Results

Since it is a continuous prediction problem, we choose Least Squares Regression as our model. In addition, we performed two Shrinkage Methods (Ridge Regression and Lasso Regression) and two Dimension Reduction Methods (Principal Components Regression and Partial Least Squares Regression) to compare the results with our base model.

Then we used 10-fold Cross Validation to choose minimum lambda for Shrinkage Methods and minimum validation components for Dimension Reduction Methods. With the chosen parameter, we compare results by calculating Mean Square Error on test dataset.

The following table is the test Mean Square Error for the four methods.

row.names	Test MSE Values
Ridge	0.05417609
Lasso	0.05414370
PCR	0.05814687
PLS	0.05717022

From the table, we can see that the Lasso Regression has the minimum test Mean Squared Error in this case, followed by Ridge Regression. The Dimension Reduction Methods in this case do not perform as well as Shrinkage Methods.

Then to see which predictors matter most in prediction, we visualize the coefficients of predictors within each method as a table.

row.names	OLS	Ridge	Lasso	PCR	PLS
Income	-0.598171486	-0.568097219	-5.514258e-01	-0.598171486	-0.598137940
Limit	0.958438722	0.702916676	7.815746e-01	0.958438722	0.957819003
Rating	0.382478949	0.608183179	5.110649e-01	0.382478949	0.383140840
Cards	0.052864969	0.043630894	3.884469e-02	0.052864969	0.052269035
Age	-0.023033397	-0.025445764	-1.676893e-02	-0.023033397	-0.023401547
Education	-0.007469459	-0.005797352	0.000000e+00	-0.007469459	-0.007590278
GenderFemale	-0.011593092	-0.010665388	-1.522737e-05	-0.011593092	-0.011926799
StudentYes	0.278154853	0.273079108	2.660849e-01	0.278154853	0.278181378
MarriedYed	-0.009054196	-0.011143321	0.000000e+00	-0.009054196	-0.008649141
EthnicityAsian	0.015950671	0.016447244	0.000000e+00	0.015950671	0.015944759
EthnicityCaucasian	0.011005286	0.011027402	0.000000e+00	0.011005286	0.011062746

From the table, we can see that in all five methods we used for prediction, Limit is the single most important predictor for Balance which corresponds to our common sense that amount of limit is usually highly correlated to amount of Balance.

For Ridge Regression, the second and third important factors are Rating and Income. For Lasso Regression, the second and third important factors are Income and Rating. For Pricipal Components Regression, the second and third important factors are Income and Rating. For Parse Least Squares Regression, the second and third important factors are Income and Rating. Hence we can conclude that the three leading factors are Limit, Income, and Rating.

To get a better intuition of the importance of each factor, we visualize the coefficients of each predictor in each five methods in bar plots.