CREDIT RISK ANALYSIS:

APPLICATION OF MULTIVARIATE ANALYSIS

PRESENTED BY:

ANANYA GUPTA (544)
ISHA AGRAWAL (469)
ISHITA AGARWAL (528)
KRITI MAHESHWARI(474)
MUSKAN AGARWAL (618)

AGENDA

1. Introduction: Understanding Credit Risk
2. Objective: Managing Credit Risk
3. About the Data
4. Analysis: Discriminant Analysis
5. Analysis: Binomial Logistic Regression
6. Analysis: Survival Analysis
7. Result: Drawing Inferences from the Analysis
8. Next Steps: How to minimize Credit Risk?

ACKNOWLEDGEMENT

We would like to extend our gratitude to Dr. Anuradha Sarkar for giving us the opportunity to explore the real-world applications of Multivariate Analysis and constantly guiding us through the course of the project.

INTRODUCTION

UNDERSTANDING CREDIT RISK

INTRODUCTION TO THE TOPIC

Credit risk is the potential risk of loss that arises when a borrower fails to repay a loan or meet their financial obligations. It is the risk that a borrower or a counterparty will default on their contractual obligations and fail to make required payments.

Credit risk is a major concern for lenders, as defaults can result in significant financial losses. By evaluating credit risk, lenders can determine whether to approve a loan application, set appropriate interest rates, and take appropriate measures to mitigate the risk of default

MANAGING CREDIT RISK

OBJECTIVE OF THE RESEARCH AND ANALYSIS

Financial institutions use various methods to assess and manage credit risk, which helps to reduce the potential for losses and maintain the overall financial health of the institution.

Identification and Prediction	Understanding the different factors and their impact on credit risk to predict whether a borrower will fall in the category of a "**Defaulter**" or "**Non-Defaulter**"
Estimation of Expected Time	Estimating the expected time for default by a borrower based on their **loan's interest rate**
Devise Mitigation Strategies	Interpreting the findings to devise mitigation strategies in order to **minimize credit risk** for the financial institutions

ABOUT THE DATA

DEFINING THE VARIABLES

Loan intent	The purpose for which an individual or a business is seeking a loan from a financial institution or lender
Loan grade	Classification assigned to a loan based on the creditworthiness of the borrower and other factors. Loans with higher grades are considered less risky, and qualify for more favorable interest rates and terms
Loan status	0 is used for non defaulters and 1 is used for defaulters (A loan defaulter is a borrower who fails to repay their loan or meet their financial obligations according to the terms of the loan agreement)
Historical default	Historical default refers to the rate at which borrowers have defaulted on loans in the past. It on the type of loan, borrower creditworthiness, and economic conditions

Home ownership	Home ownership refers to the state of owning a home, whether it be a house, condominium, or any other type of residential property
Loan interest rate	The amount of money charged by a lender to a borrower for the use of money over a specified period of time. It is based on borrower's credit score, the amount of the loan & the length of the loan term
Loan percent income	The % of a borrower's income that is used to make loan payments (DTI)- (dividing the borrower's total monthly debt payments by their monthly income). A high loan percent income indicates that the borrower may struggle to make payments on the loan, while a low loan percent income indicates that the borrower may have a lower risk of defaulting on the loan
Credit history length	The length of time a borrower has had credit accounts or loans open. A longer credit history length is viewed as more positive, as it provides lenders with a better understanding of a borrower's credit behavior over time. It demonstrates that the borrower has a track record of managing credit and making payments on time

ANALYSIS

1. Discriminant Analysis
2. Binomial Logistic Regrression
3. Survival Analysis

DISCRIMINANT ANALYSIS

Discriminant analysis is a multivariate analysis technique used to analyze and classify data based on the values of one or more predictor variables that discriminate between groups or categories.

In the context of credit risk, discriminant analysis can be used to identify the factors that are most important in predicting whether a borrower will default on a loan.

