Project Summary and Analytic Plan

The Cross-Industry Standard Process (CRISP-DM) is a widely recognized framework for guiding data mining projects (IBM, 2021). It consists of six phases, Business Understanding, Data Understanding, Modeling, Data Preparation, Evaluation, and Deployment, all of which have distinct tasks included that ensure data projects are completed successfully. The phases do not have to follow a fixed order, rather projects often move back and forth between phases to incorporate adjustments and newly gained insights along during the length of the project (2021). Utilizing CRISP-DM, the credit branch will be able to re-evaluate the current methodology for credit risk and implement a more reliable, modern system.

Business Understanding Phase

Business Problem

After the financial crisis of 2008-2009, the credit branch is re-evaluating the current methodology for determining credit risk. The business goal is to assess whether a new credit applicant is likely to default on their loan based on historical data from applicants in the past. When a customer defaults, there is a financial impact of 150% of the remaining credit balance. With this information, it is important to identify higher risk applicants faster and more accurately in order for the branch to avoid major financial losses. The objective of this project is to create a predictive model that can estimate the likelihood a customer will default based on numerous variables.

Research Question

The research question is “Can a predictive model be created that can accurately determine the likelihood of customer default during loan application based on various financial and demographic variables?”.

Helping the Business

Implementing the predictive model as a solution to the problem can help the credit branch in many ways. Firstly, it will modernize the current methodology, provided a more reliable and accurate way to predict customer defaults. This will help the business users make informed decisions when extending credit, which will reduce the financial losses that could occur. By incorporating the predictive model into the daily workflow, high risk applicants will be identified, and further steps can be taken to help minimize risk for the credit branch. In turn, this will provide improved lending practices and overall business performance of the branch.

Data Understanding Phase

The Credit_RiskData dataset contains 800 records with 31 variables, such as CHK_ACCT, AMOUNT, EMPLOYEMENT, and AGE, with the target variable being DEFAULT. This phase involves exploring the dataset to verify data format and number of records, identifying any relationships between data, and verifying how dirty the data is before the next phase (Hotz, 2024).

To get the most out of this phase, an exploratory data analysis will be performed to understand the general structure of the data and to identify any trends and/or outliers. Missing data or flawed data, such as duplicate values and inconsistencies, that could cause poor model performance will also be identified for the next phase.

Data Preparation

The data preparation phase prepares data for modeling. Roughly 80% of the project is spent in this phase and consists of five tasks: selecting data, cleaning data, constructing data, integrating data, and formatting data (2024). Since the dataset has over 30 variables, it's important to choose the most relevant and remove the irrelevant to ensure proper function within the model. Variables such as CHK_ACCT, DURATION, HISTORY, INSTALL_RATE may be kept in the dataset for the financial aspects, while other such as RADIO/TV, FOREIGN, and OBS# may be removed since they may not impact the performance of the model and don't offer any valuable insight.

After the variable are selected, cleaning the data by handling missing values, outliers, and formatting issues can be performed. This will confirm that the data is in appropriate form and ready to be passed through the model. If necessary, new variables can be created, such as combining CHK_ACCT and SAV_ACCT, in order to establish the total amount of cash on hand an individual has, and data can be transformed to help balance the data to improve model performance.

Modeling Phase

The goal of the modeling phase is to determine the best modeling techniques that would be suitable for the research questions and dataset, split the data into training and testing data, building the model, and assessing the model (2024). In this phase, various models such as random forest, multiple regression, and clustering, will be tested to see which model best fits with the available data. The dataset will be split in order to train and test the data, and the results will be assed, such as metrics such as accuracy, precision, F1 score, and recall to determine

validity. Different iterations of the model will also be completed to compare different techniques to see which has a stronger outcome.

Evaluation Phase

In the evaluation phase, the model's performance is carefully evaluated to determine if it meets the goals of the business (2024). It's important to assess the accuracy of the predictions and verify that the model can work with both historical and new data in order to meet the objectives. To accomplish this, the results of the model will be evaluated based on their business relevance, ensuring the original questions is being answered. If the question is not answered and the metrics are too low, the model may need to be tuned and retested. This may require cleaning the data further, creating a new model, or changing some parameters of the current model in order to meet the business needs.

Deployment Phase

If the model results meet the needs of the business, it will be integrated into the production environment where it will be used to predict the customer defaults. An automation will be created that runs the model in a predetermine timeframe, typically at night after the branch is closed, in order to pass the currents days data through. If a customer is identified as being high risk for default, a report will be created, and the business user will be able to reference this report and inform the customer. The predictive model will be continuously monitored and adjustment will be made if the accuracy decreases or errors are found within the results.

By utilizing CRISP-DM during the predictive modeling project for the credit branch, the framework will guide the development and implementation of the dataset and model. With this

approach, we can ensure the model meets the business needs and objectives, while delivering actionable insights around high-risk customers and preventing financial loss for the branch.

