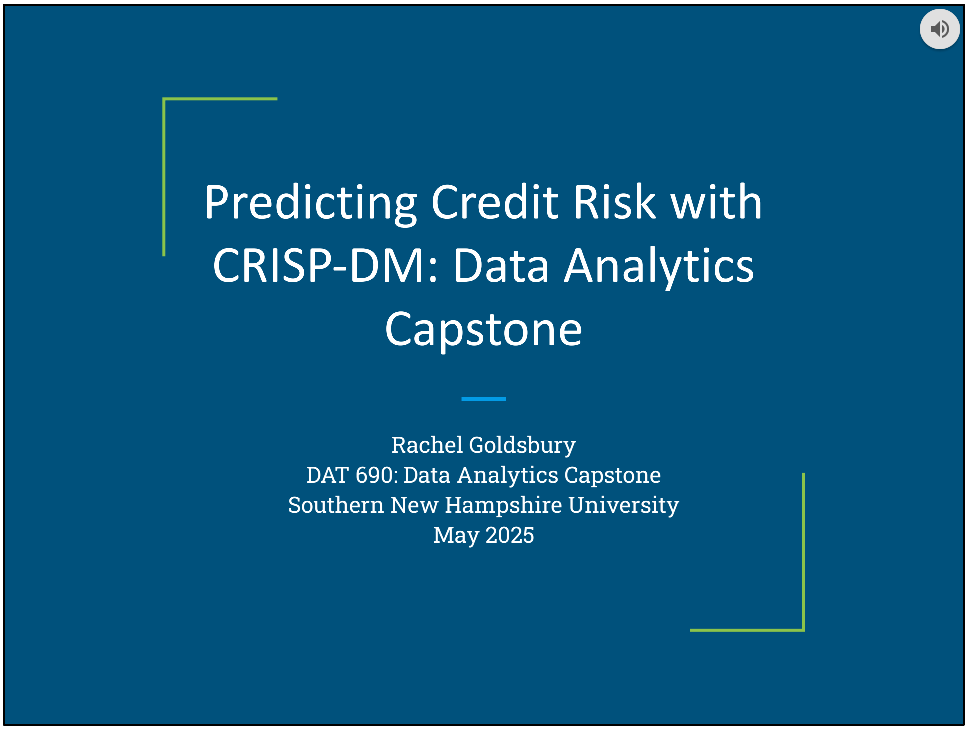


Predicting Credit Risk with CRISP-DM: Data Analytics Capstone

Rachel Goldsbury
DAT 690: Data Analytics Capstone
Southern New Hampshire University
May 2025

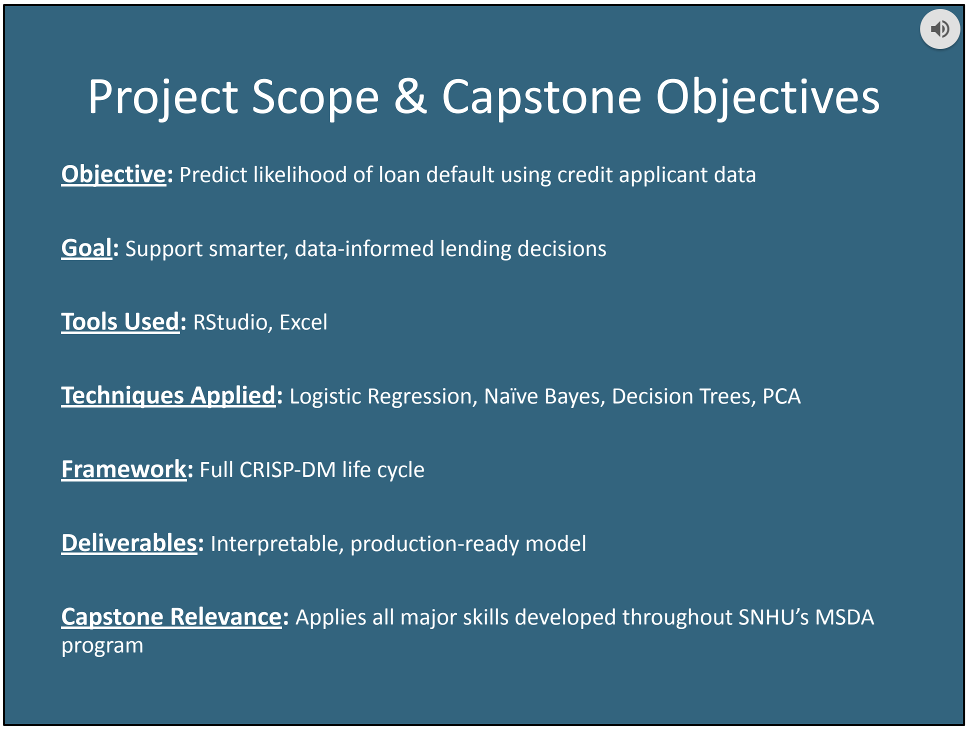
“Hello, and thank you for listening today.

My name is Rachel Goldsbury, and I’m excited to present my final capstone project.

This project is titled *‘Predicting Credit Risk with CRISP-DM’*, and it showcases how I applied the full CRISP-DM framework to develop a predictive model aimed at assessing loan default risk.

Over the course of this presentation, I’ll walk you through each stage of the data science lifecycle, from business understanding through deployment, using a real-world example of data and showcasing many techniques I’ve developed throughout this program.

With that, Let’s dive in.”



Project Scope & Capstone Objectives

Objective: Predict likelihood of loan default using credit applicant data

Goal: Support smarter, data-informed lending decisions

Tools Used: RStudio, Excel

Techniques Applied: Logistic Regression, Naïve Bayes, Decision Trees, PCA

Framework: Full CRISP-DM life cycle

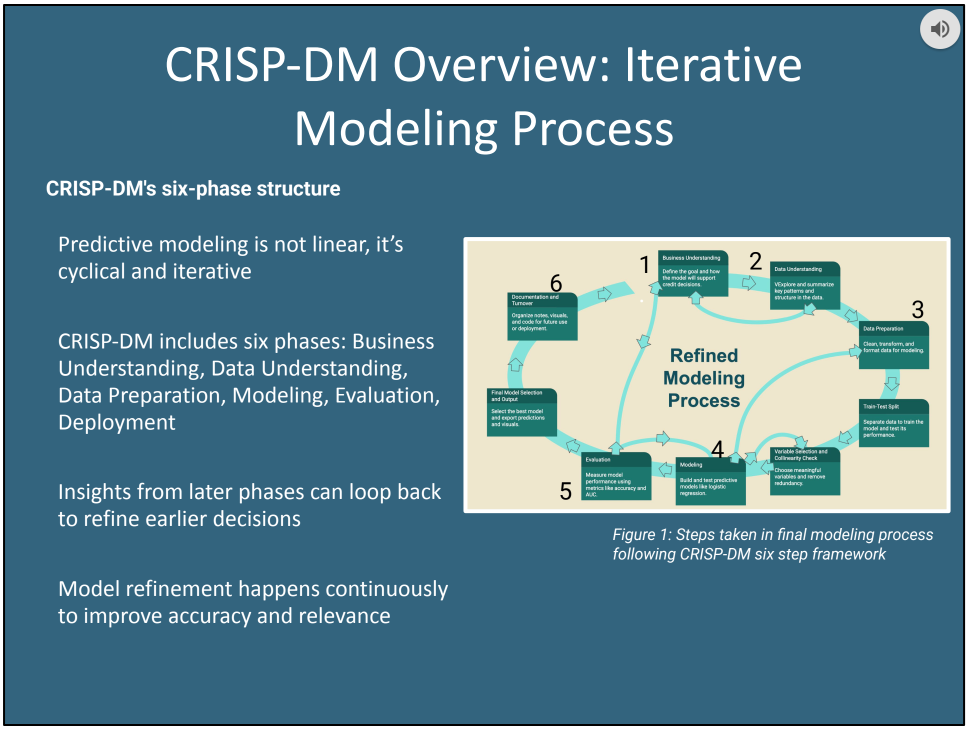
Deliverables: Interpretable, production-ready model

Capstone Relevance: Applies all major skills developed throughout SNHU's MSDA program

"This project was designed to the business challenge of predicting whether or not an applicant is likely to default on a loan. This is especially important for lenders who need to reduce risk and make more informed decisions.

I used RStudio and Excel to work with the dataset and apply modeling techniques like logistic regression, Naïve Bayes, decision trees, and PCA. The project follows the full CRISP-DM framework from business understanding to deployment, and the final deliverable is a model that is both interpretable and scalable.

This capstone has allowed me to put into practice everything I've learned in this program from data cleaning to evaluation and communication."



CRISP-DM Overview: Iterative Modeling Process

CRISP-DM's six-phase structure

Predictive modeling is not linear, it's cyclical and iterative

CRISP-DM includes six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, Deployment

Insights from later phases can loop back to refine earlier decisions

Model refinement happens continuously to improve accuracy and relevance

1 Business Understanding
Define the goal and how the model will support credit decisions.

2 Data Understanding
VExplore and summarize key patterns and structure in the data.

3 Data Preparation
Clean, transform, and format data for modeling.

Train-Test Split
Separate data to train the model and test its performance.

Variable Selection and Collinearity Check
Choose meaningful variables and remove redundancy.

4 Modeling
Build and test predictive models like logistic regression.

5 Evaluation
Measure model performance using metrics like accuracy and AUC.

Final Model Selection and Output
Select the best model and export predictions and visuals.

6 Documentation and Turnover
Organize notes, visuals, and code for future use or deployment.

Refined Modeling Process

Figure 1: Steps taken in final modeling process following CRISP-DM six step framework

"Before we dive into the details of each phase, I want to give a quick overview of the process I followed.

This project uses the CRISP-DM framework, short for *Cross-Industry Standard Process for Data Mining*. It's a six-phase, iterative process widely used in data analytics to guide projects from start to finish.