ASSUMPTIONS OF DISCRIMINANT ANALYSIS

- **Normality:** This means that the values of the variables should follow a bell-shaped curve.
- **Equal covariance:** It is important because it ensures that the discriminant function is equally effective for all groups.
- **Independence:** The predictor variables should not be highly correlated with each other.
- **Homoscedasticity:** Variance of the residuals should be the same across groups.
- **Adequate sample size:** Sufficient number of observations in each group

DESCRIPTIVE STATISICS

TABLE 1.1

Group Statistics

status		Mean	Std. Deviation	Valid N (listwise)	
				Unweighted	Weighted
.00	age	27.5106	5.59827	21437	21437.000
	income	68094.3951	37392.11631	21437	21437.000
	emp_length	4.8660	3.89998	21437	21437.000
	loan_amt	9092.9818	5791.32298	21437	21437.000
	int_rate	10.2512	2.76282	21437	21437.000
1.00	age	27.0127	5.38420	4971	4971.000
	income	46412.9306	28385.13950	4971	4971.000
	emp_length	4.0266	3.63419	4971	4971.000
	loan_amt	10574.6329	6663.37543	4971	4971.000
	int_rate	12.1518	2.74994	4971	4971.000
Total	age	27.4168	5.56191	26408	26408.000
	income	64013.1106	36857.11665	26408	26408.000
	emp_length	4.7080	3.86524	26408	26408.000
	loan_amt	9371.8854	5993.14333	26408	26408.000
	int_rate	10.6089	2.85859	26408	26408.000

Dependent Variable:

Loan Status (where 0 denotes "non-defaulters" and 1 denotes "defaulters")

Independent Variables:

1. Person's Age (in years)
2. Person's Income (in INR)
3. Person's Employment Length (in years)
4. Loan Amount (in INR)
5. Loan's Interest Rate (in %)

NOTE: In this case, the dependent variable is categorical and the independent variables are numeric in nature.

PRELIMINARY TESTS FOR VALIDITY

TABLE 1.2

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
age	.611	32.375	1	26406	.000
income	.794	1474.297	1	26406	.009
emp_length	.784	191.694	1	26406	.007
loan_amt	.882	248.951	1	26406	.055
int_rate	.724	1913.095	1	26406	.002

- Indicates whether the people in the two groups (non-defaulters and defaulters) have significantly different means.
- For all the predictor variables except **loan amount**, the two groups are significantly different from each other.

TABLE 1.3

Pooled Within-Groups Matrices

		age	income	emp_length	loan_amt	int_rate
Correlation	age	1.000	.107	.151	.044	.012
	income	.107	1.000	.173	.415	.033
	emp_length	.151	.173	1.000	.111	-.057
	loan_amt	.044	.415	.111	1.000	.079
	int_rate	.012	.033	-.057	.079	1.000

- Indicates that the correlation between the different predictor variables is very low i.e. the factors are **independent**

PRELIMINARY TESTS FOR VALIDITY

TABLE 1.4: BOX'S TEST

Test Results

Box's M		1176.859
F	Approx.	78.428
	df1	15
	df2	320579635.9
	Sig.	.000

Tests null hypothesis of equal population covariance matrices.

Since p-value < 0.001, we conclude that the group variances are **unequal**. This poses as a limitation for our analysis.

SUMMARY OF CANONICAL DISCRIMINANT FUNCTION

TABLE 1.5

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	1.030[a]	100.0	100.0	.673

a. First 1 canonical discriminant functions were used in the analysis.

The larger the eigenvalue, the more of the variance in the dependent variable is explained by that function.

correlation between the discriminant function and the dependent variable.

TABLE 1.6

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.456	4194.077	5	.000

good discriminating power of the model

SUMMARY OF CANONICAL DISCRIMINANT FUNCTION

TABLE 1.7

Standardized Canonical Discriminant Function Coefficients

	Function
	1
age	-.017
income	-.790
emp_length	-.089
loan_amt	.523
int_rate	.629

relative importance of the independent variables in predicting the default status. Coefficients with large absolute values correspond to variables with greater discriminating ability.

