Retail Credit Scoring

Commercial Banking, Corp.

RFP #: RC - S1.H1

February 1, 2019

Orange Team #1
Morgan Groves, Bill Jenista, Mia Wu, Pierce Secola & Dave Hiltbrand

Executive Summary

This report is a response to Commercial Banking, Corp's (hereafter the Company) request for Proposal RC – S1.H1 for analytics services to create a scorecard and mapping system to support its banking services including all retail credit applications. Using the information in the RFP and the data provided, the analytics team developed a scorecard that suggests a single cutoff score of 500 and both decreases the rejection rate to 2.4% and increases the acceptance rate to 78% from the previous baseline rates provided by the Company. Furthermore, the team recommends the Company make future decisions based on the results from the scorecard which will increase profits by an average of $375 per customer allowing the Company to offer more lines of credit with less accounts defaulting.

Methodology

Lending institutions mitigate risks by offering resources based on information provided by the customer or gathered by the firm. Our analytics team develops credit scorecards to improve clients' returns on lending risks. We build a mathematical model that takes information about potential customers such as an applicant's income and oldest line of credit. Using advanced techniques we create new variables that contain the greatest information value in predicting the probability that the applicant will default on the loan. Our goal is to use data provided by a firm using default and acceptance rates for loans they have already issued. We also infer the possibility that an applicant who was originally denied a loan would have defaulted to build a more robust dataset. Using this methodology we successfully ingest a customer profile and return a score which can be used to help banks and lenders determine if the risk presented by the customer is outweighed by the return on interest.

Analysis

The team was given two datasets to create the scorecard: one set with information about customers whose credit card applications were accepted and the other about those who were rejected. To account for the different ratios between good and bad loans as well as the size of the accepted and rejected datasets, a weight variable was utilized in the dataset. In order to validate the performance of the scorecard, the accepted loans dataset was split into 70% training data and 30% validation data. Then, we implemented a binning technique to group variables. This technique also discovered the seven most important variables for the model based on their weights of evidence, an informative statistic which discovers which variables are most important at identifying loan defaults.

Variable Importance	
Variable	**Information Value**
Age	0.391
Status	0.256
Time at Job	0.239
Income	0.217
Number of Persons in Household	0.193
Credit Cards	0.153
EC Card Holder	0.127

Table 1 - Variable Importance by ranking using Information Value to determine best variables to build the scorecard.

Using guidelines in the proposal to build the initial scorecard, we assigned a score of 500 to applicants with odds-ratio 20/1 and let the doubling the odds be associated with a change of 50 points. In addition, we accounted for the existing acceptance rate of 75%, the current event rate of 3.23%, expected revenue of accepted good customers, and the expected cost of accepted bad customers into this scorecard. The area under the curve of this initial scorecard model is 0.724, which indicates our scorecard accurately predicted 72% of the defaults in the validation data. Based on the trade-off chart, the optimal cut off point for a credit score is 516 for the maximization of profit. To make sure that our scorecard has complied with the FDIC, we used the rejected dataset as our reject inference, and we chose the fuzzy inference to remove the bias resulting from the exclusion of rejects.

After implementing the reject inference technique, we built our final scorecard using both the accepted and rejected datasets. We again partitioned the data with a 70% training and 30% validation split. The same seven input variables were included in the model and the area under the curve statistic is 0.731 which indicates similar prediction performance to the original scorecard. The optimal cut-off point for credit score has decreased to 500. The trade off plots from Figure 1 below depicts the changes in default versus acceptance rates which help determine the best cutoff score to use. At credit score of 500, the cumulative event rate is 0.83% lower than the previous one and the approval rate is increased by 3% of the previous rate.

