Predicting Credit Card Default Risk

Ganesh Sai Uttej Kayala

Michele Perina

Robert Mikkelson

Sharifur Rahman

Ranjith Kumar Saila

Department of Business, University of Central Oklahoma

MSBA 5314

Dr. Ho-Chang Chae

April 29, 2024

Table Of Contents

Abstract

Banks and credit card companies have always strived to identify profitable customers while minimizing risks of defaults, as these directly impact their bottom line. Despite years of work, predicting credit card defaults remains challenging, largely due to the dynamic nature of economic shifts and the multiplicity of factors involved, ranging from individual financial behaviors to broad socio-economic trends. This study examines a dataset of approximately 45,000 American Express customers in the United States. Variables encompassing credit history, personal finances, demographics, and employment are analyzed using machine learning techniques such as decision trees, neural networks, and logistic regression. We discovered the significance of credit score, and credit limit usage in increasing one's chances of defaulting on their credit card debt. Furthermore, demographic and socio-economic factors—including gender, car and home ownership, migrant worker status, and length of consecutive employment—also contribute significantly, albeit to a lesser extent. The findings suggest that credit card companies like American Express should prioritize their approval efforts on the person's credit history if they want to reduce the risk of default at the minimum. The non credit-related variables still influence credit card defaults, and provide actionable insights to credit card providers.

Introduction

Banks and credit card companies have always strived to identify profitable customers while minimizing risks of defaults, as these directly impact their bottom line. American Express is a global payments company offering credit cards to consumers, small and mid-sized companies as well as corporations. Despite their efforts and resources allocated to predict credit card default, American Express has reported losses of about $1.4 billion in the last quarter of 2023. Therefore, it is crucial to predict future defaults to decrease losses and possibly increase revenues. This is not an easy task since American Express not only wants to minimize defaults, but also increase its customer base.

Despite years of work, predicting credit card defaults remains challenging, largely due to the dynamic nature of economic shifts and the multiplicity of factors involved, ranging from individual financial behaviors to broad socioeconomic trends. Historically, many researches about this topic utilized the Taiwan Credit Data dataset provided in UCI Machine Learning Repository (Arora et al., 2022; Sayjadah et al., 2018; Subasi & Cankurt, 2019; Teng & Lee, 2019; Yeh & Lien, 2009; Yu, 2020). As a consequence, the results of these researches were similar and did not contribute to developing new insights about this topic.

Our research differs from these previous studies for the use of a novel dataset containing information specifically about American Express customers from the United States. According to Beck et al. (2008), differences in customer protection laws, financial education and credit reporting practices influence default propensities across various countries. Additionally, while factors like income may exert similar influences on default behavior, the magnitude of these influences could differ significantly across countries (Jappelli & Pagano, 2002).

Additionally, American Express and credit card companies in general are always looking to expand into new markets and demographics. Our research will try to answer the question of whether or not migrant workers, a demographic facing unique challenges in this regard, and a variable never studied before in literature, can play a significant role in predicting credit card default. By investigating whether migrant workers significantly contribute to credit card default prediction, our aim is to inform whether they pose additional risk to credit card companies. If our findings indicate otherwise, it suggests an opportunity for American Express to expand into this demographic market, potentially fostering financial inclusion and market growth.

Finally, our study endeavors to forecast credit card defaults without relying on credit-related variables. It is widely acknowledged that individuals with high credit scores pose lower default risks, prompting fierce competition among banks to attract these customers. We believe that analyzing often overlooked factors holds promise not only for advancing our understanding of credit card defaults but also for aiding credit card providers in enhancing their credit risk assessment methodologies.

Literature Review

An assessment of the existing literature reveals that previous research papers about credit card default utilized the Taiwan Credit Data dataset provided in UCI Machine Learning Repository. Yeh and Lien (2009) were the first to compare the accuracy of machine learning methods in predicting credit card default utilizing this dataset. Other researchers utilized this same dataset for similar purposes (Arora et al., 2022; Sayjadah et al., 2018; Subasi & Cankurt, 2019; Teng & Lee, 2019; Yu, 2020). T.M. Alam et al. (2020) expanded this scope by incorporating datasets from countries like Germany and Belgium. Li et al. (2019) predicted credit card defaults utilizing a dataset from one of the largest banks in China.

These and other studies have shown that credit card default is influenced by different variables and factors. According to Li et al. (2019), extensive research demonstrated that customers under 30 years of age are more likely to default. The literature consistently shows how younger individuals are more likely to default on credit cards due to their lack of financial knowledge and lower income levels (F.K. Kiarie et al., 2015; T.M. Alam et al.,2020). However, Jain and Jayabalan (2022) present a contradictory perspective, suggesting that older individuals are more likely to default compared to younger people. This divergence in findings presents an intriguing avenue for further exploration within our own research.

Similarly, gender emerges as another variable with conflicting findings. Li et al. (2019) observed that male customers are more likely to default due to the fact that they are more willing to take the risk of default. On the other hand, other studies observed that women, especially single ones, tend to struggle repaying their credit card debt (Li, 2018; Dunn & Mirzaie, 2022).

Beyond age and gender, the literature highlights the significance of additional factors in predicting credit card default. For instance, educational level has been identified as an important factor (Wang et al., 2011). People with better education tend to manage their money better, and are less likely to default. Additionally, variables like credit score and credit limit appear to have the largest impact on predicting credit card default. Individuals with higher credit scores and lower revolving credit utilization are less likely to default (Dunn & Mirzaie, 2022; Li, 2018).

In terms of predictive modeling, many machine learning models were utilized in previous studies. Logistic regression (Li et al., 2019; Sayjadah, Y. et al., 2018) and neural networks (Neema & Soibam, 2017; Sayjadah, Y. et al., 2018; Yeh & Lien, 2008) are common models used across the existing literature, as well as decision trees and random forests (Sayjadah, Y. et al., 2018; T.M. Alam et al., 2020; Yeh & Lien, 2008). Multiple studies also employed gradient boosting (Li et al., 2019; T.M. Alam et al., 2020) and LASSO and adaptive LASSO (F.K. Kiarie et al., 2015). The literature shows that the machine learning approaches with the highest accuracy in predicting credit card default are neural networks and random forest (Neema & Soibam, 2017; Sayjadah, Y. et al., 2018; Yeh & Lien, 2008). In our research, we will utilize neural networks and decision trees, the basic component of random forest. In future versions of our work, random forests will be utilized as well.

Data Understanding

The data for our project was collected from Kaggle website that was previously used for AmExpert 2021 – Code Lab competition, which was a collaborated hackathon competition conducted by American Express and Hackerearth for all the individuals in India. The dataset was uploaded by one of the contestants of the competition for education purposes.

The training dataset consists of following variables:

Table 1: Data Dictionary

Data Dictionary			
Variable	**Description**	**Measurement Level**	**Raw/Created**
Credit Card Default	Did this customer default? Yes or No?	Binary	Raw
Name	This is the first and last name of the customer.	Nominal	Raw
Age	Age of Customer	Numeric	Raw
Gender	Gender of Customer.	Nominal	Raw
Owns Car	Does the customer own a car? Y or N?	Binary	Raw
Owns House	Does the customer own a house? Y or N?	Binary	Raw
No. of Children	Number of children the customer has.	Numeric	Raw
Yearly Income	Yearly income the customer earns.	Numeric	Raw
No. of days Employed	Number of days the customer has been employed.	Numeric	Raw
Occupation	What's the job?	Nominal	Raw

Data Dictionary			
Total Family Members	Total number of family members in the customer's family.	Numeric	Raw
Migrant Worker	Is the customer a migrant worker? Yes or No.	Binary	Raw
Yearly Debt Payment	The amount of money the customer is paying per year on their debt.	Numeric	Raw
Credit Limit	The maximum amount of money the customer can spend on their credit card.	Numeric	Raw
Credit Limit Used	This is the percentage of the credit limit that is being used by the customer.	Numeric	Raw
Credit Score	This is the credit score of the customer.	Numeric	Raw
Previous Defaults	How many previous defaults does the customer have?	Numeric	Raw
Default in last 6 Months	Has the customer defaulted in the last 6 months? Yes? No?.	Binary	Raw
Customer ID	Unique ID for each customer	Nominal	Raw

The initial dataset contains 45,528 observations and 19 variables. Of the initial 19 variables, only 17 were initially selected to build the predictive models. The Customer ID will be assigned in SAS Miner to the role of ID. The Name variable will be rejected in SAS Miner due to the modifications made by American Express to protect Personal Identifiable Information (PII) by altering the names.

Table 2: Initial Excluded Variables

Excluded Variable		
Variable	**Reason for Exclusion**	**Role**
Name	Customer's name was modified due to privacy concerns.	Rejected
Customer ID	Given the role of ID	ID

Data Cleaning

Many variables in our dataset contain missing values. The following tables provides a detailed overview of the number of missing values for the numeric variables:

Table 3: Missing Numeric Values

Number of Children		
	Frequency	**Percent**
Non-missing	44,754	98.3
Missing	774	1.7

Number of Days Employed		
	Frequency	**Percent**
Non-missing	45065	98.98
Missing	463	1.02

Total Family Members		
	Frequency	**Percent**
Non-missing	45,445	99.82
Missing	83	0.18

Yearly Debt Payments		
	Frequency	**Percent**
Non-missing	45,433	99.79
Missing	95	0.21

Credit Score		
	Frequency	**Percent**
Non-missing	45,520	99.8
Missing	8	0.02

Numeric variables that contain missing values are: no_of_children, no_of_days_employed, total_family_members, yearly_debt_payments and credit_score. The percentage of missing values is relatively low (the highest is 1.7%) compared to the total number of observations, and, for this reason, one option could have been to delete the rows with missing value. However, we have opted to impute the missing values using the mean or median depending on each variable's distribution. The Impute Node was used in SAS Miner 15.2 and the Default Input Method for interval variables was set to mean. Since we plan to apply the log transformation to every variable with a skewed distribution before imputing, the mean will be used for all the variables.

Among the nominal variables, only two have missing values: owns_car and migrant_worker.

Table 4: Missing Nominal Values

Owns Car		
	Frequency	**Percent**
Non-missing	44,981	98.8
Missing	547	1.20

Migrant Worker		
	Frequency	**Percent**
Non-missing	45,441	98.81
Missing	87	0.19

Similarly to the numeric variables discussed above, we are not planning on deleting rows containing missing values for these two variables. We have opted to impute them using the mode (most frequently occurring value). Also in this case, the Impute Node was utilized to deal with missing values and the Default Input Method was set to Count for class variables.

The variable gender doesn't contain any null value, but could potentially contain a missing value.

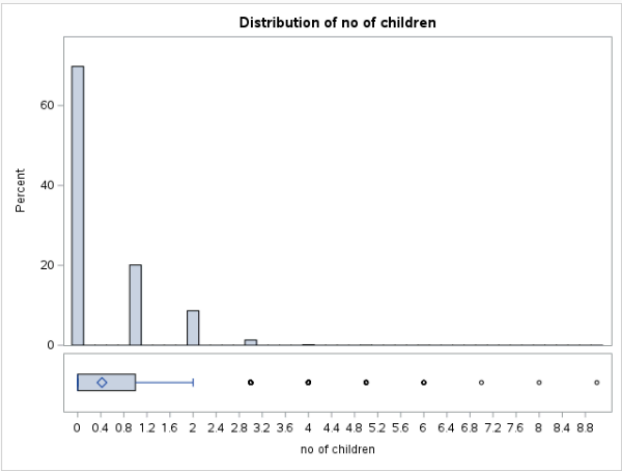
Table 5: Gender Values

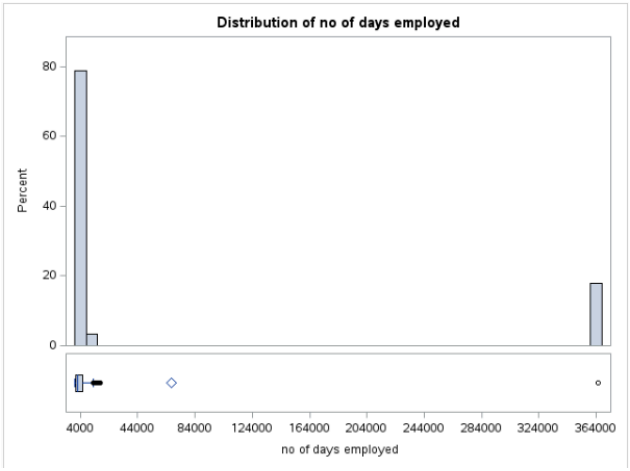
Gender		
	Frequency	**Percent**
F	29,957	65.80
M	15,570	34.20
X	1	0.00

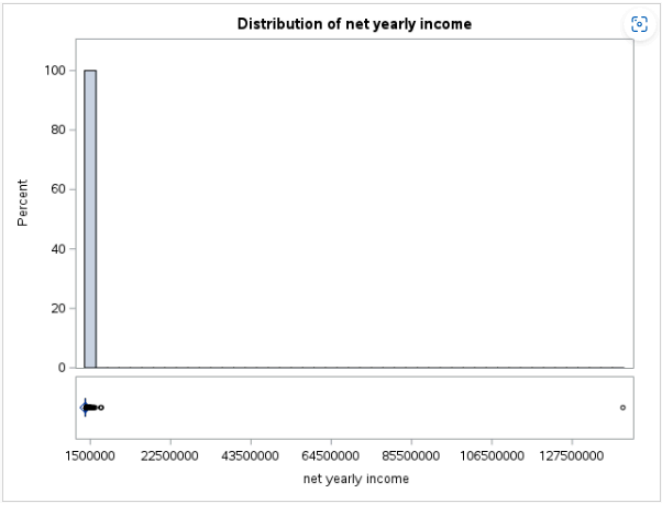
It is common to have more than two values for gender for individuals who prefer not to specify it. However, in this case there is only one different value ("X"), and it appears to be a data entry mistake. For this reason, we will treat the "X" value as a missing value and substitute it with the most frequently occurring value, that is female. The Replacement Node was employed to substitute the value of "X" for a new value of "N".