Data Understanding and Data Preparation

Data Understanding: Select Data

The CreditRisk_Data dataset contains 32 variables, both categorical and numerical. The categorical variables are CHK_ACCT, HISTORY, NEW_CAR, USED_CAR, EMPLOYEMENT, etc., while the numerical variables are DURATION, AMOUNT, INSTALL_RATE, AGE, NUM_CREDITS, and NUM_DEPENDENTS. These variables all contribute to the likelihood of a customer defaulting on their loan, which is the DEFAULT variable in the dataset.

Understanding which variables to include and exclude from a dataset is important when conducting an analysis. Datasets may include variables that do not significantly influence the analysis and excluding them may help the analyst focus on more significant and impactful data (Banghart, 2019). In order to determine which variables have a strong impact on the analysis, a correlation matrix or chart can be created. This will allow the variables that have a high correction to the target variable to be identified and kept within the analysis, while the variables with a low correlation can be removed

Data Understanding: Descriptive Statistics Analysis and Correlation

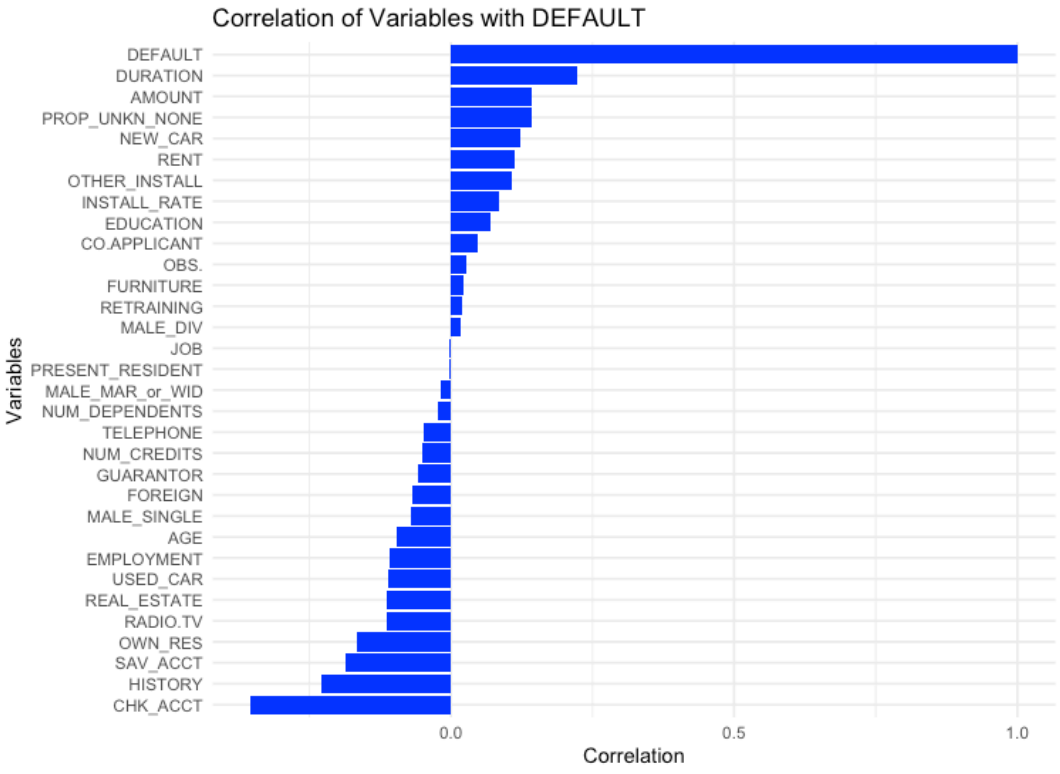
OBS.	CHK_ACCT	DURATION	HISTORY	NEW_CAR	USED_CAR	FURNITURE
Min. : 1.0	Min. :0.000	Min. : 4.00	Min. :0.000	Min. :0.0000	Min. :0.0000	Min. :0.0000
1st Qu.: 400.8	1st Qu.:0.000	1st Qu.:12.00	1st Qu.:2.000	1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:0.0000
Median : 600.5	Median :1.000	Median :18.00	Median :2.000	Median :0.0000	Median :0.0000	Median :0.0000
Mean : 576.0	Mean :1.567	Mean :21.01	Mean :2.539	Mean :0.2437	Mean :0.1037	Mean :0.1787
3rd Qu.: 800.2	3rd Qu.:3.000	3rd Qu.:24.00	3rd Qu.:4.000	3rd Qu.:0.0000	3rd Qu.:0.0000	3rd Qu.:0.0000
Max. :1000.0	Max. :3.000	Max. :72.00	Max. :4.000	Max. :1.0000	Max. :1.0000	Max. :1.0000

