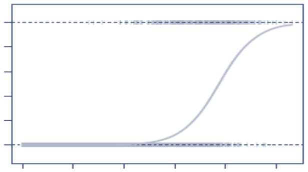
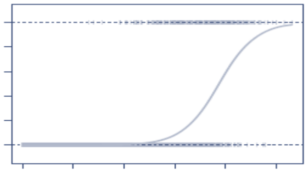


Making Loan Decisions Based on Credit Data

Nathan Grossman

Objectives



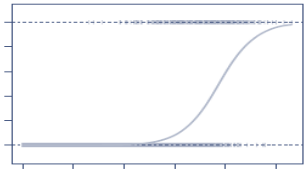
Business Objective

Maximize profits by maximizing the number (and amounts) of loans while minimizing the number (or percentage) of defaults.

Technical Objective

Estimate the probability that a loan applicant would be a good credit risk and not default if given a loan.

Data Set



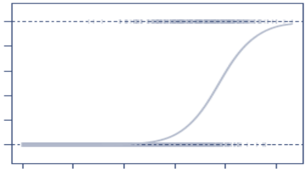
This study uses the well-known German Credit data set * to train and test an algorithm for predicting the probability that a loan applicant would be a good credit risk.

The data set comprises observations of 30 variables on 1000 credit applicants, with 700 of the applicants rated as “good” credit risks and 300 of the applicants rated as “bad” credit risks.

Var. #	Variable Name	Description	Variable Type	Code Description
1.	OBS#	Observation No.	Categorical	Sequence Number in data set
2.	CHK_ACCT	Checking account status	Categorical	0 : < 0 DM 1: 0 <= ...< 200 DM 2 : => 200 DM 3: no checking account
3.	DURATION	Duration of credit in months	Numerical	
4.	HISTORY	Credit history	Categorical	0: no credits taken 1: all credits at this bank paid back duly 2: existing credits paid back duly till now 3: delay in paying off in the past 4: critical account
5.	NEW_CAR	Purpose of credit	Binary	car (new) 0: No, 1: Yes
6.	USED_CAR	Purpose of credit	Binary	car (used) 0: No, 1: Yes
7.	FURNITURE	Purpose of credit	Binary	furniture/equipment 0: No, 1: Yes
8.	RADIO/TV	Purpose of credit	Binary	radio/television 0: No, 1: Yes
9.	EDUCATION	Purpose of credit	Binary	education 0: No, 1: Yes
10.	RETRAINING	Purpose of credit	Binary	retraining 0: No, 1: Yes
11.	AMOUNT	Credit amount	Numerical	
12.	SAV_ACCT	Average balance in savings account	Categorical	0 : < 100 DM 1 : 100<= ... < 500 DM 2 : 500<= ... < 1000 DM 3 : =>1000 DM 4 : unknown/ no savings account
13.	EMPLOYMENT	Present employment since	Categorical	0 : unemployed 1: < 1 year 2 : 1 <= ... < 4 years 3 : 4 <=... < 7 years

* This data set is available at
https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

Model Type Selection



Since our approach involves estimating the probability that a loan applicant would be a good credit risk and not default, a logistic regression model is chosen—instead of, for example, a linear regression model—since a logistic regression is better suited to modeling probabilities which range between 0 and 1.

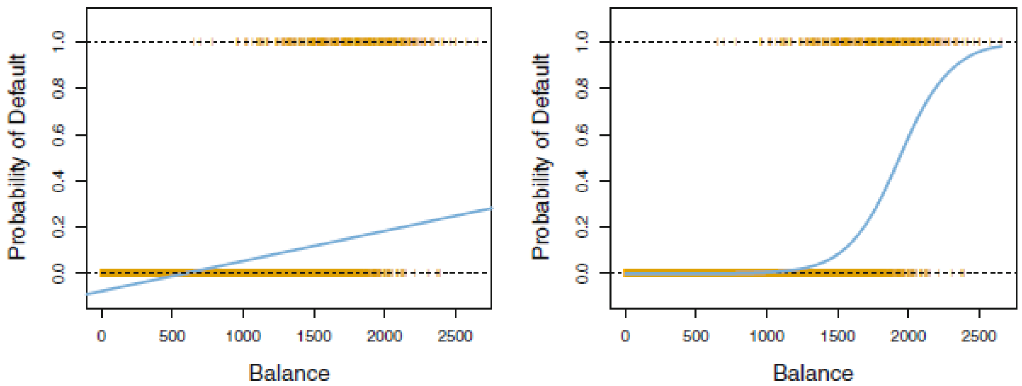
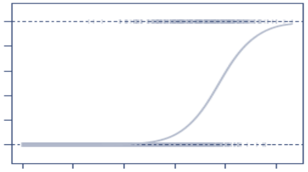


FIGURE 4.2. *Classification using the* `Default` *data.* Left: *Estimated probability of* `default` *using linear regression. Some estimated probabilities are negative! The orange ticks indicate the 0/1 values coded for* `default`*(*`No` *or* `Yes`*).* Right: *Predicted probabilities of* `default` *using logistic regression. All probabilities lie between* 0 *and* 1.