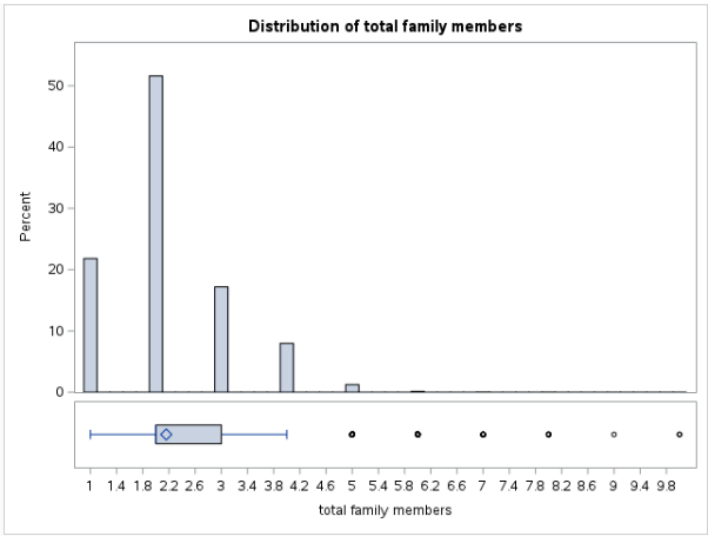
Looking at the distribution plots for numeric variables below, we can see how multiple variables have right skewed distribution and have outliers:

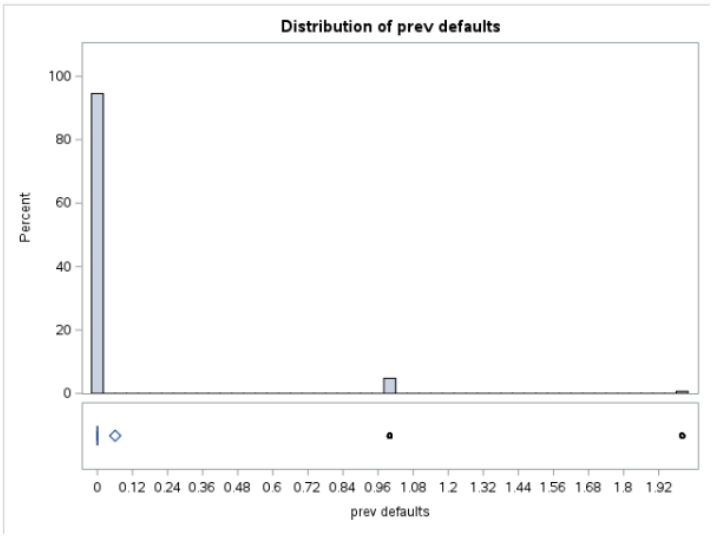
Chart 1: Numeric Variable Distribution Plots

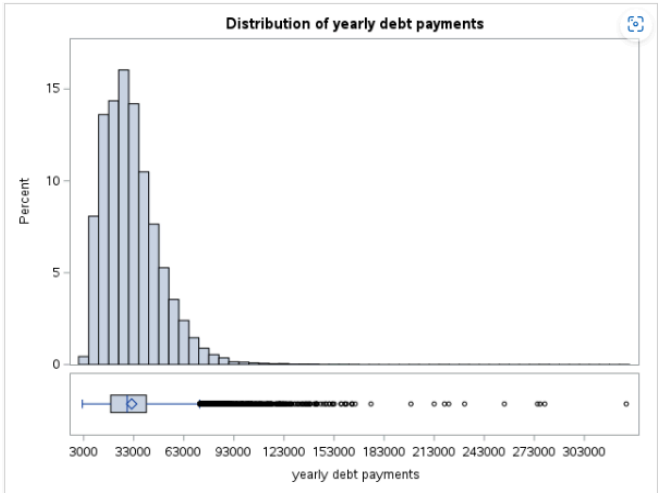


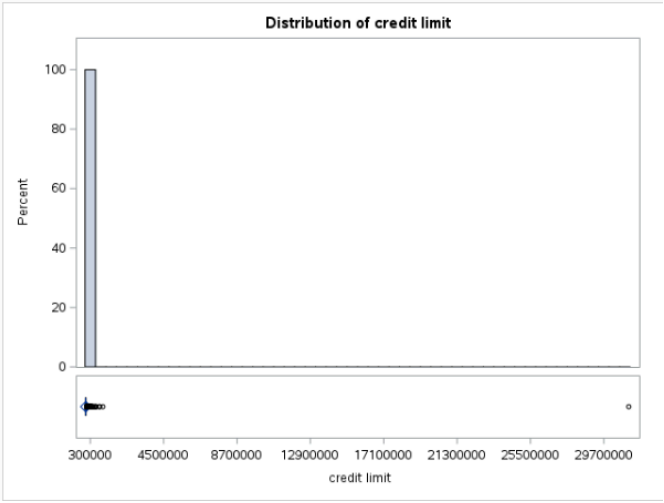

Distribution of no of days employed
Percent
80
60
40
20
0
4000
44000
84000
124000
164000
204000
244000
284000
324000
364000
no of days employed


Distribution of net yearly income
Percent
100
80
60
40
20
0
1500000
22500000
43500000
64500000
85500000
106500000
127500000
net yearly income


Distribution of total family members
Percent
50
40
30
20
10
0
1
1.4
1.8
2.2
2.6
3
3.4
3.8
4.2
4.6
5
5.4
5.8
6.2
6.6
7
7.4
7.8
8.2
8.6
9
9.4
9.8
total family members


Distribution of prev defaults
Percent
100
80
60
40
20
0
0 0.12 0.24 0.36 0.48 0.6 0.72 0.84 0.96 1.08 1.2 1.32 1.44 1.56 1.68 1.8 1.92
prev defaults


Distribution of yearly debt payments
Percent
15
10
5
0
3000 33000 63000 93000 123000 153000 183000 213000 243000 273000 303000
yearly debt payments


Distribution of credit limit
Percent
100
80
60
40
20
0
300000 4500000 8700000 12900000 17100000 21300000 25500000 29700000
credit limit

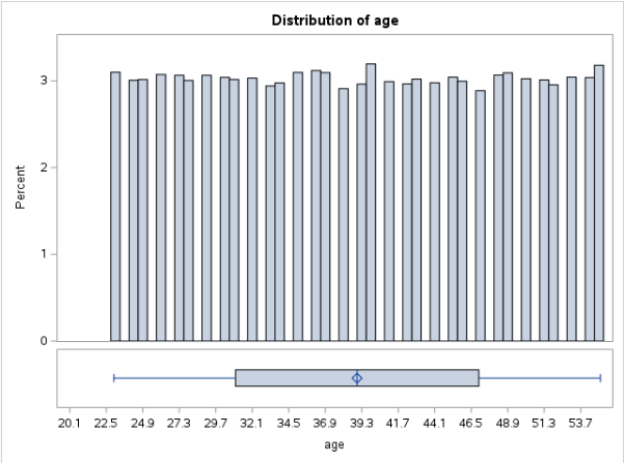
To deal with outliers, we are planning on applying the log transformation to every variable that has skewness higher than three. The only outlier we plan on deleting is the extreme outlier present in both net yearly income and credit limit. This is a single row with a clear data entry mistake, with values of 140 million for net income and 31 million for credit limit.

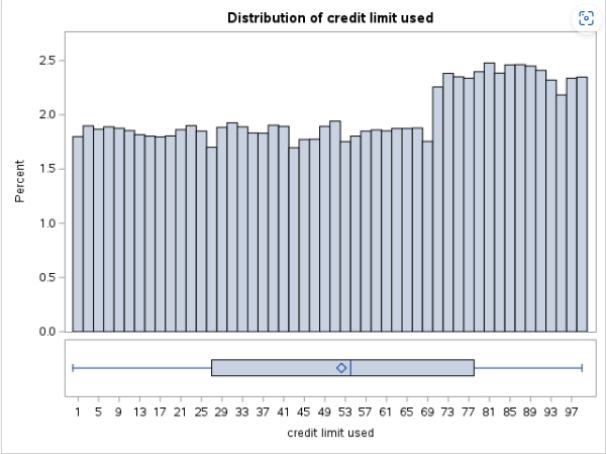
This outlier is the reason why both variables have skewness of more than 200. After deleting the outlier, the variables with skewness higher than three are net_yearly_income, prev_defaults and credit_limit. The Transform Variable node was used in SAS Miner 15.2, and after applying the Log transformation, all the variables had skewness of less than one.

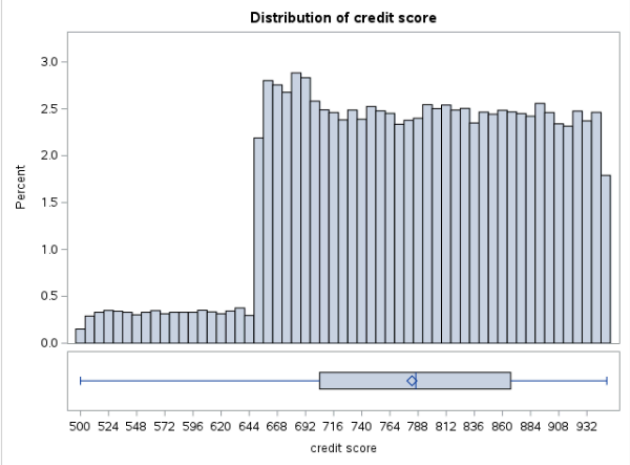
Another variable that needs cleaning is no_of_days_employed. About 20% of the dataset has values of around 364,000. This is not possible because if we divide this number by 365 days, we would get that individuals have worked for almost 1000 years. The option of deleting these rows is not really possible because they are a big chunk of the dataset. For this reason, we opted to apply the Log transformation to this variable as well to lower the effect of these outliers on predicting credit card default. After using the Transform Variable node, the skewness of no_of_days_employed is very close to one from an initial value of almost two.

In the other distribution plots for numeric variables, no outliers are observed.

Chart 2: Numeric Value Distribution Excluding Outlier







Exploratory Data Analysis

Our dataset contains 10 interval variables. The following table contains summary statistics for these variables.

Table 6: Summary Statistics for Interval Variables

Summary Statistics for Interval Variable				
Variable	**Mean**	**Std Dev**	**Minimum**	**Maximum**
Age	38.9931469	9.5439292	23	55
No_Of_Children	0.4206422	0.7241001	0	9
Net_Yearly_Income	197568.26	117030.84	27170.61	4433825.02
No_Of_Days_Employed	67610.77	139324.72	2	365252
Total_Family_Members	2.1580627	0.9115739	1	10
Yearly_Debt_Payments	31796.94	17269.92	2237.47	328112.86
Credit_Limit	42865.6	30169.97	4003.14	1015611.88
Credit_Limit_Used(%)	52.2343664	29.376901	0	99
Credit_Score	782.796898	100.6136518	500	949
Prev_Defaults	0.0607112	0.2646322	0	2

The tables below are the frequency tables for categorical variables. The last frequency table represents our target binary variable, Credit Card Default.

Table 7: Frequency Tables for Categorical Variables

Gender		
	Frequency	**Percent**
F	29,957	65.80
M	15,570	34.20
X	1	0.00

Owns Car		
	Frequency	Percent
0	29,743	66.12
1	15,238	33.88

Owns House		
	Frequency	Percent
0	13,886	30.50
1	31,642	69.50

Migrant Worker		
	Frequency	Percent
0	37,302	82.09
1	8,139	17.91

Default in Last 6 Months		
	Frequency	Percent
0	43,227	94.95
1	2,301	5.05

Occupation Type		
	Frequency	Percent
Accountants	1474	3.24
Cleaning staff	665	1.46
Cooking staff	902	1.98
Core staff	4062	8.92
Drivers	2747	6.03
HR staff	78	0.17
High skill tech staff	1682	3.69
IT staff	66	0.14

Occupation Type		
Laborers	8134	17.87
Low-skill Laborers	336	0.74
Managers	3168	6.96
Medicine staff	1275	2.8
Private service staff	387	0.85
Realty agents	101	0.22
Sales staff	4725	10.38
Secretaries	199	0.44
Security staff	1025	2.25
Unknown	14299	31.41
Waiters/barmen staff	203	0.45

Credit Card Default		
	Frequency	**Percent**
0	41,831	91.88
1	3,697	8.12

Our target variable is credit_card_default. The percentage of individuals defaulting is 8.12%, meaning that the dataset is imbalanced. We plan to address this issue by under sampling the non-defaulting target class to restore equilibrium to the dataset. We will select all the rare primary outcomes (default) and match them with a secondary outcome case (non-default) to create a true 50-50 split of the data.

The existing literature has shown that credit card default is influenced by additional variables and factors like age and gender. According to Li et al. (2019), the research shows that people under the age of 30 are more likely to default on a credit card bill. We looked at our data to see how this holds up. We grouped our data into two categories. Under 30 and over 30 years

old and then compared to see how much more likely someone is to default. We found that in our data the default rate is roughly the same regardless of age. This is different then the previous research and could likely be due to our dataset being from American Express. They typically cater to a wealthier client so this could be a reason why you are less likely to default based on age. However, Jain and Jayabalan (2022) suggest that older people are more likely to default than a younger person. This doesn't appear to be the case in our data. The percentage of customers under age 30 who default (8.29%) is very similar to the percentage of customers over age 30 that defaulted (8.06%). The chi-square statistic of 0.24 confirms this. At a significance level of 0.05, there is no significant difference between the two groups.

Table 8: Credit Card Default By Age Over/Under 30

Credit Card Default by Age Over/Under 30			
Credit Card Default	**Less than 30 Years**	**More than 30 Years**	**Total**
0	10,174	31,657	41,831
	91.71%	91.94%	
1	920	2777	3,697
	8.29%	8.06%	
Total	11,094	34,434	45,528

We also wanted to determine if gender plays a role in the expected likelihood of defaulting. In two particular studies, it is noted that women, especially single women, struggle to pay their bills more than men (Li, 2018; Dunn & Mirzaie, 2022). As explained before, we substituted the only value of X for F.`We found that females are less likely to default. 10.30% of males default compared to only 6.99% of females. The Chi Square test also shows this to be statistically significant as the Chi-Square result is less than .05 (<0.0001).

Table 9: Credit Card Default by Gender

Credit Card Default by Gender			
Credit Card Default	**Female**	**Male**	**Total**
0	27,865	13,966	41,831
	93.01%	89.70%	
1	2,093	1,604	3,697
	6.99%	10.30%	
Total	29,958	15,570	45,528

Multicollinearity is one of the biggest issues when creating predictive models. For this reason, we analyzed the correlation between the interval variables.

