

# **INSY 5339 - Data mining**

## **Project Progress**

### **Team -12**

**Darshanik Mekapati**

**Jalam Yashwanth**

**Sai Krishna Teja Adarsh Nadella**

**Bala Manikanta Ummaneni**



## **Business Problem:**

A banking Client offers loan to the eligible customers and denies the offer to customers who have not met certain criterion.

However, there are many customers who were not offered loan and they are eligible for loan approval.

## **Objective:**

This Project is taken up by our Orion Analytics and will help in finding such type of customers through exhaustive data mining techniques.

With this the banking client can classify the customers based on data driven decision and can offer the loans more precisely.

**Source: Kaggle's Loan Prediction and Approval Data sets.**

## **Data Description:**

- There are 2 datasets that were downloaded from Kaggle.
  1. Train.csv
  2. Test.csv
- Train data set has 614 rows and 13 columns. The data set has 13 Independent variables. Their names and description are given in Table1.

<b>Variable Name</b>	<b>Type</b>	<b>Description</b>
Gender	Independent	Gender of applicant male/female
Married	Independent	Yes/No
Dependents	Independent	0,1, 2, 3 and 3+
Education	Independent	Graduate or not graduated
Self- Employed	Independent	Is the applicant having business or job (Yes/No)
Applicant Income	Independent	A continuous variable depicting customer's income
Co Applicant Income	Independent	A continuous variable

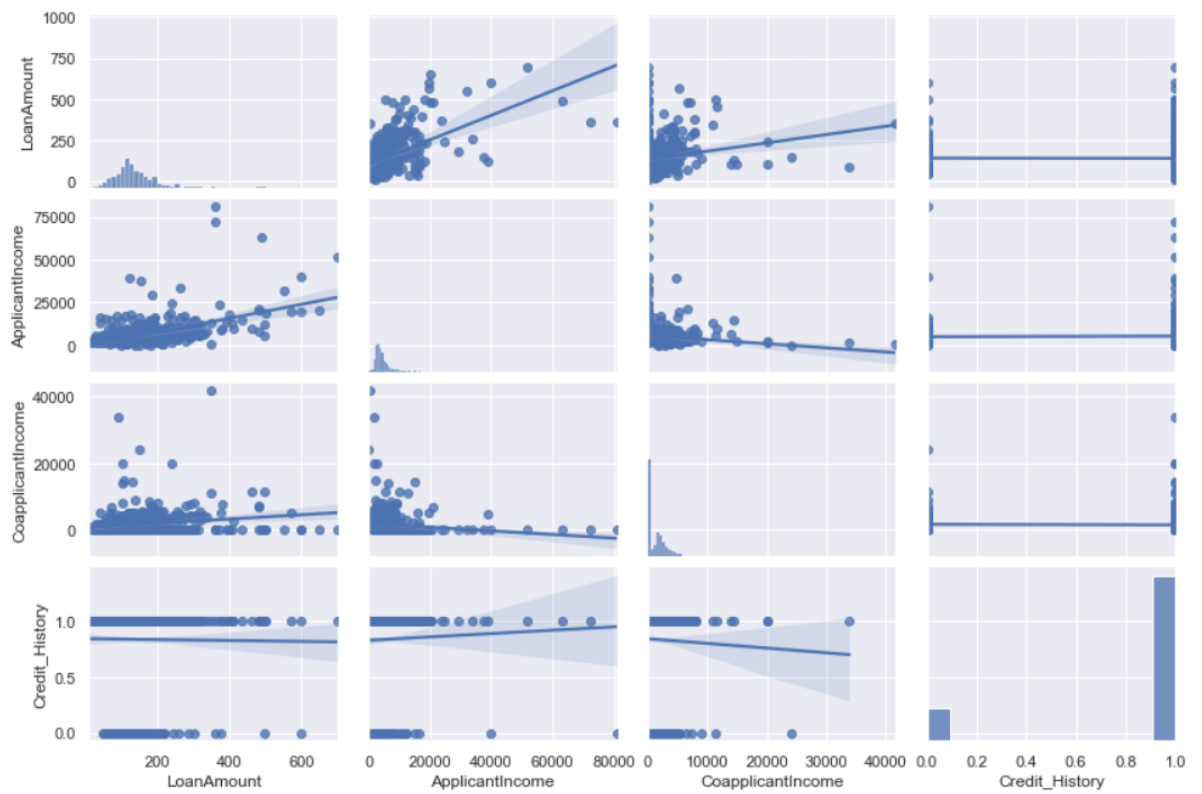
		depicting customer's dependent income
Loan Amount	Independent	A continuous variable depicting loan amount
Loan Amount Term	Independent	Time period of loan repayment in days
Credit History	Independent	A customer having good credit history, yes/No
Property Area	Independent	Applicant's collateral property locality – urban, semiurban, or rural
Loan Status	Dependent	Sanctioned(yes) or Denied (No)
Loan ID	Independent	ID of applicant

