Mortgage Default Probability using Anomaly Detection Techniques
-
Takeaway summary

Ram Subramanian

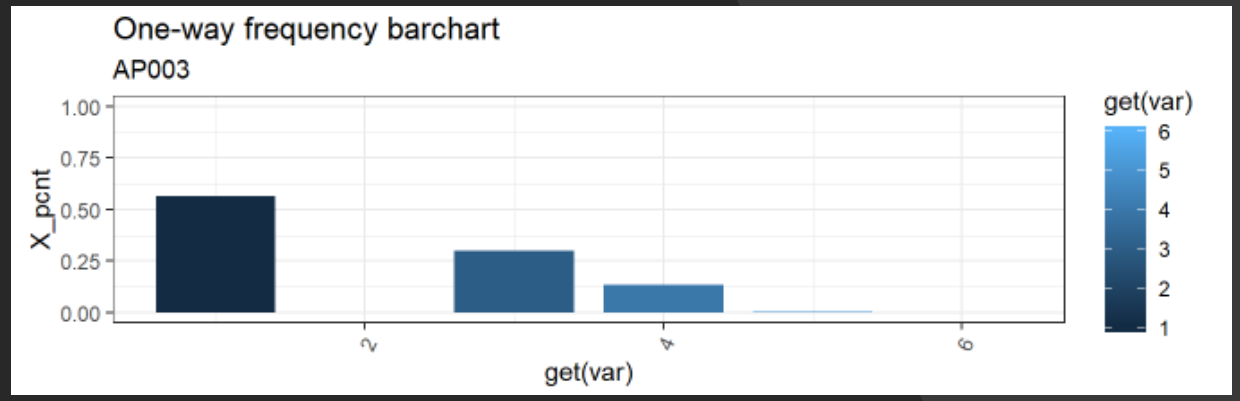
MS Applied Analytics

Columbia University, New York, NY

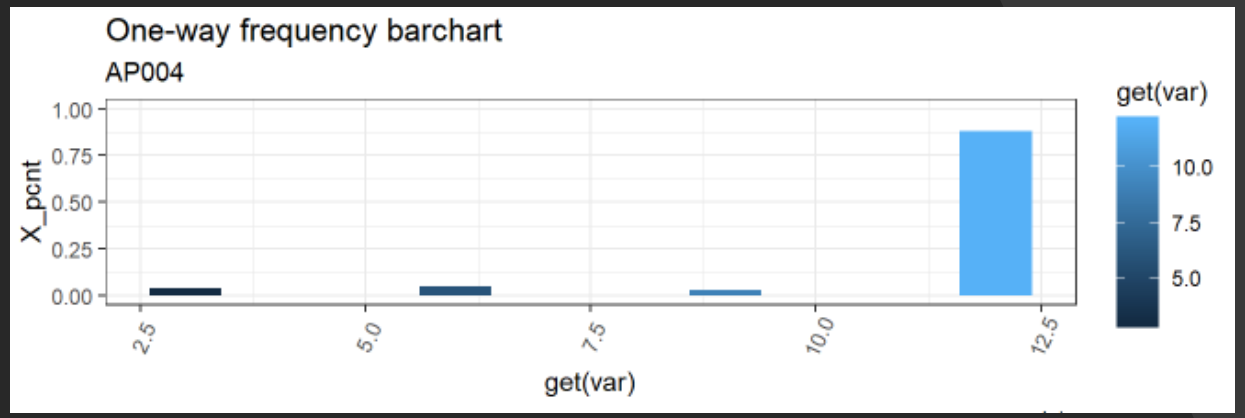
Mortgage Default Probability - Key Takeaways

- Combining the Feature selection models of Forward, Backward and Hybrid Stepwise regression and Lasso regression, we selected 31 highly statistically significant variables for logistic model
- Education(AP003) and Loan Term (AP004) turned out to be the most predictive factors of loan default based on the initial exploratory data analysis as well as the multiple variable selection methods, but other additional significant factors were also added to improve the accuracy of the model

Education vs Loan default



Loan Term vs Loan default



Mortgage Default Probability – Key Takeaways

- The accuracy of our final logistic model was approximately 80%. Though Accuracy is a good metric for balanced classes, we further evaluated the model using Gains table
- The Lift is the measure of effectiveness of our predictive model calculated as the ratio between the results obtained with and without our predictive model. Our final model has a Cumulative Lift of 2.09 which is a reasonably good index

Cum. Lift: 2.09

GAINS TABLE

Depth of file	N	Cume N	Resp Rate	Cume Resp rate	Cume Pct of Total Resp	Lift index	Cume lift	Optimal lift index	optimal cume lift	Mean model score
10	2400	2400	39.67%	39.67%	20.90%	209	209	526	526	43.28%
20	2400	4800	30.88%	35.27%	37.10%	162	185	474	500	30.66%
30	2400	7200	26.54%	32.36%	51.10%	140	170	0	333	25.51%
40	2400	9600	22.08%	29.79%	62.70%	116	157	0	250	21.95%
50	2400	12000	19.12%	27.66%	72.70%	101	145	0	200	19.03%
60	2400	14400	14.92%	25.53%	80.60%	78	134	0	167	16.43%
70	2400	16800	13.54%	23.82%	87.70%	71	125	0	143	13.98%
80	2400	19200	11.12%	22.23%	93.50%	59	117	0	125	11.47%
90	2400	21600	8.21%	20.68%	97.90%	43	109	0	111	8.64%
100	2400	24000	4.08%	19.02%	100.00%	21	100	0	100	4.60%