Table 10: Correlation Matrix Between Interval Variables

Pearson Correlation Coefficients, N = 44115										
	age	no_of_children	net_yearly_income	no_of_days_employed	total_family_members	yearly_debt_payments	credit_limit	credit_limit_used(%)	credit_score	prev_defaults
age	1.00000	-0.00946	0.00423	0.00172	-0.01185	-0.00244	0.00452	-0.00565	0.00215	0.00101
no_of_children	-0.00946	1.00000	0.00894	-0.24323	0.88001	0.02930	0.00929	0.00831	-0.01500	0.02014
net_yearly_income	0.00423	0.00894	1.00000	-0.02919	0.01043	0.07618	0.99354	0.00281	-0.00988	-0.00472
no_of_days_employed	0.00172	-0.24323	-0.02919	1.00000	-0.23045	-0.10532	-0.02757	-0.01824	0.03615	-0.03586
total_family_members	-0.01185	0.88001	0.01043	-0.23045	1.00000	0.07964	0.01019	0.00250	-0.01126	0.01056
yearly_debt_payments	-0.00244	0.02930	0.07618	-0.10532	0.07964	1.00000	0.07431	-0.00695	0.00536	-0.01267
credit_limit	0.00452	0.00929	0.99354	-0.02757	0.01019	0.07431	1.00000	0.00316	-0.00972	-0.00448
credit_limit_used(%)	-0.00565	0.00831	0.00281	-0.01824	0.00250	-0.00695	0.00316	1.00000	-0.17275	0.25263
credit_score	0.00215	-0.01500	-0.00988	0.03615	-0.01126	0.00536	-0.00972	-0.17275	1.00000	-0.47163
prev_defaults	0.00101	0.02014	-0.00472	-0.03586	0.01056	-0.01267	-0.00448	0.25263	-0.47163	1.00000

The variables with the highest correlation are net_yearly_income and credit_limit with an almost perfect positive correlation (0.99354). The variables with the second highest correlation are no_of_children and total_family_members (0.88001). The variables with the third highest correlation are prev_defaults and credit_limit_used(%). The strongest negative correlation is between credit_score and credit_limit_used(%). This makes sense because when the percentage of credit limit used goes up, the credit score tends to go down.

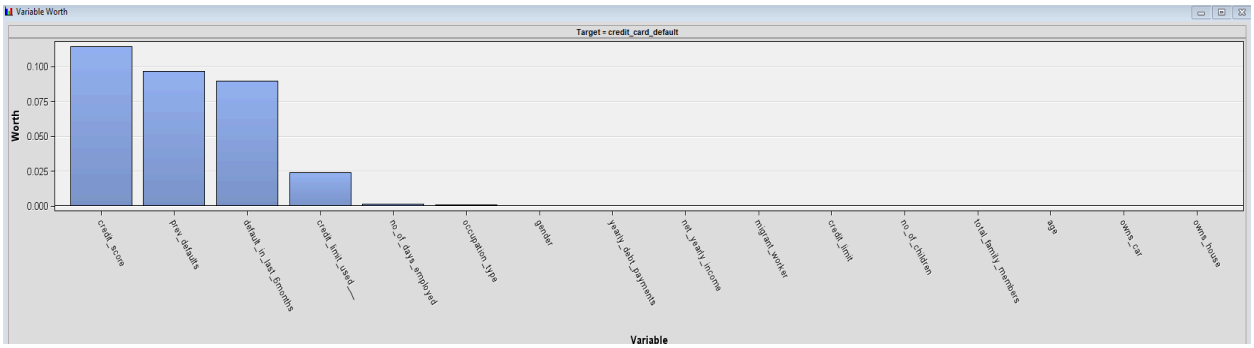
When creating machine learning models, it is important to pay close attention to these variables with high correlation to avoid multicollinearity issues. The models that we are going to use do not have internal methods to deal for multicollinearity. For this reason, we opted to reject credit_limit and no_of_children to avoid having biased predictions.

Table 11: Excluded Variables by Correlation

Excluded Variable		
Variable	**Reason for Exclusion**	**Role**
Credit Limit	High correlation with Net Yearly Income	Rejected
Number of Children	High correlation with Total Family Members	Rejected

According to the variable worth plot in SAS Miner, the most important variables for predicting credit card default are credit_score, prev_defaults, default_in_last_6months and credit_limit_used. Among the variables that are not credit-related, the most important ones are no_of_days_employed and occupation_type.

Chart 3: Variable Importance Chart



After analyzing closely the most important variable, Credit_Score, an interesting insight was discovered. No individual who has a credit score higher than 699 has defaulted in the dataset. This detail clearly shows the importance of credit score in predicting future defaults.

The second and third most important variables, Prev_Defaults and Default_In_Last_6Months were also analyzed in greater detail. In the dataset, every individual who defaulted at least once before has defaulted on its credit card debt. Similarly, among those who defaulted in the last 6 months, all have defaulted on their credit card debt. These two variables are clearly too good at predicting credit card default. There would be the risk of incredibly high odds ratio and standardized estimates that do not provide any practical insight. For these reasons, both variables were rejected in SAS Miner and not utilized to create predictive models.

Table 12: Previous Defaults by Credit Card Default

Previous Defaults by Credit Card Default		
Previous Defaults	**0**	**1**
0	41,831	1228
1	0	2172
2	0	296.00

Table 13: Previous Defaults by Credit Card Default

Default Last 6 Months by Credit Card Default		
Default Last 6 Months	**0**	**1**
0	41,831	1396
1	0	2300

Statistical Analysis

Multiple T-Tests, ANOVA tests and Chi-Square tests were performed to assess the relationships among variables.

T-test: Mean Credit Score vs Default

Table 14: T-Test of Mean Credit Score vs. Default

Variable: credit_score

credit_card_default	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
0		41824	799.0	86.4968	0.4229	650.0	949.0
1		3696	598.9	56.9395	0.9366	500.0	699.0
Diff (1-2)	Pooled		200.1	84.4840	1.4498		
Diff (1-2)	Satterthwaite		200.1		1.0277		

credit_card_default	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0		799.0	798.2	799.9	86.4968	85.9146	87.0870
1		598.9	597.1	600.8	56.9395	55.6705	58.2682
Diff (1-2)	Pooled	200.1	197.3	202.9	84.4840	83.9387	85.0364
Diff (1-2)	Satterthwaite	200.1	198.1	202.1			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	45518	138.02	<.0001
Satterthwaite	Unequal	5336.1	194.71	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	41823	3695	2.31	<.0001

An examination of the test for the equality of variance (folded F test) shows that the p-value is smaller than 0.05. Consequently, there is enough evidence to reject H0, suggesting that the variances of the two groups are statistically different. For this reason, the Satterthwaite was used.

The p-value of the Satterthwaite is less than 0.0001, and, for this reason, we can reject H0. At a significance level of 0.05, there is a statistically significant difference between the average credit score of people who defaulted and those who didn't. It appears that the credit score impacts whether a person defaults on their credit card debt or not. More precisely, people with lower credit scores are more likely to default.

T-test: Mean Credit Limit Used (%) vs Default

Table 15: Mean Credit Limit Used (%) vs. Default

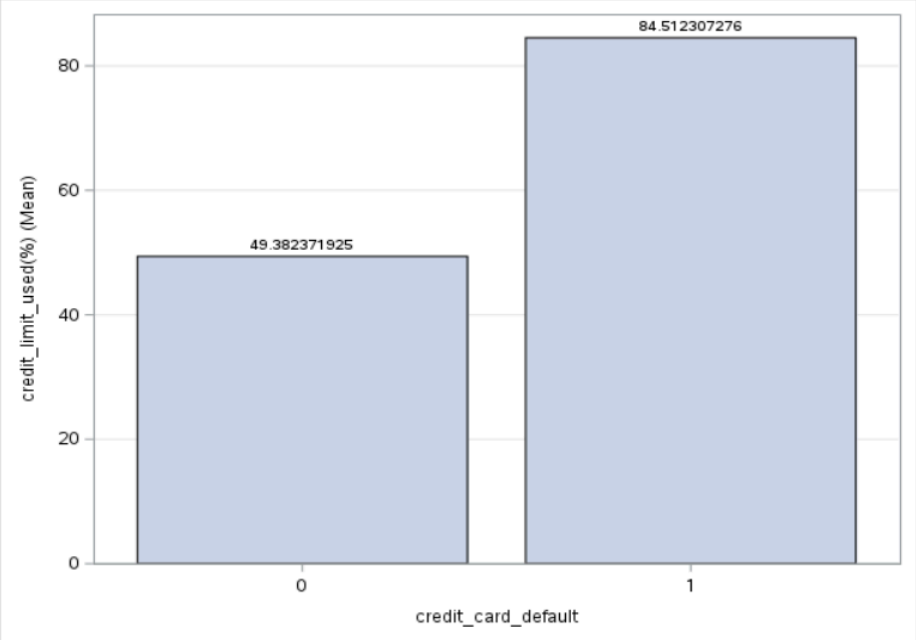
Variable: credit_limit_used(%)

credit_card_default	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
0		41831	49.3824	28.8523	0.1411	0	99.0000
1		3697	84.5123	8.6478	0.1422	70.0000	99.0000
Diff (1-2)	Pooled		-35.1299	27.7658	0.4764		
Diff (1-2)	Satterthwaite		-35.1299		0.2003		

credit_card_default	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0		49.3824	49.1059	49.6589	28.8523	28.6581	29.0491
1		84.5123	84.2335	84.7912	8.6478	8.4551	8.8496
Diff (1-2)	Pooled	-35.1299	-36.0637	-34.1962	27.7658	27.5867	27.9474
Diff (1-2)	Satterthwaite	-35.1299	-35.5226	-34.7373			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	45526	-73.74	<.0001
Satterthwaite	Unequal	13399	-175.37	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	41830	3696	11.13	<.0001



An examination of the test for the equality of variance (folded F test) shows that the p-value is smaller than 0.05. Consequently, there is enough evidence to reject H0, suggesting that the variances of the two groups are statistically different. For this reason, the Satterthwaite was used.

The p-value of the Satterthwaite is less than 0.0001, and, for this reason, we can reject H0. At a significance level of 0.05, there is a statistically significant difference between the average percentage of credit used of people who defaulted and those who didn't. It appears that the percentage of credit limit used impacts whether a person defaults on their credit card debt or not. More precisely, people who default on average have higher percentages of credit limit used. Credit_limit_used (%) is likely to be an important variable when building models to predict credit card default

T-test: Average Credit Score vs. Migrant Worker

Table 16: Average Credit Score vs. Migrant Worker

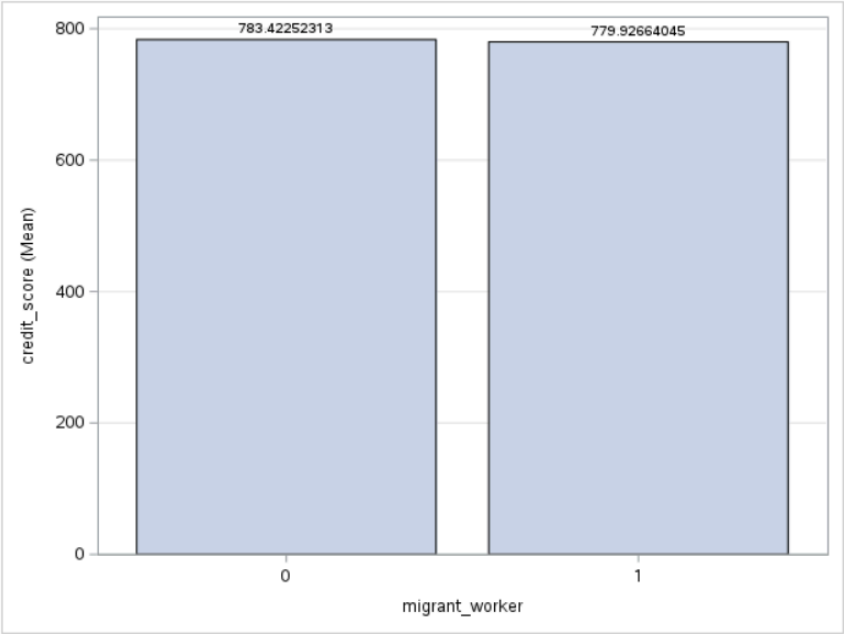
Variable: credit_score

migrant_worker	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
0		37295	783.4	99.9295	0.5175	500.0	949.0
1		8138	779.9	103.8	1.1508	500.0	949.0
Diff (1-2)	Pooled		3.4959	100.6	1.2313		
Diff (1-2)	Satterthwaite		3.4959		1.2618		

migrant_worker	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0		783.4	782.4	784.4	99.9295	99.2175	100.7
1		779.9	777.7	782.2	103.8	102.2	105.4
Diff (1-2)	Pooled	3.4959	1.0826	5.9092	100.6	99.9862	101.3
Diff (1-2)	Satterthwaite	3.4959	1.0226	5.9692			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	45431	2.84	0.0045
Satterthwaite	Unequal	11656	2.77	0.0056

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	8137	37294	1.08	<.0001



An examination of the test for the equality of variance (folded F test) shows that the p-value is smaller than 0.05. Consequently, there is enough evidence to reject H0, suggesting that the variances of the two groups are statistically different. For this reason, the Satterthwaite was used.

The p-value of the Satterthwaite test is 0.0056, and, for this reason, we can reject H0. At a significance level of 0.05, there is a statistically significant difference between the average credit score of people who are migrants and those who are not. More precisely, by looking at the mean values we can see that migrant workers on average have lower credit scores.

T-test: Mean Number of Days Employed vs Default

Table 17: T-Test of Mean Number of Days Employed vs Default

Variable: no_of_days_employed

credit_card_default	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
0		41403	69558.5	140867	692.3	2.0000	365252
1		3661	45583.7	118328	1955.6	13.0000	365252
Diff (1-2)	Pooled		23974.8	139172	2399.7		
Diff (1-2)	Satterthwaite		23974.8		2074.6		

credit_card_default	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0		69558.5	68201.6	70915.4	140867	139914	141833
1		45583.7	41749.4	49417.9	118328	115679	121103
Diff (1-2)	Pooled	23974.8	19271.4	28678.2	139172	138270	140087
Diff (1-2)	Satterthwaite	23974.8	19907.7	28041.9			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	45062	9.99	<.0001
Satterthwaite	Unequal	4628.4	11.56	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	41402	3660	1.42	<.0001

An examination of the test for the equality of variance (folded F test) shows that the p-value is smaller than 0.05. Consequently, there is enough evidence to reject H0, suggesting that the variances of the two groups are statistically different. For this reason, the Satterthwaite was used.

The p-value of the Satterthwaite is less than 0.0001, and, for this reason, we can reject H0. At a significance level of 0.05, there is a statistically significant difference between the average number of days an individual has been employed for those who defaulted and those who didn't. It appears that the number of days a person was employed for impacts whether a person defaults on their credit card debt or not. More precisely, people with a lower number of days employed are more likely to default.

