IOSR Journal of Economics and Finance (IOSR-JEF)
e-ISSN: 2321-5933, p-ISSN: 2321-5925.Volume 8, Issue 2 Ver. II (Mar. - Apr. 2017), PP 69-81
www.iosrjournals.org

Credit Risk Analysis & Modeling: A Case Study

Mr Prashanta Kumar Behera
PhD Research Scholar at Singhania University

Abstract: *Credit risk analysis and credit risk management is important to financial institutions which provide loans to businesses and individuals. Credit risk can occur for various reasons such as bank mortgages (or home loans), motor vehicle purchase finances, credit card purchases, installment purchases, and so on. Credit loans and finances have risk of being defaulted. To understand risk levels of credit users, credit providers normally collect vast amount of information on borrowers. Some predictive analytic techniques can be used to analyze or to determine risk levels involved on credits, finances, and loans, i.e., default risk levels. We are trying to find default probability of Cumulative Accuracy Profile (CAP), the Receiver Operating Characteristic (ROC), and the Kolmogorov-Smirnov (K-S) statistic.*

Key words: *Credit Risk, Probability of Default, Cumulative Accuracy Profile (CAP), the Receiver Operating Characteristic (ROC), and the Kolmogorov-Smirnov (K-S) statistic.*

I. Introduction

In this paper we are considering credit card data for Credit risk analysis and predictive modeling. Personal credit scores are normally computed from information available in credit reports collected by external credit bureaus and ratings agencies. Credit scores may indicate personal financial history and current situation. However, it does not tell us exactly what constitutes a "good" score from a "bad" score. More specifically, it does not tell us the level of risk for the lending you may be considering. Furthermore, in many countries, credit rating system is not available. Internal credit scoring methods described in this page address the problem. It is noted that internal credit scoring techniques can be applied to commercial credits as well.

Credit Risk Analysis and Modeling

In this paper, the following credit risk analysis methods are described;

- Credit risk factors profiling and analysis.
- Credit risk predictive modeling or default predictive modeling.
- Credit risk modeling or finance risk modeling.
- Internal credit risk scoring.

Credit Risk Profiling

Credit risk profiling (finance risk profiling) is very important. The principle suggests that 80% to 90% of the credit defaults may come from 10% to 20% of the lending segments. Profiling the segments can reveal useful information for credit risk management. Credit providers often collect a vast amount of information on credit users. Information on credit users (or borrowers) often consists of dozens or even hundreds of variables, involving both categorical and numerical data with noisy information. Hotspot profiling is to identify factors or variables that best summarize the segments.

Credit Risk Predictive Modeling

If past is any guide for predicting future events, predictive modeling is an excellent technique for credit risk management. Predictive models are developed from past historical records of credit loans, containing financial, demographic, psychographic, geographic information, etc. From the past credit information, predictive models can learn patterns of different credit default ratios, and can be used to predict risk levels of future credit loans. It is important to note that statistical process requires a substantially large number of past historical records (or customer loans) containing useful information. Useful information is something that can be a factor that differentially affects credit default ratios.

Credit Risk Scoring

Credit risk score is a risk rating of credit loans. It measures the level of risk of being defaulted. The level of default risk can be best predicted with predictive modeling. Credit scores can be measured in term of default probability and/or relative numerical ratings. A credit scoring model is a tool that is typically used in the decision-making process of accepting or rejecting a loan. A credit scoring model is the result of a statistical model which, based on information about the borrower (e.g. age, number of previous loans, etc.), allows one to

distinguish between "good" and "bad" loans and give an estimate of the probability of default. The fact that this model can allocate a rating on the credit quality of a loan implies a certain number of possible applications: Application area Description Health score: The model provides a score that is related to the probability that the client misses a payment. This can be seen as the "health" of the client and allows the company to monitor its portfolio and adjust its risk. New clients The model can be used for new clients to assess what is their probability of respecting to their financial obligations. Subsequently the company can decide to grant or not the requested loan. What drives default The model can be used to understand what the driving factors behind default are. The bank can utilize this knowledge for its portfolio and risk assessment. A credit scoring model is just one of the factors used in evaluating a credit application. Assessment by a credit expert remains the decisive factor in the evaluation of a loan. The history of developing credit-scoring models goes as far back as the history of borrowing and repaying. It reflects the desire to issue an appropriate rate of interest for undertaking the risk of giving away one's own money. With the advent of the modern statistics era in the 20th century appropriate techniques have been developed to assess the likelihood of someone's default on the payment, given the resemblance of his/her characteristics to those who have already defaulted in the past. In this document we will focus on one of the most prominent methods to do credit scoring, the logistic regression. Despite being one of the earliest methods of the subject, it is also one of the most successful, owing to its transparency. Although credit scoring methods are linked to the aforementioned applications in banking and finance, they can be applied to a large variety of other data analytics problems, such as: Which factors contribute to a consumer's choice? Which factors generate the biggest impact to a consumer's choice? What is the profit associated with a further boost in each of the impact factors? How likely is that a customer likes to adopt a new service? What is the likelihood that a customer will go to a competitor? Such questions can all be answered within the same statistical framework. A logistic regression model can, for example, provide not only the structure of dependencies of the explanatory variables to the default but also the statistical significance of each variable.