**Table 1. variables Legend**

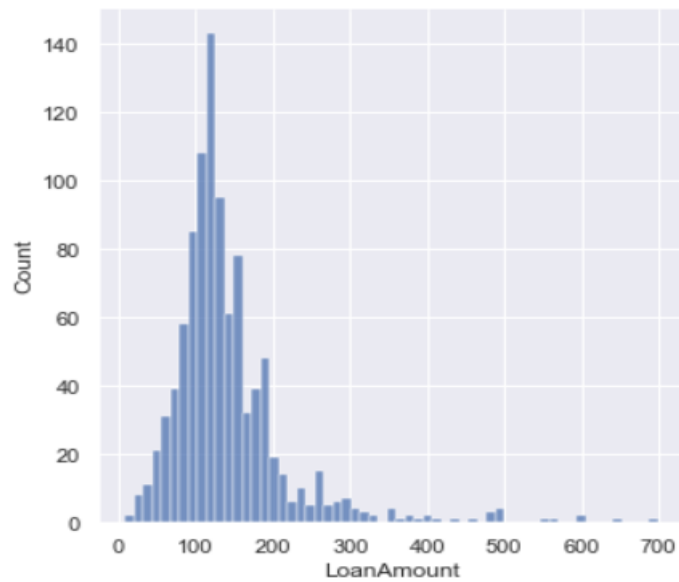
**Data Visualization:**

- Matrix graph is plotted between 4 Independent variables (Loan Amount, Applicant Income, Co-Applicant Income and Credit History) to visualize if there is any linearity among them.
- The other variables are of classification type.
- From **Fig.1**, there is visual evidence of linearity between Applicant Income and Loan Income. However, the credit history is of Binomial type and linear trend is absent. The best method to incorporate credit history is **maximum likelihood estimate (MLE)**
- To further understand the relation among the Loan Amount, Applicant Income and Co-Applicant Income variables, we use probability distribution plot of Seaborn shown in **Fig.2**.
- From **Fig.2** the Distribution is positively skewed.
- Boxplots are used to visualize outliers and the corresponding distribution for the variables. The plots are showed in **Fig.3**.
- Additionally, in Applicant Income and loan Status scatter plot (top left) in Fig.4, there is an outlier and is denoted after the vertical red line. **It says the applicant has high income (80000\$) but the loan was not approved.**

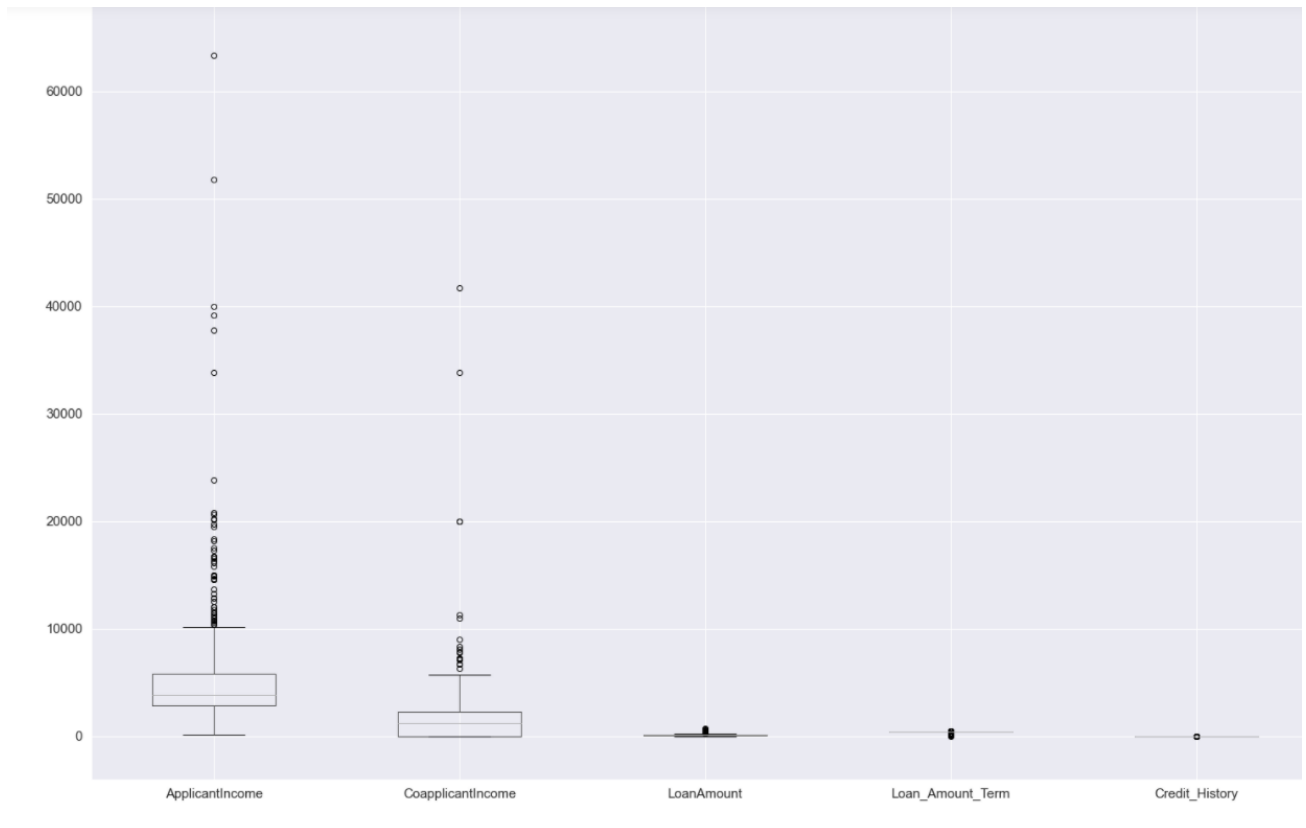
Out[13]: <seaborn.axisgrid.PairGrid at 0x192e91bd4f0>



