1).

Logit(p) = Intercept + Balance Coefficient X balance + Income Coefficient X Income + Student Indicator Coefficient X Student

Given estimates from the table:

- Intercept: -11.1924
- Balance Coefficient: 0.00574
- Income Coefficient: 3.033e-6
- Student Indicator Coefficient: 0.3234

Values:

- Balance: $2500
- Income: $40000
- Student = 1

For a student with a balance of $2500 and income of $40,000 we substitute and calculate values for logit(p):

-11.1924 + (0.00574 x 2500) + (3.003e – 6 x 4000) + (0.3234 x1)

Logit(p): 3.60232

Next, we need to convert the logit to probability:

1 / 1 + e ^-logit(p)

P = 1 / 1 + 0.272 = 0.973

The estimated probability is *97.3%*

2).

Confusion matrix:

(TP) True Positives: Predicted Yes and Actual Yes = 90

(FN) False Negatives: Predicted No but Actual Yes = 210

(FP) False Positives: Predicted yes but Actual No = 140

(TN) True Negatives: Predicted No and Actual No = 9560

Misclassification: 1 – Accuracy (TP + TN / Total):

- 90 + 9560 / 10000 = 0.965
- Misclassification: 1 – 0.965 = *0.035*

True Positive Rate (TPR):

TPR = TP / TP + FN

TPR = 90 / 90 + 210 = 0.3

True Positive Rate (TPR) = <u>*0.3*</u>

<u>False Positive Rate (FPR):</u>

FPR = 1 – TN / TN + FP

FPR = 1 – 9560 / 9560 + 140

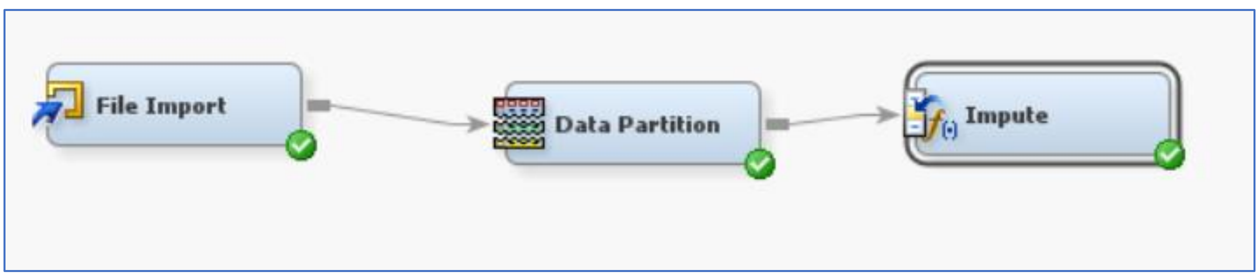
1 – 9560 / 9700 = 0.0144

False Positive Rate (FPR) = <u>*0.0144*</u>

3A).

<u>*Data Partition:*</u>

Property	Value
General	
Node ID	Part
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Output Type	Data
Partitioning Method	Default
Random Seed	36951810
Data Set Allocations	
Training	70.0
Validation	30.0
Test	0.0
Report	
Interval Targets	Yes



<u>*Impute:*</u>

Imputation Summary

Variable Name	Impute Method	Imputed Variable	Impute Value	Role	Measurement Level	Label	Number of Missing for TRAIN
CLAGE	MEAN	IMP_CLAGE	179.68906032	INPUT	INTERVAL		209
CLNO	MEAN	IMP_CLNO	21.237123663	INPUT	INTERVAL		153
DEBTINC	MEAN	IMP_DEBTINC	33.725587621	INPUT	INTERVAL		881
DELINQ	MEAN	IMP_DELINQ	0.4582668793	INPUT	INTERVAL		410
DEROG	MEAN	IMP_DEROG	0.2660675381	INPUT	INTERVAL		500
JOB	COUNT	IMP_JOB	Other	INPUT	NOMINAL		187
MORTDUE	MEAN	IMP_MORTDUE	73188.766338	INPUT	INTERVAL		352
NINQ	MEAN	IMP_NINQ	1.1600314301	INPUT	INTERVAL		354
REASON	COUNT	IMP_REASON	DebtCon	INPUT	NOMINAL		173
VALUE	MEAN	IMP_VALUE	100747.45973	INPUT	INTERVAL		80
YOJ	MEAN	IMP_YOJ	8.8981763317	INPUT	INTERVAL		361