Quality of Data

Before statistics can take over and provide answers to the above questions, there is an important step of preprocessing and checking the quality of the underlying data. This provides a first insight into the patterns inside the data, but also an insight on the trustworthiness of the data itself. The investigation in this phase includes the following aspects: What is the proportion of defaults in the data? In order for the model to be able to make accurate forecasts it needs to see enough examples of what constitutes a default. For this reason it is important that there is a sufficiently large number of defaults in the data. Typically in practice, data with less than 5% of defaults pose strong modeling challenges. What is the frequency of values in each variable in the data? This question provides valuable insight into the importance of each of the variables. The data can contain numerical variables (for example, age, salary, etc.) or categorical ones (education level, marital status, etc.). For some of the variables we may notice that they are dominated by one category, which will render the remaining categories hard to highlight in the model. Typical tools to investigate this question are scatter plots and pie charts. What is the proportion of outliers in the data? Outliers can play an important role in the model's forecasting behavior. Although outliers represent events that occur with a small probability and a high impact, it is often the case that outliers are a result of system error. For example, a numerical variable that is assigned to the value 999, can represent a code for a missing value, instead of a true numerical variable. That aside, outliers can be easily detected by the use of box plots. How many missing values are there and what is the reason? Values can be missing for various reasons, which range from missing due to no response, due to drop out of the clients, or due to censoring of the answers, or simply missing at random. Missing values pose the following dilemma: On one hand they refer to incomplete instances of data and therefore treatment or imputation may not reflect the exact state of affairs. However, avoiding handling missing values and simply ignoring them may lead to loss of valuable information. There exists a number of ways to impute missing values, such as the expectation-maximization algorithm. Quality assurance there is a standard framework around QA which aims to provide a full view on the data quality in the following aspects: Inconsistency, Incompleteness, Accuracy, Precision, Missing / Unknown.

Research Objective:

The following objectives have been outlined.

- To estimate credit risk factors profiling.
- To know default probability from credit score data.
- To examine internal credit risk scoring.
- Validate the credit scorecard model using the CAP, ROC, and Kolmogorov-Smirnov statistic

Research Methodology:

This study shows how to create a credit score card object, bin data, display and plot binned data information through MATLAB. This study also shows how to fit a logistic regression model, obtain a score for the scorecard model, and determine the probabilities of default and validate the credit scorecard model using three different metrics.

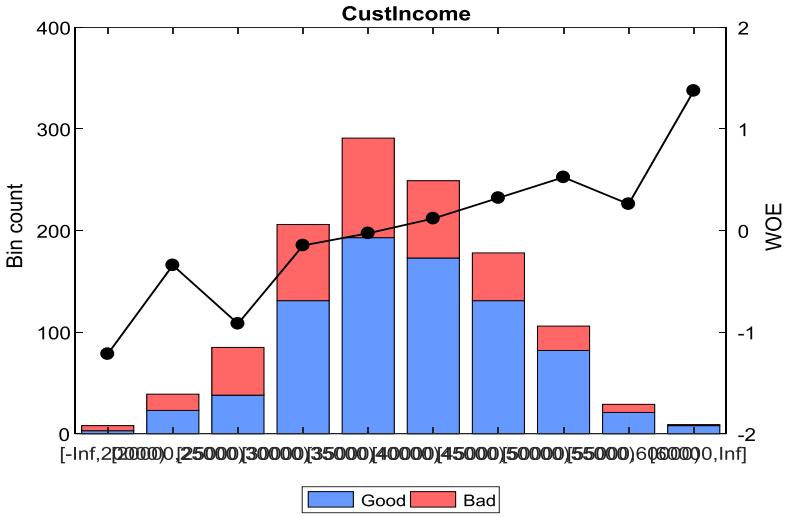
- Step 1. Create a credit score card object.
- Step 2a. Automatically bin the data.
- Step 2b. Fine-tune the bins using manual binning.
- Step 3. Fit a logistic regression model.
- Step 4. Review and format scorecard points.
- Step 5. Score the data.
- Step 6. Calculate the probability of default.
- Step 7. Validate the credit scorecard model using the CAP, ROC, and Kolmogorov-Smirnov statistic

MATLAB Code Result or Output

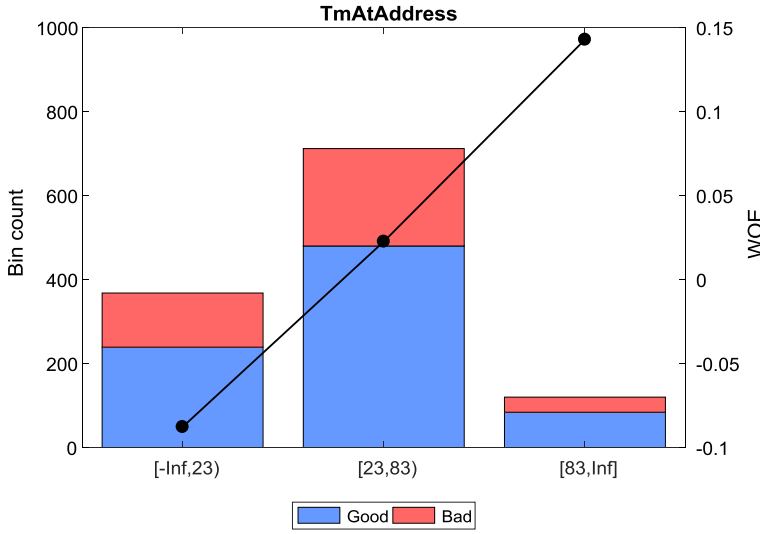
Step 1: We create a credit score card Object

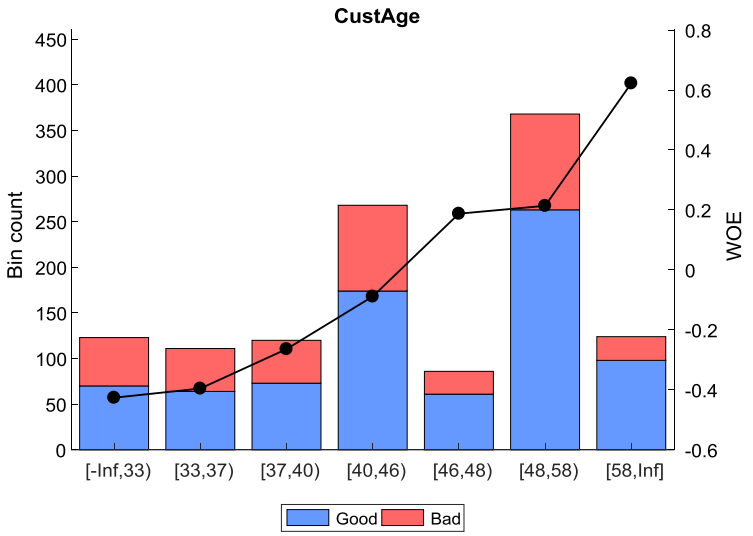
Bin	Good	Bad	Odds	WOE	Info Value
'Home Owner'	365	177	2.0621	0.019329	0.0001682
'Tenant'	307	167	1.8383	-0.095564	0.0036638
'Other'	131	53	2.4717	0.20049	0.0059418
'Totals'	803	397	2.0227	NaN	0.0097738