Source: *An Introduction to Statistical Learning with Applications in R*, James et. al., Springer 2013

Dependent Variable



The dependent variable is *good_bad*, which takes on the value *good* if a borrower has not defaulted, and takes on the value *bad* if a borrower has defaulted.

Descriptive Statistics of Categorical Variables

The FREQ Procedure

good_bad	Frequency	Percent	Cumulative Frequency	Cumulative Percent
bad	300	30.00	300	30.00
good	700	70.00	1000	100.00

Logistic Regression

Estimate the probability that a categorical variable Y takes on a certain value, or equivalently is in a certain category, as

$$P(Y = y|X = x) = p(X) = \frac{e^{\beta_0+\beta_1X_1+\cdots+\beta_pX_p}}{1 + e^{\beta_0+\beta_1X_1+\cdots+\beta_pX_p}}$$

by performing a regression on the logit

$$log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1X_1 + \cdots + \beta_pX_p$$

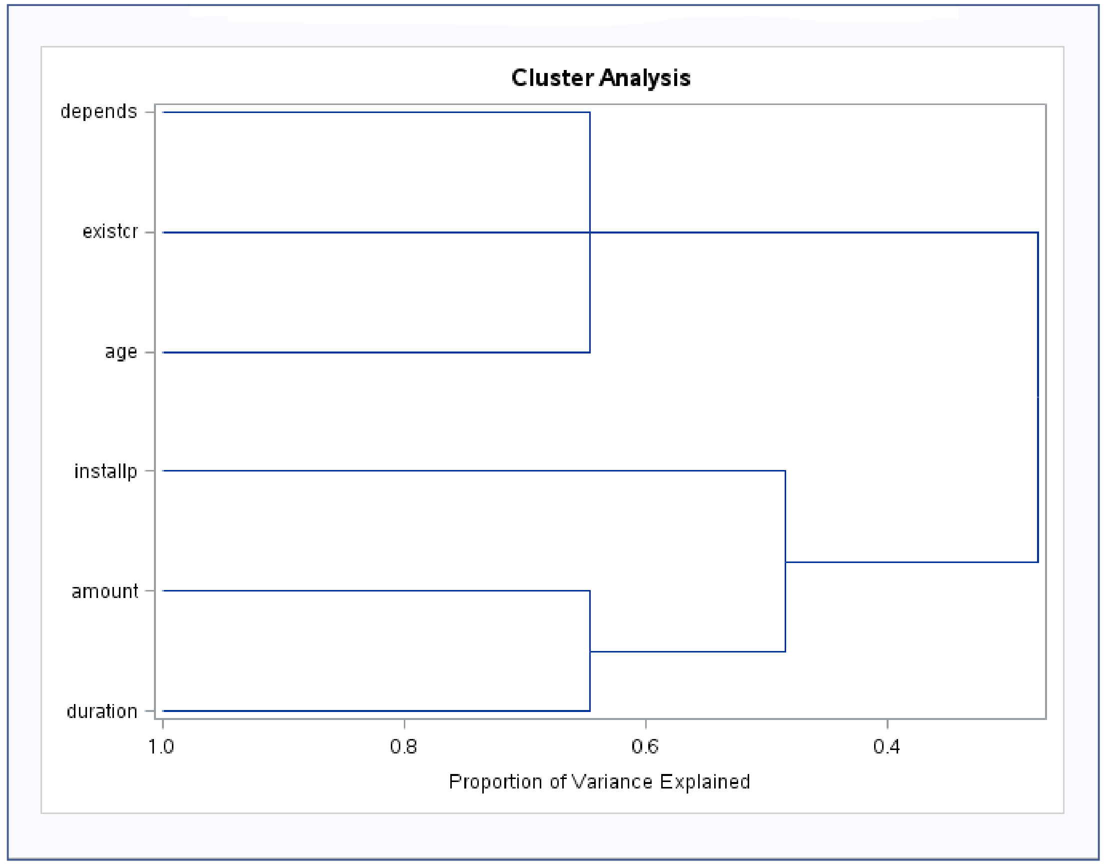
to estimate the parameters $\{\beta_i\}$ with the training data. Then evaluate the expression for P(Y = y | X = x) with the estimated parameters for the test data, and select the category that has the highest probability.