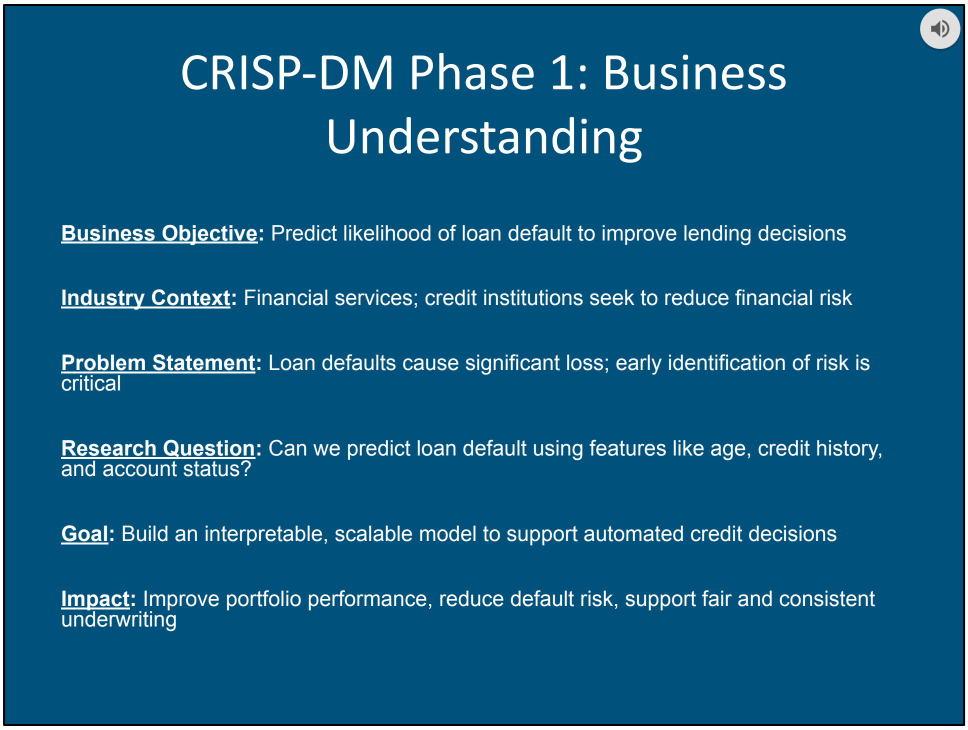
What's important to remember is that CRISP-DM isn't linear, it's designed to be flexible and cyclical. As you progress through a phase, you often uncover insights that send you back to revisit earlier steps, whether to clean data differently, redefine the business goal, or reselect features.

I followed this iterative structure closely in this project clear steps, in practice, many of the phases overlapped or looped back on each other. For example, business understanding and data understanding flowed together, linked closely, and modeling was deeply tied to evaluation.

The flowchart shown above is a closer representation to the actual progression. I built in time to revisit decisions, assess model performance, and ensure everything stayed aligned with the original business need. That flexibility made it possible to create a model that’s not only accurate, but also practical, explainable, and production ready.”

CRISP-DM Phase 1: Business Understanding

“Now that we have a good grasp on the flow of the project, Let's get into the business understanding phase.”



CRISP-DM Phase 1: Business Understanding

Business Objective: Predict likelihood of loan default to improve lending decisions

Industry Context: Financial services; credit institutions seek to reduce financial risk

Problem Statement: Loan defaults cause significant loss; early identification of risk is critical

Research Question: Can we predict loan default using features like age, credit history, and account status?

Goal: Build an interpretable, scalable model to support automated credit decisions

Impact: Improve portfolio performance, reduce default risk, support fair and consistent underwriting

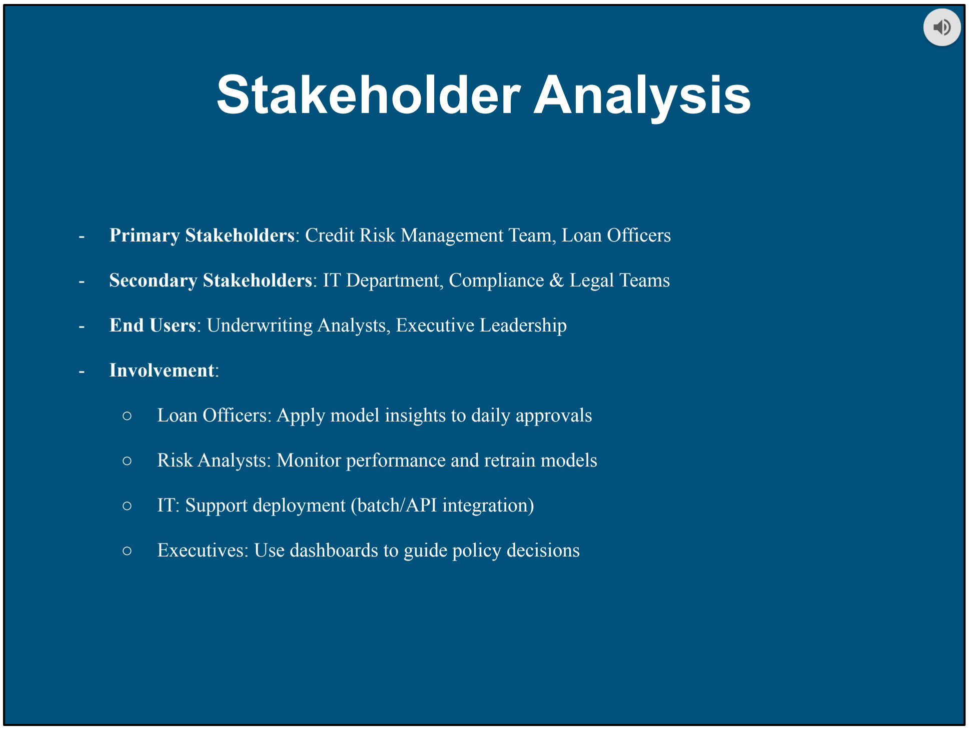
“In the first phase of the CRISP-DM framework, the focus is on understanding the business problem and aligning the data science goals with the organization’s strategic objectives.

For this project, I focused on the credit risk challenge. Loan defaults present both financial and reputational risks to lenders, so being able to predict whether an applicant is likely to default is extremely valuable.

The core research question I sought to answer was: *Can we accurately predict the likelihood of loan default using features like age, credit history, loan amount, and account type?*

My goal was to develop a predictive model that’s not only accurate but also interpretable, so stakeholders, like risk

Ultimately, this model is designed to support smarter, data-backed lending decisions that reduce losses, improve portfolio health, and promote fair access to credit.”



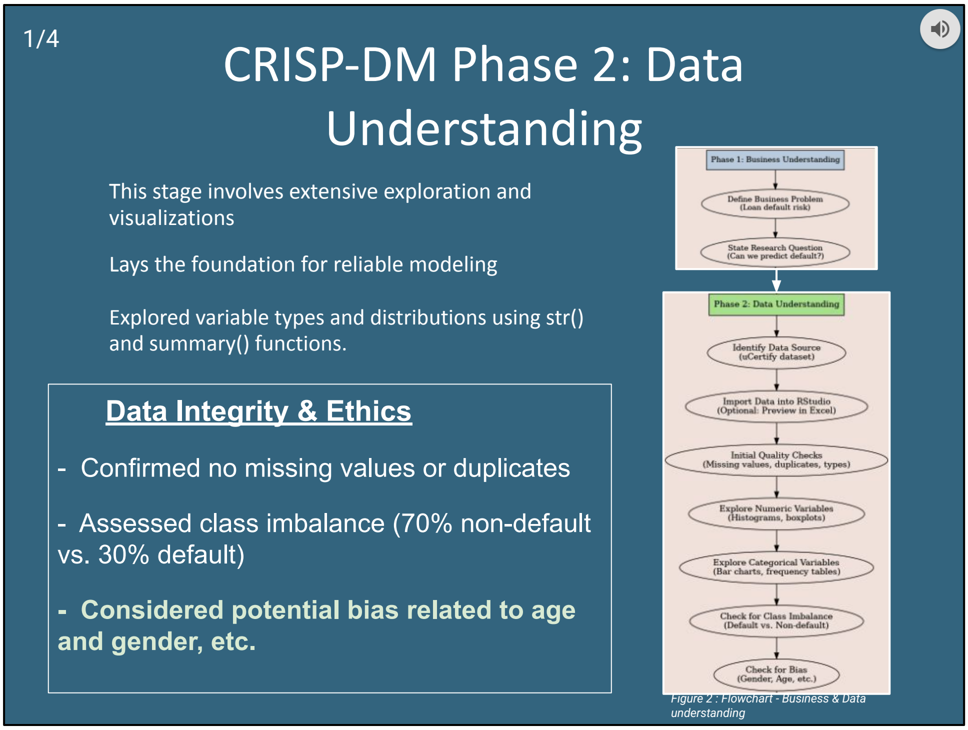
Stakeholder Analysis

- **Primary Stakeholders**: Credit Risk Management Team, Loan Officers
- **Secondary Stakeholders**: IT Department, Compliance & Legal Teams
- **End Users**: Underwriting Analysts, Executive Leadership
- **Involvement**:
  - ○ Loan Officers: Apply model insights to daily approvals
  - ○ Risk Analysts: Monitor performance and retrain models
  - ○ IT: Support deployment (batch/API integration)
  - ○ Executives: Use dashboards to guide policy decisions

This slide outlines the key stakeholders impacted by the predictive model. Primary stakeholders like credit risk analysts and loan officers rely directly on the outputs for decision-making. Secondary stakeholders such as IT and compliance ensure the model integrates with systems and aligns with regulations. Understanding how each group interacts with the model helped shape decisions from data privacy to deployment strategy.

CRISP-DM Phase 2: Data Understanding

“Now that we’ve defined the business challenge, let’s explore the data behind it.”



1/4

CRISP-DM Phase 2: Data Understanding

This stage involves extensive exploration and visualizations

Lays the foundation for reliable modeling

Explored variable types and distributions using str() and summary() functions.

Data Integrity & Ethics

- Confirmed no missing values or duplicates
- Assessed class imbalance (70% non-default vs. 30% default)
- **Considered potential bias related to age and gender, etc.**

Phase 1: Business Understanding
Define Business Problem (Loan default risk)
State Research Question (Can we predict default?)
Phase 2: Data Understanding
Identify Data Source (uCertify dataset)
Import Data into RStudio (Optional: Preview in Excel)
Initial Quality Checks (Missing values, duplicates, types)
Explore Numeric Variables (Histograms, boxplots)
Explore Categorical Variables (Bar charts, frequency tables)
Check for Class Imbalance (Default vs. Non-default)
Check for Bias (Gender, Age, etc.)