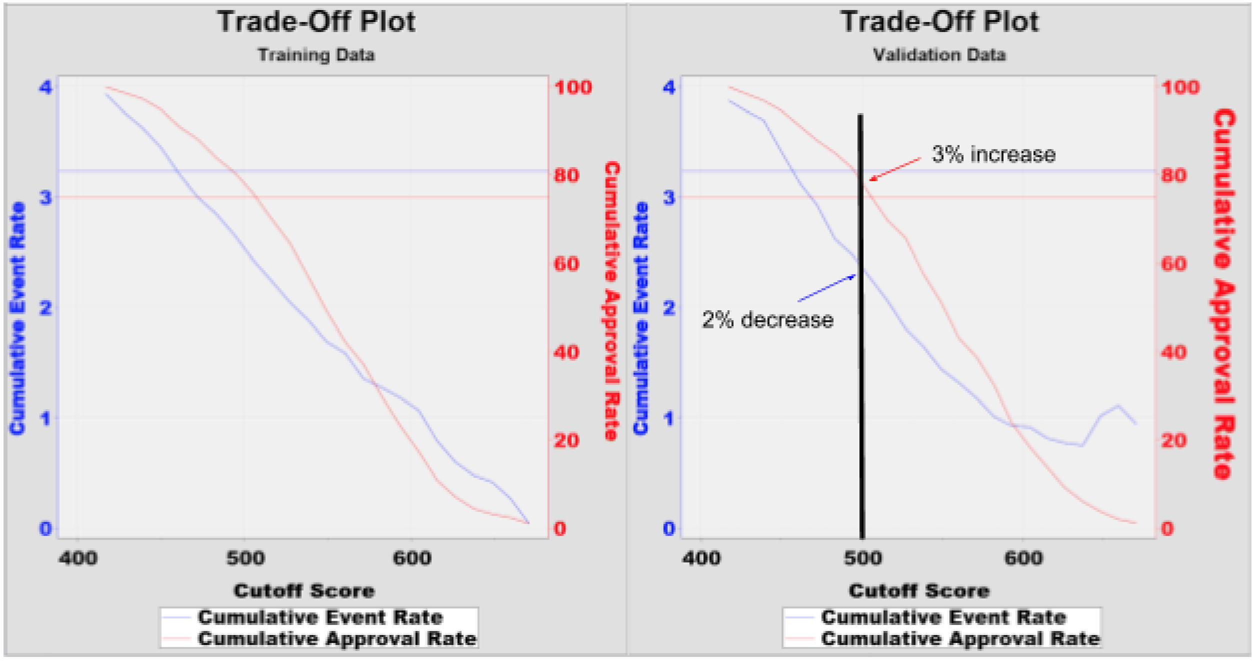


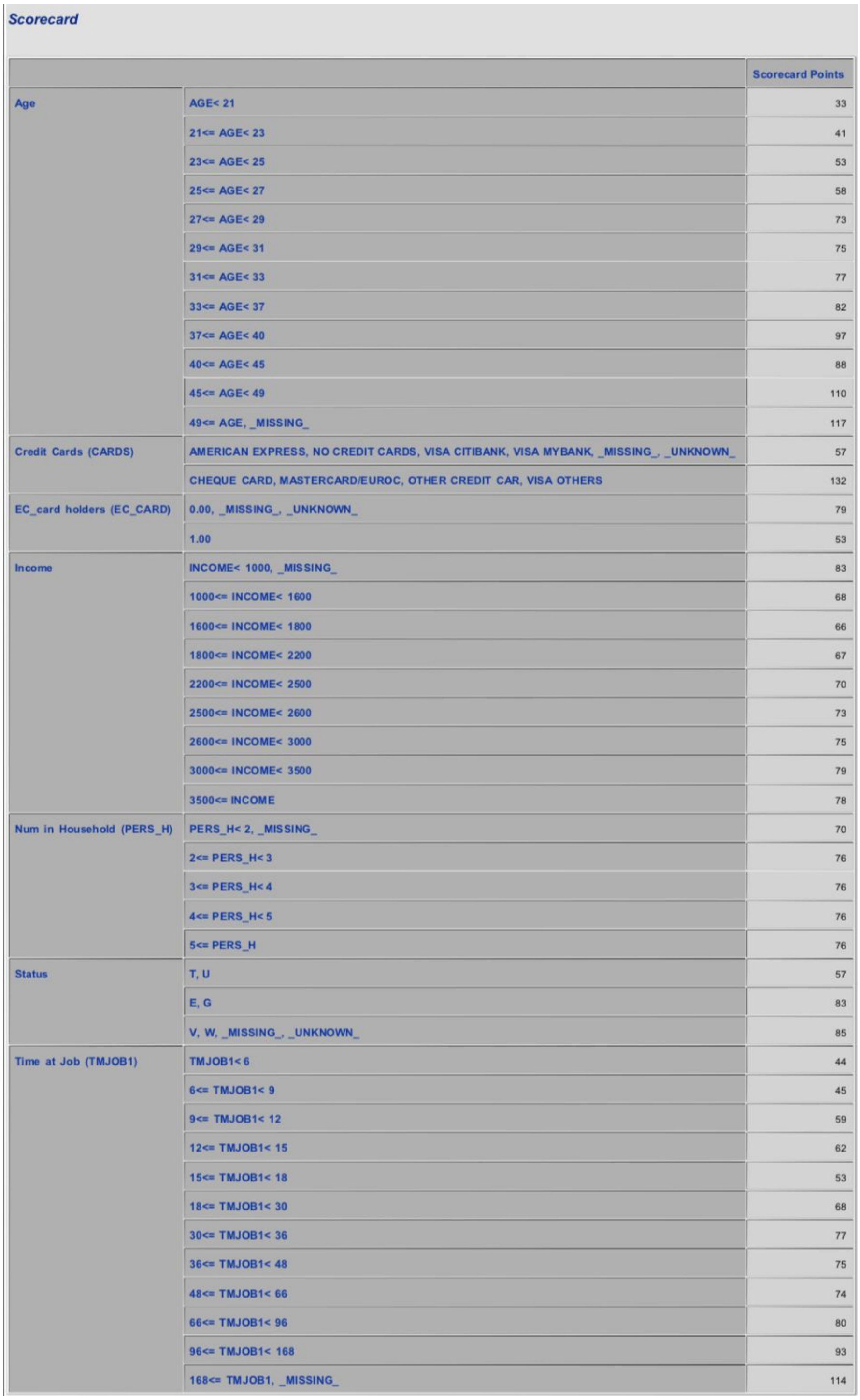
Figure 1 - Trade off plots for training data on the left and validation data on the right showing acceptance and event rates.