ANOVA test: Occupation Type vs Mean Credit Score

Table 18: ANOVA Test Occupation Type vs Mean Credit Score

Dependent Variable: credit_score

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	18	1366429.7	75912.8	7.52	<.0001
Error	45500	459417151.7	10097.1		
Corrected Total	45518	460783581.3			

R-Square	Coeff Var	Root MSE	credit_score Mean
0.002965	12.83656	100.4842	782.7969

Source	DF	Type I SS	Mean Square	F Value	Pr > F
occupation_type	18	1366429.654	75912.759	7.52	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
occupation_type	18	1366429.654	75912.759	7.52	<.0001

Level of occupation_type	N	credit_score	
		Mean	Std Dev
Accountants	1474	791.821574	97.062532
Cleaning st	665	775.866165	104.155620
Cooking sta	902	775.419069	105.219994
Core staff	4062	785.433776	96.344869
Drivers	2746	774.222141	105.045974
HR staff	78	783.948718	107.349641
High skill	1682	786.895363	98.551887
IT staff	66	808.151515	102.457382
Laborers	8131	777.540401	103.316515
Low-skill L	336	769.532738	113.463751
Managers	3168	784.429609	98.419452
Medicine st	1275	790.730196	100.470882
Private ser	387	772.777778	100.842550
Realty agen	101	768.009901	98.553792
Sales staff	4724	777.888865	102.357544
Secretaries	199	781.371859	98.552465
Security st	1025	781.275122	104.105943
Unknown	14295	787.407695	98.362133
Waiters/bar	203	775.596059	103.174055

The p-value of the ANOVA is less than 0.0001. At a significance level of 0.05, we can reject H0. Consequently, there is a statistical difference between the average credit score and at least one of the occupation types. IT staff has the highest while realty agents have the lowest one.

Chi-Square Test: Default vs. Migrant Worker

Table 19: Chi-Square Test Credit Card Default vs Migrant Worker

Table of credit_card_default by migrant_worker

credit_card_default	migrant_worker 0	1	Total
0	34434 92.31	7316 89.89	41750
1	2868 7.69	823 10.11	3691
Total	37302	8139	45441
Frequency Missing = 87			

Statistics for Table of credit_card_default by migrant_worker

Statistic	DF	Value	Prob
Chi-Square	1	52.5694	<.0001
Likelihood Ratio Chi-Square	1	49.8617	<.0001
Continuity Adj. Chi-Square	1	52.2452	<.0001
Mantel-Haenszel Chi-Square	1	52.5683	<.0001
Phi Coefficient		0.0340	
Contingency Coefficient		0.0340	
Cramer's V		0.0340	

The p-value of the chi-square (<0.0001) is smaller than the significance level of 0.05. Therefore, we can reject the null hypothesis (Ho: whether or not a person defaults is independent of their status as migrant worker). This suggests that the two variables are not independent, and the fact that a person is a migrant worker or not does have a statistically significant influence on whether or not a person defaults. As a consequence, being a migrant worker will likely be an important variable when predicting credit card default.

After conducting data exploration and statistical analysis, we have a clear understanding of the dataset. It is obvious that credit-related values have a much higher predictive power than the other variables. This is great because we are able to build models that can predict credit card defaults very precisely. However, at the same time, the predominance of

credit-related variables can inhibit the influence that other variables have on credit card default. To obviate this problem, we created two different datasets, one containing credit-related variables and the other without. Every predictive model will be estimated twice, once for each dataset. The final data dictionary of the two datasets is as follows:

Table 20: Data Dictionary Final Datasets

Data Dictionary			
Variable	**Measurement Level**	**Role Credit Variables**	**Role Without Credit Variables**
Credit Card Default	Binary	Target	Target
Name	Nominal	Rejected	Rejected
Age	Numeric	Input	Input
Gender	Nominal	Input	Input
Owns Car	Binary	Input	Input
Owns House	Binary	Input	Input
No. of Children	Numeric	Rejected	Rejected
Yearly Income	Numeric	Input	Input
No. of days Employed	Numeric	Input	Input
Occupation	Nominal	Input	Input
Total Family Members	Numeric	Input	Input
Migrant Worker	Binary	Input	Input
Yearly Debt Payment	Numeric	Input	Rejected
Credit Limit	Numeric	Rejected	Rejected
Credit Limit Used	Numeric	Input	Rejected
Credit Score	Numeric	Input	Rejected
Previous Defaults	Numeric	Input	Rejected
Default in last 6 Months	Binary	Input	Rejected
Customer ID	Nominal	ID	ID

Models

Before running the models, we plan to address the issue of having an imbalance dataset by separate sampling the non-defaulting target class to restore equilibrium to the dataset. We will select all the rare primary outcomes (default) and match them with a secondary outcome case (non-default) to create a true 50-50 split of the data. However, employing separate sampling presents certain challenges. The majority of model fit statistics may exhibit bias. For this reason, the cost-sensitive approach is implemented to allocate varying weights to different classes. This ensures that the minority target class (Default) receives a higher weight, consequently incurring a higher misclassification cost. By doing so, bias toward the majority class (non-Default) can be mitigated. We applied the inverse of the class distribution, giving a weight of 12.5 to the minority class. The sample is then divided using the Data Partition node such that 50% of the data would be utilized for training and the other 50% for validation.

Each model is run using both the dataset with credit-related variables as well as the one without. The same model with the two different datasets is reported consequently. This way, comparison between the two models is easier.

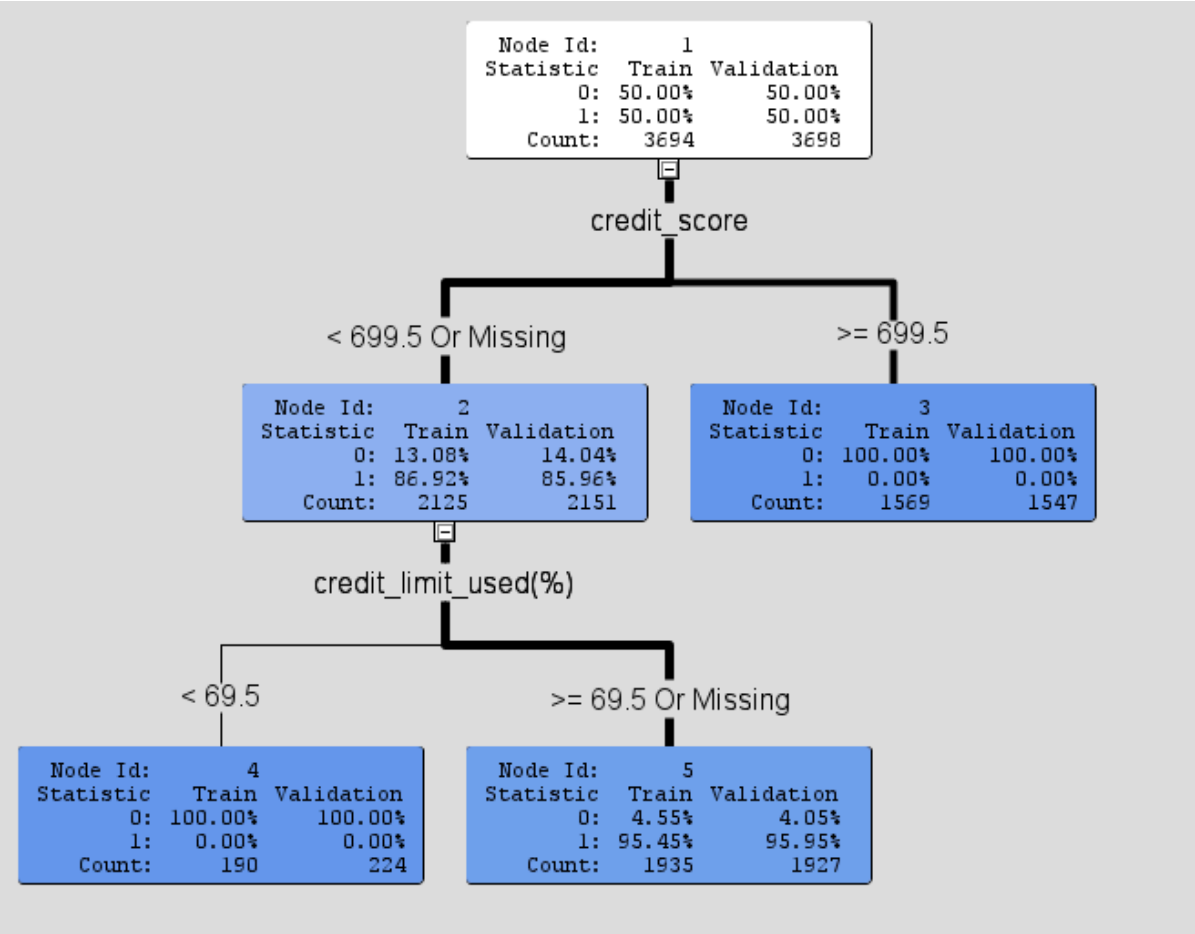
Decision Tree

A decision tree is a versatile machine learning tool used for classification tasks. It recursively divides data based on predictor variables using decision rules. This process creates a tree-like structure where each node represents a decision based on a predictor variable. Decision trees can overfit data, so techniques like pruning are used to prevent this. Despite their simplicity, decision trees can capture complex and non-linear relationships between variables and are easily interpretable.

The decision trees were developed in SAS Miner 15.2 using Decision as the Subtree Assessment Measure. Since we previously defined a profit loss matrix, the tree with the largest profit/smallest loss was selected. Since decision trees can handle missing values, the node was connected directly to the Data Partition node and no previous imputation was applied.

Decision Tree with Credit Variables

Chart 4: Decision Tree Credit Variables



The decision tree has only 3 leaves. The validation average profit is 1.957815. Credit_Score was used for the first split. The value of the first split, 699.5, is a significant one. As already discovered in the exploratory data analysis, no individual with credit scores of 699 or higher has defaulted in the dataset. For this reason, the leaf with Node ID of 3 has correctly identified all the cases as 0 or non-default. The only other variable used for a split is credit_limit_used(%).

Table 21: Decision Tree with Credit Variables Confusion Matrix Validation

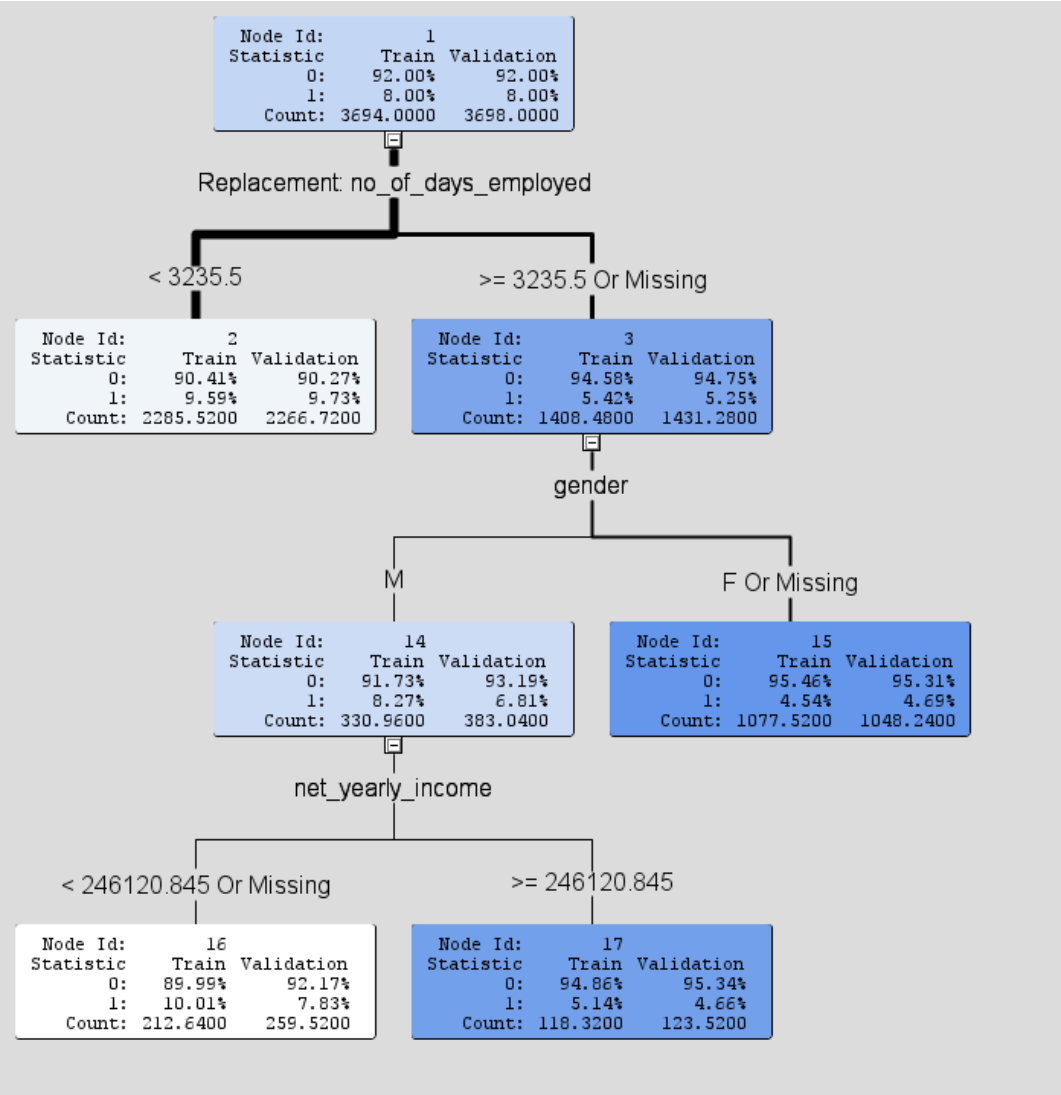
Confusion Matrix			
True Negative	**False Negative**	**False Positive**	**True Positive**
1759	0	88	1847

Table 22: Decision Tree with Credit Variables Variable Importance

Variable Importance			
Variable Name	**Number of Splitting Rules**	**Importance**	**Validation Importance**
Credit_Score	1	1	1
Credit_Limit_Used(%)	1	0.4808	0.5274
*variable not included have an importance of zero			

Decision Tree without Credit-Related Variables

Chart 5: Decision Tree without Credit Variables



Before running the decision tree, a modification was made to the variable No_Of_Days_Employed. The variable has 20% of the values higher or equal to 364,000. This is clearly an error and needs to be addressed before running the models. As mentioned above, for models like regression and neural networks that cannot be estimated with missing values, the log transformation will be applied to the variable. However, since decision trees

don't present this issue, the Replacement node was used and these unusually high variables were replaced with missing values.

The No_Of_Days_Employed with missing values was used for the first split. The validation average profit is 1.14278, much lower than the value for the decision tree created using the credit scores variables. However, not using credit-related variables, we were able to discover that variables like number of days employed, gender of the individual and yearly income are important in predicting credit card defaults.

Table 23: Decision Tree without Credit Variables Confusion Matrix Validation

Confusion Matrix			
True Negative	**False Negative**	**False Positive**	**True Positive**
620	344	1227	1503

Table 24: Decision Tree without Credit Variables Variable Importance

Variable Importance			
Variable Name	**Number of Splitting Rules**	**Importance**	**Validation Importance**
No_Of_Days_Employed	1	1	1
Gender	1	0.5189	0.1328
Net_Yearly_Income	1	0.3463	0.0455
*variable not included have an importance of zero			

Logistic Regression

Logistic regression is a fundamental statistical method used for binary classification tasks. Logistic regression is a parametric model that assumes an association between inputs and target. It models the probability of the target belonging to a particular class based on predictor variables. Unlike linear regression, logistic regression uses a logistic function to constrain the predicted values between 0 and 1, representing probabilities.