Cluster Analysis

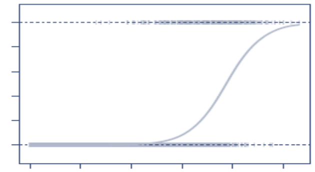
Cluster analysis was performed on the numerical variables in the data, which consisted of 1000 observations comprising categorical as well as numerical variables.

The clusters were found, associated with: (1) *duration* and *amount*; (2) *age*, *existcr* and *depends*; and (3) *installp*.

3 Clusters		R-squared with		
Cluster	Variable	Own Cluster	Next Closest	1-R**2 Ratio
Cluster 1	duration	0.8125	0.0056	0.1886
	amount	0.8125	0.0738	0.2024
Cluster 2	age	0.4531	0.0034	0.5488
	existcr	0.4360	0.0005	0.5642
	depends	0.3631	0.0051	0.6402
Cluster 3	installp	1.0000	0.0119	0.0000



Independent Variable Selection

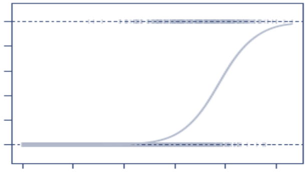


Variable selection was performed with logistic regression using a stepwise algorithm, which both iteratively adds and subtracts independent variables to the model in order to maximize the predictive power thereof.

The stepwise algorithm was run with three different numbers of independent variables specified: 12, 11 and 10.

```


/*
   Logistic Regression with Variable Selection
*/

title "Logistic Regression with Selection of 12 Variables";
proc logistic data=training_data;
  class checking history purpose savings employed marital coapp property
        other housing job;
  model good_bad = checking duration history purpose amount savings
                   employed installp marital coapp resident property
                   age other housing existcr job depends telephon foreign
                   / SELECTION=stepwise INCLUDE=12 DETAILS;
run;

title "Logistic Regression with Selection of 11 Variables";
proc logistic data=training_data;
  class checking history purpose savings employed marital coapp property
        other housing job;
  model good_bad = checking duration history purpose amount savings
                   employed installp marital coapp resident property
                   age other housing existcr job depends telephon foreign
                   / SELECTION=stepwise INCLUDE=11 DETAILS;
run;

title "Logistic Regression with Selection of 10 Variables";
proc logistic data=training_data;
  class checking history purpose savings employed marital coapp property
        other housing job;
  model good_bad = checking duration history purpose amount savings
                   employed installp marital coapp resident property
                   age other housing existcr job depends telephon foreign
                   / SELECTION=stepwise INCLUDE=10 DETAILS;
run;
```

Independent Variable Selection



The model with 10 independent variables was selected, because:

- In the other cases tested, some of the independent variables did not appear to be statistically relevant to the dependent variable (i.e. the p-values associated with some of the variables coefficients were too high).
- As will be shown in the following slide, the 10-variable model achieved a goodness-of-fit comparable to the 20-variable, 12-variable and 11-variable models.

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
checking	3	44.3229	<.0001
duration	1	6.7876	0.0092
history	4	19.2135	0.0007
purpose	9	24.4786	0.0036
amount	1	4.2798	0.0386
savings	4	12.3340	0.0150
employed	4	8.3131	0.0808
installp	1	10.9141	0.0010
marital	3	15.6666	0.0013
coapp	2	7.3337	0.0256
resident	1	0.0176	0.8945
property	3	3.7412	0.2908
age	1	1.4589	0.2271
other	2	6.2153	0.0447
housing	2	2.4251	0.2974
existcr	1	2.3131	0.1283
job	3	0.2710	0.9654
depends	1	2.5296	0.1117
telephon	1	1.4967	0.2212
foreign	1	2.9979	0.0834

20 Independent Variables

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
checking	3	45.1806	<.0001
duration	1	8.5685	0.0034
history	4	24.0414	<.0001
purpose	9	22.0328	0.0088
amount	1	2.8726	0.0901
savings	4	11.8768	0.0183
employed	4	8.4819	0.0754
installp	1	10.0479	0.0015
marital	3	13.6801	0.0034
coapp	2	8.1446	0.0170
resident	1	0.0155	0.9010
property	3	1.9200	0.5892

12 Independent Variables

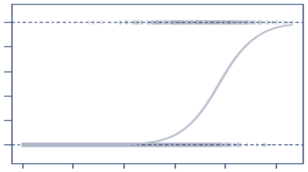
Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
checking	3	46.1285	<.0001
duration	1	9.2944	0.0023
history	4	24.7839	<.0001
purpose	9	22.4185	0.0076
amount	1	3.5464	0.0597
savings	4	11.9593	0.0177
employed	4	8.8140	0.0659
installp	1	10.5347	0.0012
marital	3	13.2397	0.0041
coapp	2	8.4509	0.0146
resident	1	0.0013	0.9712