**Fig.1. Matrix plot of independent variables**

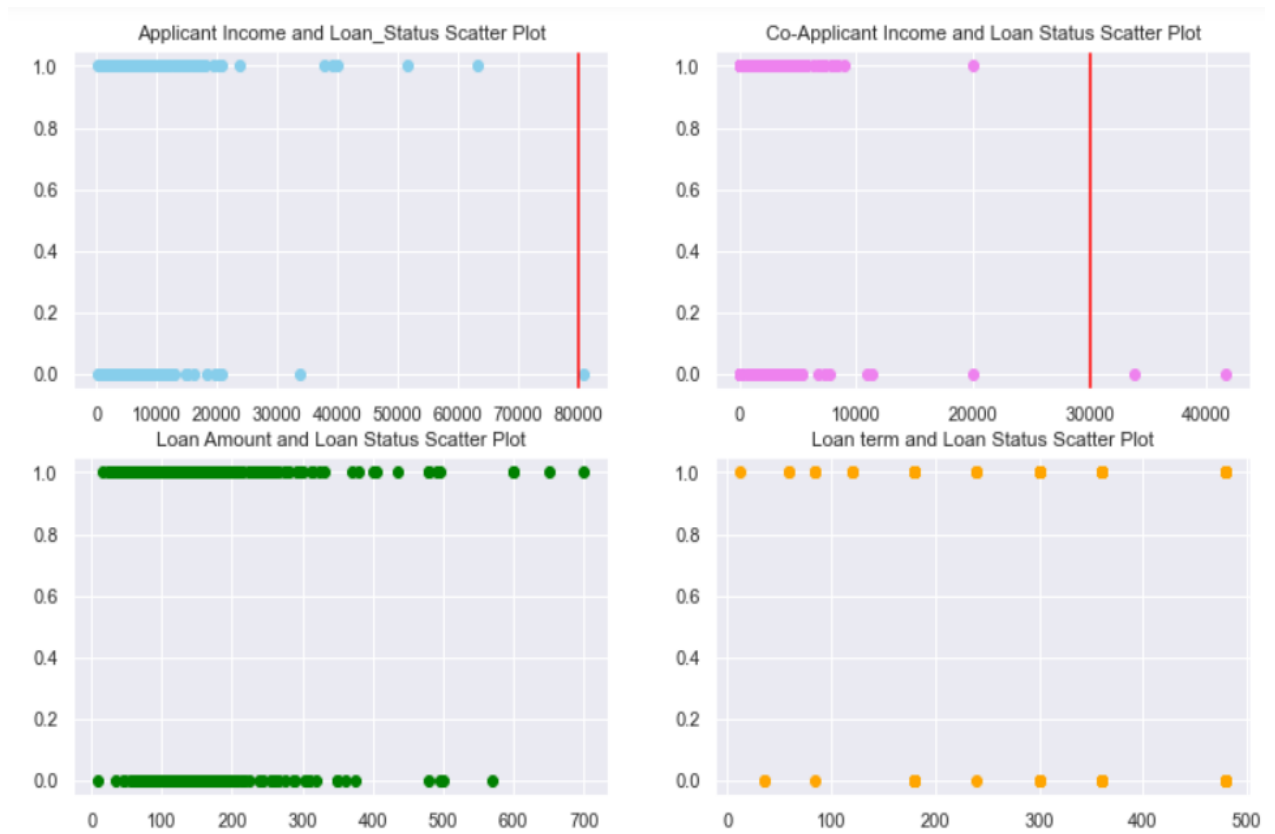


**Fig.2. Probability plot of loan amount**



**Fig.3. Box Plots for continuous independent variables**

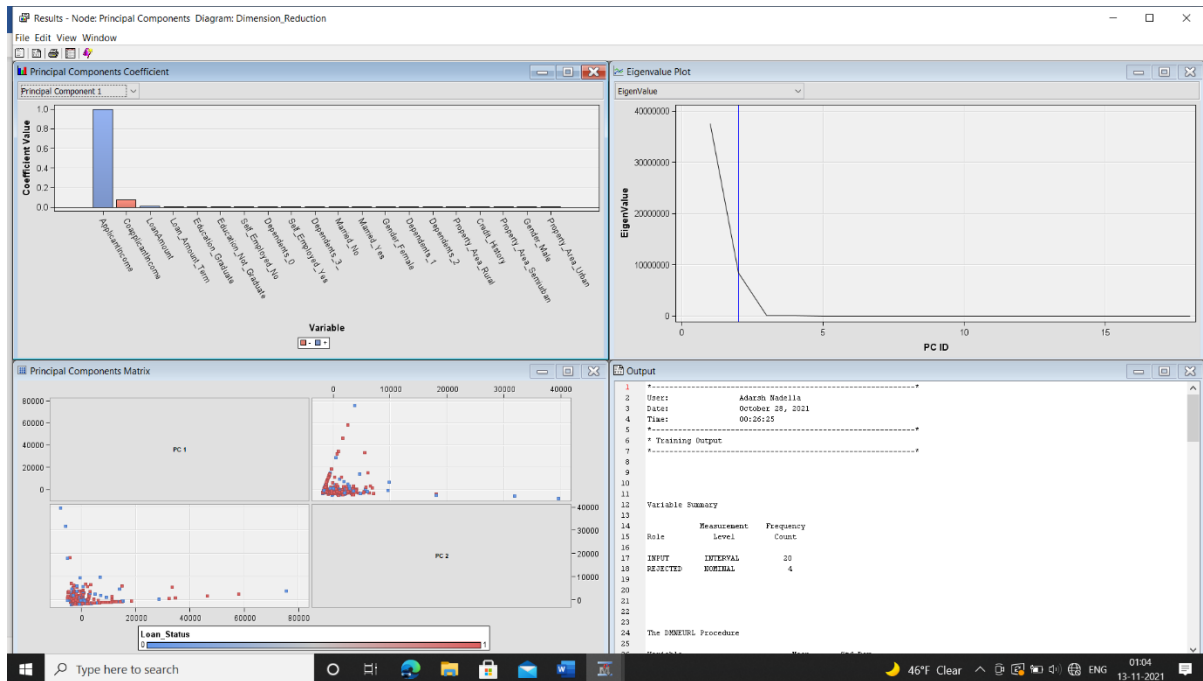
- Similarly, for Co-applicant and Loan status plot (top right) of Fig.4, there are **2 outliers** where the co-applicant income is above 30000\$ and the loan was not approved for their applicants.



**Fig.4. Potential Outliers**

### Dimensionality Reduction:

- Initially, the data set has 13 features or independent variables of which 3 were continuous and 10 were categorical. Categorical variables with missing values are treated by using One-hot encoding.
- One-hot encoding assigns dummy values for categorical variables and in this process, the missing values were also replaced with their corresponding Mode.
- As a result, there is significant outburst of features and the data set after one hot encoding has 22 features or columns and 614 rows.
- Hence the appropriate solution for minimizing the dimension of data set with minimal loss of explainable variability is Principal component analysis (**PCA**).
- Fig.5** shows the PCA of the data set in SAS EM with Eigen value cut off of 95%.



**Fig.5. Principal Component Analysis**

- 2 PCAs were selected as shown in **Fig 6** explaining 99.98% of variability compared with original 22 or 13 features.

```

52
53 The DMNEURL Procedure
54
55           Eigenvalues of Covariance Matrix
56
57           Eigenvalue   Difference   Proportion   Cumulative
58
59      1    37472895.1    29059716.3      0.8165      0.8165