Unlike decision trees, linear regression cannot use training data cases with missing values, and cannot score cases with missing values. For this reason, before running the model, the Impute node was used to replace missing values as explained in Section 6. Additionally, regressions cannot handle skewed distribution and result in biased estimates. For this reason, the Transform Variable node was used and the log transformation applied to variables with highly skewed distributions.

Stepwise Logistic Regression with Credit Variables

The first logistic regression created selected the input variables using stepwise selection. We achieved this by switching the Model Selection to Stepwise in SAS Miner. The selection criterion was set to Validation Profit/Loss.

The variables included in the final model are: Credit_Limit_Used(%) and Credit_Score. The average profit is 1.918875.

Table 25: Stepwise Regression with Credit Variables Estimates

Stepwise Regression Estimates		
Parameter	**Estimate**	**Standardized Estimate**
Credit_Limit_Used	0.1038	1.5507
Credit_Score	-0.0664	-4.5446

Credit_Score has the highest absolute standardized parameter. Credit_score is roughly three times more important than Credit_Limit_Used(%).

Table 26: Stepwise Regression with Credit Variables Odds Ratio

Stepwise Regression Odds Ratio	
Effect	**Odds Ratio**
Credit_Limit_Used	1.109
Credit_Score	0.936

For Credit_Limit_Used(%), the odds ratio estimate equals 1.109. This means that for each additional percentage of credit limit used, the odds of defaulting change by a factor of 1.109, a 10.9% increase. For Credit_Score, the odds ratio estimate equals 0.936. This means

that for each additional credit score point, the odds of defaulting change by a factor of 0.936, a 6.4% decrease. Both odds ratios agree with what we would expect.

Table 27: Stepwise Regression with Credit Variables Confusion Matrix

Confusion Matrix			
True Negative	**False Negative**	**False Positive**	**True Positive**
1755	59	92	1788

Stepwise Logistic Regression without Credit Variables

The variables included in the final model are: LOG_No_Of_Days_Employed, Owns_Car, LOG_Net_Yearly_Income, Gender, Occupation_Type. The average profit is 1.159546.

Table 28: Stepwise Regression without Credit Variables Estimates

Stepwise Regression Estimates		
Parameter	**Estimate**	**Standardized Estimate**
LOG_No_Of_Days_Employed	-0.0989	-0.122
LOG_Net_Yearly_Income	-0.1709	-0.0459

LOG_No_Of_Days_Employed has the highest absolute standardized parameter. This variable is roughly 2.5 times more important than LOG_Net_Yearly_Income.

Table 29: Stepwise Regression Without Credit Variables Odds Ratio

Odds Ratio			
Effect			**Odds Ratio**
LOG No_Of_Days_Employed			0.906
Owns_Car	0 vs 1		1.243
LOG_Net_Yearly_Income			0.843
Gender	F vs M		0.623
Occupation_Type	Accountants vs. Waiters		0.571
Occupation_Type	Cleaning Staff vs. Waiters		1.560
Occupation_Type	Cooking Staff vs. Waiters		1.188
Occupation_Type	Core Staff vs. Waiters		0.752
Occupation_Type	Drivers vs. Waiters		0.921
Occupation_Type	HR Staff vs. Waiters		0.492
Occupation_Type	High Skill vs. Waiters		0.452
Occupation_Type	IT Staff vs. Waiters		0.221

Odds Ratio		
Occupation_Type	Laborers vs. Waiters	0.877
Occupation_Type	Low Skill vs. Waiters	1.268
Occupation_Type	Managers vs. Waiters	0.639
Occupation_Type	Medicine Staff vs. Waiters	0.828
Occupation_Type	Private Service vs. Waiters	1.011
Occupation_Type	Realty Agent vs. Waiters	4.489
Occupation_Type	Sales Staff vs. Waiters	1.013
Occupation_Type	Secretaries vs. Waiters	0.532
Occupation_Type	Security Staff vs. Waiters	0.852
Occupation_Type	Unknown vs. Waiters	0.800

For LOG No_Of_Days_Employed, the odds ratio estimate equals 0.906. This means that for each additional unit, the odds of defaulting change by a factor of 0.906, a 9.4% decrease. For LOG_Net_Yearly_Income, the odds ratio estimate equals 0.843. This means that for each additional unit, the odds of defaulting change by a factor of 0.843, a 15.7% decrease. Both odds ratios agree with what we would expect.

For Owns_Car, the odds ratio (0 versus 1) estimate equals 1.243. This means that for cases of individuals who don't own a car, the odds of defaulting are 1.243 times higher than the odds of defaulting for individuals who own a car. For Gender, the odds ratio (M vs F) estimate equals 1.605. This means that for cases with male individuals, the odds of defaulting are 1.605 times higher than the odds of defaulting for female individuals. This agrees with what was already predicted by the decision tree, and would confirm the findings of Li et al. (2019).

Looking at the odds ratio for Occupation_Type can reveal some interesting facts. For instance, for cases with an occupation of Waiters, the odds of defaulting are 4.525 times

higher than the odds of defaulting for cases with an occupation of IT Staff. The odds of defaulting for cases with an occupation of Realty Agent are 4.489 higher than the odds of cases with occupation of Waiters. This might be due to the fact that Realty Agents have a much less stable income since they get paid mainly in commissions.

Table 30: Stepwise Regression without Credit Variables Confusion Matrix

Confusion Matrix			
True Negative	**False Negative**	**False Positive**	**True Positive**
1048	711	799	1136

Polynomial Regression with Credit Variables

One of the disadvantages of using a logistic regression is that it only accounts for linear relationships. If the true association is more complicated, such an assumption might result in biased predictions. The Regression Node in SAS Miner 15.2 has a Term Editor that allows us to add polynomial terms. Since only two variables were used in the above regression, we included a quadratic term for each of the two variables and one interaction term between the two (Credit_Score*Credit_Limit_Used(%)). Perhaps, a change in the percentage of credit limit used is affected by an individual's credit score.

The regression model only employed one of the three polynomial terms, the quadratic term of Credit_Limit_Used(%), together with Credit_Score and Credit_Limit_Used(%). The validation average profit for credit card default is 1.939427.

The main disadvantage of adding polynomial terms to the logistic regression is interpretability. Standardized estimates and odds ratios are not calculated for the quadratic term of Credit_Limit_Used(%), making it hard to understand how this new variable influences credit card default.

Table 31: Polynomial Regression with Credit Variables Confusion Matrix

Confusion Matrix			
True Negative	**False Negative**	**False Positive**	**True Positive**
1752	23	95	1824

Polynomial Regression Without Credit Variables

A similar approach was utilized to build a polynomial regression with the dataset not including the credit related variables. Since only two numeric variables were selected by the stepwise regression, we included a quadratic term for each of the two variables and one interaction term between the two variables (LOG_No_Of_Days_Employed*LOG_Net_Yearly_Income).

The regression model selected the interaction term between LOG_No_Of_Days_Employed and LOG_Net_Yearly_Income. This might suggest that perhaps a change in the number of consecutives days an individual has been employed influences the net yearly income. The other variables selected by the model are Owns_Car, Gender and Occupation_type. The validation average profit for credit card default is 1.159546, the same as the stepwise regression without polynomial terms.

Table 32: Polynomial Regression without Credit Variables Confusion Matrix

Confusion Matrix			
True Negative	**False Negative**	**False Positive**	**True Positive**
1044	724	803	1123

Variable Selection Logistic Regression with Credit Variables

SAS Miner has multiple other ways to select variables to be used in the regression model. One of these is the Variable Selection Node, which looks for values that maximize the model R-square value. Additionally, the AOV16 Variables, Group Variables and Interactions parameters were all set to Yes.

Table 33: Variable Selection Regression with Credit Variables

Variable Selection	
Variable Name	**Role**
AOV16 LOG_No_Of_Days_Employed	Input
AOV16 Yearly_Debt_Payments	Input
AOV16 LOG Net_Yearly_Income	Input
AOV16 Credit_Limit_Used	Input
AOV16 Credit_Score	Input
GI Occupation_Type and Owns_House	Input
GI LOG_No_Of_Days_Employed and Occupation Type	Input

The validation average profit is 1.940508. Similarly to the polynomial terms, the variables created by the Variable Selection Node are hard to interpret and cannot be used for practical insights.

Table 34: Variable Selection with Credit Variables Confusion Matrix

Confusion Matrix			
True Negative	**False Negative**	**False Positive**	**True Positive**
1762	17	85	1830

Variable Selection Logistic Regression Without Credit Variables

The Variable Node was utilized to select the variables for the dataset not containing the credit related variables as well. Also in this case, the AOV16 Variables, Group Variables and Interactions parameters were all set to Yes. The validation average profit is 1.142239.

Table 35: Variable Selection Regression without Credit Variables

Variable Selection	
Variable Name	**Role**
AOV16 LOG_No_Of_Days_Employed	Input
AOV16 LOG Net_Yearly_Income	Input
GI Migrant Worker and Occupation_Type	Input
GI LOG_No_Of_Days_Employed and Occupation_Type	Input
GI Total_Family_Member and Occupation_Type	Input
GI Owns_Car and Occupation_Type	Input
GI Gender and Occupation_Type	Input
GI Occupation_Type and Owns_House	Input
G Occupation_Type	Input

Table 36: Variable Selection Without Credit Variables Confusion Matrix

Confusion Matrix			
True Negative	**False Negative**	**False Positive**	**True Positive**
1116	719	731	1128

Partial Least Square Logistic Regression with Credit Variables

Another way to perform input selection is to utilize partial least squares, a merging of multiple and principal components regression. The goal in partial least squares is to have linear combinations of the inputs that account for both the target and the inputs. The variables selected for the dataset with credit related values are listed in the table below.

Table 37: Partial Least Square Variable Selection with Credit Variables

Variable Selection PLS	
Variable Name	**Role**
LOG_Prev_Defaults	Input
Credit_Limit_Used	Input
Credit_Score	Input

The partial least squares selected the same two variables as the stepwise regression, Credit_Limit_Used(%) and Credit_Score. For this reason, the validation average profit is the same (1.918875), as well as the confusion matrix.

Partial Least Square Logistic Regression without Credit Variables

In this case, the Partial Least Squares Node selected different variables from the ones chosen by the stepwise regression. Migrant_Worker, one of the variables we are most interested in our project, was selected for the first time by one of the models. The validation average profit is 1.143321

Table 38: Partial Least Square Variable Selection Without Credit Variables

Variable Selection PLS	
Variable Name	**Role**
LOG_No_Of_Days_Employed	Input
Migrant_Worker	Input
Gender	Input
Occupation_Type	Input

Table 39: Partial Least Square Odds Ratio without Credit Variables

Odds Ratio			
Effect			**Odds Ratio**
LOG No_Of_Days_Employed			0.914
Migrant_Worker	0 vs 1		0.891
Gender	F vs M		0.687
Occupation_Type	Accountants vs. Waiters		0.555
Occupation_Type	Cleaning Staff vs. Waiters		1.673
Occupation_Type	Cooking Staff vs. Waiters		1.25
Occupation_Type	Core Staff vs. Waiters		0.743
Occupation_Type	Drivers vs. Waiters		0.875
Occupation_Type	HR Staff vs. Waiters		0.532
Occupation_Type	High Skill vs. Waiters		0.446
Occupation_Type	IT Staff vs. Waiters		0.223

Odds Ratio		
Occupation_Type	Laborers vs. Waiters	0.885
Occupation_Type	Low Skill vs. Waiters	1.409
Occupation_Type	Managers vs. Waiters	0.591
Occupation_Type	Medicine Staff vs. Waiters	0.854
Occupation_Type	Private Service vs. Waiters	1.002
Occupation_Type	Realty Agent vs. Waiters	4.378
Occupation_Type	Sales Staff vs. Waiters	1.024
Occupation_Type	Secretaries vs. Waiters	0.486
Occupation_Type	Security Staff vs. Waiters	0.884
Occupation_Type	Unknown vs. Waiters	0.809

For Migrant_Worker, the odds ratio (1 versus 0) estimate equals 1.122. This means that for cases with a migrant individual, the odds of defaulting are 1.122 times higher than the odds of defaulting for individuals who are not migrant workers. The remaining odds ratios are very similar to the ones estimated in the stepwise logistic regression and don't require further analysis.

Table 40: Partial Least Square without Credit Variables Confusion Matrix

Confusion Matrix			
True Negative	**False Negative**	**False Positive**	**True Positive**
1021	729	826	1118

Selection Tree Regression with Credit Variables

One final way input selection was performed is by using a decision tree. To make sure enough inputs were selected, the Number of Surrogate Rules was changed to 1 and Subtree Method to Largest in SAS Miner. The validation average profit of the regression run using the variables listed below is 1.916171.

Table 41: Selection Tree Variable with Credit Variables

Variable Selection Tree		
Variable Name	**Importance**	**Role**
Credit_Limit_Used	1	Input
Credit_Score	0.9817	Input
Yearly_Debt_Payments	0.447	Input
LOG_Net_Yearly_Income	0.1049	Input

Table 42: Selection Tree Estimates with Credit Variables

Selection Tree Estimates		
Parameter	**Estimate**	**Standardized Estimate**
Yearly_Debt_Payments	0.000026	0.024
LOG_Net_Yearly_Income	-0.148	-0.0398
Credit_Limit_Used	0.1037	1.5496
Credit_Score	-0.0665	-4.5471

Among the variables, Credit_Score has the highest absolute standardized parameter. Credit_Score is roughly three times more important than Credit_Limit_Used(%). The second most important numeric variable is Credit_Limit_Used(%). Credit_Limit_Used(%) is respectively 65 times more important than Yearly_Debt_Payments and 39 times more

important than LOG_Net_Yearly_Income. This shows one more time how much better credit related variables are at predicting variable credit card defaults, and explains why the models were also run without them.

Table 43: Selection Tree Odds Ratios with Credit Variables

Selection Tree Odds Ratio	
Effect	**Odds Ratio**
Yearly_Debt_Payments	1
LOG_Net_Yearly_Income	0.862
Credit_Limit_Used	1.109
Credit_Score	0.936

The odds ratio of Yearly_Debt_Payments is exactly 1. This means that an additional dollar of yearly debt payments doesn't increase or decrease the odds of defaulting. This basically means that there is no association between the variable Yearly_Debt_Payments and the target variable.

Table 44: Selection Tree Credit Variables Confusion Matrix

Confusion Matrix			
True Negative	**False Negative**	**False Positive**	**True Positive**
1757	60	90	1787

Selection Tree Regression without Credit Variables

A Decision Tree Node was also used to find important variables in the dataset without credit related variables. The variables selected are below and the validation average profit of the regression run using these variables is 1.162791.