11 Independent Variables

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
checking	3	46.2027	<.0001
duration	1	9.2961	0.0023
history	4	24.8254	<.0001
purpose	9	22.4402	0.0076
amount	1	3.5486	0.0596
savings	4	11.9613	0.0176
employed	4	8.9094	0.0634
installp	1	10.5365	0.0012
marital	3	13.2556	0.0041
coapp	2	8.4577	0.0146

10 Independent Variables

Goodness-of-Fit Assessment



The model fit statistics AIC and SC decrease, and the statistic -2 Log L increases, for the intercept and covariates (the intercept-only statistics are generally ignored) as the number of variables decreases.

However the changes appear relatively small, indicating that the fit achieved by the 10-variable model is for practical purposes as the fit for the full 20-variable model.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	852.065	710.645
SC	856.616	933.648
-2 Log L	850.065	612.645

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	237.4199	48	<.0001
Score	201.9413	48	<.0001
Wald	142.3777	48	<.0001

20 Independent Variables

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	852.065	707.637
SC	856.616	876.027
-2 Log L	850.065	633.637

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	216.4283	36	<.0001
Score	189.5488	36	<.0001
Wald	138.4854	36	<.0001

12 Independent Variables

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	852.065	703.576
SC	856.616	858.312
-2 Log L	850.065	635.576

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	214.4893	33	<.0001
Score	187.4476	33	<.0001
Wald	137.4647	33	<.0001

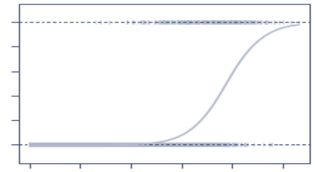
11 Independent Variables

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	852.065	701.577
SC	856.616	851.762
-2 Log L	850.065	635.577

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	214.4880	32	<.0001
Score	187.4069	32	<.0001
Wald	137.4610	32	<.0001

10 Independent Variables

Goodness-of-Fit Assessment



The p-value associated with the Hosmer and Lemeshow goodness-of-fit test is 0.0970.

This relatively low value indicates that the fit might be improved by adding non-linear and/or interaction terms to the model.

That said, reliability of the Hosmer and Lemeshow goodness-of-fit test is not universally accepted.

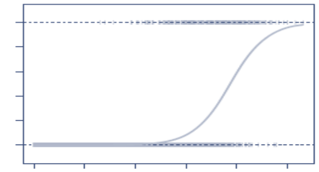
Partition for the Hosmer and Lemeshow Test

		good_bad = bad		good_bad = good	
Group	Total	Observed	Expected	Observed	Expected
1	70	4	1.54	66	68.46
2	70	2	3.89	68	66.11
3	70	4	6.17	66	63.83
4	70	11	9.02	59	60.98
5	70	14	13.01	56	56.99
6	70	16	18.07	54	51.93
7	70	20	24.49	50	45.51
8	70	40	33.01	30	36.99
9	70	37	42.23	33	27.77
10	70	59	55.56	11	14.44

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
13.4578	8	0.0970

Model Interpretation



The maximum likelihood estimates are comparisons between each category of a categorical variable and the overall average (roughly speaking) of the log-odds of defaulting, adjusting for other variables in the model.

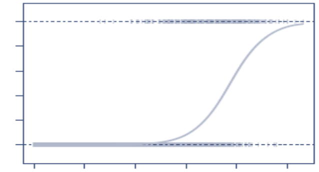
For example, the log-odds of default for a borrower in category 1 of *checking* is 0.7232 above average, while the log-odds of default for a borrower in category 3 of checking is -0.3211 below average.

When there are two alternatives, $x = 1$ and $x = 0$, with $\Pr[x = 1] = p$ (a Bernoulli), it is common to interpret $x = 1$ as a success and $x = 0$ as a failure. In such common circumstances, statisticians and people who go to the track, usually not economists, like to talk about the odds of a success rather than the probability of a success, where the odds, O, are