RADIO.TV	EDUCATION	RETRAINING	AMOUNT	SAV_ACCT	EMPLOYMENT	INSTALL_RATE
Min. :0.0000	Min. :0.00000	Min. :0.00000	Min. : 250	Min. :0.000	Min. :0.000	Min. :1.000
1st Qu.:0.0000	1st Qu.:0.00000	1st Qu.:0.00000	1st Qu.: 1374	1st Qu.:0.000	1st Qu.:2.000	1st Qu.:2.000
Median :0.0000	Median :0.00000	Median :0.00000	Median : 2326	Median :0.000	Median :2.000	Median :3.000
Mean :0.2725	Mean :0.05375	Mean :0.09375	Mean : 3272	Mean :1.109	Mean :2.356	Mean :2.965
3rd Qu.:1.0000	3rd Qu.:0.00000	3rd Qu.:0.00000	3rd Qu.: 3960	3rd Qu.:2.000	3rd Qu.:3.000	3rd Qu.:4.000
Max. :1.0000	Max. :1.00000	Max. :1.00000	Max. :18424	Max. :4.000	Max. :4.000	Max. :4.000

MALE_DIV	MALE_SINGLE	MALE_MAR_or_WID	CO.APPLICANT	GUARANTOR	PRESENT_RESIDENT
Min. :0.00000	Min. :0.0000	Min. :0.00000	Min. :0.0000	Min. :0.0000	Min. :1.000
1st Qu.:0.00000	1st Qu.:0.0000	1st Qu.:0.00000	1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:2.000
Median :0.00000	Median :1.0000	Median :0.00000	Median :0.0000	Median :0.0000	Median :3.000
Mean :0.04375	Mean :0.5475	Mean :0.09375	Mean :0.0425	Mean :0.0525	Mean :2.868
3rd Qu.:0.00000	3rd Qu.:1.0000	3rd Qu.:0.00000	3rd Qu.:0.0000	3rd Qu.:0.0000	3rd Qu.:4.000
Max. :1.00000	Max. :1.0000	Max. :1.00000	Max. :1.0000	Max. :1.0000	Max. :4.000

REAL_ESTATE	PROP_UNKN_NONE	AGE	OTHER_INSTALL	RENT	OWN_RES	NUM_CREDITS
Min. :0.0000	Min. :0.000	Min. :19.00	Min. :0.0000	Min. :0.0000	Min. :0.0000	Min. :1.000
1st Qu.:0.0000	1st Qu.:0.000	1st Qu.:27.00	1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:1.000
Median :0.0000	Median :0.000	Median :33.00	Median :0.0000	Median :0.0000	Median :1.0000	Median :1.000
Mean :0.2725	Mean :0.155	Mean :35.56	Mean :0.1862	Mean :0.1837	Mean :0.7063	Mean :1.399
3rd Qu.:1.0000	3rd Qu.:0.000	3rd Qu.:42.00	3rd Qu.:0.0000	3rd Qu.:0.0000	3rd Qu.:1.0000	3rd Qu.:2.000
Max. :1.0000	Max. :1.000	Max. :75.00	Max. :1.0000	Max. :1.0000	Max. :1.0000	Max. :4.000

JOB	NUM_DEPENDENTS	TELEPHONE	FOREIGN	DEFAULT
Min. :0.000	Min. :1.000	Min. :0.000	Min. :0.00	Min. :0.0000
1st Qu.:2.000	1st Qu.:1.000	1st Qu.:0.000	1st Qu.:0.00	1st Qu.:0.0000
Median :2.000	Median :1.000	Median :0.000	Median :0.00	Median :0.0000
Mean :1.893	Mean :1.155	Mean :0.395	Mean :0.03	Mean :0.3063
3rd Qu.:2.000	3rd Qu.:1.000	3rd Qu.:1.000	3rd Qu.:0.00	3rd Qu.:1.0000
Max. :3.000	Max. :2.000	Max. :1.000	Max. :1.00	Max. :1.0000



	Variable	Correlation
32	DEFAULT	1.000000000
3	DURATION	0.223939097
11	AMOUNT	0.142992186
22	PROP_UNKN_NONE	0.142561064
5	NEW_CAR	0.121786920
25	RENT	0.111906662
24	OTHER_INSTALL	0.107057074
14	INSTALL_RATE	0.084158629
9	EDUCATION	0.070119683
18	CO.APPLICANT	0.048227919
1	OBS.	0.027004360
7	FURNITURE	0.022693795
10	RETRAINING	0.018898363
15	MALE_DIV	0.016987505
28	JOB	-0.002721079
20	PRESENT_RESIDENT	-0.003778863
17	MALE_MAR_or_WID	-0.018316875
29	NUM_DEPENDENTS	-0.022292729
30	TELEPHONE	-0.048678967
27	NUM_CREDITS	-0.051086375
19	GUARANTOR	-0.059123593
31	FOREIGN	-0.069153351
16	MALE_SINGLE	-0.071578376
23	AGE	-0.094862006
13	EMPLOYMENT	-0.108571691
6	USED_CAR	-0.110443386
21	REAL_ESTATE	-0.114277681
8	RADIO.TV	-0.114277681
26	OWN_RES	-0.166895552
12	SAV_ACCT	-0.185712104
4	HISTORY	-0.228524646
2	CHK_ACCT	-0.353295568