Table 45: Selection Tree Variable without Credit Variables

Variable Selection Tree		
Variable Name	**Importance**	**Role**
LOG_No_Of_Days_Employed	1	Input
Occupation_Type	0.9817	Input
Gender	0.447	Input
LOG_Net_Yearly_Income	0.1049	Input
Total_Family_Members	0.447	Input
Owns_House	0.1049	Input
Owns_Car	0.447	Input
Migrant_Worker	0.1049	Input

Table 46: Selection Tree Estimates without Credit Variables

Selection Tree Estimates		
Parameter	**Estimate**	**Standardized Estimate**
LOG_No_Of_Days_Employed	-0.0925	-0.1141
Total_Family_Members	0.0411	0.021
LOG_Net_Yearly_Income	-0.17	-0.0457

Among the variables, LOG_No_Of_Days_Employed has the highest absolute standardized parameter. LOG_No_Of_Days_Employed is roughly five times more important than Total_Family_Members and 2.5 times more important than LOG_Net_Yearly_Income.

Table 47: Selection Tree Odds Ratios Without Credit Variables

Odds Ratio			
Effect			**Odds Ratio**
LOG No_Of_Days_Employed			0.912
Migrant_Worker	0 vs 1		0.89
Owns_Car	0 vs 1		1.256
Total_Family_Members			1.042
LOG_Net_Yearly_Income			0.844
Gender	F vs M		0.627
Owns_House	0 vs 1		1.016
Occupation_Type	Accountants vs. Waiters		0.573
Occupation_Type	Cleaning Staff vs. Waiters		1.578
Occupation_Type	Cooking Staff vs. Waiters		1.19
Occupation_Type	Core Staff vs. Waiters		0.75
Occupation_Type	Drivers vs. Waiters		0.926
Occupation_Type	HR Staff vs. Waiters		0.496
Occupation_Type	High Skill vs. Waiters		0.453
Occupation_Type	IT Staff vs. Waiters		0.193
Occupation_Type	Laborers vs. Waiters		0.874
Occupation_Type	Low Skill vs. Waiters		1.26
Occupation_Type	Managers vs. Waiters		0.64
Occupation_Type	Medicine Staff vs. Waiters		0.831
Occupation_Type	Private Service vs. Waiters		1.01
Occupation_Type	Realty Agent vs. Waiters		4.644
Occupation_Type	Sales Staff vs. Waiters		1.018
Occupation_Type	Secretaries vs. Waiters		0.533
Occupation_Type	Security Staff vs. Waiters		0.851
Occupation_Type	Unknown vs. Waiters		0.805

For Total_Family_Members, the odds ratio estimate equals 1.042. This means that for each additional member in the family, the odds of defaulting change by a factor of 1.042, a 4.2% increase. For Owns_House, the odds ratio (0 versus 1) estimate equals 1.016. This means that for cases with individuals that don't own a house, the odds of defaulting are 1.016 times higher than for individuals who own a house.

Table 48: Selection Tree Without Credit Variables Confusion Matrix

Confusion Matrix			
True Negative	**False Negative**	**False Positive**	**True Positive**
1039	706	808	1141

Neural Network

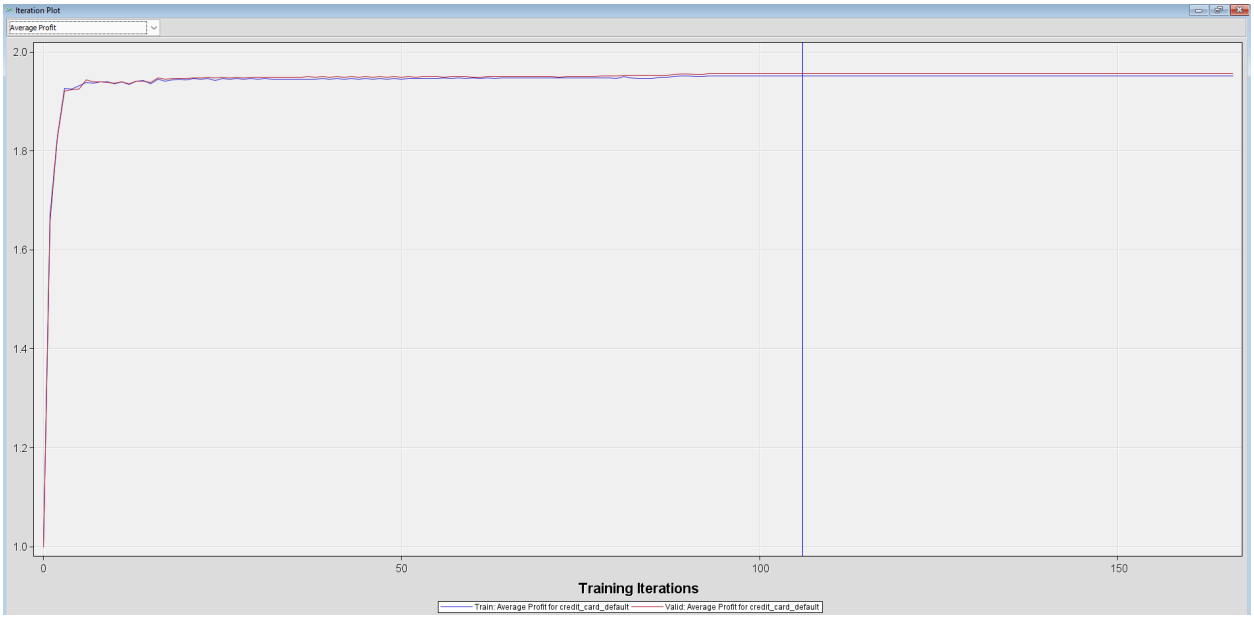
Neural networks are sophisticated models inspired by the brain's structure. They consist of interconnected nodes organized into layers: input, hidden, and output. Neural networks are a natural extension of a regression model. Unlike regression models, neural networks are more flexible and can model any type of association between inputs and target variables.

However, one of the issues of neural networks is that they don’t have internal processes to select input variables in SAS Miner. For this reason, we utilized external processes to select useful inputs.

Stepwise Regression Neural Network with Credit Variables

The first neural network was created using the variables selected using stepwise regression. The variables are: Credit_Limit_Used(%) and Credit_Score. The Model Selection Criterion was set to Profit/Loss, the Maximum Iterations to 300, and the number of Hidden Units to 3. Preliminary training was disabled.

Chart 6: Stepwise Neural Network 3HU with Credit Variables Iteration Plot



The model estimated 13 parameters. Iteration 106 was selected. The validation average profit is 1.957274. Unfortunately, unlike regression models, neural networks can be hard to interpret and there is no easy solution.

Table 49: Stepwise Network 3HU with Credit Variables Confusion Matrix

Confusion Matrix			
True Negative	**False Negative**	**False Positive**	**True Positive**
1758	0	89	1847

We built another neural network model using the variables selected by the stepwise regression. However this time the number of hidden units was increased to 6 to increase network flexibility. The model estimated 25 parameters. Iteration 51 was selected. The validation average profit is basically the same and is 1.957275.

Chart 7: Stepwise Neural Network 6HU with Credit Variables Iteration Plot

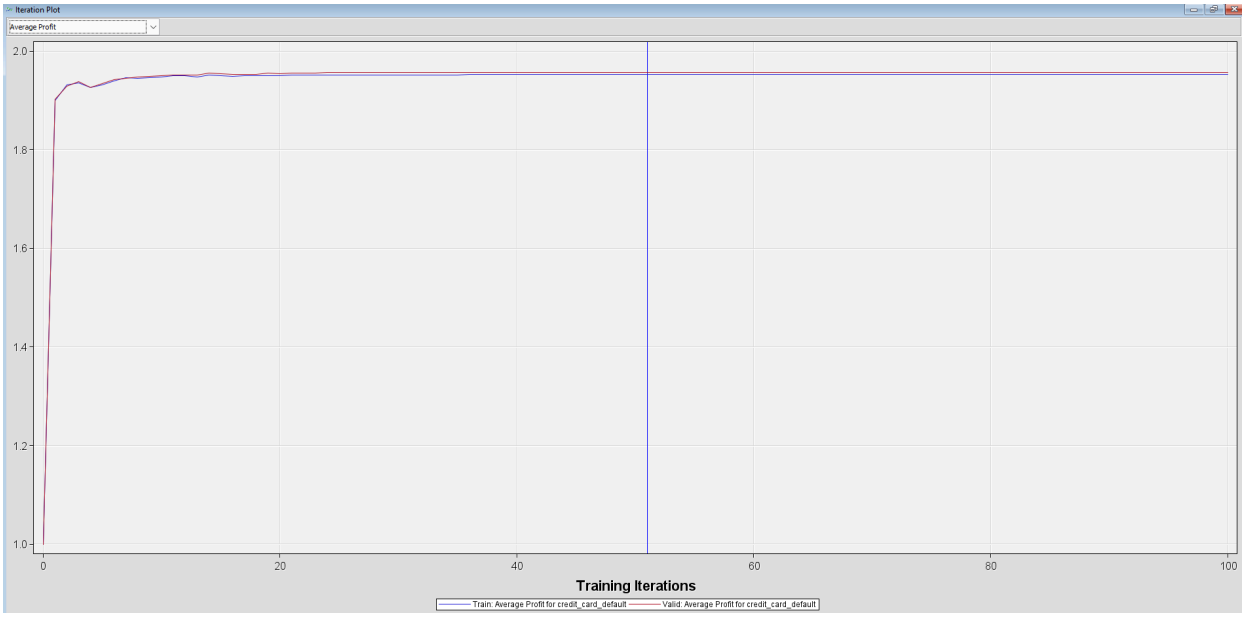


Table 50: Stepwise Neural Network 6HU Credit Variables Confusion Matrix

Confusion Matrix			
True Negative	**False Negative**	**False Positive**	**True Positive**
1759	0	88	1847

Stepwise Regression Neural Network without Credit Variables

The neural network was created using the variables selected using stepwise regression. The variables are: LOG_No_Of_Days_Employed, Owns_Car, LOG_Net_Yearly_Income, Gender, Occupation_Type. The Model Selection Criterion was set to Profit/Loss, the Maximum Iterations to 300, and the number of Hidden Units to 3. Preliminary training was disabled. The model estimated 73 parameters. Iteration 24 was selected. The validation average profit is 1.164413.

Chart 8: Stepwise Neural Network 3HU without Credit Variables Iteration Plot

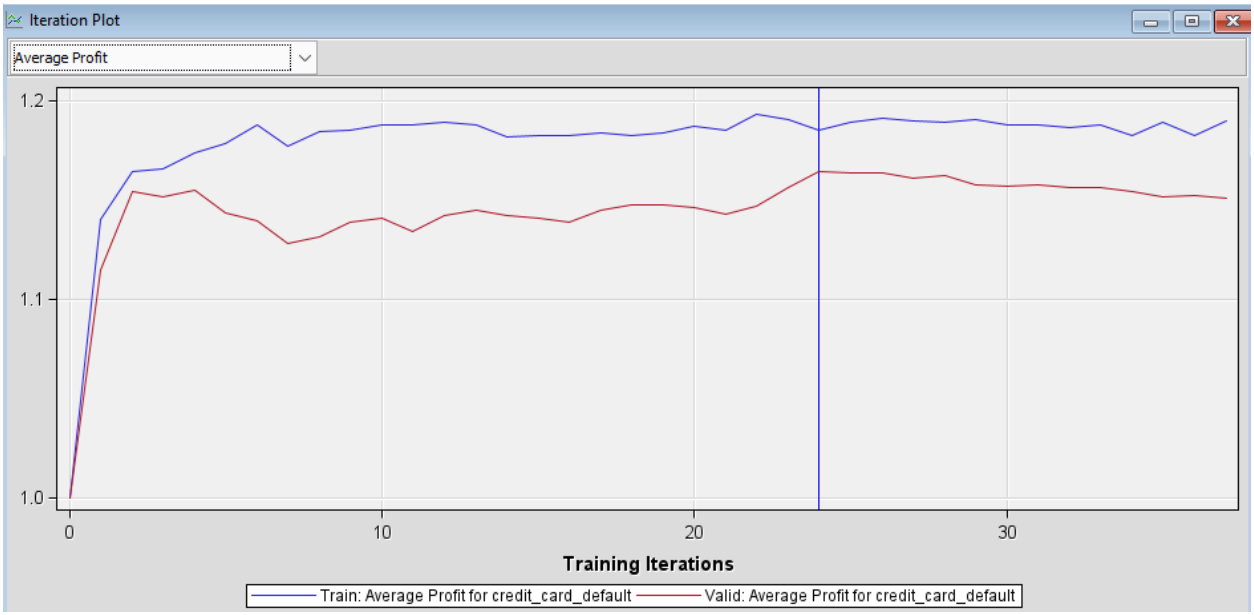


Table 51: Stepwise Network 3HUwithout Credit Variables Confusion Matrix

Confusion Matrix			
True Negative	**False Negative**	**False Positive**	**True Positive**
1104	762	743	1085

Also in this case, we reran the model model increasing the number of hidden units to 6 to increase network flexibility. The model estimated 145 parameters. Iteration 10 was selected.

The validation average profit is better than the model with three hidden units and is 1.170903.

Chart 9: Stepwise Neural Network 6HU without Credit Variables Iteration Plot

Table 52: Stepwise Neural Network 6HU Without Credit Variables Confusion Matrix

Confusion Matrix			
True Negative	**False Negative**	**False Positive**	**True Positive**
1057	730	790	1117

Autoneural Network with Credit Variables

The AutoNeural node offers a way to find the optimal hidden units counts without having to manually change it like we did for the neural network models above. Also in this case, the Node was run using the variables selected using the stepwise regression. Three hidden units were selected. The average profit is 1.954029.