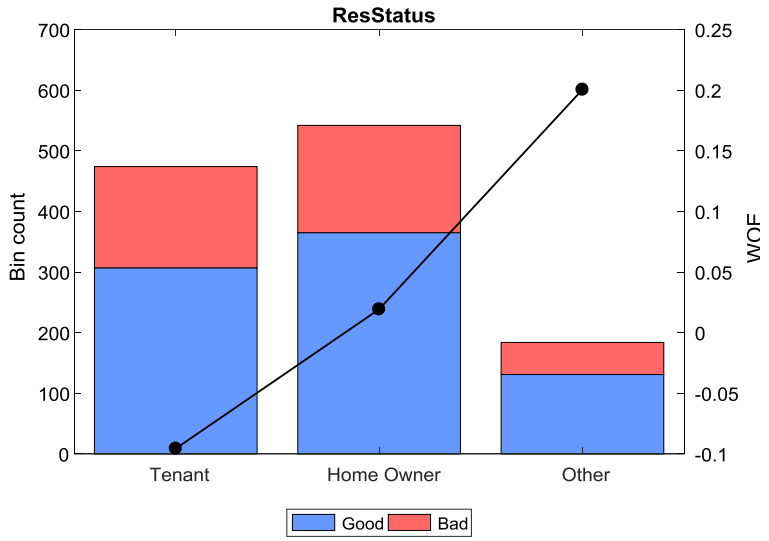
Bin	Good	Bad	Odds	WOE	InfoValue
'[-Inf,20000)'	3	5	0.6	-1.2152	0.010765
'[20000,25000)'	23	16	1.4375	-0.34151	0.0039819
'[25000,30000)'	38	47	0.80851	-0.91698	0.065166
'[30000,35000)'	131	75	1.7467	-0.14671	0.003782
'[35000,40000)'	193	98	1.9694	-0.026696	0.00017359
'[40000,45000)'	173	76	2.2763	0.11814	0.0028361
'[45000,50000)'	131	47	2.7872	0.32063	0.014348
'[50000,55000)'	82	24	3.4167	0.52425	0.021842
'[55000,60000)'	21	8	2.625	0.26066	0.0015642
'[60000,Inf]'	8	1	8	1.375	0.010235
'Totals'	803	397	2.0227	NaN	0.13469

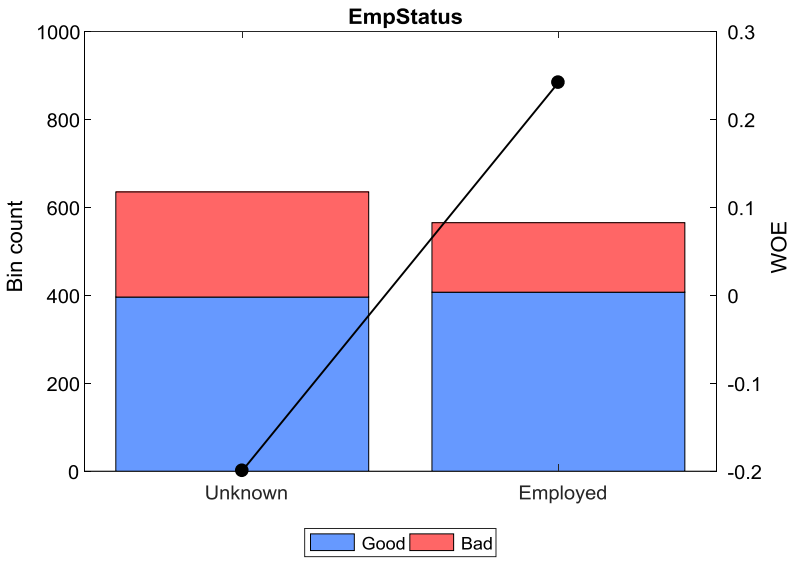


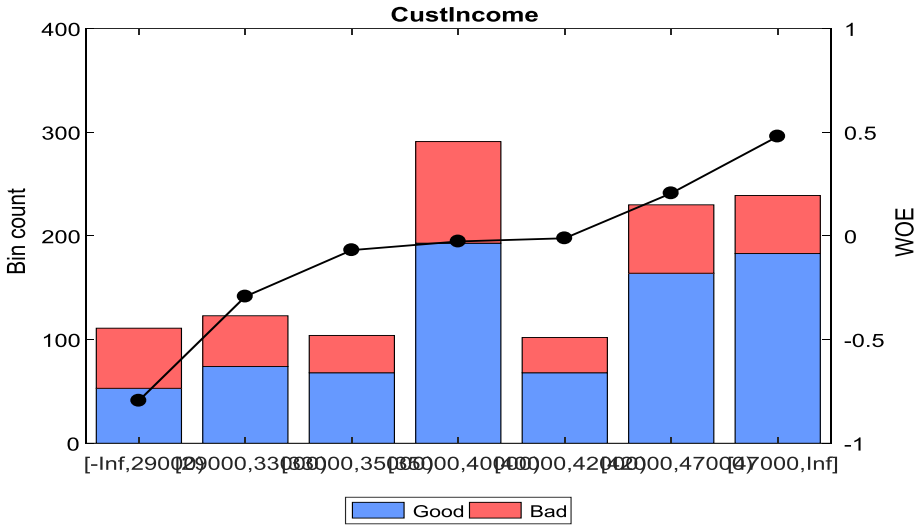
Step 2a: Automatically bin the data.