The results of the descriptive analysis and correlation analysis (above) show a high level information about the dataset. The descriptive analysis provides measurements of central tendency and measure of variability. Measures of central tendency highlight the averages of data, whereas measures of variability show the spread of data (Hayes, 2024). Both are very useful during a descriptive analysis and can provide useful data at a glance. For example, we can tell from this descriptive analysis that the average duration is about 21 months, with a minimum value of 4 and a maximum value of 72. We can also see from this analysis that the average age of customer in the dataset is about 35 years, with a minimum value of 19 and a maximum value of 75. Utilizing this information can give an analyst a high-level overview of the data being worked with.

The correlation analysis shows the relationship between the variables within the dataset to the target variable, DEFAULT. A positive correlation means that as the value of a variable increases or decreases, the likelihood of default does the same. A negative correlation means that as the variable increases or decreases, the likelihood of default does the opposite. For example, since DURATION has a positive correlation, this means that longer loan durations are associated with a higher risk of default. CHK_ACCT has a strong negative correlation, meaning that customers with a lower checking account balance are associated with a higher risk of default, and customer with a higher balance are associated with a lower risk of default. This correlation analysis indicates which variables have correlations and have a stronger impact on the overall analysis.

Data Understanding: Descriptive Statistics Results for Selected Variables

After completing the descriptive analysis and correlation analysis, it becomes more clear which variables have an impact on the likelihood of default. Keeping relevant variables in the analysis can improve the model accuracy, make it easier to understand, and improve the model efficiency/performance (Chowdhury & Turin, 2020). Within the CreditRisk_Data dataset, it would be beneficial to remove several variables that have little to no correlation to the target. This includes JOB, PRESENT_RESIDENT, MALE_DIV, RETRAINING, MALE_MAR_or_WID, and NUM_DEPENDENTS. These variables can all be found in the middle of the chart above, indicating that their correlation to DEFAULT is very low. I would also remove OBS# from the analysis, not because it has a low correlation, but because it is used for identification within the dataset, not analysis. The fact that this variable is associated with default is pure coincidence, and doesn't provide any valuable insight to the analysis.

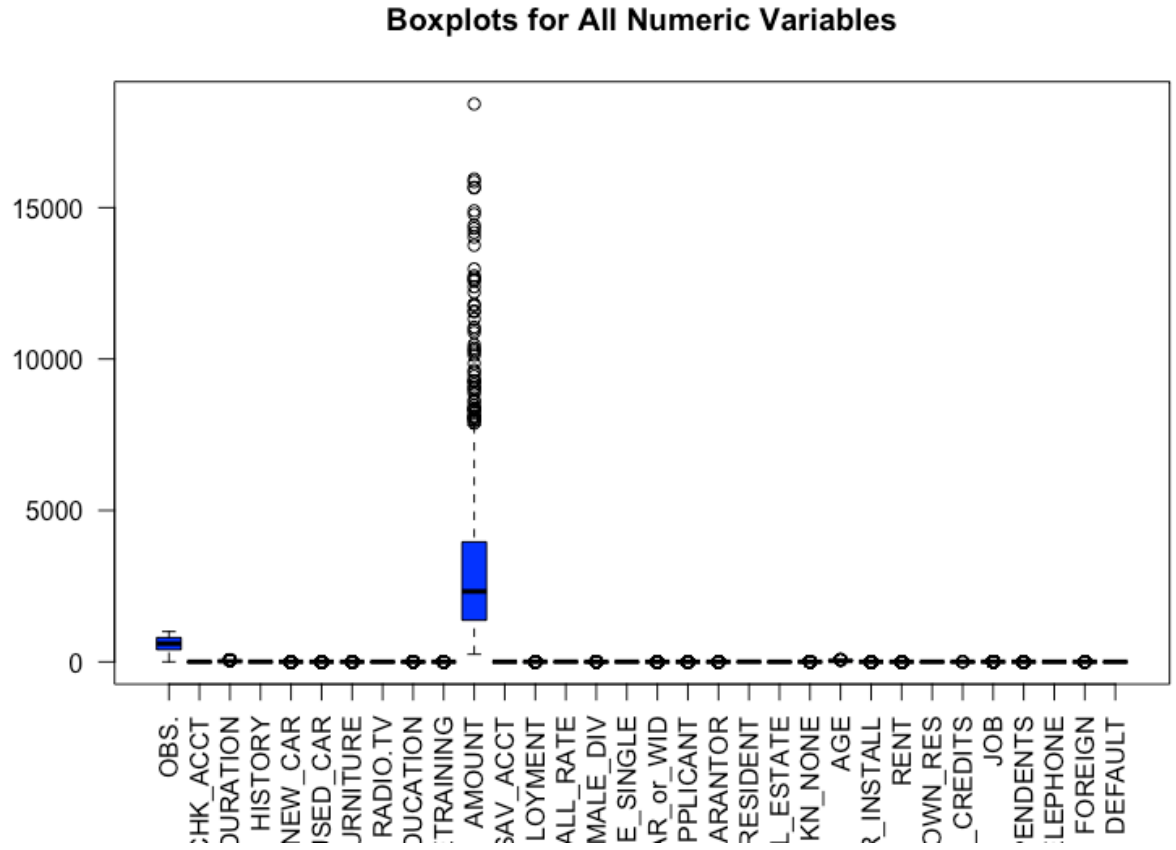
By removing these seven variables that have low correlation to the target, noise will be removed and allow the model to perform more accurately and reliably. The remaining variables have either a positive or negative correlation to default, which ensures the model can use this data for predictions. If the model isn't performing up to the desired accuracy, variables can be added back into the model to prevent overfitting or bias within the model.