Chart 10: Autoneural with Credit Variables Training Step

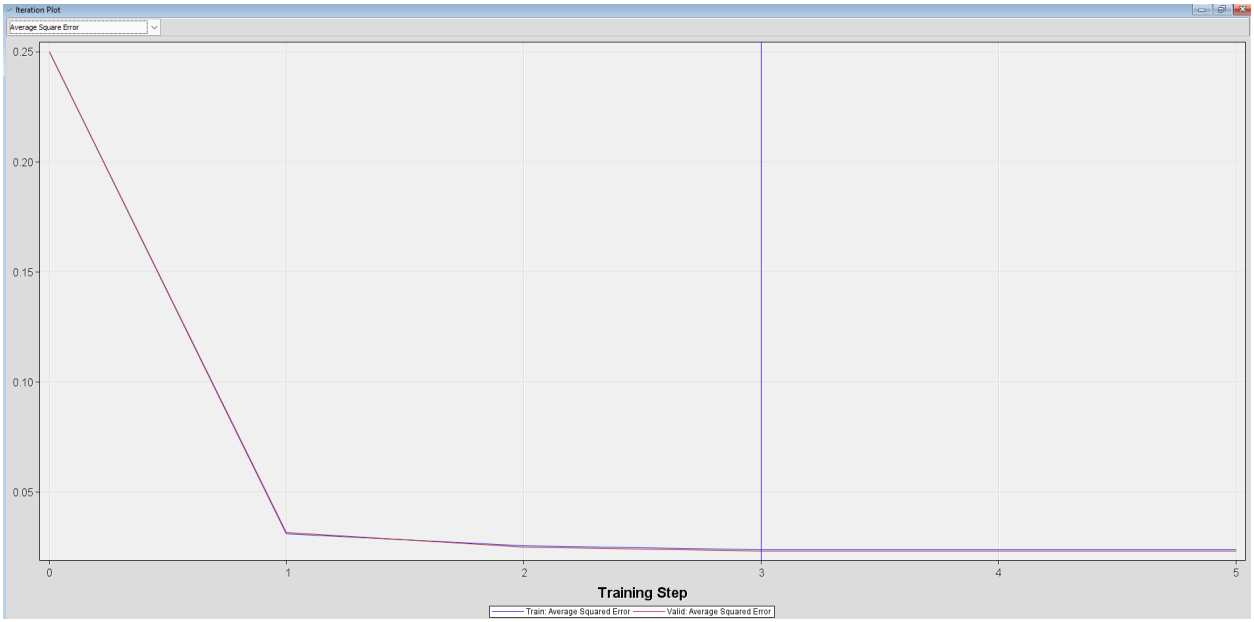


Table 53: Autoneural with Credit Variables Confusion Matrix

Confusion Matrix			
True Negative	**False Negative**	**False Positive**	**True Positive**
1758	2	89	1845

Autoneural Network without Credit Variables

For the dataset without credit related variables, only one hidden unit was selected. The average profit for the validation dataset is 1.163872.

Chart 11: Autoneural without Credit Variables Training Step

Table 54: Autoneural Credit Variables Confusion Matrix

Confusion Matrix			
True Negative	**False Negative**	**False Positive**	**True Positive**
1067	750	780	1097

Variable Selection Neural Network with Credit Variables

The following neural network was created using the variables previously selected by the Variable Selection Node. For the dataset where the credit related variables were not rejected, the variables selected were AOV16 LOG_No_Of_Days_Employed, AOV16 Yearly_Debt_Payments, AOV16 LOG Net_Yearly_Income, AOV16 Credit_Limit_Used, AOV16 Credit_Score, GI Occupation_Type and Owns_House and GI LOG_No_Of_Days_Employed and Occupation Type. The number of parameters estimated is 241. Iteration number 108 was selected. The validation average profit is 1.840454.

Chart 12: Variable Selection Neural Network with Credit Variables Iteration Plot

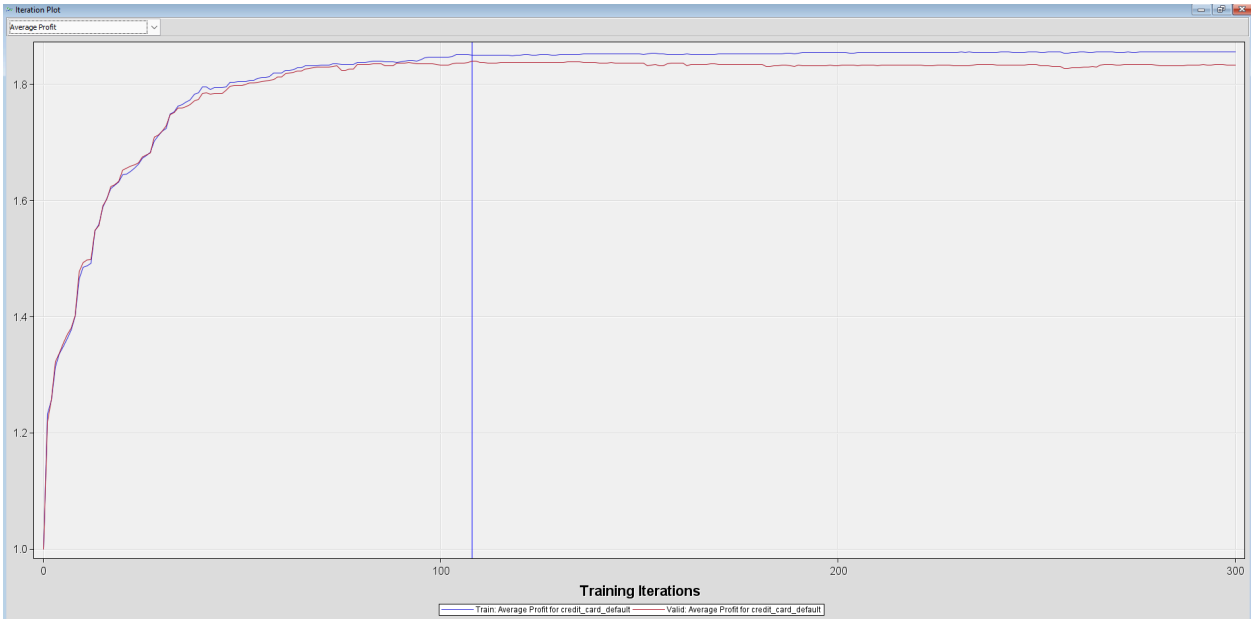


Table 55: Variable Selection Neural Network Credit Variables Confusion Matrix

Confusion Matrix			
True Negative	**False Negative**	**False Positive**	**True Positive**
1621	50	226	1797

Variable Selection Neural Network without Credit Variables

For the dataset where the credit related variables were rejected, the variables selected by the Variable Selection Node were AOV16 LOG_No_Of_Days_Employed, AOV16 LOG Net_Yearly_Income, GI Migrant Worker and Occupation_Type, GI LOG_No_Of_Days_Employed and Occupation_Type, GI Total_Family_Member and Occupation_Type, GI Owns_Car and Occupation_Type, GI Gender and Occupation_Type, GI Occupation_Type and Owns_House and G Occupation_Type. The number of parameters estimated is 274. Iteration number 30 was selected. The validation average profit is 1.16225.

Chart 13: Variable Selection Neural Network without Credit Default Iteration Plot

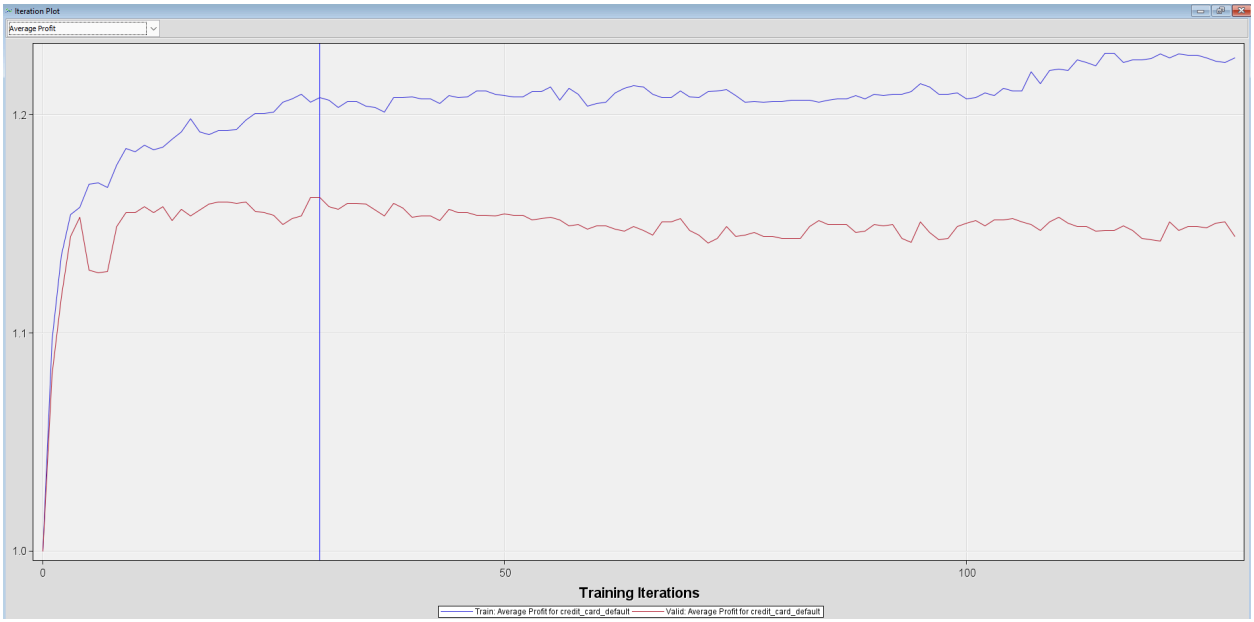


Table 56: Variable Selection Neural Network Without Credit Variables Confusion Matrix

Confusion Matrix			
True Negative	**False Negative**	**False Positive**	**True Positive**
1103	719	744	1128

Partial Least Squares Neural Network without Credit Variables

For the dataset containing credit related variables, the variables selected by the Partial Least Squares are the same as the stepwise regression (Credit_Score and Credit_Limit_Used). For this reason, the validation average profit would be the same as the Stepwise Neural Network (1.951814).

On the other hand, the variables selected for the dataset with rejected credit related variables are different and are LOG_No_Of_Days_Employed, Migrant_Worker, Gender, Occupation_Type. The model estimated 70 parameters. Iteration 6 was selected. The validation average profit is 1.153056.

Chart 14: PLS Neural Network without Credit Default Iteration Plot

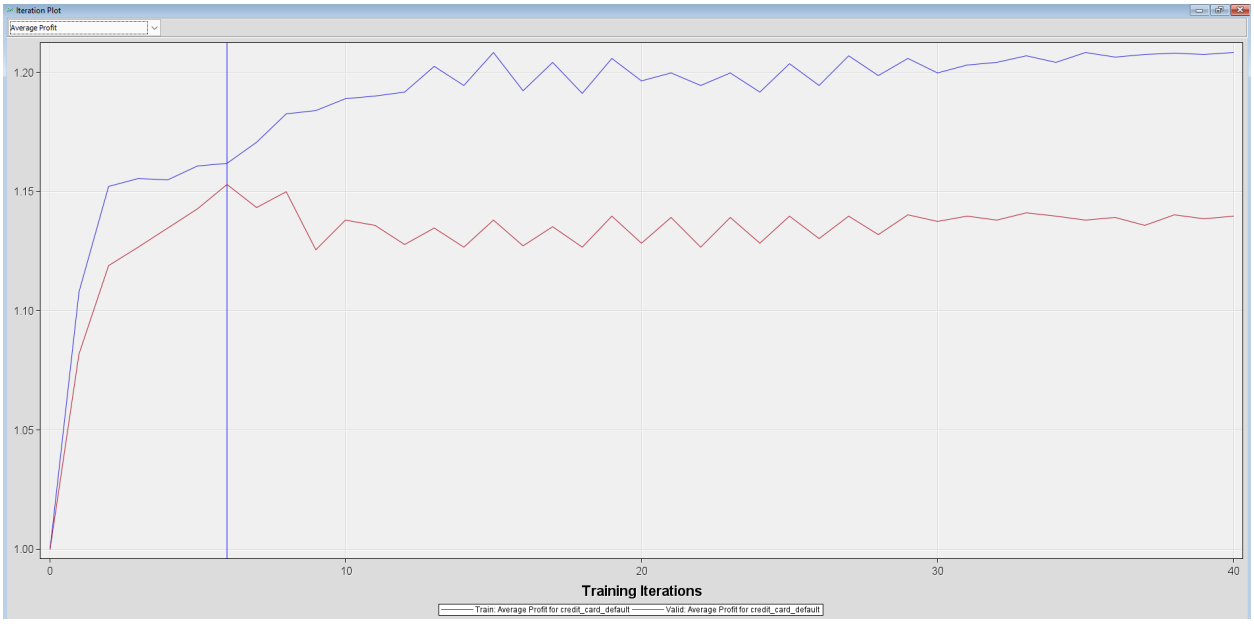


Table 57: PLS Neural Network Without Credit Variables Confusion Matrix

Confusion Matrix			
True Negative	**False Negative**	**False Positive**	**True Positive**
965	666	882	1181

Selection Tree Neural Network with Credit Variables

The final neural network was created using variable selection using a decision tree. These variables are Credit_Limit_Used, Credit_Score, Yearly_Debt_Payments and LOG_Net_Yearly_Income. The model estimated 19 parameters. Iteration 27 was selected. The validation average profit is 1.954029.

Chart 15: Selection Tree Neural Network with Credit Variables Iteration Plot

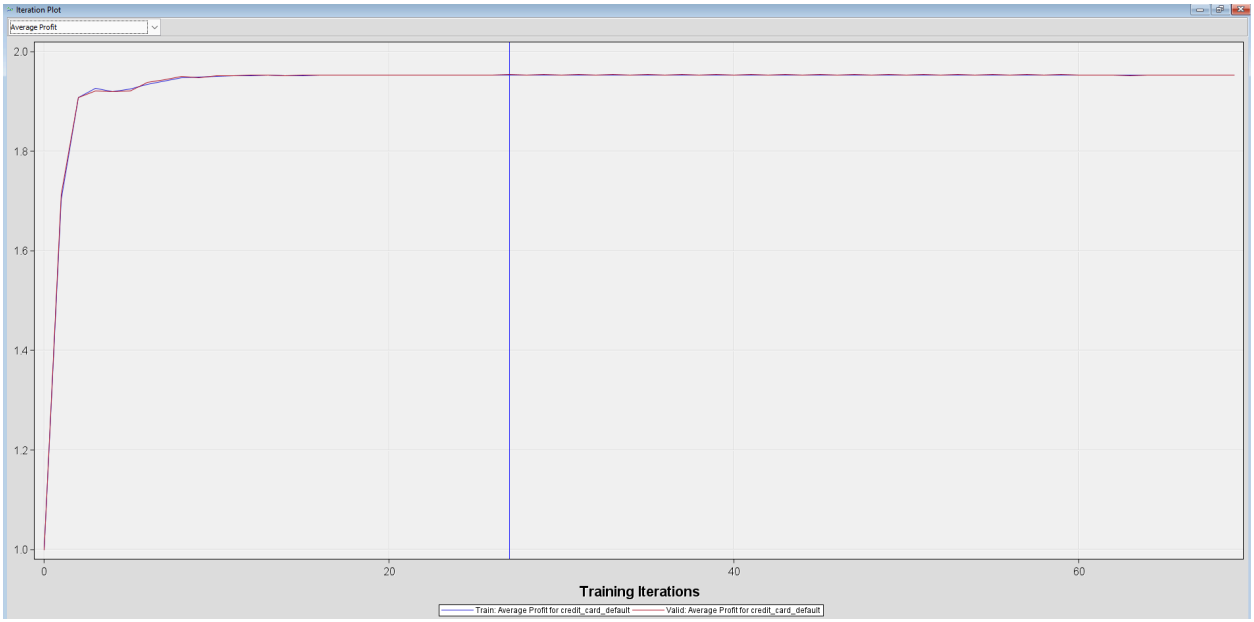


Table 58: Selection Tree Neural Network Credit Variables Confusion Matrix

Confusion Matrix			
True Negative	**False Negative**	**False Positive**	**True Positive**
1759	0	88	1847

Selection Tree Neural Network without Credit Variables

The variables selected by the selection tree for the dataset not containing credit related variables are LOG_No_Of_Days_Employed, Occupation_Type, Gender, LOG_Net_Yearly_Income, Total_Family_Members, Owns_House, Owns_Car, Migrant_Worker. The model estimated 82 parameters. Iteration 28 was selected. The validation average profit is 1.173067.