60      2     8413178.8     8408493.7      0.1833      0.9998
61      3       4685.1        896.7      0.0001      0.9999
62      4       3788.4        3787.9      0.0001      1.0000
63      5          0.6          0.2      0.0000      1.0000
64      6          0.4          0.0      0.0000      1.0000
65      7          0.3          0.0      0.0000      1.0000
66      8          0.3          0.0      0.0000      1.0000
67      9          0.3          0.0      0.0000      1.0000
68     10          0.2          0.0      0.0000      1.0000
69     11          0.2          0.0      0.0000      1.0000
70     12          0.2          0.0      0.0000      1.0000
71     13          0.1          0.0      0.0000      1.0000
72     14          0.1          0.1      0.0000      1.0000
73     15          0.0          0.0      0.0000      1.0000
74     16          0.0          0.0      0.0000      1.0000
75     17          0.0          0.0      0.0000      1.0000
76     18          0.0          0.0      0.0000      1.0000
77     19          0.0          0.0      0.0000      1.0000
78     20          0.0          0.0      0.0000      1.0000
79
80
81
82
83
84
85
86
87
88
89
90
91 *-----*
92 Summary of Exported Principal Components
93 *-----*
94
95 Total number of input variables: 20
96 Maximum number cutoff of principal components: 20
97 Cumulative proportional eigenvalue cutoff: 0.95
98 Proportional eigenvalue increment cutoff: 0.001
99 Number of the selected principal components: 2
100 Total variation explained by the selected principal components: 0.9998153104
101