TABLE 1.8

Structure Matrix

	Function
	1
int_rate	.649
income	-.569
loan_amt	.234
emp_length	-.205
age	-.084

The structure matrix table shows the correlations of each variable with each discriminant function. The correlations then serve like factor loadings in factor analysis

DISCRIMINANT FUNCTION

TABLE 1.9

Canonical Discriminant Function Coefficients

	Function
	1
age	-.003
income	-.489
emp_length	-.023
loan_amt	.021
int_rate	.228
(Constant)	-1.633

Unstandardized coefficients

Di = -1.633 - 0.003*Age - 0.489*Income - 0.023*Employment_Length + 0.021*Loan_Amount + -.228*Interest_Rate

CLASSIFICATION STATISTICS

TABLE 1.10

Classification Results[a,c]

			Predicted Group Membership		
		status	non defaulter	defaulter	Total
Original	Count	non defaulter	21192	245	21437
		defaulter	1363	3608	4971
	%	non defaulter	98.9	1.1	100.0
		defaulter	27.4	72.6	100.0
Cross-validated[b]	Count	non defaulter	21192	245	21437
		defaulter	1363	3608	4971
	%	non defaulter	98.9	1.1	100.0
		defaulter	27.4	72.6	100.0

a. 82.8% of original grouped cases correctly classified.

b. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

c. 82.8% of cross-validated grouped cases correctly classified.

- 98.9% of the non-defaulters are correctly classified as non-defaulters.
- 72.6% of the defaulters are correctly classified as defaulters
- overall 82.8% of the cases are correctly classified in both original and cross-validated grouped cases

BINARY LOGISTIC REGRESSION

Binary logistic regression is a statistical technique used to model the relationship between a binary dependent variable (also called the outcome or response variable) and one or more predictor variables (also called independent variables or covariates). The dependent variable can only have two values, typically coded as 0 and 1

The goal of binary logistic regression is to estimate the probability of a particular outcome occurring based on the values of the predictor variables. The predictor variables can be either continuous or categorical

ASSUMPTIONS OF BINARY LOGISTIC REGRESSION

- **Binary outcome**: The dependent variable in binary logistic regression should be binary, with only two possible outcomes
- **Independence of observations**: The outcome of one observation should not depend on the outcome of another observation.
- **Linearity of predictors and log odds**: The relationship between the predictor variables and the log odds of the outcome should be linear.
- **Large sample size**: Binary logistic regression performs better with larger sample sizes

ANALYSIS OF BINARY LOGISTIC REGRESSION

TABLE 2.1

Categorical Variables Codings

		Frequency	Parameter coding (1)	(2)	(3)	(4)	(5)
loan_grade	A	9401	1.000	.000	.000	.000	.000
	B	9151	.000	1.000	.000	.000	.000
	C	5695	.000	.000	1.000	.000	.000
	D	2410	.000	.000	.000	1.000	.000
	E	140	.000	.000	.000	.000	1.000
	F	12	.000	.000	.000	.000	.000
loan_intent	EDUCATION	5372	1.000	.000	.000	.000	
	MEDICAL	4957	.000	1.000	.000	.000	
	HOMEIMPROVEMENT	2958	.000	.000	1.000	.000	
	VENTURE	4681	.000	.000	.000	1.000	
	DEBTCONSOLIDATION	8841	.000	.000	.000	.000	
home_ownership	OWN	2067	1.000	.000	.000		
	MORTGAGE	11170	.000	1.000	.000		
	RENT	13487	.000	.000	1.000		
	OTHER	85	.000	.000	.000		
default_on_file	Y	4206	1.000				
	N	22603	.000				

TABLE 2.2

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	8694.744	20	.000
	Block	8694.744	20	.000
	Model	8694.744	20	.000

TABLE 2.3

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	17251.256[a]	.369	.707

Representation of adjusted R squared

TABLE 2.4

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	2.490	8	.962

Ho: No significant difference between observed and predicted observations

H1: There is a significant difference between observed and predicted observations

