Building an Ensemble Model to predict personal loan interest rates

Data Preparation

Amount Requested: Continuous, Income: Continuous, Rate: Continuous, Months: Continuous

IL $ ILL: Grouped into IL (Illinois)

TEX $ TX: Grouped into TX(Texas)

WI $ WISC: Grouped into WI (Wisconsin)

Imputations

The missing values in column „Utilization" are imputed by calculating Credit balance/ Credit limit. As the credit balance is 0 for all the missing values in „Utilization", these missing values are imputed with 0.00%. Missing values in column "Month" are imputed with Multivariate Normal Imputation.

Monthly Pmt, Amt Funded and Standing variables have not been used for modeling the data.

Transformations

Z ln - Amt Requested, z ln - Pmts to date, z ln - Income, z ln - Open LoC, zln - Total LoC,
z sqrt - Balance, z sqrt - Delinquent, z - Utilization, z – Months, z -D/I

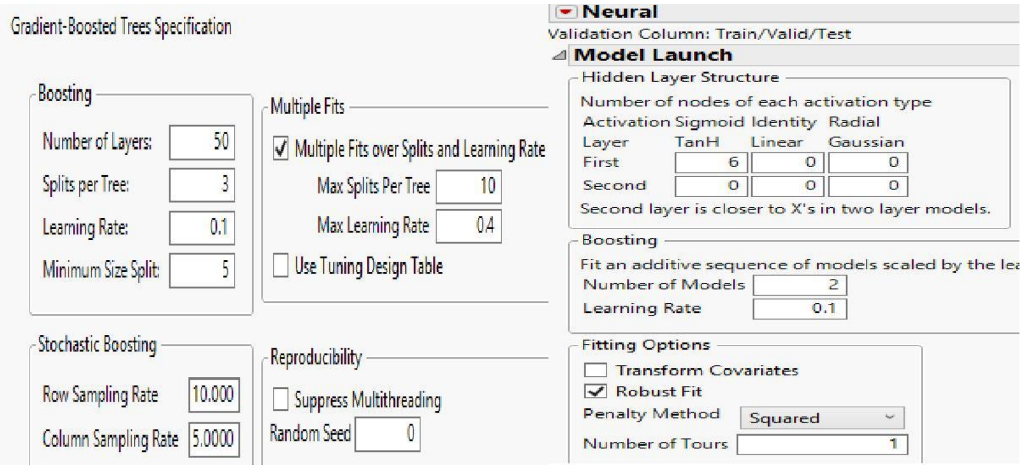
Exploratory Data Analysis

The variables Amt Requested, Pmts to date, Open LoC, Total LoC and Income are right skewed. Rate and FICO Score can be inversely proportional. Rate is observed to be high when there is zero Utilization. The interest rates for small businesses and consolidated debts are observed to be high. FICO Score has an inverse relation with Delinquencies. Greater the FICO score, lesser are the number of delinquencies. Income, State and Employment in years may not have a direct impact on the interest rates. The minimum interest rate is 5.42 % for applicants who have a FICO score of greater than 781. The maximum interest rate is 24.8 % for applicants who have a FICO score between 660-680.

Fitting the model

The chosen model for predictions uses Ensemble method (Regression/Standard least squares). The 4 Predictor variables used in the final model are output variables of 4 base models of regression, k-NN, Boosted tree and Neural net, which are denoted as Regression Rate, Rate Neural, Rate Boosted, Rate kNN in the model. Output variable is Rate.