```

**Fig.6 Principal Component Analysis Variability**



## Prediction Trials.

### I. Maximum Likelihood Estimate:

- Since the Dependent variable is not continuous and is binomial in nature with 2 possible outcomes (Loan approved = 1.0 or denied = 0.0), The Logit model is used for predictions and the method by which the **Logit or Logistic regression** is solved is called **Maximum likelihood estimate (MLE)**. The model's performance is evaluated by **Accuracy, precision and Recall metrics**.
- Apart from Logistic regression, **Classification boosted trees** will also be used as **Random Forrest classifier** is best at solving almost all types of classification models. The **Tree's performance will be tuned or pruned with Gini Index**.
- **Fig.7** shows the initial MLE model's summary run in Python with Stats model's Logit API.

```
In [30]: results_log.summary()
```

```
Out[30]:
```

Logit Regression Results

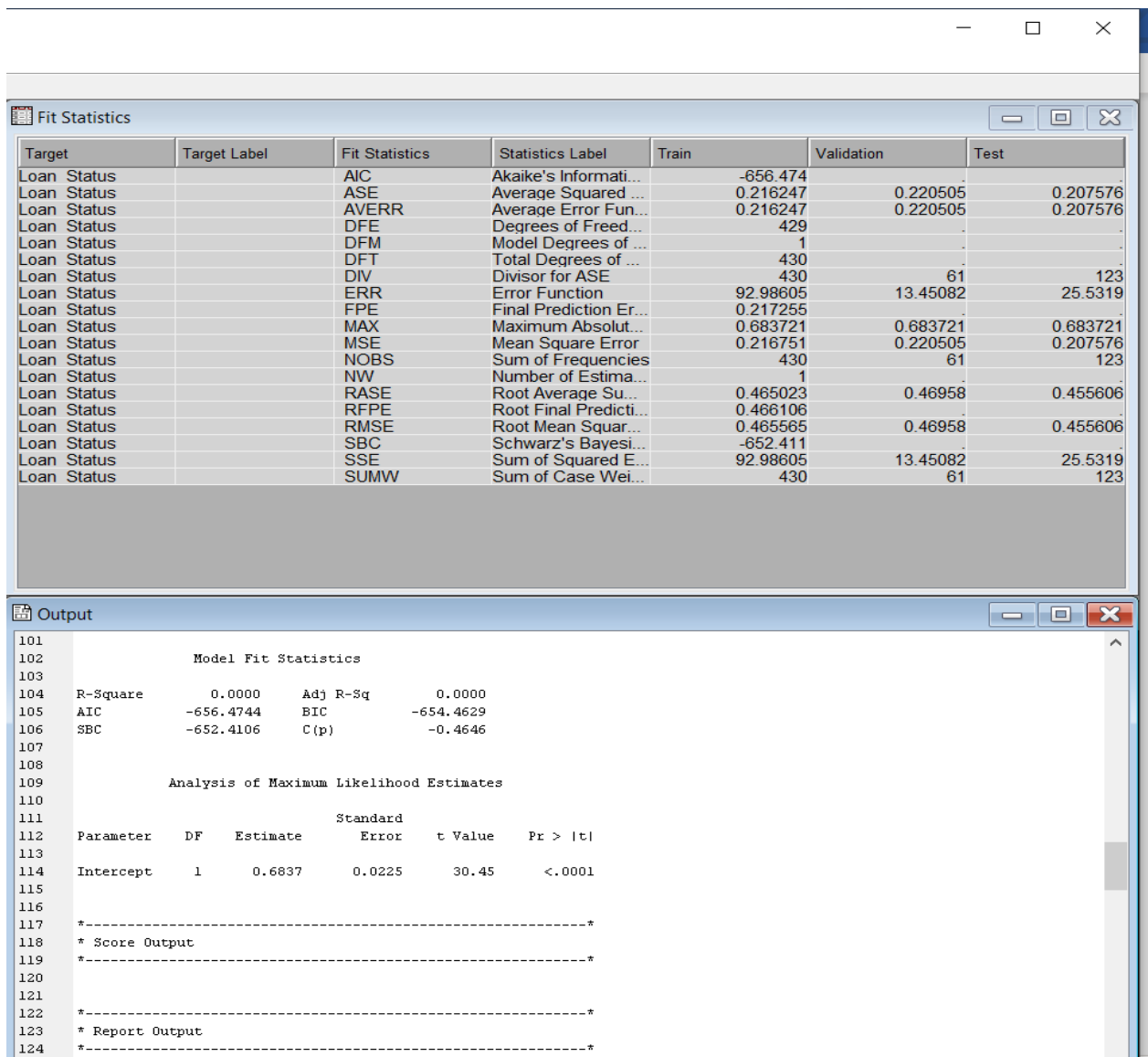
Dep. Variable:	Loan_Status	No. Observations:	614
Model:	Logit	Df Residuals:	608
Method:	MLE	Df Model:	5
Date:	Sat, 23 Oct 2021	Pseudo R-squ.:	0.2345
Time:	22:27:16	Log-Likelihood:	-292.00
converged:	True	LL-Null:	-381.45
Covariance Type:	nonrobust	LLR p-value:	9.250e-37

	coef	std err	z	P> z	[0.025	0.975]
const	-1.9135	0.722	-2.651	0.008	-3.328	-0.499
ApplicantIncome	8.807e-06	2.29e-05	0.384	0.701	-3.61e-05	5.37e-05
CoapplicantIncome	-4.755e-05	3.24e-05	-1.468	0.142	-0.000	1.59e-05
LoanAmount	-0.0011	0.002	-0.721	0.471	-0.004	0.002
Loan_Amount_Term	-0.0011	0.002	-0.633	0.527	-0.004	0.002
Credit_History	3.8152	0.409	9.317	0.000	3.013	4.618