$$O = \frac{p}{1-p}$$

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.8475	0.5342	28.4138	<.0001
checking	1	1	0.7232	0.1776	16.5843	<.0001
checking	2	1	0.5799	0.1783	10.5839	0.0011
checking	3	1	-0.3211	0.3027	1.1255	0.2887
duration		1	0.0322	0.0106	9.2961	0.0023
history	0	1	0.5678	0.3841	2.1861	0.1393
history	1	1	1.2408	0.3934	9.9501	0.0016
history	2	1	-0.3966	0.1895	4.3796	0.0364
history	3	1	-0.3642	0.2856	1.6262	0.2022
purpose	0	1	0.7871	0.2680	8.6277	0.0033
purpose	1	1	-0.9262	0.4306	4.6277	0.0315
purpose	2	1	0.0953	0.2877	0.1096	0.7406
purpose	3	1	-0.0430	0.2731	0.0248	0.8748
purpose	4	1	0.5817	0.8185	0.5052	0.4772
purpose	5	1	0.2986	0.5767	0.2681	0.6046
purpose	6	1	0.9818	0.4104	5.7237	0.0167
purpose	8	1	-1.3202	1.1881	1.2348	0.2665
purpose	9	1	0.0961	0.3676	0.0684	0.7937
amount		1	0.000098	0.000052	3.5486	0.0596
savings	1	1	0.5729	0.2028	7.9820	0.0047
savings	2	1	0.4063	0.2892	1.9735	0.1601
savings	3	1	0.1222	0.4210	0.0842	0.7716
savings	4	1	-0.8903	0.4888	3.3181	0.0685
employed	1	1	0.4598	0.3138	2.1471	0.1428
employed	2	1	0.2562	0.2149	1.4206	0.2333
employed	3	1	0.00956	0.1740	0.0030	0.9562
employed	4	1	-0.6658	0.2370	7.8919	0.0050
installp		1	0.3329	0.1026	10.5365	0.0012
marital	1	1	0.4806	0.3187	2.2740	0.1316
marital	2	1	0.0510	0.1899	0.0721	0.7883
marital	3	1	-0.6311	0.1807	12.1956	0.0005
coapp	1	1	0.0618	0.2347	0.0694	0.7922
coapp	2	1	0.9378	0.3461	7.3413	0.0067

Model Interpretation



The odds ratio estimates are comparisons between different categories of a given categorical variable.

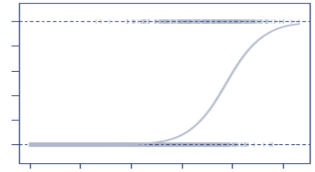
For example, the odds ratio of 1.926 indicates that the predicted odds of default for a borrower in category 3 of *checking* are 1.926 times the odds for a borrower in category 4. In other words, the odds of a default for category 3 borrowers are 93% higher than the odds for category 4 borrowers.

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
checking 1 vs 4	5.502	3.187	9.499
checking 2 vs 4	4.768	2.791	8.144
checking 3 vs 4	1.936	0.812	4.616
duration	1.033	1.012	1.054
history 0 vs 4	5.031	1.847	13.703
history 1 vs 4	9.862	3.563	27.291
history 2 vs 4	1.918	1.158	3.178
history 3 vs 4	1.981	0.930	4.220
purpose 0 vs X	3.812	0.783	18.554
purpose 1 vs X	0.687	0.123	3.839
purpose 2 vs X	1.909	0.385	9.460
purpose 3 vs X	1.662	0.338	8.166
purpose 4 vs X	3.105	0.299	32.258
purpose 5 vs X	2.339	0.339	16.129
purpose 6 vs X	4.632	0.818	26.232
purpose 8 vs X	0.463	0.023	9.233
purpose 9 vs X	1.910	0.361	10.113
amount	1.000	1.000	1.000
savings 1 vs 5	2.190	1.216	3.946
savings 2 vs 5	1.854	0.846	4.065
savings 3 vs 5	1.396	0.462	4.213
savings 4 vs 5	0.507	0.142	1.808
employed 1 vs 5	1.681	0.721	3.922
employed 2 vs 5	1.372	0.727	2.589
employed 3 vs 5	1.072	0.623	1.845
employed 4 vs 5	0.546	0.279	1.068
installp	1.395	1.141	1.706
marital 1 vs 4	1.464	0.529	4.050
marital 2 vs 4	0.953	0.462	1.964
marital 3 vs 4	0.482	0.235	0.987
coapp 1 vs 3	2.891	1.085	7.701
coapp 2 vs 3	6.941	1.880	25.629