Boosted Tree model & Neural net model specifications -



Gradient-Boosted Trees Specification

Boosting
Number of Layers: 50
Splits per Tree: 3
Learning Rate: 0.1
Minimum Size Split: 5

Multiple Fits
☑ Multiple Fits over Splits and Learning Rate
Max Splits Per Tree 10
Max Learning Rate 0.4
□ Use Tuning Design Table

Stochastic Boosting
Row Sampling Rate 10.000
Column Sampling Rate 5.0000

Reproducibility
□ Suppress Multithreading
Random Seed 0

Neural
Validation Column: Train/Valid/Test
Model Launch
Hidden Layer Structure
Number of nodes of each activation type

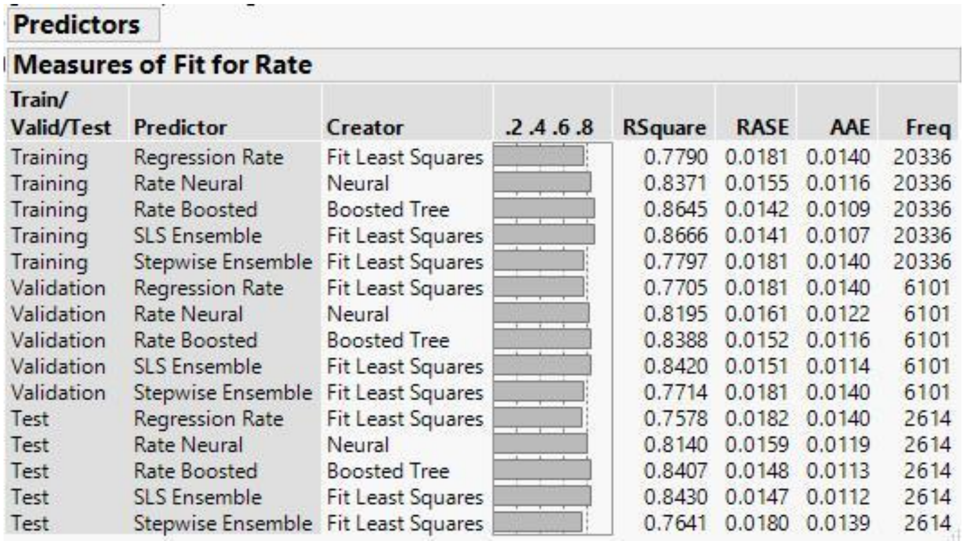
Activation Layer	Sigmoid TanH	Identity Linear	Radial Gaussian
First	6	0	0
Second	0	0	0

Second layer is closer to X's in two layer models.

Boosting
Fit an additive sequence of models scaled by the lea
Number of Models 2
Learning Rate 0.1

Fitting Options
□ Transform Covariates
☑ Robust Fit
Penalty Method Squared
Number of Tours 1

A screenshot of all generated models along with their training and validation R-square values is given below -



Predictors

Measures of Fit for Rate

Train/ Valid/Test	Predictor	Creator	.2 .4 .6 .8	RSquare	RASE	AAE	Freq
Training	Regression Rate	Fit Least Squares		0.7790	0.0181	0.0140	20336
Training	Rate Neural	Neural		0.8371	0.0155	0.0116	20336
Training	Rate Boosted	Boosted Tree		0.8645	0.0142	0.0109	20336
Training	SLS Ensemble	Fit Least Squares		0.8666	0.0141	0.0107	20336
Training	Stepwise Ensemble	Fit Least Squares		0.7797	0.0181	0.0140	20336
Validation	Regression Rate	Fit Least Squares		0.7705	0.0181	0.0140	6101
Validation	Rate Neural	Neural		0.8195	0.0161	0.0122	6101
Validation	Rate Boosted	Boosted Tree		0.8388	0.0152	0.0116	6101
Validation	SLS Ensemble	Fit Least Squares		0.8420	0.0151	0.0114	6101
Validation	Stepwise Ensemble	Fit Least Squares		0.7714	0.0181	0.0140	6101
Test	Regression Rate	Fit Least Squares		0.7578	0.0182	0.0140	2614
Test	Rate Neural	Neural		0.8140	0.0159	0.0119	2614
Test	Rate Boosted	Boosted Tree		0.8407	0.0148	0.0113	2614
Test	SLS Ensemble	Fit Least Squares		0.8430	0.0147	0.0112	2614
Test	Stepwise Ensemble	Fit Least Squares		0.7641	0.0180	0.0139	2614

Standard Least squares Ensemble : Validation R square - 0.842, Training R squared - 0.8666