Fig.7 MLE result's summary

## II. Linear Regression with PCA Input

- Fig.7 shows the Linear Regression Maximum likelihood estimate and significance level of selected principal components after linear regression.
- Data is partitioned into **70% training, 10% Validation and 20% Testing** and fed to Regression node.
- However, the output of linear regression is abnormal with Adjusted R squared as 0.00 but the Mean Squared error is 21.67% and Test Average squared error as 20.8%.
- But there is positive linear relation between PC1 and Loan status.
- Hence, PCA is not helping to achieve better prediction and explainability of Loan status in case of linear regression.



**Fig.8. Linear Regression with PCA**

### III. Linear Regression without PCA

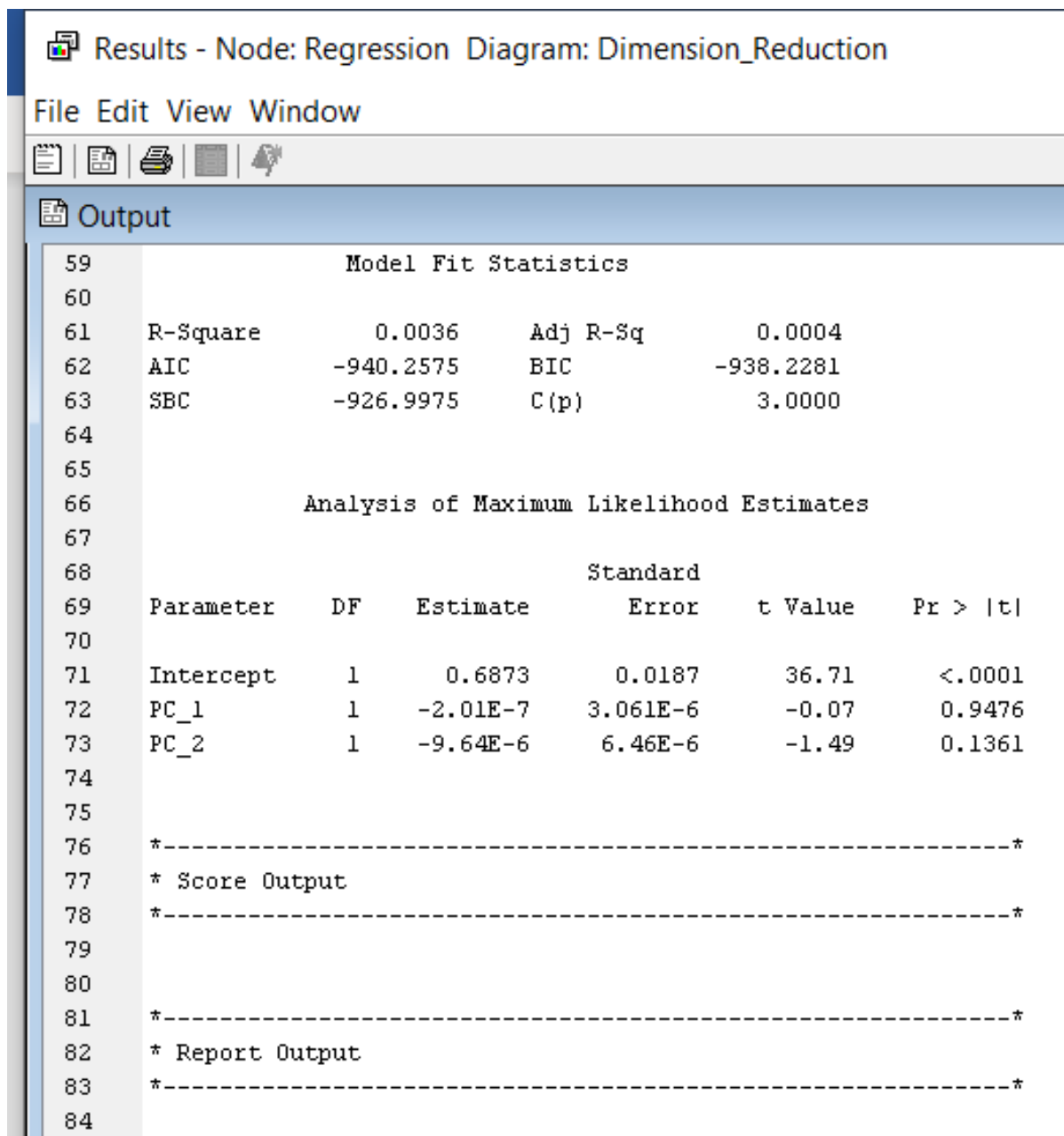
- Linear regression yielded more accurate results when with raw data with partitioning without PCA. Fig.8 shows the results and Hidden patterns.
- The root mean squared error (RMSE) is 37.91% which means the model classified 62.19% of loan status of customers accurately.
- Additionally, the hidden pattern is **non-married applicants who have collateral property in semi urban area with good credit history are very likely to get loan approval.**
- Here the **Intuition** is collateral property in semi urban area has lot of potential to get classified into high demand development areas paving the way for the banks to minimize risk of applicant's financial background in decision making.