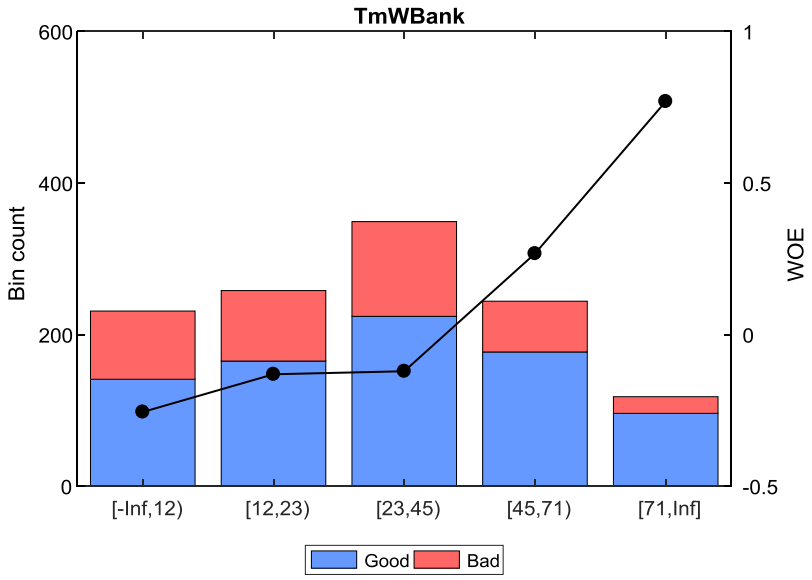


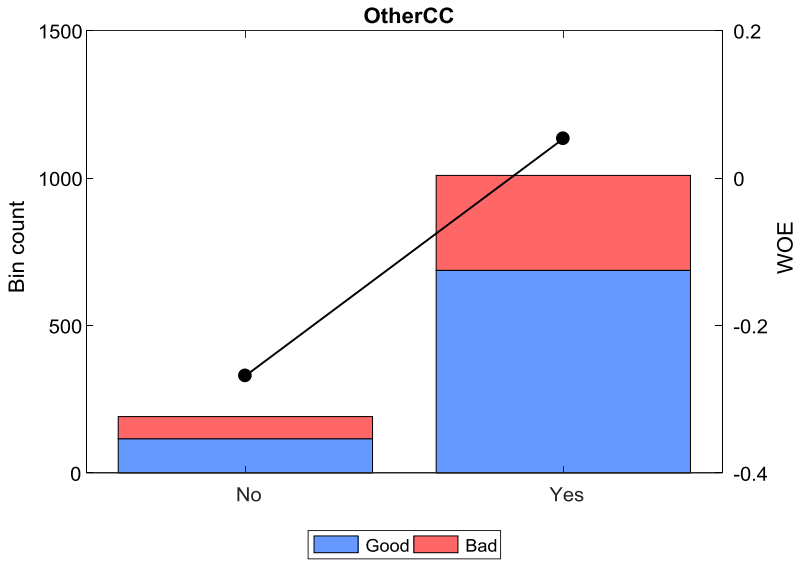


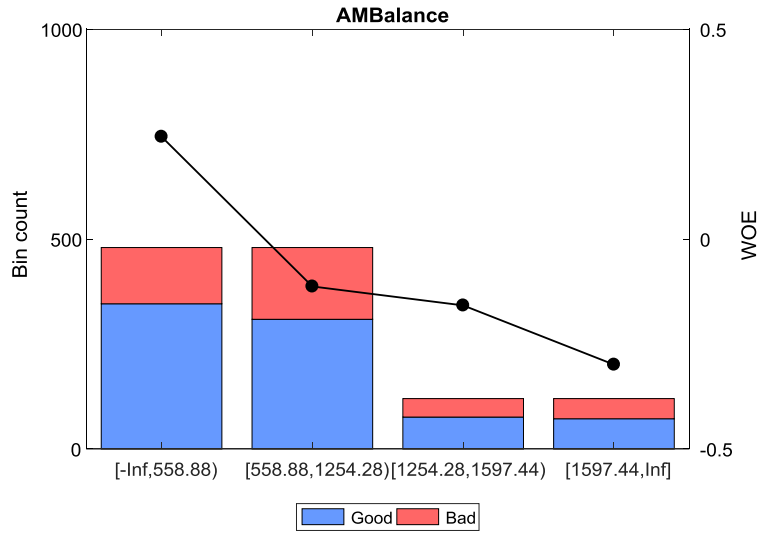

ResStatus
Bin count
WOF
Tenant
Home Owner
Other
Good
Bad


EmpStatus
Bin count
WOE
Unknown
Employed
Good
Bad

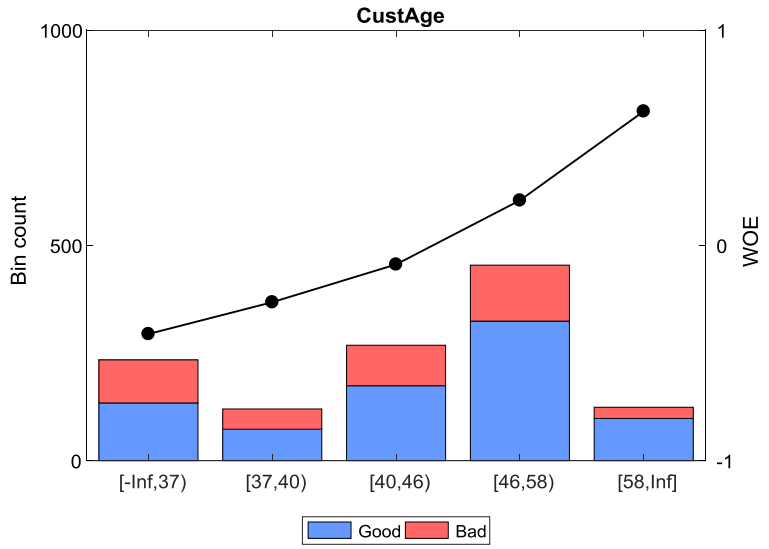

CustIncome
Bin count
WOE
[-Inf,2900
[4700,Inf]
Good
Bad

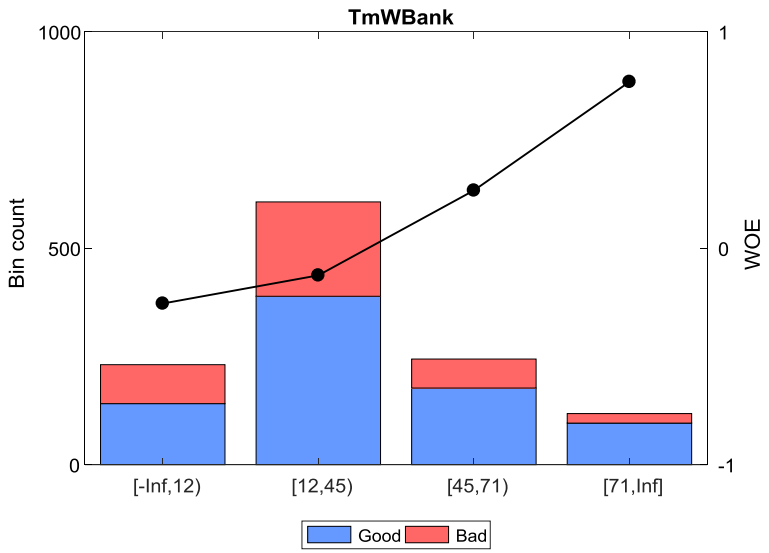

TmWBank
600
400
200
0
Bin count
1
0.5
0
-0.5
WOE
[-Inf,12)
[12,23)
[23,45)
[45,71)
[71,Inf]
Good
Bad

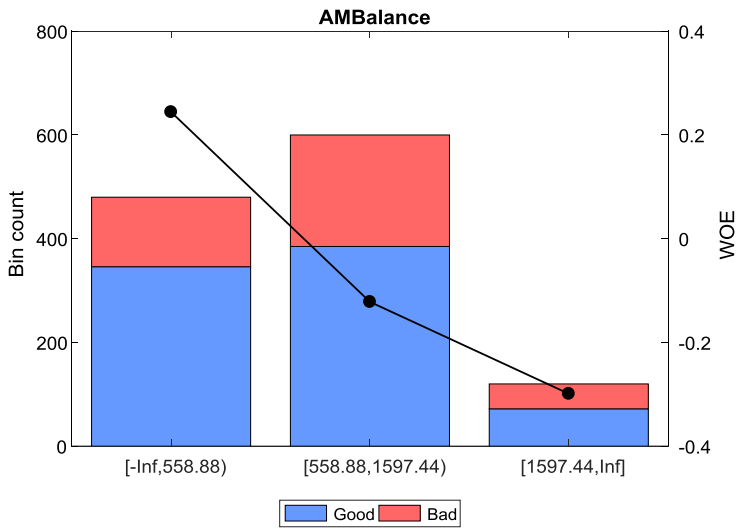

OtherCC
1500
1000
500
0
Bin count
0.2
0
-0.2
-0.4
WOE
No
Yes
Good
Bad


AMBalance
1000
500
0
Bin count
0.5
0
-0.5
WOE
[-Inf,558.88)
[558.88,1254.28)
[1254.28,1597.44)
[1597.44,Inf]
Good
Bad

Step 2b. Fine-tune the bins using manual binning.







Step 3. Fit a logistic regression model.

1. Adding CustIncome, Deviance = 1490.8954, Chi2Stat = 32.545914, PValue = 1.1640961e-08
2. Adding TmWBank, Deviance = 1467.3249, Chi2Stat = 23.570535, PValue = 1.2041739e-06
3. Adding AMBalance, Deviance = 1455.858, Chi2Stat = 11.466846, PValue = 0.00070848829
4. Adding EmpStatus, Deviance = 1447.6148, Chi2Stat = 8.2432677, PValue = 0.0040903428