Hence, as the p value is greater than 0.05, our test is significant

(Opposite to conventional tests)

TABLE 2.5

Contingency Table for Hosmer and Lemeshow Test

		Non defaulters		Defaulters		
		Observed	Expected	Observed	Expected	Total
Step 1	1	2659	2657.234	22	23.766	2681
	2	2588	2584.276	93	92.900	2681
	3	2543	2590.173	138	139.000	2681
	4	2506	2547.041	175	180.000	2681
	5	2453	2488.472	228	234.000	2681
	6	2452	2405.349	229	230.000	2681
	7	2372	2276.475	309	325.000	2681
	8	2188	2135.506	493	500.000	2681
	9	1477	1472.821	1204	1368.000	2681
	10	520	520.000	2160	2207.000	2680

No major difference between observed and expected values

Slight differences between observed and expected values

TABLE 2.6

Classification Table[a]

Observed		Predicted		
		loan status		Percentage Correct
		.00	1.00	
Step 1	loan status .00	20998	760	96.5
	1.00	970	4040	71.0
	Overall Percentage			85.6

a. The cut value is .500

← Very good predictor for non- defaulters of loans

← Good predictor for defaulters of loans

Odds is the ratio of probability= P(A)/P(B)

Concept of log odds= log (P(defaulter)/P(non defaulter))

B(beta) is the predicted change in log odds ie for 1 unit change in predictor, there is Exp (B) change in the probability of the outcome

TABLE 2.7

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	age	-.010	.007	2.372	1	.124	.990	.977	1.003
	income	.000	.000	3.545	1	.060	1.000	1.000	1.000
	employment length	-.002	.005	.098	1	.754	.998	.988	1.009
	loan amount	.109	.000	521.556	1	.000	1.000	1.000	1.000
	loan interest rate	.117	.021	31.726	1	.000	1.124	1.079	1.171
	loan_percent income	3.764	.268	2637.534	1	.000	1.560	1.400	1.700
	credit history length	.005	.010	.230	1	.631	1.005	.985	1.025
	loan_intent			327.632	4	.000			
	loan_intent(1)	-.465	.057	66.918	1	.000	.628	.562	.702
	loan_intent(2)	.052	.054	.952	1	.329	1.054	.948	1.171
	loan_intent(3)	.489	.065	56.603	1	.000	1.630	1.435	1.852
	loan_intent(4)	-.749	.064	137.364	1	.000	.473	.417	.536
	home_ownership			768.972	3	.000			
	home_ownership(1)	-2.250	.347	41.949	1	.000	.105	.053	.208
	home_ownership(2)	-.343	.329	1.084	1	.298	.710	.373	1.353
	home_ownership(3)	.499	.328	2.315	1	.128	1.646	.866	3.129
	loan_grade			1075.360	5	.000			
	loan_grade(1)	-2.309	.684	11.383	1	.001	.099	.026	.380
	loan_grade(2)	-2.302	.670	11.811	1	.001	.100	.027	.372
	loan_grade(3)	-2.236	.664	11.339	1	.001	.107	.029	.393
	loan_grade(4)	-.108	.663	.027	1	.870	.897	.245	3.291
	loan_grade(5)	-.182	.690	.070	1	.792	.833	.215	3.224
	default_on_file(1)	.025	.060	.179	1	.672	1.026	.912	1.154
	Constant	-2.146	.823	6.794	1	.009	.117		

Odds ratio= 1

Probability of falling in the group= probability of falling outside the group

Odds ratio> 1

Probability of an event occurring increases

Odds ratio< 1

Probability of an event occurring decreases

Variables significant

Loan percent income, credit history length, loan intent, home ownership, loan grade

Eg for loan interest rate:

Probability of having a high interest rate is more for defaulters than non-defaulters

WHY SURVIVAL ANALYSIS?