Chart 16: Selection Tree Neural Network without Credit Variables Iteration Plot

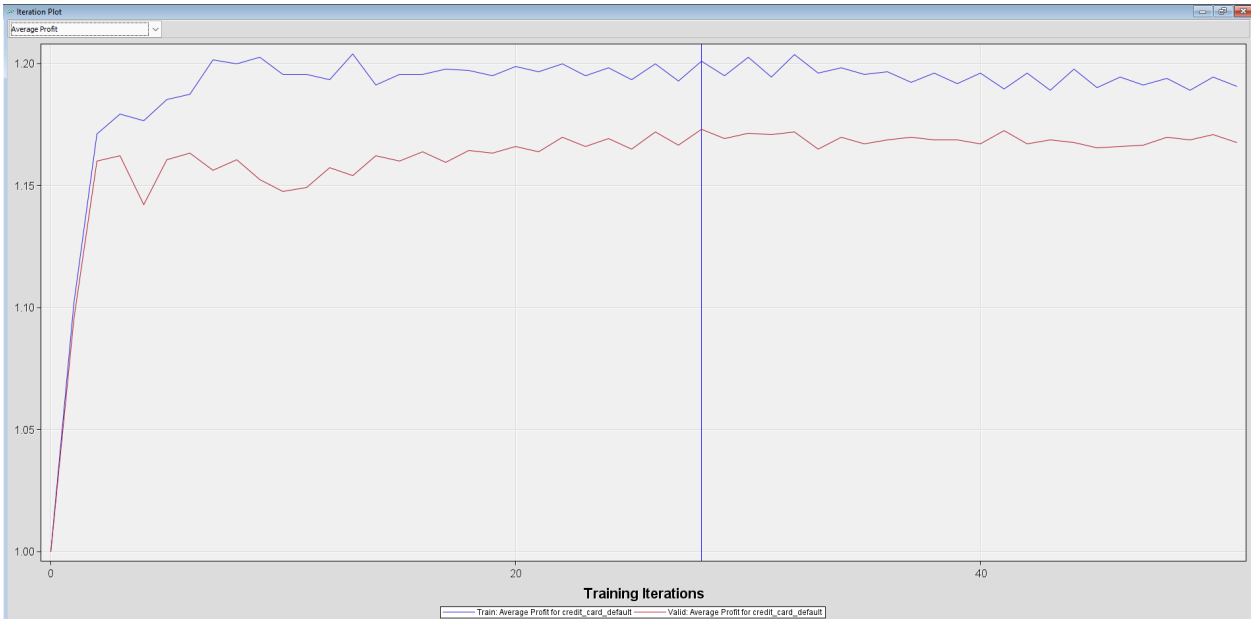


Table 59: Selection Tree Neural Network without Credit Variables Confusion Matrix

Confusion Matrix			
True Negative	**False Negative**	**False Positive**	**True Positive**
1119	748	728	1099

Ensemble

The Ensemble Node creates a new model combining the predictions from multiple models. An ensemble model will be better than the single models that compose it only if the single models disagree with each other. To combine the models, the Average function was used. The value for the validation average profit was not calculated for the Ensemble Node.

Table 60: Ensemble with Credit Variables Confusion Matrix

Confusion Matrix			
True Negative	**False Negative**	**False Positive**	**True Positive**
1756	0	91	1847

Table 61: Ensemble without Credit Variables Confusion Matrix

Confusion Matrix			
True Negative	**False Negative**	**False Positive**	**True Positive**
1024	675	823	1172

Evaluation

The cost-sensitive approach was used to adjust for separate sampling. For this reason, the Model Comparison Node was used to select the best model and the Selection Statistic was set to Average Profit/Loss.

The champion model for the dataset containing the credit related variables is the decision tree with an average validation profit of 1.957815. The decision model is not only the best at predicting credit card defaults, but it is also easily interpretable, and can be used for practical insights. Other model metrics like accuracy, sensitivity, specificity, precision and F1-score were calculated to further assess the model. Because we handled the imbalance data set correctly, all the variables have high values for all the metrics.

The best model for the dataset not containing the credit related variables is the Selection Tree Neural Network. One of the issues with this model is that it is hard to interpret. Since our main focus for this dataset is mainly on practical insights for credit card providers rather than just predictive accuracy, it is not the ideal choice. For this reason, the champion model we selected is the Selection Tree Logistic Regression with an average validation profit of 1.162791. The average validation profit is slightly lower than the one for the Selection Tree Neural Network, but it is easier to interpret, making it a better fit for our needs.

Comparing the average validation profit of models across the two datasets highlights the significant impact of credit-related variables on predicting credit card default. It's evident that models incorporating these variables outperform those that don't. The accuracy and average validation profit of models utilizing credit-related factors substantially surpass those without, underscoring the crucial role of these variables in predictive accuracy for credit card default.

Our findings align with the existing literature. In terms of predictive modeling, the approaches with the highest accuracy in predicting credit card default are neural networks and random forest (Neema & Soibam, 2017; Sayjadah, Y. et al., 2018; Yeh & Lien, 2008). Our best model for the credit variables dataset is a decision tree, the basic component of a random forest, and the best model for the dataset without credit related variables is a neural network.

Table 62: Model Comparison Credit Variables

Model Metrics						
Model	**Average Profit**	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	**F1 Score**
Decision Tree	1.957815	97.60%	100.00%	95.20%	95.50%	97.70%
Stepwise Regression	1.918875	95.90%	96.80%	95.00%	95.10%	95.90%
Polynomial Regression	1.939427	96.80%	98.80%	94.90%	95.00%	96.90%
VS Regression	1.940508	97.20%	99.10%	95.40%	95.60%	97.30%
PLS Regression	1.918875	95.90%	96.80%	95.00%	95.10%	95.90%
ST Regression	1.916171	95.90%	96.80%	95.10%	95.20%	96.00%
Stepwise Neural Network	1.957274	97.60%	100.00%	95.20%	95.40%	97.60%
Stepwise Neural Network 6HU	1.957275	97.60%	100.00%	95.20%	95.50%	97.70%
Stepwise AutoNeural	1.954029	97.50%	99.90%	95.20%	95.40%	97.60%
VS Neural Network	1.840454	92.50%	97.30%	87.80%	88.80%	92.90%
PLS Neural Network	1.957274	97.60%	100.00%	95.20%	95.40%	97.60%
ST Neural Network	1.954029	97.60%	100.00%	95.20%	95.50%	97.70%
Ensemble		97.50%	100.00%	95.10%	95.30%	97.60%

Table 63: Model Comparison Without Credit Variables

Model Metrics						
Model	**Average Profit**	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	**F1 Score**
Decision Tree	1.14278	57.50%	81.40%	33.60%	55.10%	65.70%
Stepwise Regression	1.159546	59.10%	61.50%	56.70%	58.70%	60.10%
Polynomial Regression	1.159546	58.70%	60.80%	56.50%	58.30%	59.50%
VS Regression	1.142239	60.70%	61.10%	60.40%	60.70%	60.90%
PLS Regression	1.143321	57.90%	60.50%	55.30%	57.50%	59.00%
ST Regression	1.162791	59.00%	61.80%	56.30%	58.50%	60.10%
Stepwise Neural Network	1.164413	59.30%	58.70%	59.80%	59.40%	59.00%
Stepwise Neural Network 6HU	1.170903	58.90%	60.50%	57.20%	58.60%	59.50%
Stepwise AutoNeural	1.163872	58.60%	59.40%	57.80%	58.40%	58.90%
VS Neural Network	1.16225	60.40%	61.10%	59.70%	60.30%	60.70%
PLS Neural Network	1.153056	58.10%	63.90%	52.20%	57.20%	60.40%
ST Neural Network	1.173067	60.00%	59.50%	60.60%	60.20%	59.80%
Ensemble		59.40%	63.50%	55.40%	58.70%	61.00%

Conclusion

Given the importance of predicting credit card default, more accurate predictions would be valuable information for banks and credit card companies. The dynamic nature of economic shifts and the multiplicity of factors involved, ranging from individual financial behaviors to broad socio-economic trends require a continuous improvement of existing models. Our analysis utilizing two datasets, one with credit-related variables and the other without, yielded insightful results. While the models incorporating credit-related variables demonstrated stronger predictive power, the models without such variables provided novel insights. Despite their comparatively lower predictive accuracy, these alternative models offer valuable perspectives, uncovering previously overlooked factors influencing credit card default behavior. Exploring both types of models contributes to the ongoing refinement of predictive methodologies in the realm of credit risk assessment.

Our research highlights that key predictors of credit card default include a person's credit history, encompassing mainly factors like their credit score, and utilization of credit limits. These findings align with the existing literature and domain knowledge. Even though no novel insights can be gathered from these results, the importance of our findings should not be underestimated. Economic conditions, consumer behaviors and regulatory landscapes are subject to constant change, shaping the dynamics of credit card defaults. An ongoing refinement and adaptation of current models is essential to capture how the influence of credit-related factors changes over time.

In contrast, demographic and socio-economic factors play a lesser role in predicting credit card defaults. However, these factors still influence credit card defaults and provide useful and practical insights for credit card providers. Gender is a factor with contrasting results in

existing literature. Our findings agree with what was previously observed by Li et al. (2019), that is males are more likely to default. Our findings also suggest that individuals who do not own a car or a house, have larger families, and especially have not been employed for a long period of time are more likely to default. Additionally, a unique aspect of the study is the inclusion of a variable defining migrant worker, a category of individuals that appears to be more likely to default than the rest of the population.

Credit card providers, and especially American Express, can utilize the findings of this study in multiple ways. Credit score is clearly the most important factor in predicting future defaults. For this reason, the application process for individuals with high credit scores should be heavily streamlined, improving customer satisfaction and reducing acceptance time. This way, banks can save time and improve customer experience, reinforcing brand loyalty.

Our findings regarding the influence of non credit-related variables can be utilized in different ways by credit card providers. Similarly to what was said above, the application process for certain categories of individuals can be sped up to improve customer satisfaction. Additionally, the results of our studies can be employed by American Express to enter new markets. Unlike companies like Discover, American Express requires a credit score to be considered for a credit card. Our models without credit-related variables show how, even though riskier, the company can still predict whether or not an individual who does not have a credit history will default. This would allow American Express to increase its market share and acquire customers for a much cheaper cost since most companies overlook them.

In conclusion, our study extends the credit card literature profoundly and provides American Express with actionable insights to improve its business.

References

Alam, T. M., Shaukat, K., Hameed, I. A., Luo, S., Sarwar, M. U., Shabbir, S., ... & Khushi, M. (2020). An investigation of credit card default prediction in the imbalanced datasets. *Ieee Access*, *8*, 201173-201198.

Arora, S., Bindra, S., Singh, S., & Nassa, V. K. (2022). Prediction of credit card defaults through data analysis and machine learning techniques. Materials Today: Proceedings, 51, 110-117.

Amexpert CodeLab. HackerEarth. (2021). https://www.hackerearth.com/challenges/new/competitive/amexpert-code-lab/

Beck, T., Demirgüç-Kunt, A., & Levine, R. (2006). Explaining Cross-Country Differences in Credit Market Structures. Journal of Financial Intermediation, 15(1), 32-57.

Bellotti, T., & Crook, J. (2013). Forecasting and stress testing credit card default using dynamic models. International Journal of Forecasting, 29(4), 563-574.

Dunn, L. F., & Mirzaie, I. A. (2023). Gender differences in consumer debt stress: impacts on job performance, family life and health. Journal of Family and Economic Issues, 44(3), 550-567.

Jain, S. V., & Jayabalan, M. (2022). Applying machine learning methods for credit card payment default prediction with cost savings. In Biomedical and Business Applications Using Artificial Neural Networks and Machine Learning (pp. 285-305). IGI Global.

Jappelli, T., & Pagano, M. (2002). Credit Card Defaults, Credit Growth, and the Macroeconomy: A Cross-Country Analysis. Journal of Monetary Economics, 49(8), 1639-1664.

Kiarie, F. K., Nzuki, D. M., & Gichuhi, A. W. (2015). Influence of Socio-Demographic Determinants on Credit Card Default Risk in Commercial Bank in Kenya. *International Journal of Science and Research*, *4*(5), 1611-1615.

Li, G. (2018). Gender-related differences in credit use and credit scores.

Li, Y., Li, Y., & Li, Y. (2019). What factors are influencing credit card customer's default behavior in China? A study based on survival analysis. *Physica A: Statistical Mechanics and its Applications*, *526*, 120861.

Neema, S., & Soibam, B. (2017). The comparison of machine learning methods to achieve most cost-effective prediction for credit card default. *Journal of Management Science and Business Intelligence*, *2*(2), 36-41.

Sayjadah, Y., Hashem, I. A. T., Alotaibi, F., & Kasmiran, K. A. (2018). Credit card default prediction using machine learning techniques. In *2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA)* (pp. 1-4). IEEE.

Teng, H. W., & Lee, M. (2019). Estimation procedures of using five alternative machine learning methods for predicting credit card default. Review of Pacific Basin Financial Markets and Policies, 22(03), 1950021.

Subasi, A., & Cankurt, S. (2019). Prediction of default payment of credit card clients using Data Mining Techniques. In 2019 International engineering conference (IEC) (pp. 115-120). IEEE.

Wang, L., Lu, W., & Malhotra, N. K. (2011). Demographics, attitude, personality and credit card features correlate with credit card debt: A view from China. *Journal of economic psychology*, *32*(1), 179-193.

Yeh, I. C., & Lien, C. H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert systems with applications*, *36*(2), 2473-2480.

Yu, Y. (2020). The application of machine learning algorithms in credit card default prediction. In *2020 International Conference on Computing and Data Science (CDS)* (pp. 212-218). IEEE.

Appendix

SAS Enterprise Miner Final Diagram