Data Preparation: Preparing and Cleaning Data

Preparing the data for analysis is an important step during the CRISP-DM process and commonly accounts for 80% of the project's total time (Hayes, 2024). Ensuring the data is cleaned, formatted, and organized correctly sets the project up for success when modeling. Even after this phase is completed, it can be revisited later in the project for adjustments if needed. To start preparing the data for analysis, I start by opening the spreadsheet to gain a quick overview of the data I will be working with. I make sure all the headers are formatting the same way and there is a decent amount of data available. Although this isn't the easiest way, I quickly scan to see if there are any missing values. If there are, I can try to correct them moving forward. After these initial steps, I read the data into RStudio where I continue the cleaning processes.

When in RStudio, I perform some exploratory analysis that can help with cleaning. I run head(creditrisk_data), which shows the first several rows of data, summary(creditrisk_data), which shows summary data such as central tendency for the variables, and I also check for missing values using the sum(is.na(creditrisk_data)) command. In this case, there were no missing values within the dataset, so I moved onto identifying any outliers. I created a box plot that shows all the variables and their associated values (below). Typically, outliers lie further away from the median and Q3, so a box plot is an effective visual way to see if any exist. With

the dataset, the only variables that could have an outlier looks like AMOUNT. However, after viewing the variable with the command print(sort(creditrisk_data$AMOUNT, decreasing = TRUE)), it doesn't actually seem that the variable has an outlier. It seems to follow the increase as the other values do within this variable, so I would leave this value in the dataset. The dataset wasn't as dirty as others, so there wasn't a lot of cleaning necessary.



Data Preparation: Construct Data

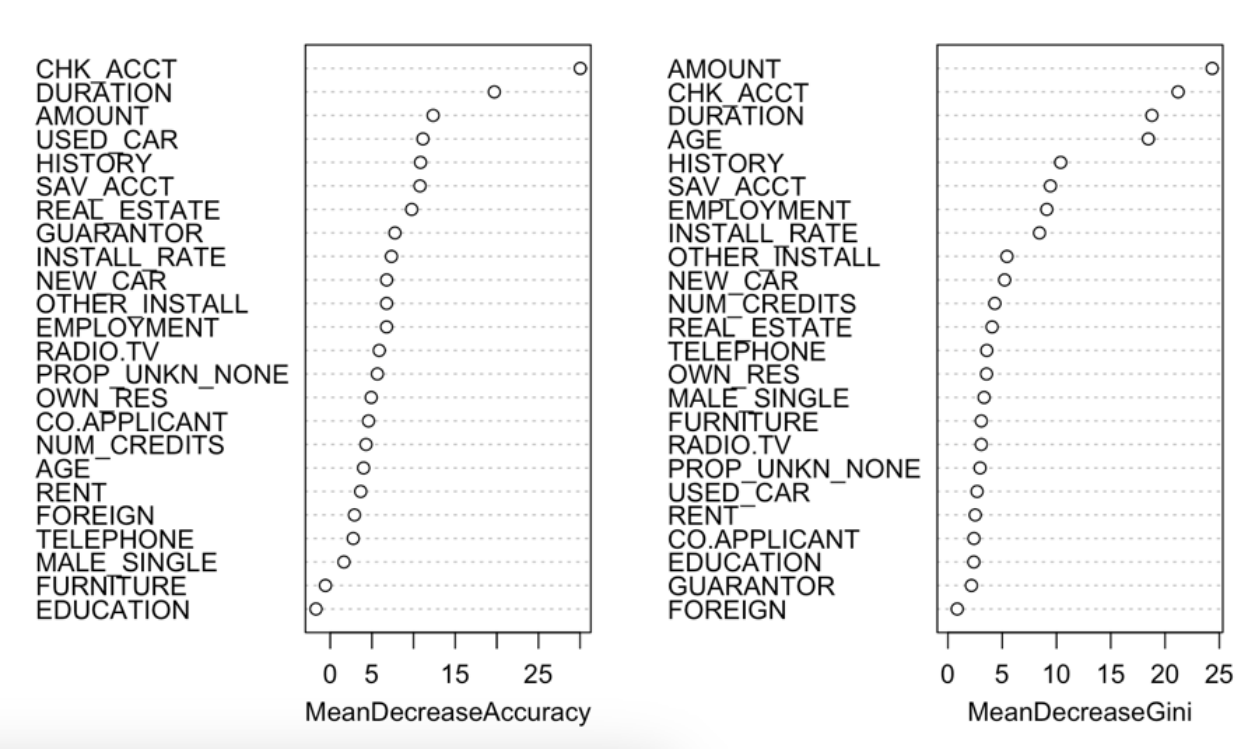
By cleaning the data, identifying and handling outliers and missing data, and ensuring the formatting is optimal for model creating, I have determined that the dataset is complete and consistent. At this point, I don't believe any new variables needed to be created, as the available data should be enough for the predictive model. There were also no other sources to merge data