Survival analysis is a statistical technique used to analyze the time until an event of interest occurs. In the context of credit risk evaluation, survival analysis can be used to model the time until a borrower defaults on a loan, which is the event of interest.

By analyzing the time to default, survival analysis can provide insight into the probability of default over time and help lenders evaluate the credit risk of a borrower.
Survival analysis can provide a more nuanced understanding of credit risk compared to traditional credit scoring models, which typically use binary indicators of default .

Survival analysis can also be used to identify the factors that are associated with higher risk of default. By including various loan characteristics such as credit score, income, and loan amount in the survival model, one can determine which factors have a significant impact on the probability of default.

KAPLAN MEIER PROCEDURE

It is a statistical method for analyzing survival data. It shows the probability that a subject will survive up to time t. This test makes no assumptions about the underlying distribution of survival times and hence can be used to compare survival times of different groups or treatments, by comparing their respective survival curves using statistical tests such as the log-rank test or the Wilcoxon test.

The Kaplan-Meier curve can be used to estimate the survival function from data that are censored, truncated, or have missing values. Therefore it has widespread applications.

SURVIVAL ANALYSIS IN OUR DATASET

TABLE 3.1

P_ID	Months	Event	Group
1	1	Default	high interest credit risk
3	2	Default	high interest credit risk
4	3	Default	high interest credit risk
7	27	Default	high interest credit risk
18	30	Non-Default	high interest credit risk
29	15	Default	Low interest credit risk

- Event- (Default)- We have coded the event of defaulting as 1 and non-defaulting as 0.
- Group- The individuals are classified into 2 groups; high credit risk borrowers (interest rate greater than 20% & low credit risk borrowers (interest rate less than 20%)
- Months- Time after which the borrowers default since the start of study

SURVIVAL ANALYSIS RESULTS

TABLE 3.2

Means and Medians for Survival Time

Group	Mean[a]				Median			
			95% Confidence Interval				95% Confidence Interval	
	Estimate	Std. Error	Lower Bound	Upper Bound	Estimate	Std. Error	Lower Bound	Upper Bound
high interest credit risk	13.550	2.465	8.718	18.382	9.000	4.472	.235	17.765
Low interest credit risk	18.933	2.320	14.385	23.481	17.000	9.533	.000	35.684
Overall	16.228	1.744	12.810	19.646	14.000	3.795	6.562	21.438

a. Estimation is limited to the largest survival time if it is censored.

- The mean and median survival times for both the groups of credit risk borrowers are shown.
- Although there is a slight difference in the values of both the groups, we see that the survival time of low interest credit risk borrowers is higher than the high interest group.
- The median is usually chosen as a better statistic than the mean.

COMPARISON BETWEEN THE TWO GROUPS

H0: There is no significant difference in the survival times of high interest and low interest credit risk borrowers.

H1: There is significant difference in the survival times of high interest and low interest credit risk borrowers.