Figure 2 : Flowchart - Business & Data understanding

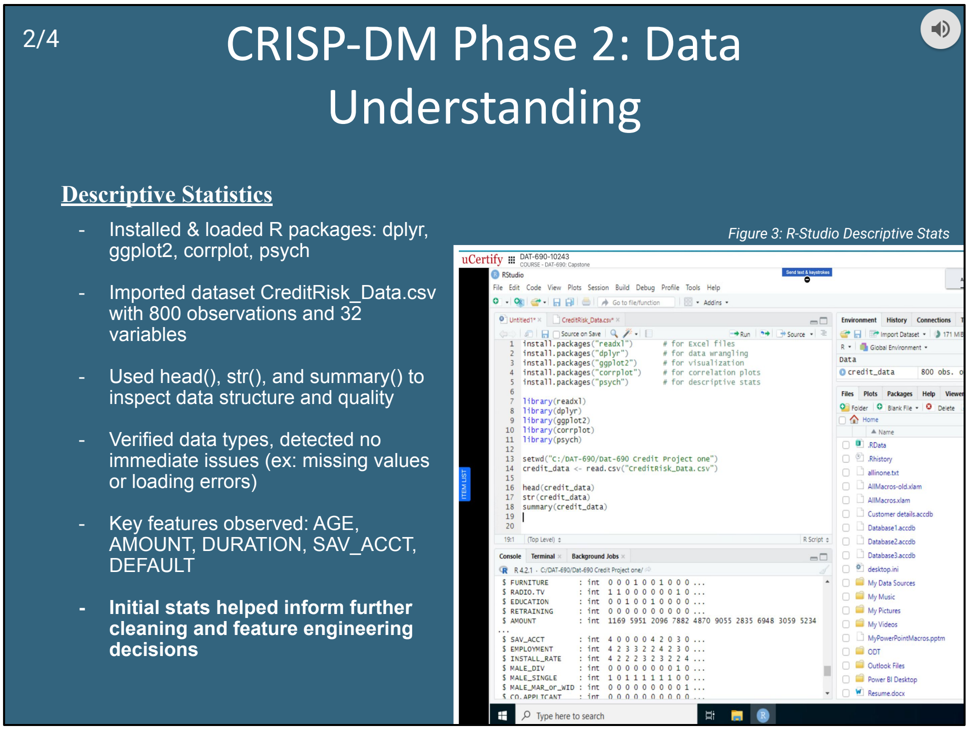
“In Phase 2 of CRISP-DM, I focused on understanding the structure and content of the credit risk dataset. This is a foundational step that directly impacts the reliability of downstream modeling.

I began by checking for basic data integrity issues, such as missing values or duplicate records and confirmed that the dataset was clean.

I also assessed class imbalance. About 70% of applicants in this dataset were classified as non-default, while only 30% were defaults. This imbalance could impact prediction performance and required careful consideration.

From an ethical standpoint, I paid attention to variables that might introduce bias, such as age or gender indicators. Ensuring fairness is key, especially in financial

You can see a snapshot of the process in the flowchart on the right. This guided my early data exploration and quality assurance checks before moving into deeper statistical analysis and visualization.”



2/4

CRISP-DM Phase 2: Data Understanding

Descriptive Statistics

- Installed & loaded R packages: dplyr, ggplot2, corrplot, psych
- Imported dataset CreditRisk_Data.csv with 800 observations and 32 variables
- Used head(), str(), and summary() to inspect data structure and quality
- Verified data types, detected no immediate issues (ex: missing values or loading errors)
- Key features observed: AGE, AMOUNT, DURATION, SAV_ACCT, DEFAULT
- **Initial stats helped inform further cleaning and feature engineering decisions**

Figure 3: R-Studio Descriptive Stats

```
install.packages("readxl")      # for Excel files
install.packages("dplyr")       # for data wrangling
install.packages("ggplot2")     # for visualization
install.packages("corrplot")    # for correlation plots
install.packages("psych")       # for descriptive stats

library(readxl)
library(dplyr)
library(ggplot2)
library(corrplot)
library(psych)

setwd("C:/DAT-690/Dat-690 Credit Project one")
credit_data <- read.csv("CreditRisk_Data.csv")

head(credit_data)
str(credit_data)
summary(credit_data)

```

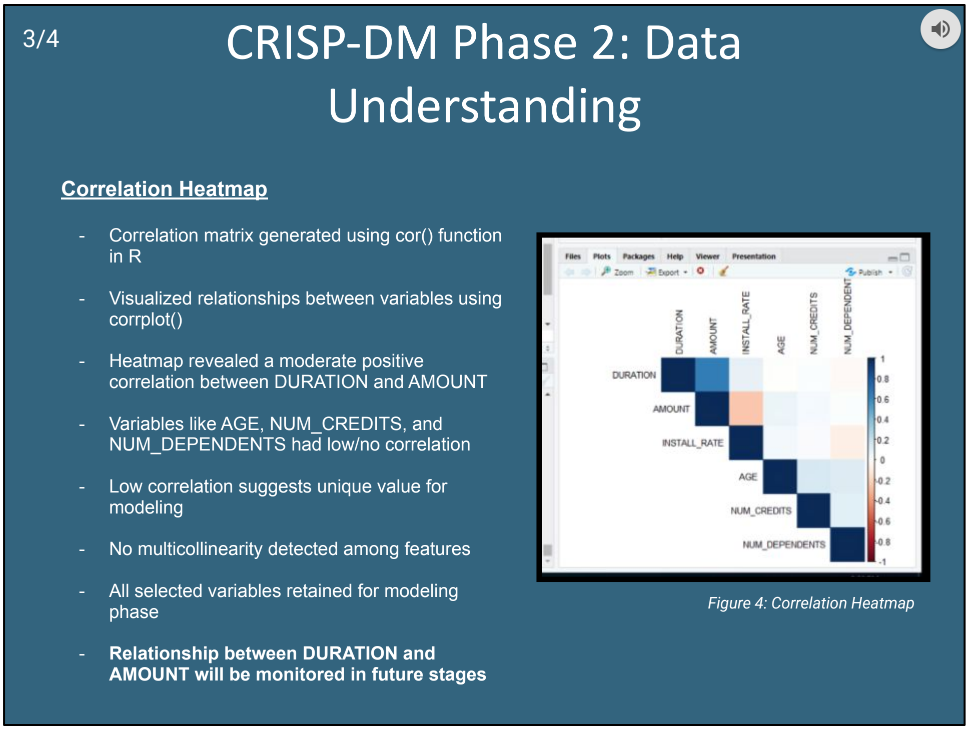
```
 $ FURNITURE       : int  0 0 0 1 0 0 1 0 0 0 ...
 $ RADIO.TV        : int  1 1 0 0 0 0 0 0 1 0 ...
 $ EDUCATION       : int  0 0 1 0 0 1 0 0 0 0 ...
 $ RETRAINING      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ AMOUNT          : int  1169 5951 2096 7882 4870 9055 2835 6948 3059 5234
...
 $ SAV_ACCT        : int  4 0 0 0 0 4 2 0 3 0 ...
 $ EMPLOYMENT      : int  4 2 3 3 2 2 4 2 3 0 ...
 $ INSTALL_RATE    : int  4 2 2 2 3 2 3 2 2 4 ...
 $ MALE_DIV        : int  0 0 0 0 0 0 0 0 1 0 ...
 $ MALE_SINGLE     : int  1 0 1 1 1 1 1 1 0 0 ...
 $ MALE_MAR_or_WID : int  0 0 0 0 0 0 0 0 0 1 ...
 $ CO.APPLICANT    : int  0 0 0 0 0 0 0 0 0 0 ...
```

In this second part of the Data Understanding phase, I focused on getting familiar with the dataset through direct exploration in R.

I used basic R functions such as `head()`, `str()`, and `summary()`, to inspect the structure and values of the dataset. This helped me confirm that the data loaded properly and gave me a quick look at the types and ranges of variables I'd be working with.

Descriptive statistics were generated for key numeric features to understand distributions, check for outliers, and verify the scale of values across fields like `AGE`, `AMOUNT`, and `DURATION`.

also allowed me to spot anything unexpected that might need to be addressed before preprocessing.


Files
Plots
Packages
Help
Viewer
Presentation
Zoom
Export
Publish
DURATION
AMOUNT
INSTALL_RATE
AGE
NUM_CREDITS
NUM_DEPENDENTS
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

3/4

CRISP-DM Phase 2: Data Understanding

Correlation Heatmap

- Correlation matrix generated using cor() function in R
- Visualized relationships between variables using corrplot()
- Heatmap revealed a moderate positive correlation between DURATION and AMOUNT
- Variables like AGE, NUM_CREDITS, and NUM_DEPENDENTS had low/no correlation
- Low correlation suggests unique value for modeling
- No multicollinearity detected among features
- All selected variables retained for modeling phase
- **Relationship between DURATION and AMOUNT will be monitored in future stages**

Figure 4: Correlation Heatmap

“In this stage of data understanding, we’re examining how variables relate to each other using a correlation heatmap.

To begin, I generated a correlation matrix using the `cor()` function in R. This allowed me to calculate the strength and direction of linear relationships between numeric variables in the dataset.

Then, I visualized this matrix using the `corrplot()` function, which produced the heatmap shown here on the right in Figure 4.

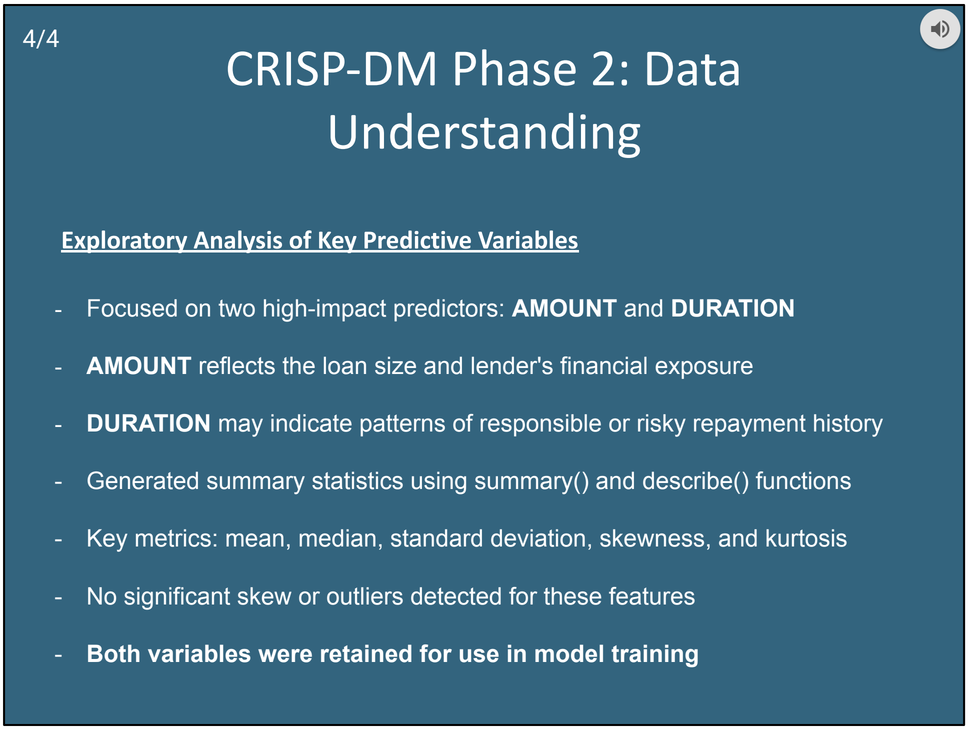
The heatmap revealed a moderate positive correlation between DURATION and AMOUNT, meaning that longer loan terms tend to be associated with higher credit amounts. This makes intuitive sense and aligns with common lending practices.

NUM_DEPENDENTS showed little to no correlation with each other or with the primary variables, suggesting they may bring unique value to the modeling process.

Crucially, no multicollinearity was detected, which means we don’t have variables that are overly redundant or strongly correlated. This helps maintain the integrity of the model’s predictions and improves interpretability.

At this stage, I chose to retain all variables for the modeling phase, but I made a note to keep an eye on the relationship between DURATION and AMOUNT in case it influences model performance down the line.