from, as the dataset is first party data from the credit branch. An important aspect of CRISP-DM is that you can revisit phases at various points in the project. If a phase wasn't completed as thoroughly as intended, you can revisit and make corrections at any time. This ensures that the project is accurate and efficient before going live into production.

Modeling and Evaluation

Modeling Phase: Artifacts

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 155  32
         1  11  41

               Accuracy : 0.8201
                 95% CI : (0.7654, 0.8666)
    No Information Rate : 0.6946
    P-Value [Acc > NIR] : 7.184e-06

                  Kappa : 0.5388

 Mcnemar's Test P-Value : 0.002289

            Sensitivity : 0.9337
            Specificity : 0.5616
         Pos Pred Value : 0.8289
         Neg Pred Value : 0.7885
             Prevalence : 0.6946
         Detection Rate : 0.6485
   Detection Prevalence : 0.7824
      Balanced Accuracy : 0.7477

       'Positive' Class : 0
```



A Random Forest model was selected for the credit risk scenario containing the hyperparameters of 1,000 trees, 4 variables considered at each split, and each node containing at least 5 samples. These hyperparameters create a starting point for the forest to be tested and evaluated, allowing for changes to be implemented to strengthen the model if needed. After running the model, a confusion matrix and variable importance plot was created. The confusion matrix contains the preliminary results and shows that the accuracy of the model is about 82%, which is above the target value of 70%. It also shows that the sensitivity is around 93%, which means the model highly accurate of predicting non defaulting customers. The specificity is about 56%, which means the model is moderately accurate at predicting defaulting customers.

The variable importance plot shows that CHK_ACCT, DURATION, and AMOUNT are highly ranked features for Mean Decrease Accuracy, indicating that the model's accuracy would significantly decrease if these variables were removed. AMOUNT, CHK_ACCT, and DURATION are highly ranked features for Mean Decrease Gini, indicating that these variables are important for determining the quality of splits within the Random Forest model. These three variables play a significant role in the model's performance when predicting the likelihood of customer default.

Modeling Phase: Data Quality

The data set originally contained 800 observations and 31 variables. There are no missing values, formatting issues, or additional data cleaning required. The variables OBS#, JOB, PRESENT_RESIDENT, MALE_DIV, RETRAINING, MALE_MAR_or_WID, and NUM_DEPENDENTS were removed from the model as they hold little importance and minimal correlation to customer default. After excluding these variables, the model performed strongly, confirming that the data was not necessary.

Modeling Phase: Data Structure

After removed multiple variables, the data set contains 24 variables. To ensure proper handling and analysis, the target variable, DEFAULT, was converted to a factor in the Random Forest model. The data was split for the model, with 70% being used for training purposes, and 30% used for testing purposes. The structure of the data was carefully inspected in order to improve the predictive power of the model. By refining the data set structure and excluding certain variables with low importance, the model achieved high accuracy and confirmed the data preparation was strong.

Evaluation Phase: Model Evaluation

The model performs very well, obtaining an accuracy of about 82% which exceeds the target of 70%. The sensitivity is about 93%, indicating an excellent ability to identify non-defaulting customers. The specificity is about 56%, which is a moderate ability for identifying customers who default. This metric could be increased in several ways, such as adjusting the hyperparameters like mtry, nodeside, or number of trees. If that doesn't help, feature engineering could take place, such as combining CHK_ACCT and SAV_ACCT to create a new variable called "Cash_on_Hand". This could improve the accuracy of the model and reduce overfitting.