5. Adding CustAge, Deviance = 1442.06, Chi2Stat = 5.5547849, PValue = 0.018430237
6. Adding ResStatus, Deviance = 1437.9435, Chi2Stat = 4.1164321, PValue = 0.042468555
7. Adding OtherCC, Deviance = 1433.7372, Chi2Stat = 4.2063597, PValue = 0.040272676

Generalized linear regression model:
logit(status) ~ 1 + CustAge + ResStatus + EmpStatus + CustIncome + TmWBank + OtherCC + AMBalance

Distribution = Binomial
Estimated Coefficients:

	Estimate	SE	tStat	pValue
(Intercept)	0.7024	0.064	10.975	5.0407e-28
CustAge	0.61562	0.24783	2.4841	0.012988
ResStatus	1.3776	0.65266	2.1107	0.034799
EmpStatus	0.88592	0.29296	3.024	0.0024946
CustIncome	0.69836	0.21715	3.216	0.0013001
TmWBank	1.106	0.23266	4.7538	1.9958e-06
OtherCC	1.0933	0.52911	2.0662	0.038806
AMBalance	1.0437	0.32292	3.2322	0.0012285

1200 observations, 1192 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 89.7, p-value = 1.42e-16

Step 4. Review and format scorecard points.

After fitting the logistic model, by default the points are un scaled and come directly from the combination of WOE values and model coefficients. The display points function summarizes the scorecard points.