This analysis confirms that the dataset is both diverse and structurally sound, laying a solid foundation as we move into data preparation and modeling.”



4/4

CRISP-DM Phase 2: Data Understanding

Exploratory Analysis of Key Predictive Variables

- Focused on two high-impact predictors: **AMOUNT** and **DURATION**
- **AMOUNT** reflects the loan size and lender's financial exposure
- **DURATION** may indicate patterns of responsible or risky repayment history
- Generated summary statistics using summary() and describe() functions
- Key metrics: mean, median, standard deviation, skewness, and kurtosis
- No significant skew or outliers detected for these features
- **Both variables were retained for use in model training**

“In this final slide of the Data Understanding phase, I zoomed in on two specific variables, AMOUNT and DURATION, because of their high relevance to credit risk. AMOUNT directly relates to loan size and the level of exposure the lender takes on, while DURATION may signal underlying trends in repayment behavior, especially when considered alongside credit defaults.

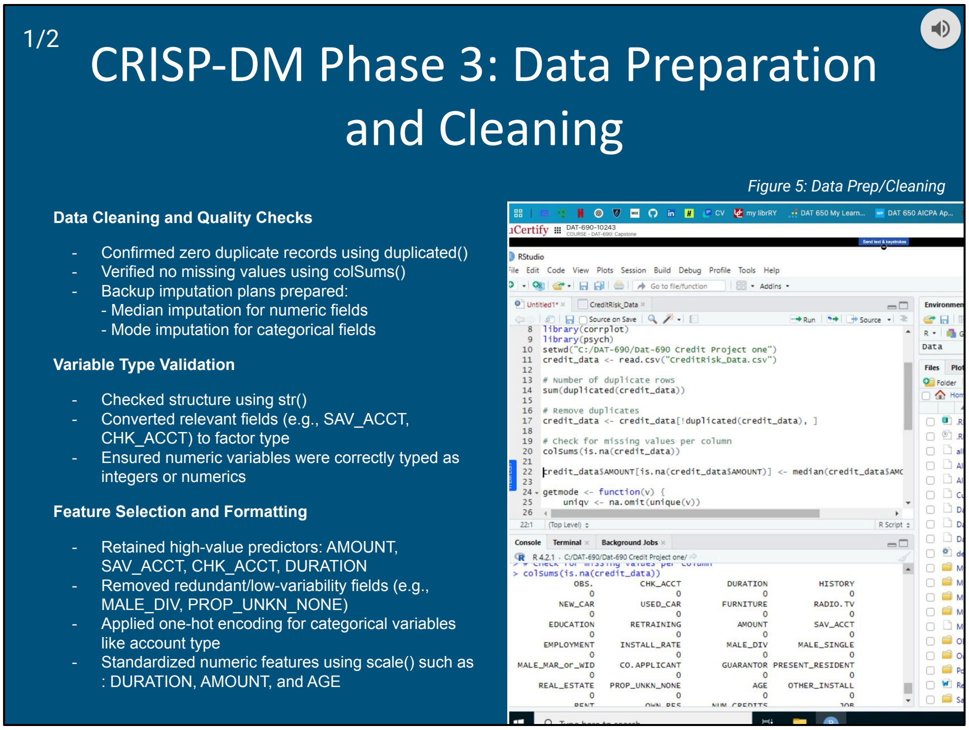
I used both `summary()` and `describe()` functions in R to generate essential descriptive statistics for these variables. We examined central tendency and spread using metrics like mean and standard deviation, and also checked skewness and kurtosis to assess symmetry and shape of distributions.

The good news: no major outliers or skewness concerns came up during this step, which confirms that these features are clean, well-behaved, and reliable for modeling. So both variables were retained and moved forward to the modeling phase.

This kind of detailed exploratory work lays the groundwork for strong, trustworthy models and ensures our inputs are both meaningful and statistically sound.”

CRISP-DM Phase 3: Data Preparation

“With a solid grasp of the data structure, we’re ready to clean and transform it.”



1/2

CRISP-DM Phase 3: Data Preparation and Cleaning

Data Cleaning and Quality Checks

- Confirmed zero duplicate records using duplicated()
- Verified no missing values using colSums()
- Backup imputation plans prepared:
  - Median imputation for numeric fields
  - Mode imputation for categorical fields

Variable Type Validation

- Checked structure using str()
- Converted relevant fields (e.g., SAV_ACCT, CHK_ACCT) to factor type
- Ensured numeric variables were correctly typed as integers or numerics

Feature Selection and Formatting

- Retained high-value predictors: AMOUNT, SAV_ACCT, CHK_ACCT, DURATION
- Removed redundant/low-variability fields (e.g., MALE_DIV, PROP_UNKN_NONE)
- Applied one-hot encoding for categorical variables like account type
- Standardized numeric features using scale() such as : DURATION, AMOUNT, and AGE

Figure 5: Data Prep/Cleaning

```
library(corrplot)
library(psych)
setwd("C:/DAT-690/Dat-690 Credit Project one")
credit_data <- read.csv("CreditRisk_Data.csv")

# Number of duplicate rows
sum(duplicated(credit_data))

# Remove duplicates
credit_data <- credit_data[!duplicated(credit_data), ]

# Check for missing values per column
colSums(is.na(credit_data))

credit_data$AMOUNT[is.na(credit_data$AMOUNT)] <- median(credit_data$AMC

getmode <- function(v) {
    uniqv <- na.omit(unique(v))
```

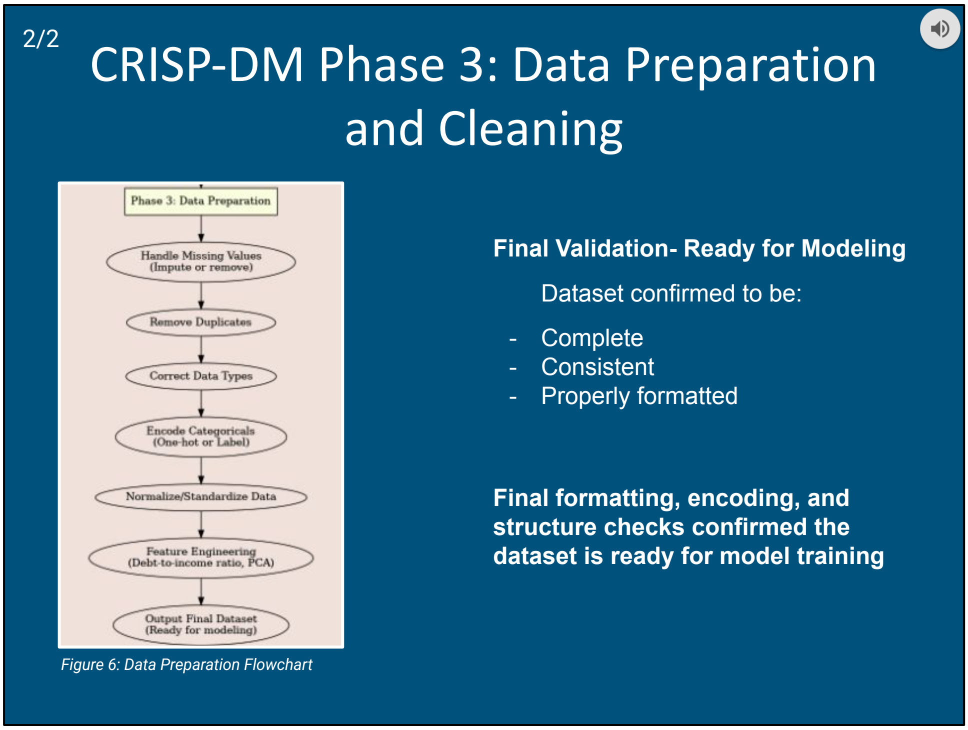
```
> # Check for missing values per column
> colSums(is.na(credit_data))
           OBS.         CHK_ACCT         DURATION          HISTORY
              0                0                0                0
        NEW_CAR         USED_CAR        FURNITURE         RADIO.TV
              0                0                0                0
      EDUCATION       RETRAINING           AMOUNT         SAV_ACCT
              0                0                0                0
     EMPLOYMENT     INSTALL_RATE         MALE_DIV      MALE_SINGLE
              0                0                0                0
MALE_MAR_or_WID     CO.APPLICANT        GUARANTOR PRESENT_RESIDENT
              0                0                0                0
    REAL_ESTATE   PROP_UNKN_NONE              AGE    OTHER_INSTALL
              0                0                0                0
           RENT          OWN_RES      NUM_CREDITS              JOB
```

In this first part of the Data Preparation phase, I focused on making sure the dataset was clean, consistent, and ready for modeling. I started by checking for quality issues, specifically, I used the duplicated() function to confirm there were no duplicate records, and colSums() to check for missing values. Even though I didn’t find any, I still created a backup plan just in case. For example, I would have used median imputation for numeric fields and mode imputation for categorical ones.

Next, I validated the structure and data types. I ran str() to review each variable’s format, and manually converted key categorical fields like SAV_ACCT and CHK_ACCT into factors so they’d work properly with R modeling functions.

For feature selection, I kept the most meaningful predictors from earlier analysis, things like loan amount, account status, and duration. I removed redundant or low-value fields, as well as sparsely populated flags. Categorical variables were one-hot encoded, and numeric ones were standardized using the scale() function to account for magnitude differences across features.

This step helped set a solid foundation for modeling by making sure the data was clean, interpretable, and technically ready for the next phase.



