large number of possible predictors, we suspected severe multicollinarity between some pairs of variables, and overfitting of our model. To alleviate this, we “penalized” regression coefficients to select a parsimonious set of variables that leads to similar prediction results.

In a normal logistic regression with a binary outcome, our model is defined by , where Y is the binary outcome, **X** = is the vector of covariate values, and = is the vector of regression parameters. In a penalized regression, estimation is generally in the form of an optimization problem that seeks to maximize the function in the form of , where is the “penalty” and is the “tuning parameter”. The penalty refers to a constraint surrounding the betas and the tuning parameter corresponds to the amount of “shrinkage” applied to the coefficient. In this analysis, an adaptive LASSO (least absolute shrinkage and selection operator) with logistic regression was employed to estimate coefficients and select the most important predictors. Adaptive LASSO is conducted in 2 parts – the first is a “ridge” regression is conducted that uses a penalty proportional to the sum of squares of the regression coefficients and efficiently handles multicollinearty (Equation 1).

A picture containing text

Description automatically generated (Equation 1)

The second step is the LASSO step that imposes a constraint on the sum of the absolute value of the regression coefficients and uses data-dependent weights (usually the inverse of the ridge regression coefficients) to penalize regression coefficients accordingly (Equation 2). Therefore strong predictors’ coefficients are shrunk less than coefficients of weak predictors.

A picture containing diagram

Description automatically generated (Equation 2).

For more details, please see appendix (add exactly where – figure XXX).