Now imagine two groups: A and B, maybe women and men, such that the probability of a success for A, p_A, is different from the probability of success for B, p_B. In which case consider the odds ratio

$$O_r = \frac{\frac{p_A}{1-p_A}}{\frac{p_B}{1-p_B}}$$

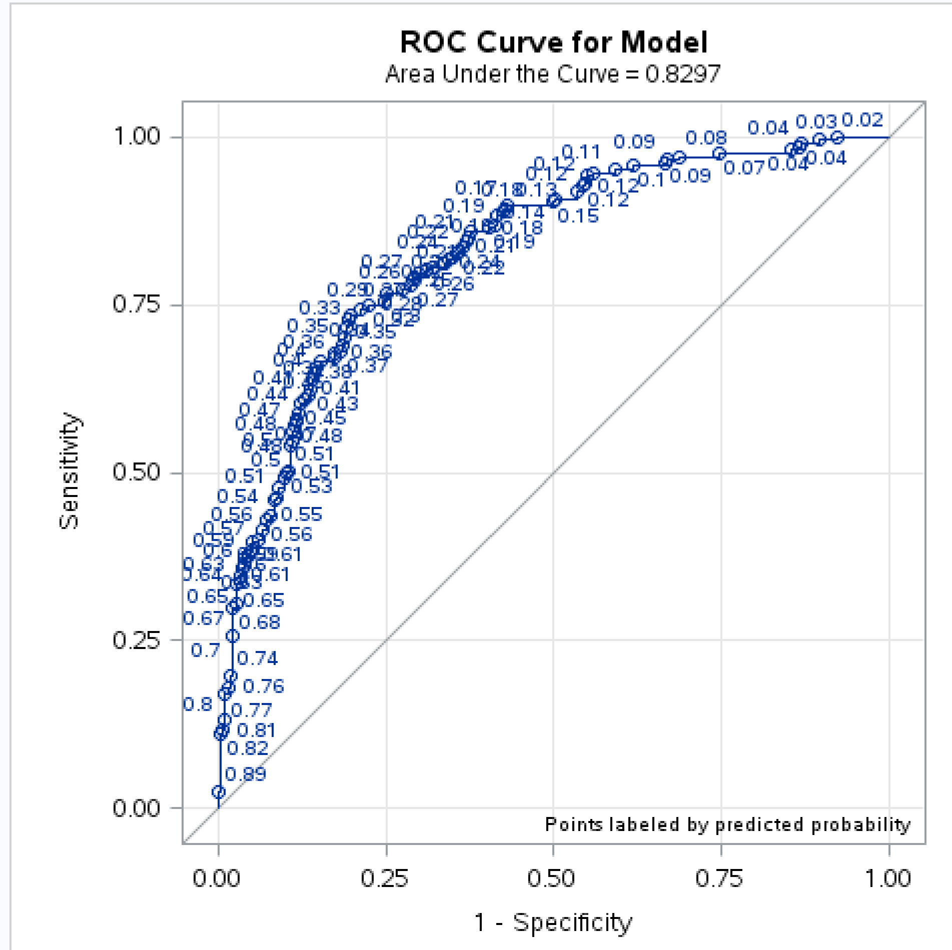
Model Limitation



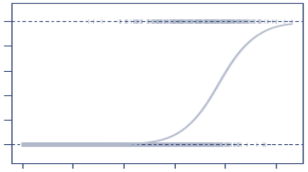
In the ROC curve diagram, the 45-degree line represents the expected ROC curve for a model with an intercept only, that is, one with no predictive power. The more the curve departs from the 45-degree line, the greater the predictive power. The standard statistic for summarizing that departure is the area under the curve, which here is reported as 0.8297.

The “generalized” R-Square calculated by the LOGISTIC procedure in SAS is a generalization of the R-squared metric from linear to logistic regression. However, it cannot be interpreted as a proportion of variance “explained” by the independent variables. Rather, it is reflective of the log-likelihood which is maximized by the algorithm used to obtain the model.

R-Square	0.2639	**Max-rescaled R-Square**	0.3754



Kolmogorov-Smirnov Analysis



Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov statistic measures the maximum deviation of the EDF within the classes from the pooled EDF. PROC NPAR1WAY computes the Kolmogorov-Smirnov statistic as

$$KS = \max_j \sqrt{\frac{1}{n} \sum_i n_i \left(F_i(x_j) - F(x_j)\right)^2} \quad \text{where} \quad j = 1, 2, \ldots, n$$

The asymptotic Kolmogorov-Smirnov statistic is computed as

$$KS_a = KS \times \sqrt{n}$$

For each class level i and overall, PROC NPAR1WAY displays the value of F_i at the maximum deviation from F and the value $\sqrt{n_i}\ (F_i - F)$ at the maximum deviation from F. PROC NPAR1WAY also gives the observation where the maximum deviation occurs.