2/2

CRISP-DM Phase 3: Data Preparation and Cleaning

Phase 3: Data Preparation
Handle Missing Values (Impute or remove)
Remove Duplicates
Correct Data Types
Encode Categoricals (One-hot or Label)
Normalize/Standardize Data
Feature Engineering (Debt-to-income ratio, PCA)
Output Final Dataset (Ready for modeling)

Figure 6: Data Preparation Flowchart

Final Validation- Ready for Modeling

Dataset confirmed to be:

- Complete
- Consistent
- Properly formatted

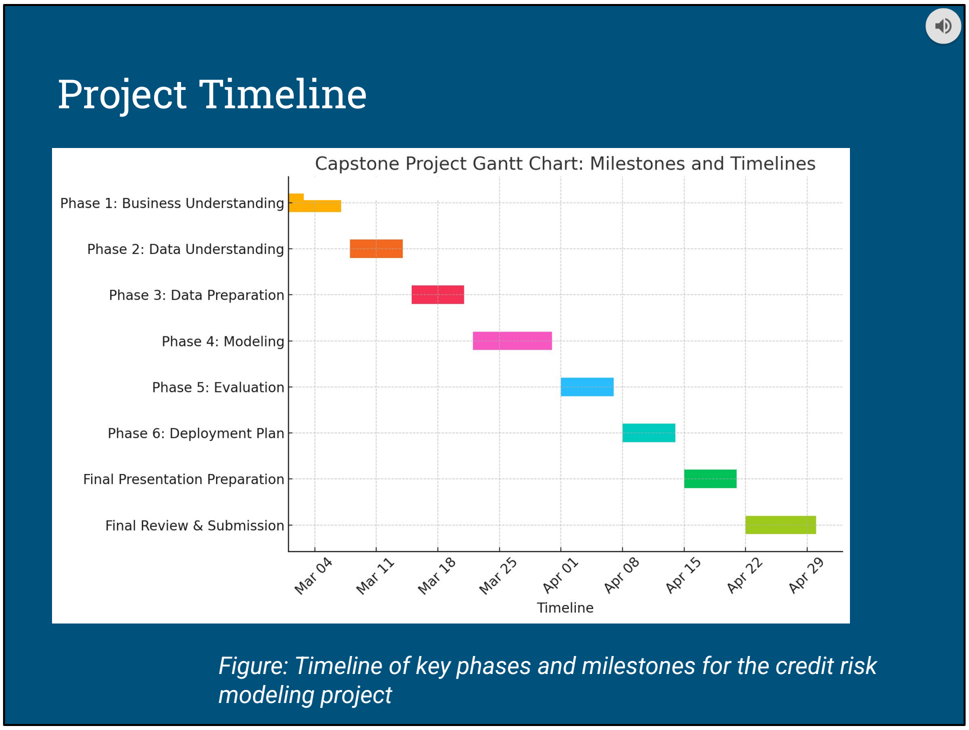
Final formatting, encoding, and structure checks confirmed the dataset is ready for model training

“This slide wraps up the data preparation phase by summarizing the final validation outcomes. At this point, I confirmed the dataset was fully ready for modeling, meaning it was complete, consistent, and properly formatted.

All data cleaning steps were successfully completed: there were no missing values or duplicates, and each variable was assigned the correct data type. Categorical fields were properly encoded using one-hot encoding, and numeric variables were standardized using the `scale()` function.

I also finalized feature selection, ensuring the dataset retained only the most predictive, high-quality variables. The flowchart on the left visually represents the entire preparation process, from handling missing values through to outputting a model-ready dataset

Completing this stage was a crucial milestone, setting a solid foundation for modeling by ensuring that data integrity, consistency, and structure were all in place.”



Project Timeline

Figure: Timeline of key phases and milestones for the credit risk modeling project

This Gantt chart outlines the full timeline and major milestones for the project, from business understanding to deployment planning.

I structured the project using the CRISP-DM framework, and as you can see, some phases, like data understanding and data preparation, overlapped, which is typical in iterative analytics work.

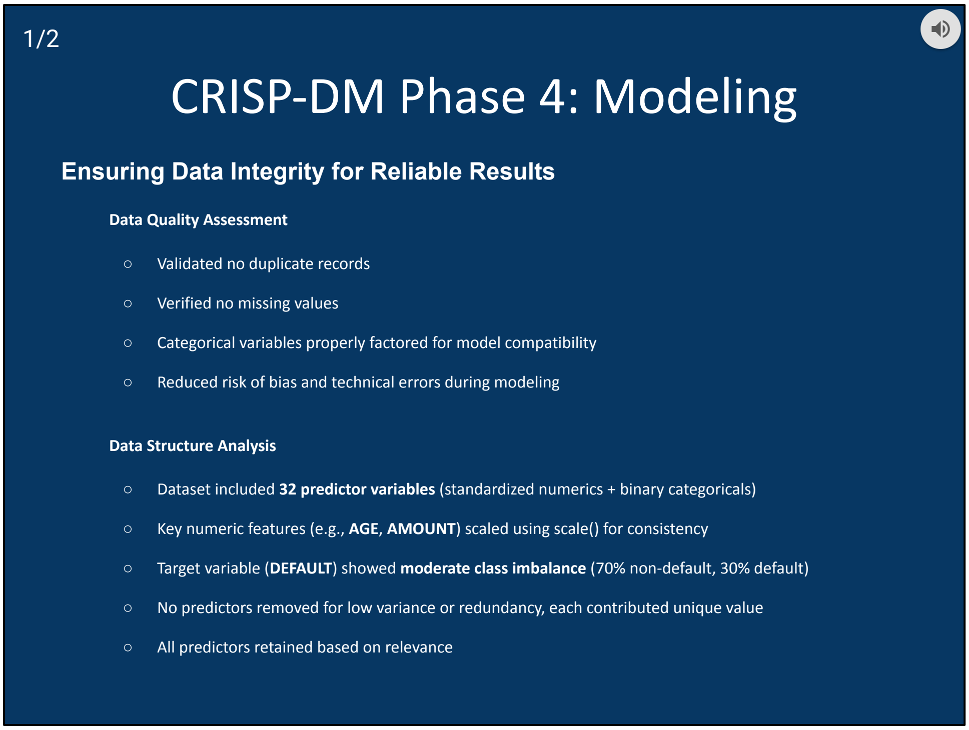
The modeling and evaluation stages were particularly dynamic, with back-and-forth adjustments based on performance metrics.

This visual highlights how each task built on the previous one, culminating in a deployable logistic regression model and documented recommendations for future use.

It also emphasizes how project planning and time management were critical in ensuring successful delivery within the course timeline.

CRISP-DM Phase 4: Modeling

“Now that the dataset is ready, we can move into modeling and test its predictive power”



1/2

CRISP-DM Phase 4: Modeling

Ensuring Data Integrity for Reliable Results

Data Quality Assessment

- Validated no duplicate records
- Verified no missing values
- Categorical variables properly factored for model compatibility
- Reduced risk of bias and technical errors during modeling

Data Structure Analysis

- Dataset included **32 predictor variables** (standardized numerics + binary categoricals)
- Key numeric features (e.g., **AGE**, **AMOUNT**) scaled using scale() for consistency
- Target variable (**DEFAULT**) showed **moderate class imbalance** (70% non-default, 30% default)
- No predictors removed for low variance or redundancy, each contributed unique value
- All predictors retained based on relevance

“In Phase 4, we began the modeling process by first ensuring the dataset's quality and structure were suitable for analysis.

We validated the dataset for duplicates and missing values, and verified that categorical variables were properly factored, which reduced the risk of bias or technical issues during modeling.

From a structure standpoint, we retained all 32 predictors, these included both standardized numeric variables and binary categorical indicators.

We also scaled key numeric fields such as AGE and AMOUNT using the `scale()` function in R for consistency.

common challenge in credit data and something we kept in mind for evaluation.

Fortunately, all features were unique and meaningful enough to retain, so no variables were dropped at this stage.”

2/2

CRISP-DM Phase 4: Modeling

Strategy final pick: Logistic Regression for Clarity and Performance

Model Summary

- **Model Type**: Logistic regression using `glm()` function with a binomial family
- **Training Dataset**: Fit model on 70% training split
- **Purpose**: Predict likelihood of loan default using applicant credit features

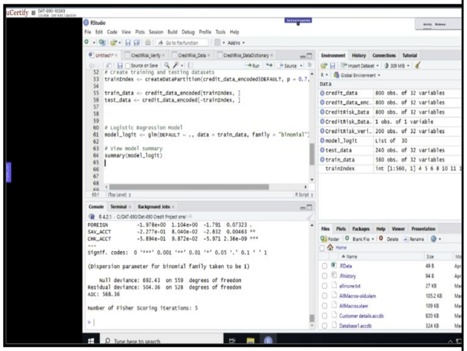
Key Outputs (See Figure 7)

- Coefficients for predictors (e.g., `SAV_ACCT`, `CHK_ACCT`)
- Standard errors, z-values, and p-values for significance testing
- Model fit statistics:
  - Null deviance & residual deviance
  - **AIC score**: Indicates model efficiency
  - Number of iterations for convergence

Interpretation

- **Helps determine which features significantly influence loan default risk**
- **Output supports feature importance evaluation and model refinement**

Figure 7: Model summary coefficients and p-values.



RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Untitled1*
CreditRisk_Verify
CreditRisk_Data
CreditRisk_DataDictionary
Source on Save
Run
Source
Environment History Connections Tutorial
Import Dataset
309 MiB
Global Environment
Console Terminal Background Jobs
Files Plots Packages Help Viewer Presentation
Folder Blank File Delete Rename
Home
Name Size
.RData 49 B
.Rhistory 94 B
allinone.txt 27 KB
Type here to search

```
# Create training and testing datasets
trainIndex <- createDataPartition(credit_data_encoded$DEFAULT, p = 0.7,

train_data <- credit_data_encoded[trainIndex, ]
test_data <- credit_data_encoded[-trainIndex, ]



# Logistic Regression Model
model_logit <- glm(DEFAULT ~ ., data = train_data, family = "binomial")

# View model summary
summary(model_logit)
```

```
Data
credit_data           800 obs. of 32 variables
credit_data_enc…      800 obs. of 32 variables
CreditRisk_Data       800 obs. of 32 variables
CreditRisk_Data…      1 obs. of 1 variable
CreditRisk_Veri…      200 obs. of 32 variables
model_logit           List of  30
test_data             240 obs. of 32 variables
train_data            560 obs. of 32 variables
trainIndex            int [1:560, 1] 4 5 6 8 10 11 1
```

```
R 4.2.1 · C:/DAT-690/Dat-690 Credit Project one/
FOREIGN         -1.978e+00  1.104e+00  -1.791  0.07323 .
SAV_ACCT        -2.277e-01  8.040e-02  -2.832  0.00463 **
CHK_ACCT        -5.894e-01  9.872e-02  -5.971 2.36e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 692.43  on 559  degrees of freedom
Residual deviance: 504.36  on 528  degrees of freedom
AIC: 568.36

Number of Fisher Scoring iterations: 5

>
```

“After preparing and validating the data, we moved on to model selection and testing.

I tested multiple models, including decision trees and Naïve Bayes, but ultimately selected **logistic regression** because it offers strong interpretability, something that’s really important for stakeholder trust and explainability.

The model was trained using a 70% training split and fit with the `glm()` function in R using the binomial family.

As shown in the summary (Figure 7), we examined key outputs including predictor coefficients, standard errors, and p-values.