Predictors	Bin	Points
'CustAge'	'[-Inf,37)'	-0.15314
'CustAge'	'[37,40)'	-0.062247
'CustAge'	'[40,46)'	0.045763
'CustAge'	'[46,58)'	0.22888
'CustAge'	'[58,Inf]'	0.48354
'ResStatus'	'Tenant'	-0.031302
'ResStatus'	'Home Owner'	0.12697
'ResStatus'	'Other'	0.37652
'EmpStatus'	'Unknown'	-0.076369
'EmpStatus'	'Employed'	0.31456
'CustIncome'	'[-Inf,29000)'	-0.45455
'CustIncome'	'[29000,33000)'	-0.1037
'CustIncome'	'[33000,42000)'	0.077768
'CustIncome'	'[42000,47000)'	0.24406
'CustIncome'	'[47000,Inf]'	0.43536
'TmWBank'	'[-Inf,12)'	-0.18221
'TmWBank'	'[12,45)'	-0.038279
'TmWBank'	'[45,71)'	0.39569
'TmWBank'	'[71,Inf]'	0.95074
'OtherCC'	'No'	-0.193
'OtherCC'	'Yes'	0.15868
'AMBalance'	'[-Inf,558.88)'	0.3552
'AMBalance'	'[558.88,1597.44)'	-0.026797
'AMBalance'	'[1597.44,Inf]'	-0.21168

This is a good time to modify the bin labels, if this is something of interest for cosmetic reasons. To do so, use modifies bins to change the bin labels.

Predictors	Bin	Points
'CustAge'	'Up to 36'	-0.15314
'CustAge'	'37 to 39'	-0.062247
'CustAge'	'40 to 45'	0.045763
'CustAge'	'46 to 57'	0.22888
'CustAge'	'58 and up'	0.48354
'ResStatus'	'Tenant'	-0.031302
'ResStatus'	'Home Owner'	0.12697
'ResStatus'	'Other'	0.37652
'EmpStatus'	'Unknown'	-0.076369
'EmpStatus'	'Employed'	0.31456
'CustIncome'	'Up to 28999'	-0.45455
'CustIncome'	'29000 to 32999'	-0.1037
'CustIncome'	'33000 to 41999'	0.077768
'CustIncome'	'42000 to 46999'	0.24406
'CustIncome'	'47000 and up'	0.43536
'TmWBank'	'Up to 11'	-0.18221
'TmWBank'	'12 to 44'	-0.038279
'TmWBank'	'45 to 70'	0.39569
'TmWBank'	'71 and up'	0.95074
'OtherCC'	'No'	-0.193
'OtherCC'	'Yes'	0.15868
'AMBalance'	'Up to 558.87'	0.3552
'AMBalance'	'558.88 to 1597.43'	-0.026797
'AMBalance'	'1597.44 and up'	-0.21168

Points are usually scaled and also often rounded. To do this, use the format points function. For example, you can set a target level of points corresponding to a target odds level and also set the required points-to-double-the-odds (PDO).

Predictors	Bin	Points
'CustAge'	'Up to 36'	53.239
'CustAge'	'37 to 39'	59.796
'CustAge'	'40 to 45'	67.587
'CustAge'	'46 to 57'	80.796
'CustAge'	'58 and up'	99.166
'ResStatus'	'Tenant'	62.028
'ResStatus'	'Home Owner'	73.445
'ResStatus'	'Other'	91.446
'EmpStatus'	'Unknown'	58.777
'EmpStatus'	'Employed'	86.976
'CustIncome'	'Up to 28999'	31.497
'CustIncome'	'29000 to 32999'	56.805
'CustIncome'	'33000 to 41999'	69.896
'CustIncome'	'42000 to 46999'	81.891
'CustIncome'	'47000 and up'	95.69
'TmWBank'	'Up to 11'	51.142
'TmWBank'	'12 to 44'	61.524
'TmWBank'	'45 to 70'	92.829
'TmWBank'	'71 and up'	132.87
'OtherCC'	'No'	50.364
'OtherCC'	'Yes'	75.732
'AMBalance'	'Up to 558.87'	89.908
'AMBalance'	'558.88 to 1597.43'	62.353
'AMBalance'	'1597.44 and up'	49.016

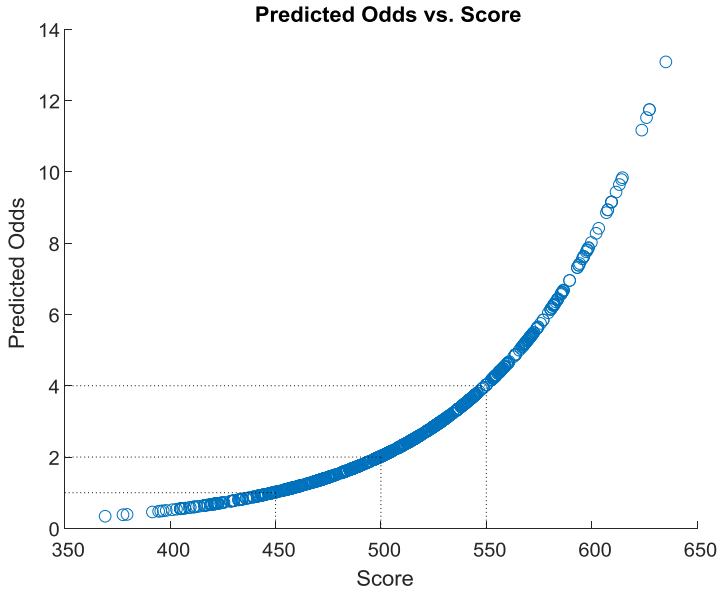
Step 5. Score the data.

The score function computes the scores for the training data. An optional data input can also be passed to score, for example, validation data. The points per predictor for each customer are provided as an optional output.