Conclusion

Using advanced techniques, the analytics team is proud to present a new scorecard that the Company can utilize immediately. Through the implementation of the scorecard, the Company will have higher acceptance rates paired with lower default rates at a cutoff score of 500. Our methodology will allow the bank to increase their customer base while safely mitigating risks with the new rates of 78% for approvals and 2.4% for defaults and an increase of $375 profit per customer.

Appendix

A.1 - Scorecard output from SAS E-Miner



Scorecard

		Scorecard Points
Age	AGE< 21	33
	21<= AGE< 23	41
	23<= AGE< 25	53
	25<= AGE< 27	58
	27<= AGE< 29	73
	29<= AGE< 31	75
	31<= AGE< 33	77
	33<= AGE< 37	82
	37<= AGE< 40	97
	40<= AGE< 45	88
	45<= AGE< 49	110
	49<= AGE, _MISSING_	117
Credit Cards (CARDS)	AMERICAN EXPRESS, NO CREDIT CARDS, VISA CITIBANK, VISA MYBANK, _MISSING_, _UNKNOWN_	57
	CHEQUE CARD, MASTERCARD/EUROC, OTHER CREDIT CAR, VISA OTHERS	132
EC_card holders (EC_CARD)	0.00, _MISSING_, _UNKNOWN_	79
	1.00	53
Income	INCOME< 1000, _MISSING_	83
	1000<= INCOME< 1600	68
	1600<= INCOME< 1800	66
	1800<= INCOME< 2200	67
	2200<= INCOME< 2500	70
	2500<= INCOME< 2600	73
	2600<= INCOME< 3000	75
	3000<= INCOME< 3500	79
	3500<= INCOME	78
Num in Household (PERS_H)	PERS_H< 2, _MISSING_	70
	2<= PERS_H< 3	76
	3<= PERS_H< 4	76
	4<= PERS_H< 5	76
	5<= PERS_H	76
Status	T, U	57
	E, G	83
	V, W, _MISSING_, _UNKNOWN_	85
Time at Job (TMJOB1)	TMJOB1< 6	44
	6<= TMJOB1< 9	45
	9<= TMJOB1< 12	59
	12<= TMJOB1< 15	62
	15<= TMJOB1< 18	53
	18<= TMJOB1< 30	68
	30<= TMJOB1< 36	77
	36<= TMJOB1< 48	75
	48<= TMJOB1< 66	74
	66<= TMJOB1< 96	80
	96<= TMJOB1< 168	93
	168<= TMJOB1, _MISSING_	114