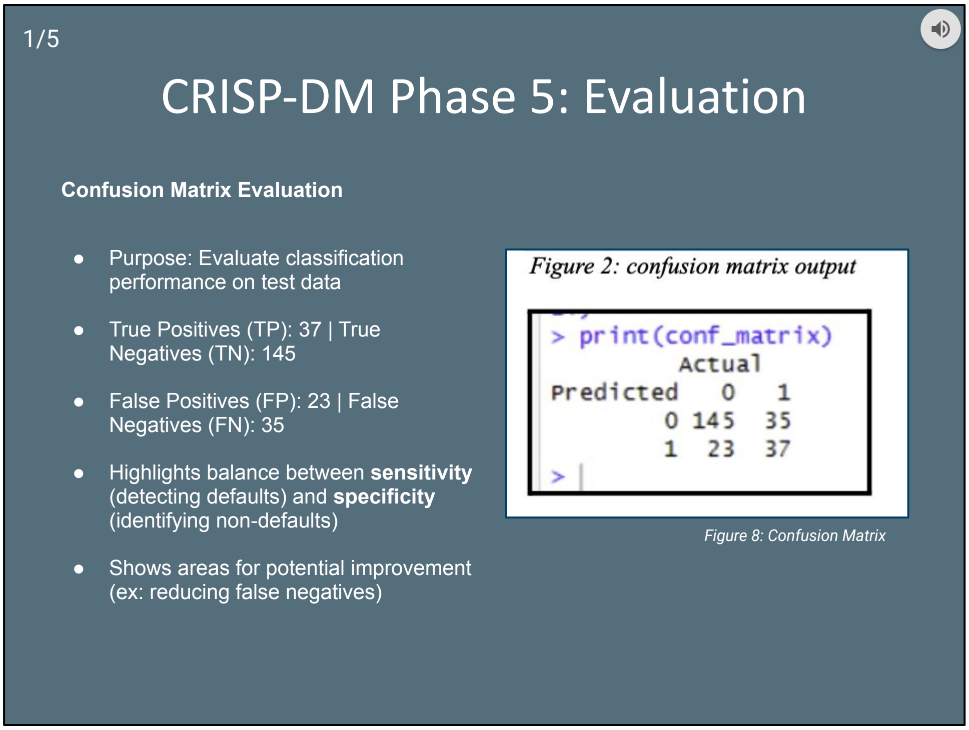
The model also provided deviance values and an AIC of 568.36, which helped evaluate model fit and complexity.

future iterations.

Overall, the model performed well in training, and we were able to move forward to formal evaluation with strong confidence in its structure and clarity.”

CRISP-DM Phase 5: Evaluation

“With a model in place, we’ll evaluate how well it performs on unseen data.”



1/5

CRISP-DM Phase 5: Evaluation

Confusion Matrix Evaluation

- Purpose: Evaluate classification performance on test data
- True Positives (TP): 37 | True Negatives (TN): 145
- False Positives (FP): 23 | False Negatives (FN): 35
- Highlights balance between **sensitivity** (detecting defaults) and **specificity** (identifying non-defaults)
- Shows areas for potential improvement (ex: reducing false negatives)

Figure 2: confusion matrix output

```
> print(conf_matrix)
         Actual
Predicted   0   1
        0 145  35
        1  23  37
> |
```

Figure 8: Confusion Matrix

On this slide, I introduce the confusion matrix, which gives us a detailed snapshot of how well the model performs on unseen test data.

We can see the model identified 37 true defaults and 145 true non-defaults. However, we had 23 false positives and 35 false negatives.

These results show that while the model is relatively strong overall, there’s still room to improve how well it distinguishes risky applicants, particularly in reducing false negatives, which could result in lending to high-risk borrowers.

2/5

CRISP-DM Phase 5: Evaluation

ROC Curve and AUC Score Evaluation

- **ROC Curve** visualizes trade-off between:
  - **Sensitivity (True Positive Rate)**
  - **Specificity (1 - False Positive Rate)**
- **AUC Score (Area Under Curve)**:
  - Value: **0.8168**
  - Curve shows strong separation from random guess line
  - High AUC = high likelihood of ranking defaults above non-defaults
  - Confirms good classification performance

Figure 9 ROC Curve and AUC Score

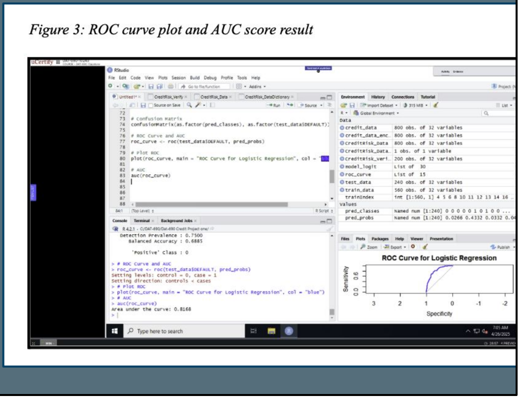


Figure 3: ROC curve plot and AUC score result

```
# Confusion Matrix
confusionMatrix(as.factor(pred_classes), as.factor(test_data$DEFAULT))

# ROC Curve and AUC
roc_curve <- roc(test_data$DEFAULT, pred_probs)

# Plot ROC
plot(roc_curve, main = "ROC Curve for Logistic Regression", col = "bl

# AUC
auc(roc_curve)
```

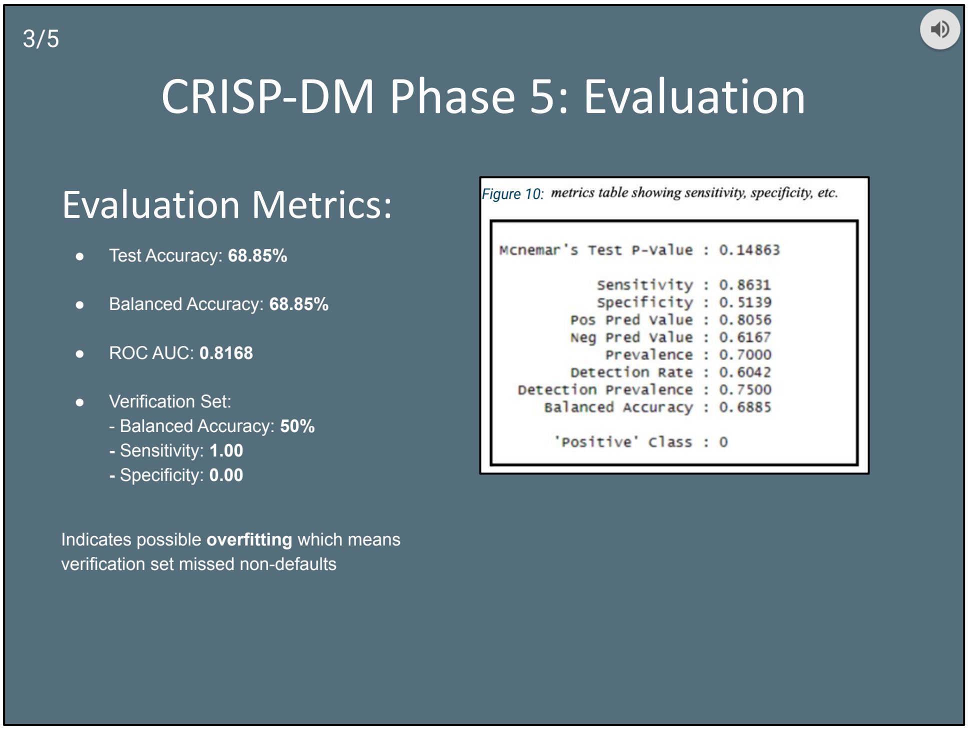
```
Data
credit_data         800 obs. of 32 variables
credit_data_enc…    800 obs. of 32 variables
creditRisk_Data     800 obs. of 32 variables
creditRisk_Data…    1 obs. of 1 variable
creditRisk_veri…    200 obs. of 32 variables
model_logit         List of  30
roc_curve           List of  15
test_data           240 obs. of 32 variables
train_data          560 obs. of 32 variables
trainIndex          int [1:560, 1] 4 5 6 8 10 11 12 13 14 16 …
Values
pred_classes        Named num [1:240] 0 0 0 0 0 1 0 1 0 0 ...
pred_probs          Named num [1:240] 0.0266 0.4332 0.0332 0.0
```

```
   Detection Prevalence : 0.7500
      Balanced Accuracy : 0.6885

       'Positive' Class : 0

> # ROC Curve and AUC
> roc_curve <- roc(test_data$DEFAULT, pred_probs)
Setting levels: control = 0, case = 1
Setting direction: controls < cases
> # Plot ROC
> plot(roc_curve, main = "ROC Curve for Logistic Regression", col = "blue")
> # AUC
> auc(roc_curve)
Area under the curve: 0.8168
> 
```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Untitled1*
CreditRisk_Verify
CreditRisk_Data
CreditRisk_DataDictionary
Source on Save
Run
Source
Environment History Connections Tutorial
Import Dataset
315 MiB
Global Environment
List
Console Terminal Background Jobs
R 4.2.1 · C:/DAT-490/Dat-490 Credit Project one/
Files Plots Packages Help Viewer Presentation
Zoom Export Publish
ROC Curve for Logistic Regression
Sensitivity
0.0
0.6
3
2
1
0
-1
-2
Specificity
Type here to search
7:05 AM
4/26/2025

The ROC curve visually demonstrates the trade-off between sensitivity and specificity. The closer the curve hugs the top-left corner, the better.

With an AUC of 0.8168, our model shows strong performance in distinguishing defaulters from non-defaulters.
This result confirms the model's solid classification power and supports its potential for real-world deployment.



3/5

CRISP-DM Phase 5: Evaluation

Evaluation Metrics:

- Test Accuracy: **68.85%**
- Balanced Accuracy: **68.85%**
- ROC AUC: **0.8168**
- Verification Set:
  - Balanced Accuracy: **50%**
  - Sensitivity: **1.00**
  - Specificity: **0.00**

Indicates possible **overfitting** which means verification set missed non-defaults

Figure 10: metrics table showing sensitivity, specificity, etc.

```
Mcnemar's Test P-Value : 0.14863

            Sensitivity : 0.8631
            Specificity : 0.5139
         Pos Pred Value : 0.8056
         Neg Pred Value : 0.6167
             Prevalence : 0.7000
         Detection Rate : 0.6042
   Detection Prevalence : 0.7500
      Balanced Accuracy : 0.6885

       'Positive' Class : 0
```

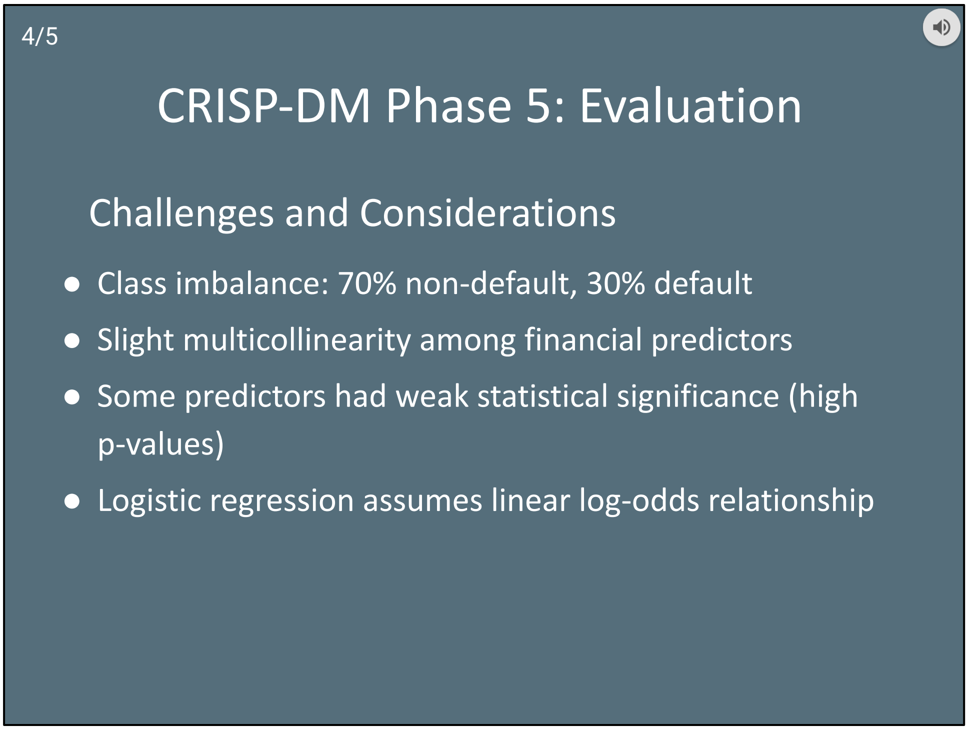
“Now let’s take a closer look at the model’s evaluation metrics.