528.2044
554.8861
505.2406
564.0717
554.8861
586.1904
441.8755
515.8125
524.4553
508.3169

CustAge	ResStatus	EmpStatus	CustIncome	TmWBank	OtherCC	AMBalance
80.796	62.028	58.777	95.69	92.829	75.732	62.353
99.166	73.445	86.976	95.69	61.524	75.732	62.353
80.796	62.028	86.976	69.896	92.829	50.364	62.353
80.796	73.445	86.976	95.69	61.524	75.732	89.908
99.166	73.445	86.976	95.69	61.524	75.732	62.353
99.166	73.445	86.976	95.69	92.829	75.732	62.353
53.239	73.445	58.777	56.805	61.524	75.732	62.353
80.796	91.446	86.976	95.69	61.524	50.364	49.016
80.796	62.028	58.777	95.69	61.524	75.732	89.908
80.796	73.445	58.777	95.69	61.524	75.732	62.353

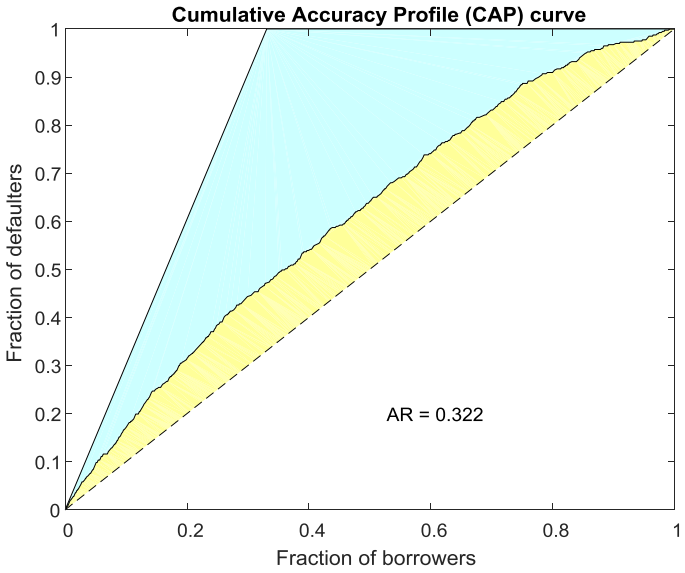
Step 6. Calculate the probability of default.

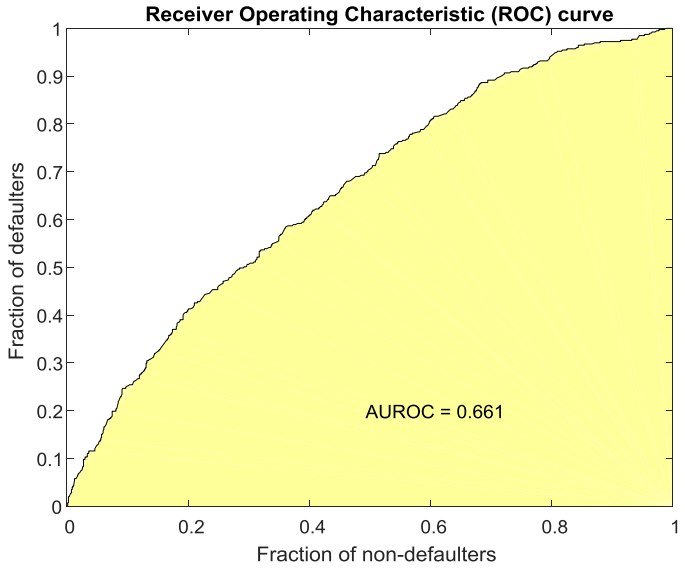


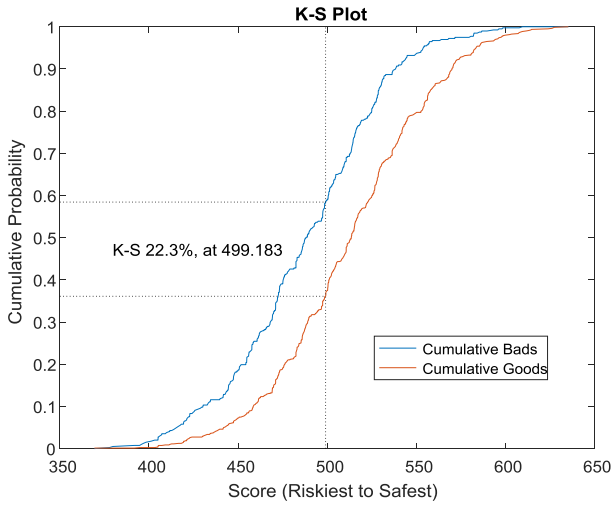
Step 7. Validate the credit scorecard model using the CAP, ROC, and Kolmogorov-Smirnov statistic

Measure	Value
'Accuracy Ratio'	0.32225
'Area under ROC curve'	0.66113
'KS statistic'	0.22324
'KS score'	499.18

Scores	ProbDefault	TrueBads	FalseBads	TrueGoods	FalseGoods	Sensitivity	FalseAlarm	PctObs
0.7535	0	1	802	397	0	0.0012453	0.00083333	369.4
377.86	0.73107	1	1	802	396	0.0025189	0.0012453	0.0016667
379.78	0.7258	2	1	802	395	0.0050378	0.0012453	0.0025
391.81	0.69139	3	1	802	394	0.0075567	0.0012453	0.0033333
394.77	0.68259	3	2	801	394	0.0075567	0.0024907	0.0041667
395.78	0.67954	4	2	801	393	0.010076	0.0024907	0.005
396.95	0.67598	5	2	801	392	0.012594	0.0024907	0.0058333
398.37	0.67167	6	2	801	391	0.015113	0.0024907	0.0066667
401.26	0.66276	7	2	801	390	0.017632	0.0024907	0.0075
403.23	0.65664	8	2	801	389	0.020151	0.0024907	0.0083333
405.09	0.65081	8	3	800	389	0.020151	0.003736	0.0091667
405.15	0.65062	11	5	798	386	0.027708	0.0062267	0.013333
405.37	0.64991	11	6	797	386	0.027708	0.007472	0.014167
406.18	0.64735	12	6	797	385	0.030227	0.007472	0.015
407.14	0.64433	13	6	797	384	0.032746	0.007472	0.015833