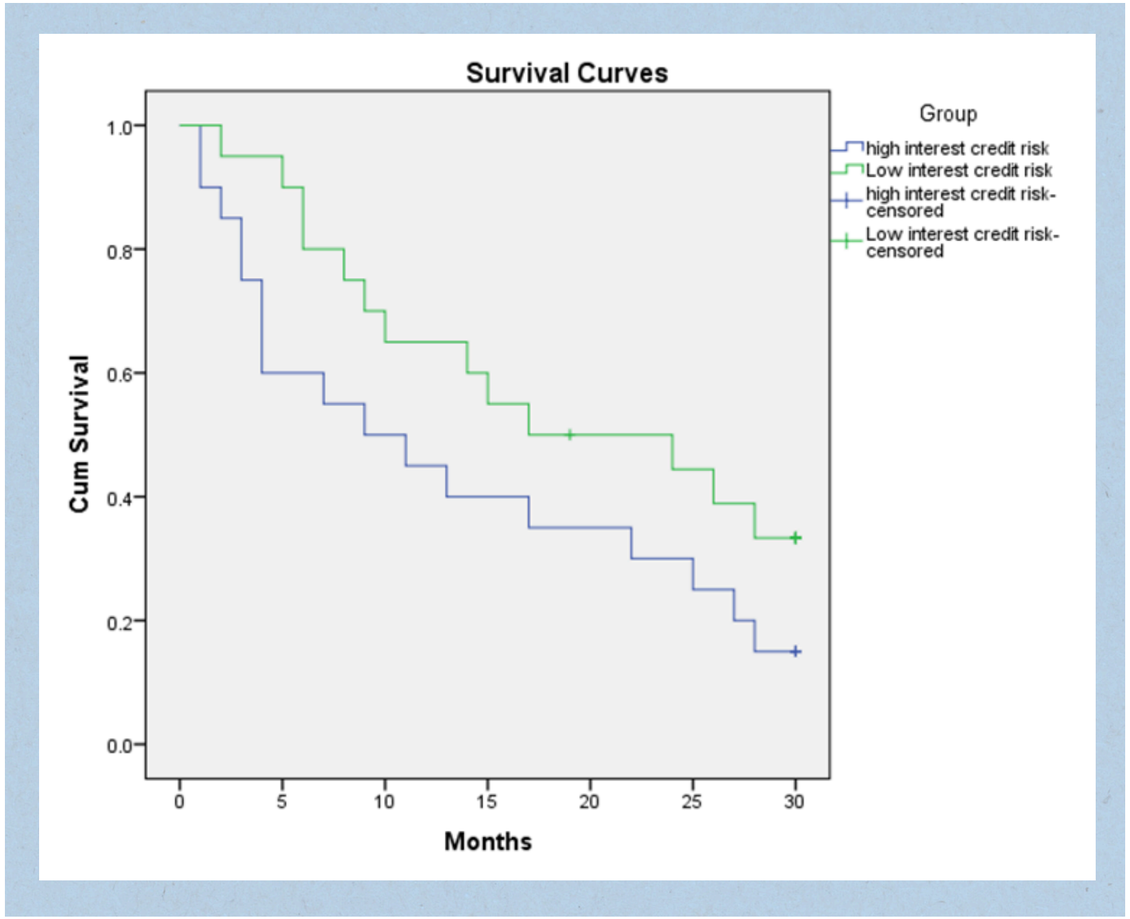
TABLE 3.3

Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	6.327	1	.013

Test of equality of survival distributions for the different levels of Group.

SURVIVAL CURVES FOR THE 2 GROUPS



From the two survival curves,

- The median survival time of high interest credit risk borrowers is 9 weeks.
- The median survival time of low interest credit risk borrowers is 17 weeks.

This indicates that high interest credit risk borrowers are prone to default at a faster rate than low interest credit risk borrowers.

AN INSTITUTION CAN TAKE SEVERAL MEASURES TO MINIMIZE ITS CREDIT RISK

- **Conduct a thorough credit analysis-**This should include analyzing the borrower's financial statements, credit history, income, and other factors that may affect their ability to repay the loan.
- **Use credit scoring models**: Credit scoring models use statistical techniques to predict the likelihood that a borrower will default on a loan.
- **Diversify the loan portfolio**: Institutions should diversify their loan portfolios across different types of loans and borrowers to reduce the impact of any single default.
- **Set appropriate loan terms**: Institutions should set appropriate loan terms, including interest rates, repayment periods, and collateral requirements, based on the borrower's credit risk profile

RESOURCE PAGE

Books:

An Introduction to Multivariate Statistical Analysis" Third edition by T. W. Anderson .

Publications:

https://www.slideshare.net/harjindal/survival-analysis-30664604

Youtube links:

https://www.youtube.com/watch?v=tT1kJhQS2Dk&t=16s

https://www.youtube.com/watch?v=7zYcMZ-61c4&t=1132s

https://www.youtube.com/watch?v=Tw1WVxiXHsk&ab_channel=Dr.ToddGrande

THANK YOU