If there are only two class levels, PROC NPAR1WAY computes the two-sample Kolmogorov-Smirnov test statistic D as

$$D = \max_j |F_1(x_j) - F_2(x_j)| \quad \text{where} \quad j = 1, 2, \ldots, n$$

The p-value for this test is the probability that D is greater than the observed value d under the null hypothesis of no difference between class levels (samples). PROC NPAR1WAY computes the asymptotic p-value for D by using the approximation

$$\text{Prob}(D > d) = 2 \sum_{i=1}^{\infty} (-1)^{(i-1)}\ e^{(-2i^2\ z^2)}$$

where

$$z = d\sqrt{n_1\ n_2 / n}$$

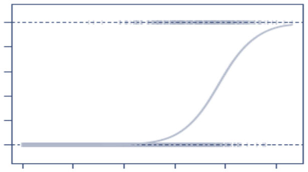
Source: *SAS/STAT® 14.1 User's Guide - The NPAR1WAY Procedure*

Kolmogorov-Smirnov Test for Variable probability Classified by Variable good_bad			
good_bad	**N**	**EDF at Maximum**	**Deviation from Mean at Maximum**
good	493	0.801217	3.516160
bad	207	0.265700	-5.426339
Total	700	0.642857	
Maximum Deviation Occurred at Observation 102			
Value of probability at Maximum = 0.342913			

Kolmogorov-Smirnov Two-Sample Test (Asymptotic)			
KS	0.244390	**D**	0.535517
KSa	6.465953	**Pr > KSa**	<.0001

```
/*
   Kolmogorov-Smirnov Analysis
*/

title "Logistic Regression to Find the Best Distribution Model";
proc logistic data=training_data;
  class checking history purpose savings employed marital coapp;
  model good_bad = checking duration history purpose amount savings
                   employed installp marital coapp;
  output out=output_prob p=probability;
run;

title "Kolmogorov-Smirnov Analysis";
proc npar1way data=output_prob;
class good_bad;
var probability;
output out=npar_an;
run;

proc print data=npar_an;
var _D_;
title 'data=npar_an';
run;

```

Scoring Populations

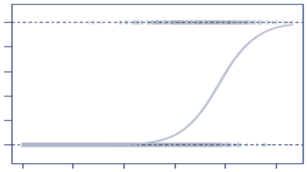


The data, which contained 1000 observations, was divided into 700 observations of training data and 300 observations of test data.

After the model was obtained using the training data, the model was run on both test data and the training data, for sake of comparison.

```

/*
   Data Scoring
*/

* title "Storing Model for Data Scoring";
proc logistic data=training_data outmodel=trained_model noprint;
  class checking history purpose savings employed marital coapp;
  model good_bad = checking duration history purpose amount savings
                   employed installp marital coapp;
run;


title "Training Data Scoring";
proc logistic inmodel=trained_model;
  score data=training_data fitstat out=training_data_score;
run;

* proc print data=training_data_score;
* run;

title "Test Data Scoring";
proc logistic inmodel=trained_model;
  score data=test_data fitstat out=test_data_score;
run;

* proc print data=test_data_score;
* run;

```

Scoring Populations



The error rate was 21.3% for the training data and 24.0% for the test data.

Thus the model does not appear to suffer from over fitting.

Training Data Scoring

The LOGISTIC Procedure

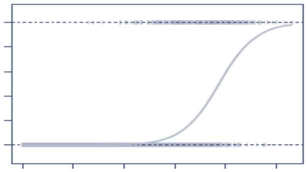
Fit Statistics for SCORE Data											
Data Set	Total Frequency	Log Likelihood	Error Rate	AIC	AICC	BIC	SC	R-Square	Max-Rescaled R-Square	AUC	Brier Score
WORK.TRAINING_DATA	700	-317.8	0.2129	701.5768	704.9462	851.7625	851.7625	0.263916	0.375358	0.829654	0.145846

Test Data Scoring

The LOGISTIC Procedure

Fit Statistics for SCORE Data											
Data Set	Total Frequency	Log Likelihood	Error Rate	AIC	AICC	BIC	SC	R-Square	Max-Rescaled R-Square	AUC	Brier Score
WORK.TEST_DATA	300	-152.0	0.2400	370.0516	378.4876	492.2764	492.2764	0.201241	0.2834	0.793933	0.168305

Conclusion



The model appears to achieve the objectives of estimating the probability of default—and identifying probable defaulters—reasonably well.

Improvements might be made by:

- Including non-linear and interaction terms in the model.
- Altering the threshold for prediction of default in order to reflect the cost of increasing the number of defaults versus the benefit of increasing the number of loans granted.