II. Analysis & Finding

Credit scoring is one of the most widely used credit risk analysis tools and modeling. The goal of credit scoring is ranking borrowers by their credit worthiness. In the context of retail credit (credit cards, mortgages, car loans, etc.), credit scoring is performed using a credit scorecard. Credit scorecards represent different characteristics of a customer (age, residential status, time at current address, time at current job, and so on) translated into points and the total number of points becomes the credit score. The credit worthiness of customers is summarized by their credit score; high scores usually correspond to low-risk customers, and conversely. Scores are also used for corporate credit analysis of small and medium enterprises, and, large corporations. In this case study, our objectives was

- To estimate credit risk factors profiling.
- To know default probability from credit score data.
- To examine internal credit risk scoring.
- Validate the credit scorecard model using the Cumulative Accuracy Profile (CAP), Receiver Operating Charactestic (ROC), and Kolmogorov-Smirnov statistic.

Finally, we find the result Accuracy Ratio = 0.32225, Area under ROC curve = 0.66113
KS statistic = 0.22324 and KS score = 499.18

III. Conclusion

To understand risk levels of credit users, credit providers normally collect vast amount of information on borrowers. Some predictive analytic techniques can be used to analyze or to determine risk levels involved on credits, finances, and loans, i.e., default risk levels. We are trying to find default probability of Cumulative Accuracy Profile (CAP), the Receiver Operating Characteristic (ROC), and the Kolmogorov-Smirnov (K-S) statistic. The goal of this case study was predictive analytic technique we used to analyze to find default probability and accuracy of score.

References

[1]. Altman, E. "Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy." *Journal of Finance.* Vol. 23, No. 4, (Sep., 1968), pp. 589–609.
[2]. Basel Committee on Banking Supervision, *International Convergence of Capital Measurement and Capital Standards: A Revised Framework, Bank for International Settlements (BIS).* comprehensive version, June 2006.
[3]. Hanson, S. and T. Schuermann. "Confidence Intervals for Probabilities of Default." *Journal of Banking & Finance.* Vol. 30(8), Elsevier, August 2006, pp. 2281–2301.
[4]. Jafry, Y. and T. Schuermann. "Measurement, Estimation and Comparison of Credit Migration Matrices." *Journal of Banking & Finance.* Vol. 28(11), Elsevier, November 2004, pp. 2603–2639.
[5]. Löffler, G. and P. N. Posch. *Credit Risk Modeling Using Excel and VBA.* West Sussex, England: Wiley Finance, 2007.
[6]. Schuermann, T. "Credit Migration Matrices." in E. Melnick and B. Everitt (eds.), *Encyclopedia of Quantitative Risk Analysis and Assessment.* Wiley, 2008.
[7]. Agarwal, Sumit; Yan Chang; and Abdullah Yavas (2010). ìAdverse Selection in Mortgage Securitizationî, Paolo Ba¢ Centre Research Paper. No. 2010-67.
[8]. Altman, Edward I. and Anthony Saunders (1997). ìCredit Risk Measurement: Developments over the last 20 Years.î Journal of Banking & Finance, 21(11-12), pp. 1721-1742
[9]. Anderson, Gordon (1996). ìNonparametric Tests of Stochastic Dominance in Income Distributions.î Econometrica, 64(5), pp. 1183-93.
[10]. Andriotis A., (2011) Riskier Loans Make a Comeback, as Private Firms Take the Field, Wall Street Journal, Homes, July 12.

[11]. Avery, Robert B.; Raphael W. Bostic; Paul S. Calem; and Glenn B. Canner (2000). ìCredit Scoring: Statistical Issues and Evidence from Credit-Bureau Files.îReal Estate Economics, 28(3), pp. 523-547.
[12]. Avery, Robert B.; Paul S. Calem; and Glenn B. Canner (2003). ìAn Overview of Consumer Data and Credit Reporting,îBoard of Governors Federal Reserve Bulletin, February, pp. 47-73.
[13]. Avery, Robert B.; Paul S. Calem; and Glenn B. Canner (2004). ìConsumer Credit Scoring: Do Situational Circumstances Matter?î Journal of Banking & Finance, 28(4), pp. 835-856.
[14]. Bennardo, Alberto; Marco Pagano; and Salvatore Piccolo, (2009). ìMultiple-Bank Lending, Creditor Rights and Information Sharing.îCEPR Discussion Papers DP7186.
[15]. Bhattacharya; Anand, William Berliner; and Jonathan Lieber, (2006) "Alt-A Mortgages and MBS." In: Frank Fabozzi, Ed. The Handbook of Mortgage-Backed Securities. McGraw-Hill: New York. pp. 187-206.
[16]. Brooks, Rick and Ruth Simon ìSubprime Debacle Traps Even Very Credit-Worthy. As Housing Boomed, Industry Pushed Loans to a Broader Marketî. Wall Street Journal, Leader, December 3, 2007.
[17]. Brown, Martin and Christian Zehnder (2007). ìCredit Reporting, Relationship Banking, and Loan Repayment.î Journal of Money, Credit and Banking, 39(8), pp. 1883-1918.
[18]. Foote, Christopher L.; Kristopher Gerardi; Lorenz Goette; and Paul S. Willen (2008). ìJust the Facts: An Initial Analysis of Subprimeís Role in the Housing Crisis.î Journal of Housing Economics, 17(4), pp. 291-305.
[19]. Foust, Dean and Aaron Pressman ìCredit Scores: Not-So-Magic Numbers.îBusinessWeek, February 7, 2008.
[20]. Vercammen, James A (1995). ìCredit Bureau Policy and Sustainable Reputation E§ects in Credit Markets.îEconomica, 62(248), pp. 461-78.