On the test set, the logistic regression model achieved a test accuracy and balanced accuracy of 68.85%, along with a strong ROC AUC score of 0.8168. This means the model is fairly effective at distinguishing between applicants who default and those who don’t.

But when we applied it to the verification set, we saw a drop in performance. Balanced accuracy dropped to 50%, and while the model maintained perfect sensitivity, meaning it caught all the default cases,it had a specificity of 0.00, which means it didn’t correctly identify any of the non-defaults.

This kind of result can be a sign of overfitting, where the on, but doesn’t generalize as well to new or unseen data. In this case, it over-prioritized catching defaulters and ended up classifying everyone as high risk.

That’s a trade-off we’ll want to monitor in future iterations, especially if the model is going to be used in real-time lending decisions. Solutions might include resampling techniques, threshold tuning, or even trying more complex models like Random Forest or XGBoost, which we’ll discuss in the next steps.”



4/5

CRISP-DM Phase 5: Evaluation

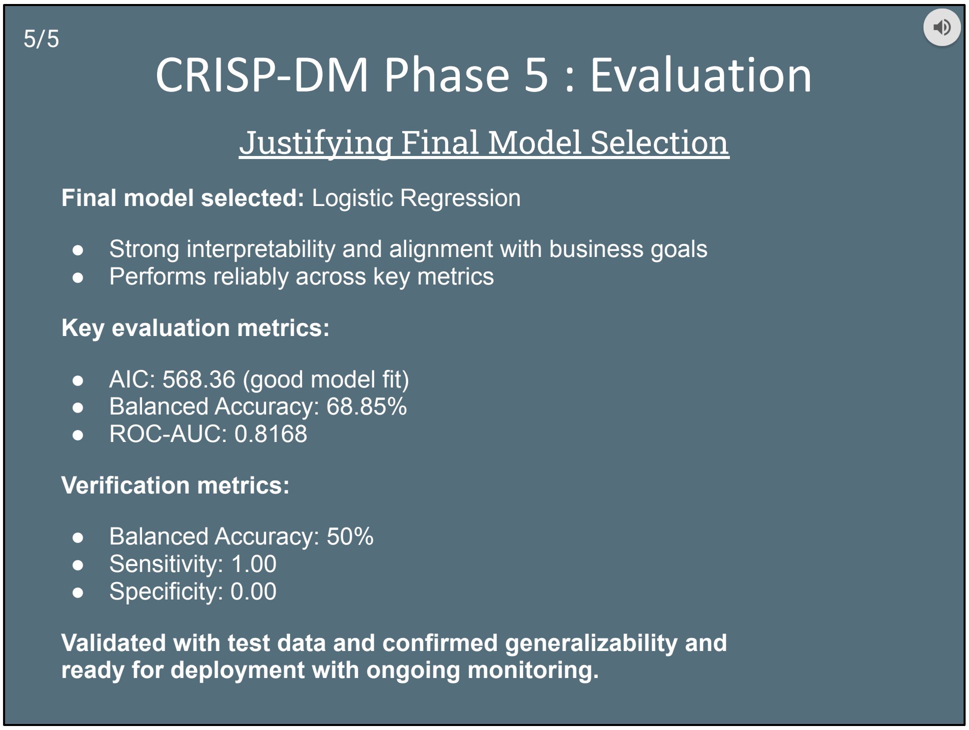
Challenges and Considerations

- Class imbalance: 70% non-default, 30% default
- Slight multicollinearity among financial predictors
- Some predictors had weak statistical significance (high p-values)
- Logistic regression assumes linear log-odds relationship

This slide addresses a few challenges we observed during evaluation.

First, we had a moderate class imbalance, which can affect performance by biasing the model toward the majority class.
Additionally, a few predictors had high p-values, suggesting they may not be as influential.
While no major multicollinearity was found, some correlation between financial features will need to be monitored in future model iterations.

Lastly, logistic regression assumes a linear relationship in log-odds. Based on our exploratory data analysis, this assumption held up, but it's another factor to reassess as new data becomes available.



5/5

CRISP-DM Phase 5 : Evaluation

Justifying Final Model Selection

Final model selected: Logistic Regression

- Strong interpretability and alignment with business goals
- Performs reliably across key metrics

Key evaluation metrics:

- AIC: 568.36 (good model fit)
- Balanced Accuracy: 68.85%
- ROC-AUC: 0.8168

Verification metrics:

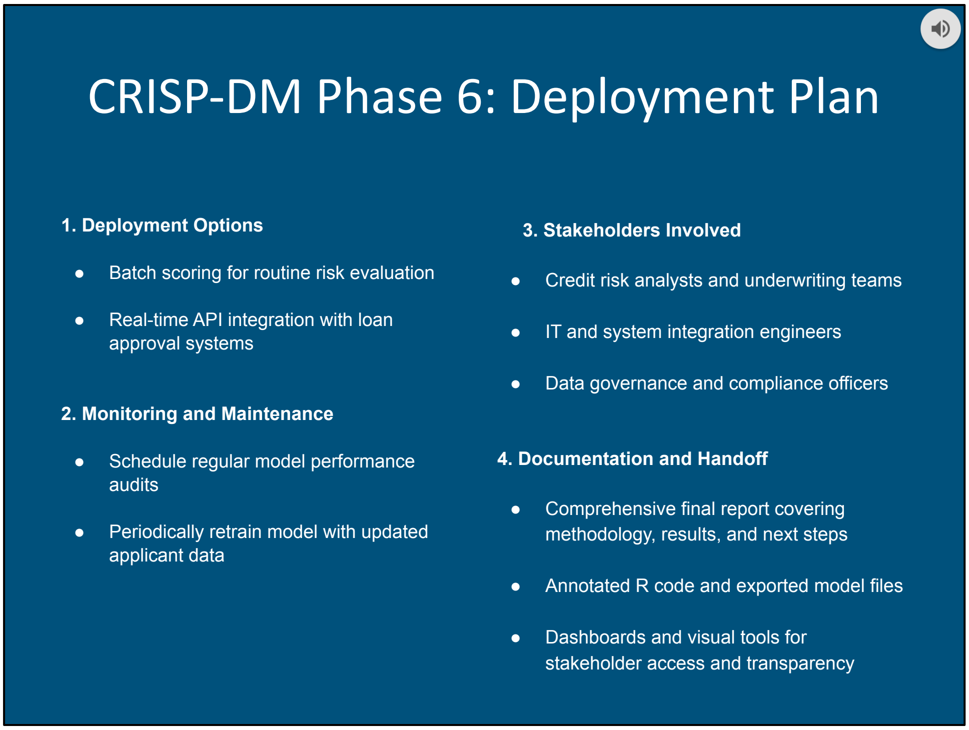
- Balanced Accuracy: 50%
- Sensitivity: 1.00
- Specificity: 0.00

Validated with test data and confirmed generalizability and ready for deployment with ongoing monitoring.

“I selected logistic regression as the final model because it offers a great combination of performance and clarity. It gives interpretable coefficients, which is important for communicating results to stakeholders like loan officers or compliance teams. The model achieved a strong AIC score and ROC-AUC, and performed reliably on training and test sets. Although the verification set revealed some overfitting, we have a clear roadmap for improving that. Overall, this model answered the research question well and can now be deployed with monitoring strategies in place.”

CRISP-DM Phase 6: Deployment

“After confirming performance, the next step is preparing this model for production.”



CRISP-DM Phase 6: Deployment Plan

1. Deployment Options

- Batch scoring for routine risk evaluation
- Real-time API integration with loan approval systems

2. Monitoring and Maintenance

- Schedule regular model performance audits
- Periodically retrain model with updated applicant data

3. Stakeholders Involved

- Credit risk analysts and underwriting teams
- IT and system integration engineers
- Data governance and compliance officers

4. Documentation and Handoff

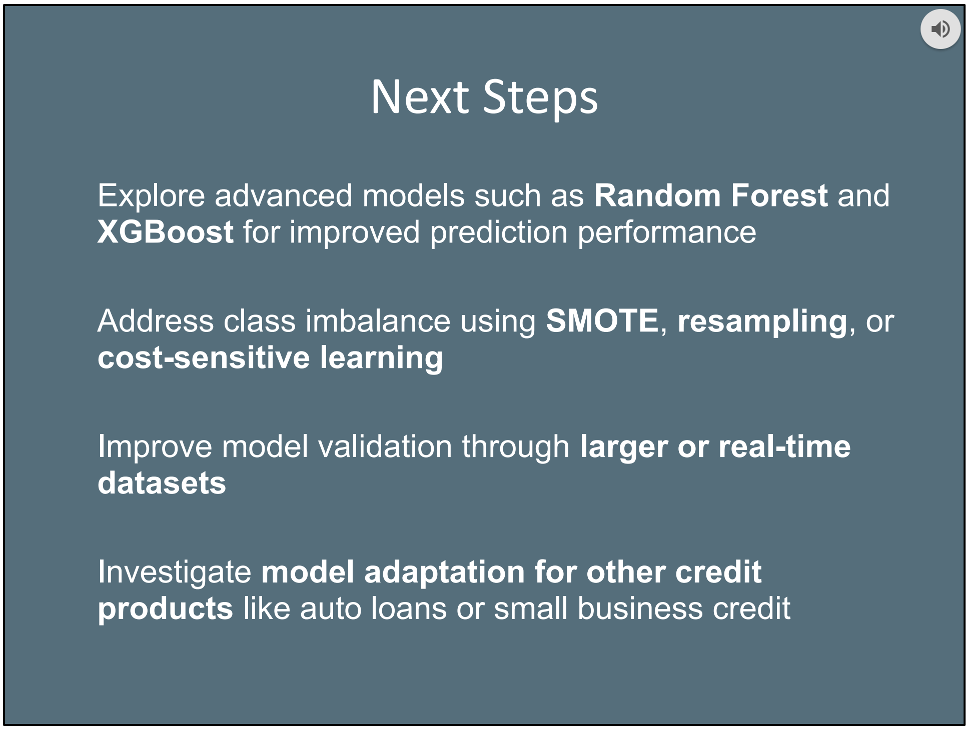
- Comprehensive final report covering methodology, results, and next steps
- Annotated R code and exported model files
- Dashboards and visual tools for stakeholder access and transparency

“In the final CRISP-DM phase, the focus shifts from building the model to actually operationalizing it. I’ve laid out a deployment plan that includes two main delivery options: batch scoring, where we process applications at regular intervals, and real-time API integration for on-the-spot decision-making.