Evaluation Phase: Areas of Concern

An area of concern with the model is its specificity performance. Currently, the model has a specificity of about 56%, which indicates room for improvement. This suggests that the model struggles to correctly identify defaulting customers. If not addressed, the results could be bias towards predicting non-defaulting customers, potentially caused by a class imbalance within the dataset.

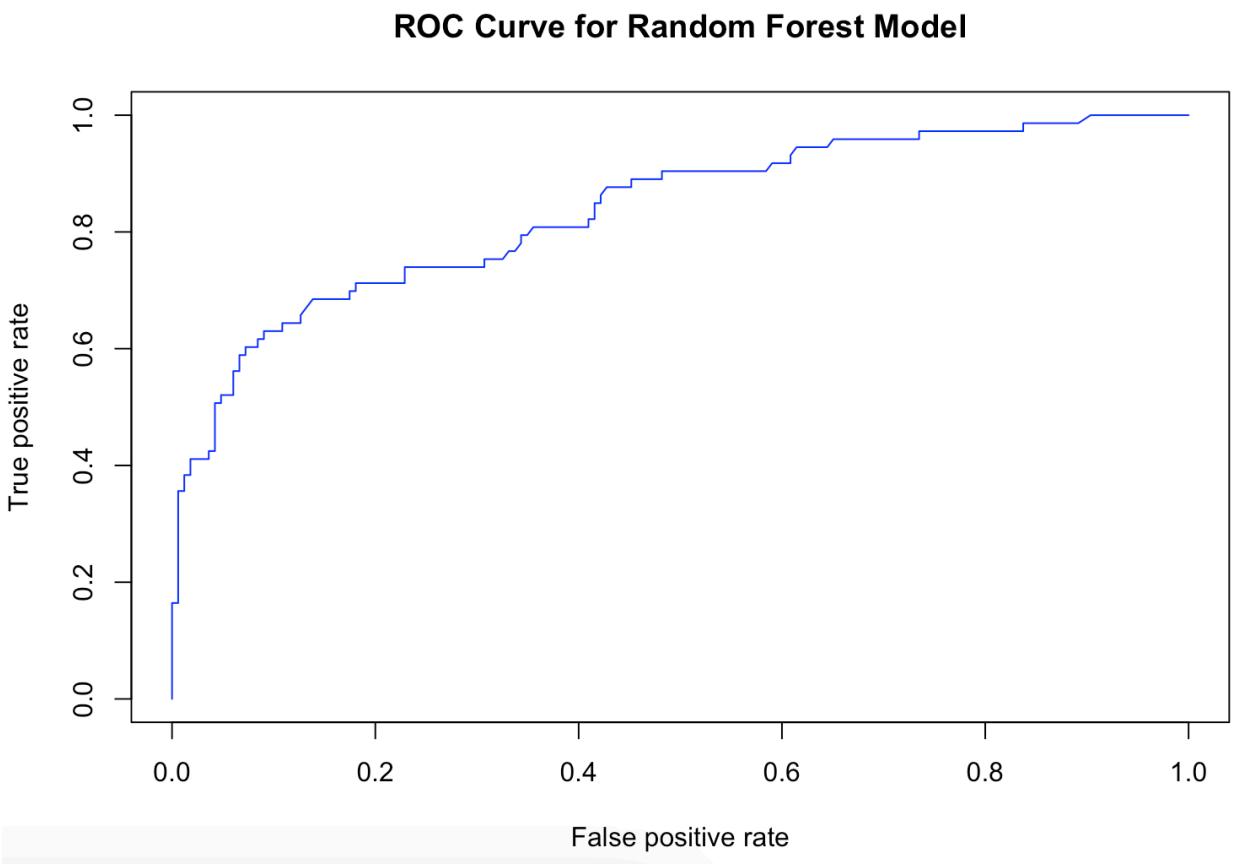
Evaluation Phase: Fit of Model

During the model building process, various statistics were used to evaluate the fit and predictive accuracy of the Random Forest model. The confusion matrix showed that overall, the model could make correct predictions about 82% of the times (accuracy). The model matched the actual result while accounting for chance about 54% of the time (kappa), correctly identified non-defaulting customers about 93% of the time (sensitivity), and correctly identified defaulting customers about 56% of the time (specificity).

The variable importance plot shows how much the accuracy of the model would decrease if certain variables were removed (mean decrease accuracy), and how much each variable contributes to the impurity in the data set (mean gini accuracy). CHK_ACCT, DURATION, and AMOUNT were the top variables in both and indicates that these variables have a strong impact on the performance and accuracy of the model. The Random Forest model achieved strong overall performance, surpassing the target accuracy of 70%. Although some areas could use improvement, such as specificity, the model can still produce reliable results. If needed, hyper tuning can take place to increase lacking areas.

Evaluation Phase: Model Results

The Random Forest model results were evaluated using several metrics, which can show the strengths and weaknesses. When predicting non-defaulting customers, the model correctly identified non-defaulting customers about 83% of the time (true negative). When predicting defaulting customers, the model was able to correctly identify defaulting customers about 79% of the time (true positive). These results could suggest that the model may struggle with false positive predictions, incorrectly predicting non-defaulting customer as defaulting.



The ROC curve above shows the performance of the model across different thresholds. Since the goal of this model is to predict customers who will default, achieving a True Positive Rate of about 79% is a worthwhile trade off, even if the False Positive Rate is slightly higher. This approach enables the model to identify higher risk customers accurately, and if a customer is incorrectly classified as high risk when they aren't, further analysis can be conducted for the specific situation and a decision can be made regarding the customer.

Ethical Consideration

With the Credit Risk dataset already being completed for this project, there weren't many ethical considerations that were needed to be completed on my end. The dataset did not include any sensitive information, such as social security number, name, date of birth, or any other

identifiable variables that could link the data to a specific person. There were several variables that could be identified as bias, such as MALE_SINGLE, MALE_DIV and MALE_MAR_or_WID, since these only apply to the male population and not the female. Most of these variables were left out of the analysis due to their low correlation to the target, but all of them could have been left out completed to avoid any risk of bias in the analysis.

When working with the dataset, it was assumed that we had permission to use the information for training purposes. However, this was never confirmed. If a customer didn't know their data was being used, or didn't give their permission, this could create a lack of trust between the bank and their customers, leading to the customers leaving the bank for another institute. Moving forward, I would want to confirm that the customers understand their data is being used to avoid any miscommunication.