Thank you for your time

```
libname mydata '/folders/myfolders/';
options label=no;

/*
   Display data
*/

* title "Data";
* proc print data=mydata.dmagecr;
* run;


/*
   Divide data into training and test data
*/

data training_data;
  set mydata.dmagecr (firstobs=1 obs=700);
run;

* title "Training Data";
* proc print data=training_data;
* run;

data test_data;
  set mydata.dmagecr (firstobs=701 obs=1000);
run;

* title "Test Data";
* proc print data=test_data;
* run;


/*
   Descriptive Statistics of Variables
*/

title "Descriptive Statistics of Numerical Variables";
proc means data=mydata.dmagecr n nmiss mean std median;
var duration amount installp age existcr depends;
run;

title "Descriptive Statistics of Categorical Variables";
proc freq data=mydata.dmagecr;
tables good_bad checking history purpose savings employed marital coapp property
       other housing job;
run;
```

```
/*
   Cluster Analysis of Independent Variables
*/

title "Cluster Analysis of Numerical Independent Variables";
proc varclus data=mydata.dmagecr;
var duration amount installp age existcr depends;
run;


/*
   Logistic Regression without Variable Selection
*/

title "Logistic Regression without Variable Selection";
proc logistic data=training_data;
  class checking history purpose savings employed marital coapp property
        other housing job;
  model good_bad = checking duration history purpose amount savings
                   employed installp marital coapp resident property
                   age other housing existcr job depends telephon foreign;
run;


/*
   Logistic Regression with Variable Selection
*/

title "Logistic Regression with Selection of 12 Variables";
proc logistic data=training_data;
  class checking history purpose savings employed marital coapp property
        other housing job;
  model good_bad = checking duration history purpose amount savings
                   employed installp marital coapp resident property
                   age other housing existcr job depends telephon foreign
                   / SELECTION=stepwise INCLUDE=12 DETAILS;
run;

title "Logistic Regression with Selection of 11 Variables";
proc logistic data=training_data;
  class checking history purpose savings employed marital coapp property
        other housing job;
  model good_bad = checking duration history purpose amount savings
                   employed installp marital coapp resident property
                   age other housing existcr job depends telephon foreign
                   / SELECTION=stepwise INCLUDE=11 DETAILS;
run;

title "Logistic Regression with Selection of 10 Variables";
proc logistic data=training_data;
```

```
  class checking history purpose savings employed marital coapp property
        other housing job;
  model good_bad = checking duration history purpose amount savings
                   employed installp marital coapp resident property
                   age other housing existcr job depends telephon foreign
                   / SELECTION=stepwise INCLUDE=10 DETAILS;
run;


/*
   Logistic Regresson with Hard-Coded Variable Selection
*/

title "Logistic Regression with 10 Hard-Coded Variables";
proc logistic data=training_data;
  class checking history purpose savings employed marital coapp;
  model good_bad = checking duration history purpose amount savings
                   employed installp marital coapp;
run;


/*
   Outliers and Influential Observations
*/

title "Outliers / Influencers";
proc logistic data=training_data
              plots(label)=(influence dfbetas leverage);
  class checking history purpose savings employed marital coapp;
  model good_bad = checking duration history purpose amount savings
                   employed installp marital coapp;
run;


/*
   Goodness of Fit
*/

title "Hosmer-Lemeshow (HL) Statistic";
proc logistic data=training_data;
  class checking history purpose savings employed marital coapp;
  model good_bad = checking duration history purpose amount savings
                   employed installp marital coapp / lackfit;
run;


/*
   Predictive Power
*/
```

```
title "Generalized R-Squared & ROC Curves";
proc logistic data=training_data
              plots(only)=roc(id=cutpoint);
  class checking history purpose savings employed marital coapp;
  model good_bad = checking duration history purpose amount savings
                   employed installp marital coapp / rsq;
run;


/*
   Kolmogorov-Smirnov Analysis
*/

title "Logistic Regression to Find the Best Distribution Model";
proc logistic data=training_data;
  class checking history purpose savings employed marital coapp;
  model good_bad = checking duration history purpose amount savings
                   employed installp marital coapp;
  output out=output_prob p=probability;
run;

title "Kolmogorov-Smirnov Analysis";
proc npar1way data=output_prob;
class good_bad;
var probability;
output out=npar_an;
run;

proc print data=npar_an;
var _D_;
title 'data=npar_an';
run;


/*
   Data Scoring
*/

* title "Storing Model for Data Scoring";
proc logistic data=training_data outmodel=trained_model noprint;
  class checking history purpose savings employed marital coapp;
  model good_bad = checking duration history purpose amount savings
                   employed installp marital coapp;
run;


title "Training Data Scoring";
proc logistic inmodel=trained_model;
  score data=training_data fitstat out=training_data_score;
run;
```

```
* proc print data=training_data_score;
* run;

title "Test Data Scoring";
proc logistic inmodel=trained_model;
  score data=test_data fitstat out=test_data_score;
run;

* proc print data=test_data_score;
* run;
```