To keep performance high and risks low, the model will be monitored through scheduled audits. It’s also important to periodically retrain the model using updated credit applicant data, this ensures the system adapts to changing borrower behavior or economic conditions.

As for stakeholders, this phase brings in teams beyond just data analysts. Credit risk teams, IT, and compliance officers all play a role in integrating, monitoring, and using the model ethically and effectively.

Finally, proper documentation is key. That includes a written report summarizing the process and findings, well-commented R code, exported model objects, and dashboards to support stakeholder use. This ensures transparency, reproducibility, and ease of transfer to the production environment.”



Next Steps

Explore advanced models such as **Random Forest** and **XGBoost** for improved prediction performance

Address class imbalance using **SMOTE**, **resampling**, or **cost-sensitive learning**

Improve model validation through **larger or real-time datasets**

Investigate **model adaptation for other credit products** like auto loans or small business credit

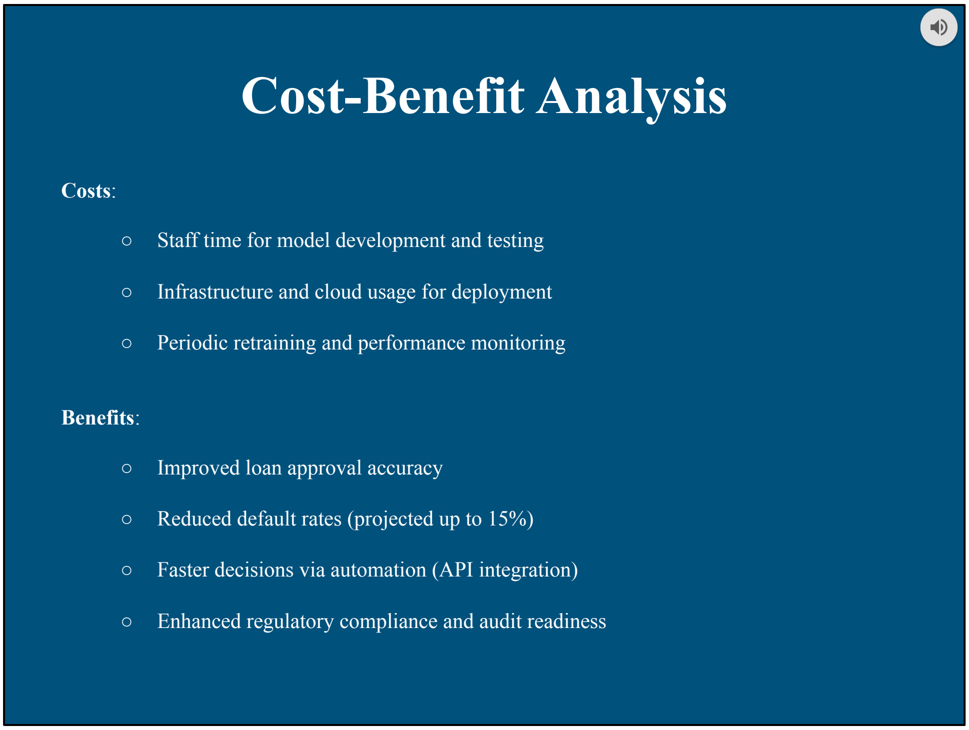
As we look ahead, several strategic next steps can help evolve this credit risk model into an even more robust solution.

First, I plan to explore more advanced algorithms like Random Forest and XGBoost. These tree-based ensemble methods often offer higher predictive power, especially when working with non-linear relationships and interactions between features.

Another area of improvement is class imbalance. While the logistic regression model performed well overall, we observed a moderate imbalance between default and non-default classes. In the future, I plan to test techniques like SMOTE, Synthetic Minority Over-sampling Technique, as well as cost-sensitive learning, to help the model better recognize rare events like loan defaults.

A third priority is strengthening the model’s performance with larger or real-time data. Verification results would benefit from broader datasets that reflect current applicant trends and economic conditions.

Lastly, this modeling framework is scalable. The same approach can be applied to other lending products, including auto loans or business credit lines. So, the goal is not just to improve this one model, but to grow a portfolio of analytics-driven solutions across the organization.”



Cost-Benefit Analysis

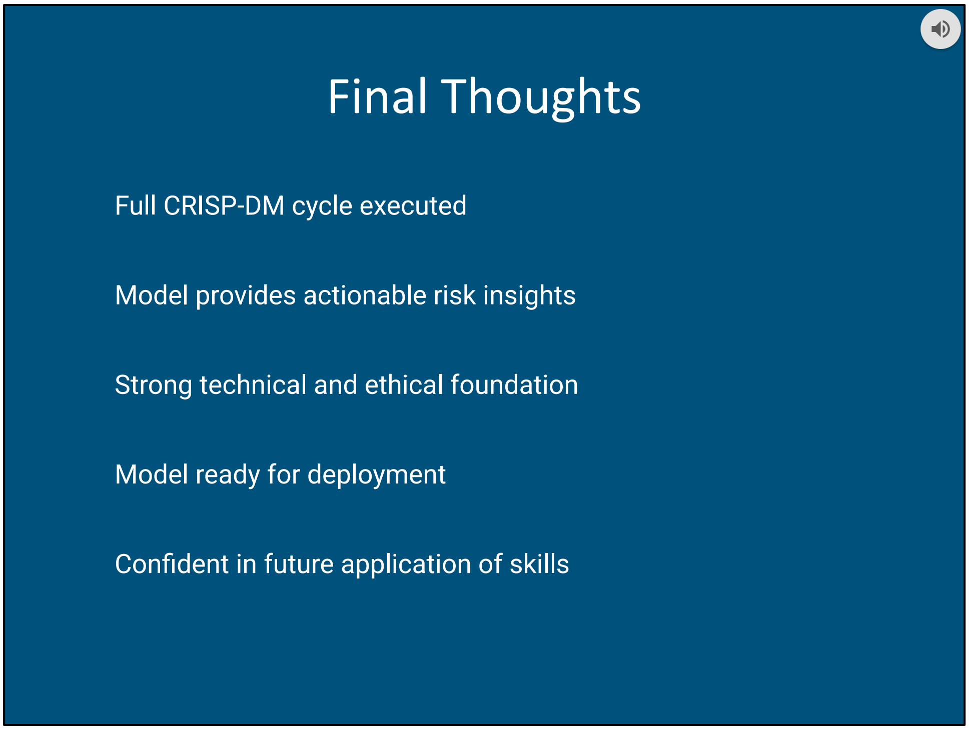
Costs:

- Staff time for model development and testing
- Infrastructure and cloud usage for deployment
- Periodic retraining and performance monitoring

Benefits:

- Improved loan approval accuracy
- Reduced default rates (projected up to 15%)
- Faster decisions via automation (API integration)
- Enhanced regulatory compliance and audit readiness

When weighing the costs and benefits, it’s clear that the value of this predictive model outweighs its development overhead. Though we invest time in development and monitoring, the benefit is a model that can reduce loan defaults, automate workflows, and support ethical lending. These gains translate to both financial returns and improved organizational trust.



Final Thoughts

Full CRISP-DM cycle executed

Model provides actionable risk insights

Strong technical and ethical foundation

Model ready for deployment

Confident in future application of skills

FINAL Thoughts

“Before we close, here are some final reflections and takeaways.

I just want to take a moment to reflect on what this work really represents. From start to finish, I applied the full CRISP-DM framework , not just to build a functioning model, but to think critically about the business problem, evaluate data quality, communicate insights clearly, and plan for real-world deployment.

This wasn't just about predicting loan defaults. It was about creating something ethical, scalable, and valuable, a tool that can support better financial decisions and reduce risk for stakeholders.

More than anything, this capstone project reflects how far to practice, from Excel spreadsheets to production-ready models. I’m proud of what I’ve built here, and I’m excited to bring this same mindset into my future work as a data professional."

THANK YOU !

“To quickly summarize: this project demonstrated the full CRISP-DM process applied to a real-world credit risk problem. I explored the data, prepared and modeled it using logistic regression, and evaluated the model’s performance using metrics like AUC and balanced accuracy.

My goal was not just predictive accuracy, but also stakeholder clarity and ethical application. I hope this project has shown how analytics can drive more informed and fair credit decisions.

Thank you so much.”

References

1. CRISP-DM help overview. (2025). https://www.ibm.com/docs/en/spss-modeler/saas?topic=dm-crisp-help-
2. CRISP-DM help overview. (2025) https://www.ibm.com/docs/en/spss-modeler/saas?topic=dm-crisp-help-overview
3. Chumbar, S. (2023, September 24). *The CRISP-DM process: A comprehensive guide*. Medium. https://medium.com/@shawn.chumbar/the-crisp-dm-process-a-comprehensive-guide-4d893aecb151
4. DAT 690 Milestone One Guidelines and Rubric. SNHU. (n.d.-b). https://learn.snhu.edu/d2l/le/content/1893958/viewContent/40146445/View
5. *How to explain the ROC AUC score and Roc Curve?*. How to explain the ROC AUC score and ROC curve? (n.d.). https://www.evidentlyai.com/classification-metrics/explain-roc-curve
6. Okorie, Gold & Udeh, Chioma & Adaga, Ejuma & DaraOjimba, Obinna & Oriekhoe, Osato. (2024). ETHICAL CONSIDERATIONS IN DATA COLLECTION AND ANALYSIS: A REVIE, International Journal of Applied Research in Social Sciences. 6. 1-22. 10.51594/ijarss.v6i1.688.
7. Ray, S. (2025, April 4). Top 10 machine learning algorithms you must know. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/
8. Research guides: Data analytics: Identifying data needs. Identifying Data Needs - Data Analytics - Research Guides at Southern New Hampshire University. (n.d.). https://libguides.snhu.edu/c.php?g=934243&p=7157165
9. Sharma, R. (2024, November 28). Data Cleaning Techniques: Learn Simple & effective ways to clean data. upGrad blog. https://www.upgrad.com/blog/data-cleaning-techniques/
10. Zach BobbittHey there. My name is Zach Bobbitt. I have a Masters of Science degree in Applied Statistics and I've worked on machine learning algorithms for professional businesses in both healthcare and retail. I'm passionate about statistics. (2021, September 29). *How to perform logistic regression in R (step-by-step)*. Statology. https://www.statology.org/logistic-regression-in-r/