Data integrity, protection, and appropriate use are all important when working with data. Many pieces of data tend to be sensitive to many people, and adhering to proper ethical guideless is of the upmost importance. This capstone project has helped me create a framework that promotes ethical approaches because of the data type that was being worked with. Financial information is very sensitive to a lot of people, so handling the data properly is a requirement. This means only giving access to those who need it, ensure personal information is not included for anonymity, and treating the data as it was my own.

Production Turnover Report

Date: December 8, 2024

Business Department: Credit Department

Project Name: Predictive Credit Risk Analysis System

Project Description: This project is designed to develop and deploy a Random Forest predictive model to assess the likelihood of customer loan default, with each default posing a risk of 150% of the remaining loan balance. The model uses historical customer data for training and testing, while current customer data is used for verification, allowing the identification of key variables that influence defaults. This project aims to enhance data driven decision making within the credit branch, reduce the potential financial loss, and ensure compliance with regulatory standards. By implementing this model, the credit branch will achieve a higher level of reliability and operational efficiency.

Model Baselines:

Training Model	Verification Model
Accuracy: 82.01% **Sensitivity**: 93.37% **Specificity**: 56.16% **Balanced Accuracy**: 74.77% **Lift**: 2.58 **AUC**: 84.01%	**Accuracy**: 73.00% **Sensitivity**: 91.72% **Specificity**: 23.64% **Balanced Accuracy**: 57.68% **Lift**: 1.89 **AUC**: 78.46%

Model Performance: As data volumes increase, it will be important to monitor both the data sources and model variables to maintain consistent performance. The Oracle database must efficiently handle increased volumes without delay, and dataset should be monitored for quality issues such as missing values, duplicates, and outliers. Variables with a high impact on the outcomes should be reviewed to detect and avoid data drift, which may occur over time due to changing market conditions. With an increase of data volume, the model's performance needs to be monitored and tracked in order to identify patterns that may reflect poor performance. If this occurs, the model should be retrained or tweaked to uphold integrity of the outcomes.

Comments: The business users will utilize the model's output to determine if a loan should be extended to a customer. Each night, a batch process will be run through the model and likelihood of default will be predicted for each customer. Customers identified as high risk will not be offered a loan, while those calculated as low risk will be approved. Dashboards containing visualizations and key metrics will be created using programs such as Power BI or Tableau, enabling the credit branch to monitor patterns and make data-driven decisions. Regular resting and maintenance of the model will be performed to ensure accuracy and reliability when new data is added over time.

References

Banghart, M. (2019, November 8). *Dropping unneeded variables*. Social Science Computing Cooperative | WISC. https://ssc.wisc.edu/sscc/pubs/DWE/book/4-4-dropping-unneeded-variables.html

Chowdhury, M. Z. I., & Turin, T. C. (2020). Variable selection strategies and its importance in clinical prediction modelling. *Family medicine and community health*, *8*(1), e000262. https://doi.org/10.1136/fmch-2019-000262

Hayes, A. (2024, June 27). *Descriptive Statistics: definition, overview, types, and examples*. Investopedia. https://www.investopedia.com/terms/d/descriptive_statistics.asp

Hotz, N. (2024, April 28). *What is CRISP DM?* Data Science Process Alliance. https://www.datascience-pm.com/crisp-dm-2/

IBM. (2021, August 17). *CRISP-DM Overview*. https://www.ibm.com/docs/en/spss-modeler/saas?topic=dm-crisp-help-overview