Number Of Observations

Variable Name	Impute Method	Imputed Variable	Impute Value	Role	Measurement Level	Label	Number of Missing for TRAIN
CLAGE	MEAN	IMP_CLAGE	179.68906032	INPUT	INTERVAL		209
CLNO	MEAN	IMP_CLNO	21.237123663	INPUT	INTERVAL		153
DEBTINC	MEAN	IMP_DEBTINC	33.725587621	INPUT	INTERVAL		881
DELINQ	MEAN	IMP_DELINQ	0.4582668793	INPUT	INTERVAL		410
DEROG	MEAN	IMP_DEROG	0.2660675381	INPUT	INTERVAL		500
JOB	COUNT	IMP_JOB	Other	INPUT	NOMINAL		187
MORTDUE	MEAN	IMP_MORTDUE	73188.766338	INPUT	INTERVAL		352
NINQ	MEAN	IMP_NINQ	1.1600314301	INPUT	INTERVAL		354
REASON	COUNT	IMP_REASON	DebtCon	INPUT	NOMINAL		173
VALUE	MEAN	IMP_VALUE	100747.45973	INPUT	INTERVAL		80
YOJ	MEAN	IMP_YOJ	8.8981763317	INPUT	INTERVAL		361

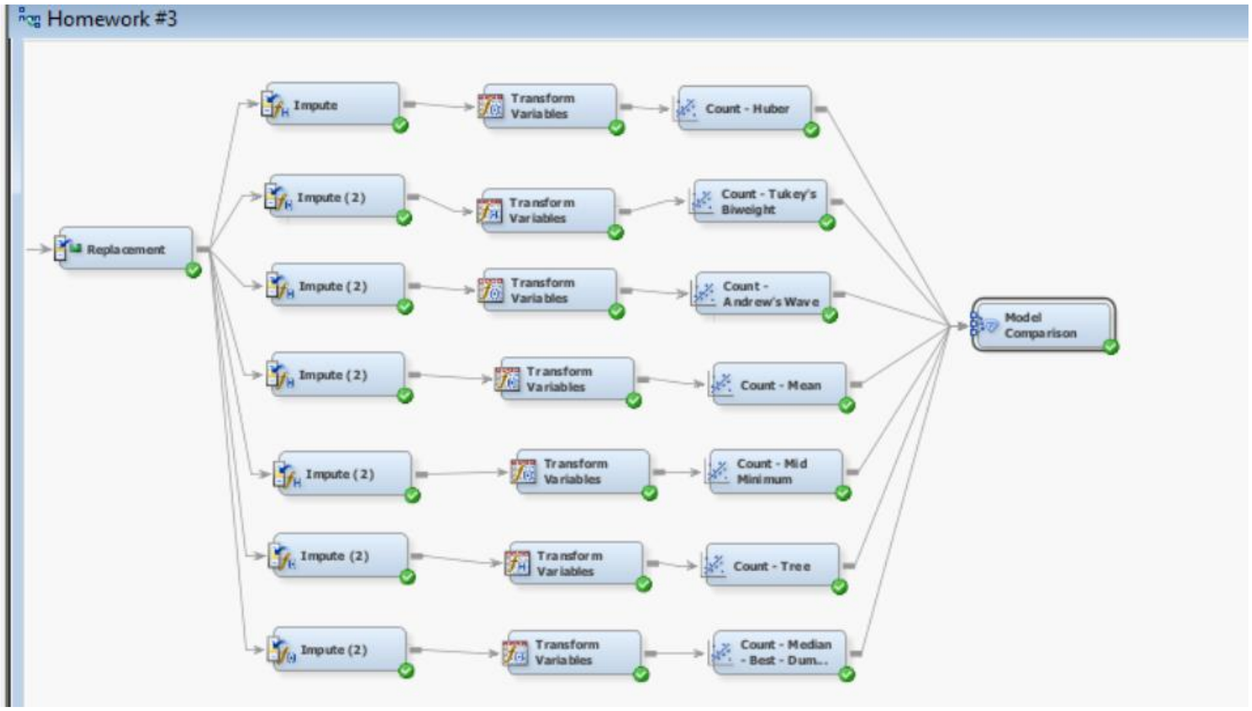
Regression:

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept		1	0.1535	0.0307	4.99	<.0001
IMP_CLAGE		1	-0.00069	0.000070	-9.82	<.0001
IMP_CLNO		1	-0.00199	0.000624	-3.19	0.0014
IMP_DEBTINC		1	0.00580	0.000695	8.35	<.0001
IMP_DELINQ		1	0.1104	0.00518	21.30	<.0001
IMP_DEROG		1	0.0856	0.00689	12.43	<.0001
IMP_JOB	Mgr	1	-0.0408	0.0156	-2.61	0.0091
IMP_JOB	Office	1	-0.0953	0.0148	-6.45	<.0001
IMP_JOB	Other	1	-0.0289	0.0120	-2.41	0.0159
IMP_JOB	ProfExe	1	-0.0370	0.0138	-2.68	0.0074
IMP_JOB	Sales	1	0.1627	0.0346	4.71	<.0001
IMP_MORTDUE		1	-7.03E-7	2.306E-7	-3.05	0.0023
IMP_NINQ		1	0.0267	0.00355	7.53	<.0001
IMP_REASON	DebtCon	1	-0.0219	0.00626	-3.51	0.0005
IMP_VALUE		1	5.311E-7	1.869E-7	2.84	0.0045
IMP_YOJ		1	-0.00108	0.000791	-1.36	0.1725
LOAN		1	-1.86E-6	5.576E-7	-3.34	0.0008

The logistic regression analysis reveals that several variables are statistically significant in predicting default (BAD) status. Key financial indicators, including IMP_CLAGE, IMP_CLNO, IMP_DEBTINC, IMP_DELINQ, IMP_DEROG, IMP_MORTDUE, IMP_NINQ, IMP_VALUE, and LOAN, are significant, highlighting their impact on credit risk. Additionally, categorical factors such as IMP_JOB (with categories like Mgr, Office, ProfExe, Sales) and IMP_REASON: DebtCon are also significant, indicating that occupation and loan purpose influence default likelihood. Imputing missing values appropriately ensured that these variables could be effectively analyzed without data gaps.

3B).



Models	Input - Interval	Input – Class	Transformation - Interval	Transformation – Class
Count-Huber	Count	Huber	NA	NA
Count – Tukey's Biweight	Count	Tukey's Biweight	NA	NA
Count – Andrew's Wave	Count	Andew's Wave	NA	NA
Count - Mean	Count	Mean	NA	NA
Count – Mid Minimum Spacing	Count	Mid Minimum Spacing	NA	NA
Count - Tree	Count	Tree	NA	NA
Count - Median	Count	Median	Best	Dummy Indicators

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Average Squared Error	Train: Akaike's Information Criterion	Train: Average Squared Error
Y	Reg3	Reg3	Count - Andrew's Wave	BAD		0.092755	-10212.1	0.083152
	Reg7	Reg7	Count - Median - Best - Dummy Indicators	BAD		0.094852	-10217.9	0.084889
	Reg	Reg	Count - Huber	BAD		0.095951	-10172.4	0.085901
	Reg2	Reg2	Count - Tukey's Biweight	BAD		0.121245	-8710.49	0.122948
	Reg4	Reg4	Count - Mean	BAD		0.123627	-8655.78	0.124571
	Reg5	Reg5	Count - Mid Minimum	BAD		0.128241	-8531.66	0.128332
	Reg6	Reg6	Count - Tree	BAD		0.13392	-8480.12	0.129928

Based on the fit statistics in the provided results, different imputation and transformation methods were explored to optimize the logistic regression model's performance. Among the various configurations, the model with Andrew's Wave imputation and Count for class variables achieved the lowest validation misclassification error, with a value of 0.092755. This suggests that using Andrew's Wave for imputation and Count for categorical variables provided the best model fit in terms of minimizing misclassification error on the validation set.

The next best-performing model utilized Median Imputation combined with the Best transformation and Dummy Indicators for categorical variables, yielding a validation misclassification error of 0.094852. Other imputation methods like Huber, Tukey's Biweight, Mean, Mid Minimum, and Tree imputation resulted in higher misclassification errors, indicating less optimal performance.

In summary, Andrew's Wave Imputation with Count transformation was identified as the best configuration to predict the target variable "BAD," based on the lowest validation misclassification error in this analysis.