Output							
244							
245	1	Credit_History	1	1	176.61	<.0001	66.2879
246	2	Property_Area_Semiurban	1	2	9.09	0.0027	65.1944
247	3	Married_No	1	3	9.67	0.0020	64.0772
248							
249							
250	The selected model, based on the cross-validation error rate, is the model trained in Step 3. It consists of the following effects:						
251							
252	Intercept	Credit_History	Married_No	Property_Area_Semiurban			
253							
254							
255	Analysis of Variance						
256							
257							
258	Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	
259							
260	Model	3	29.963790	9.987930	67.51	<.0001	
261	Error	426	63.022256	0.147940			
262	Corrected Total	429	92.986047				
263							
264							
265	Model Fit Statistics						
266							
267	R-Square	0.3222	Adj R-Sq	0.3175			
268	AIC	-817.7278	BIC	-815.6426			
269	SBC	-801.4727	C(p)	3.4552			
270							
271							
272	Analysis of Maximum Likelihood Estimates						
273							
274							
275	Parameter	DF	Estimate	Standard Error	t Value	Pr >  t	
276							
277	Intercept	1	0.0664	0.0515	1.29	0.1984	
278	Credit_History	1	0.7129	0.0534	13.36	<.0001	
279	Married_No	1	-0.1237	0.0398	-3.11	0.0020	
280	Property_Area_Semiurban	1	0.1208	0.0385	3.14	0.0018	
281							
282							
283	Estimated Covariance Matrix						
284							
285							
286	Parameter	Intercept	Credit_History	Married_No	Property_Area_Semiurban		
287							
288	Intercept	0.002656	-.002360	-.000428	-.000399		
289	Credit_History	-.002360	0.002848	-.000073	-.000163		
290	Married_No	-.000428	-.000073	0.001582	-.000048		
291	Property_Area_Semiurban	-.000399	-.000163	-.000048	0.001480		
292							
293							

Fig.9. Hidden Pattern in Data

#### IV. Logistic Regression with PCA

- Similar to linear regression with PCA logistic regression yielded unsatisfactory results with both PC1 and PC2 having very low statistical significance.
- Additionally, from Fig.9, the Adjusted R squared is very low (0.04%) showing the evidence that the model underfitted the data.



The screenshot displays the SAS Output window for a Logistic Regression model. The title bar reads "Results - Node: Regression Diagram: Dimension\_Reduction". The menu bar includes "File", "Edit", "View", and "Window". The toolbar contains icons for opening, saving, printing, and other functions. The "Output" pane shows the following statistics:

Model Fit Statistics						
R-Square	0.0036	Adj R-Sq	0.0004			
AIC	-940.2575	BIC	-938.2281			
SBC	-926.9975	C(p)	3.0000			

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	t Value	Pr >  t	
Intercept	1	0.6873	0.0187	36.71	<.0001	
PC_1	1	-2.01E-7	3.061E-6	-0.07	0.9476	
PC_2	1	-9.64E-6	6.46E-6	-1.49	0.1361	

Below the parameter estimates, there are three sections separated by dashed lines and asterisks:

- \* Score Output
- \* Report Output

**Fig.10. Under fitted Logistic regression**

## V. Logistic Regression without PCA

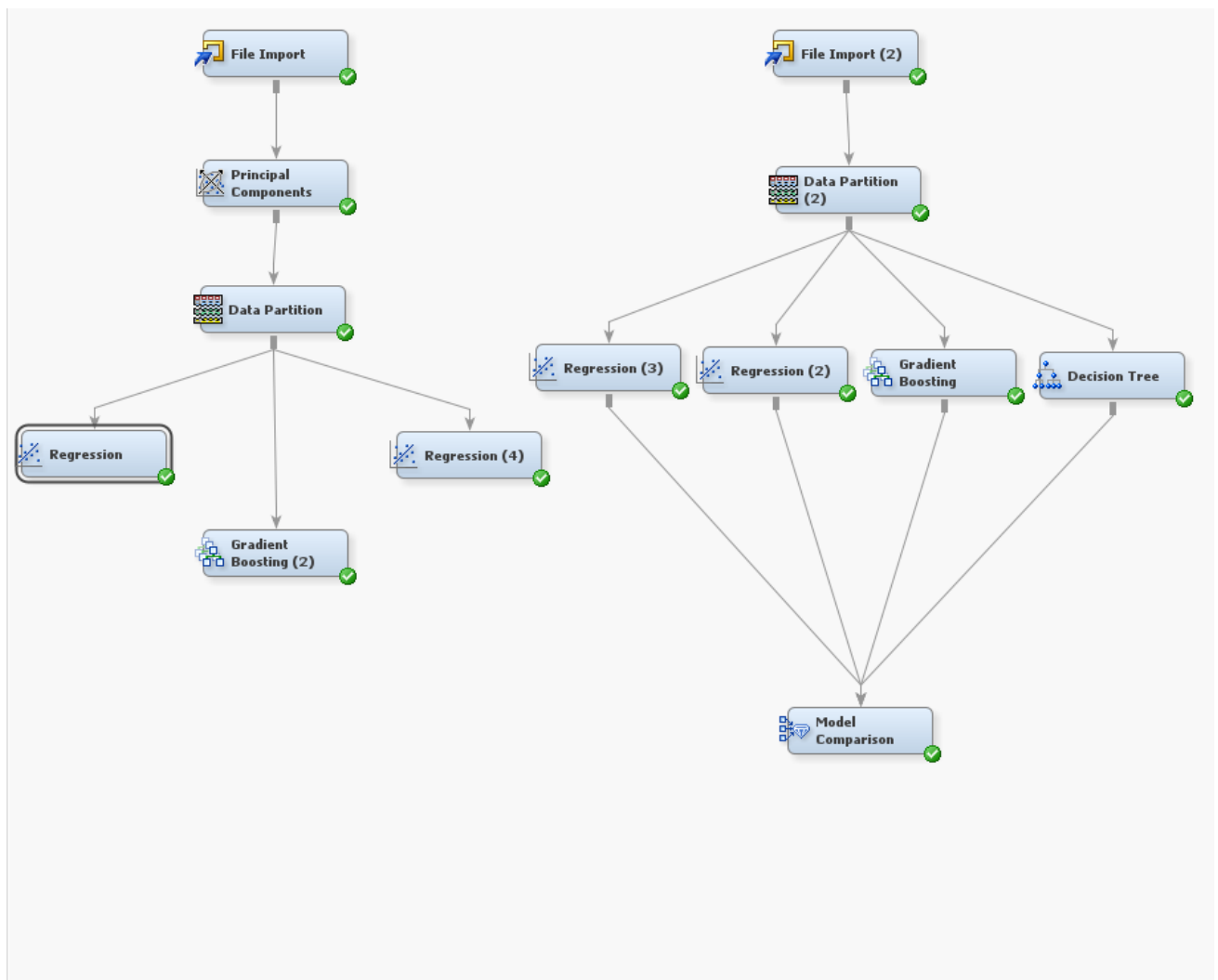
- Similar to linear regression without PCA, logit also gave same results with Adjusted R-squared of 31.66%.
- Logistic regression also gave the same hidden pattern criteria given by previous linear regression results further reinforcing our evidence that non married applicant with collateral property in semi-urban area and good credit history are more likely to get loan approval. This is shown in Fig.10.

Output

60	Model Fit Statistics					
61						
62	R-Square	0.3437	Adj R-Sq	0.3166		
63	AIC	-803.5465	BIC	-799.9775		
64	SBC	-730.3983	C(p)	18.0000		
65						
66						
67	Analysis of Maximum Likelihood Estimates					
68						
69				Standard		
70	Parameter	DF	Estimate	Error	t Value	Pr >  t
71						
72	Intercept	1	0.1952	0.2421	0.81	0.4204
73	ApplicantIncome	1	3.569E-6	3.855E-6	0.93	0.3551
74	CoapplicantIncome	1	0.000010	0.000011	0.98	0.3265
75	Credit_History	1	0.7191	0.0542	13.26	<.0001
76	Dependents_0	1	-0.0237	0.1204	-0.20	0.8442
77	Dependents_1	1	-0.0461	0.1259	-0.37	0.7143
78	Dependents_2	1	0.0584	0.1270	0.46	0.6458
79	Dependents_3_	1	0.0637	0.1333	0.48	0.6330
80	Education_Graduate	1	0.0527	0.0460	1.14	0.2532
81	Education_Not_Graduate	0	0	.	.	.
82	Gender_Female	1	0.0187	0.1582	0.12	0.9060
83	Gender_Male	1	-0.0304	0.1503	-0.20	0.8399
84	LoanAmount	1	-0.00048	0.000271	-1.79	0.0750
85	Loan_Amount_Term	1	-0.00039	0.000311	-1.24	0.2158
86	Married_No	1	-0.1184	0.0457	-2.59	0.0100
87	Married_Yes	0	0	.	.	.
88	Property_Area_Rural	1	-0.0700	0.0483	-1.45	0.1483
89	Property_Area_Semiurban	1	0.0833	0.0455	1.83	0.0678
90	Property_Area_Urban	0	0	.	.	.
91	Self_Employed_No	1	0.0574	0.0913	0.63	0.5295
92	Self_Employed_Yes	1	0.0670	0.1028	0.65	0.5152
93						
94						
95	*-----*					
96	* Score Output					
97	*-----*					
98						
99						
100	*-----*					
101	* Report Output					
102	*-----*					
103						

Fig.11. Logistic regression summary output.

## Tree Diagram:



How many observations in the dataset?	614
How many binary/categorical variables?	9
How many continuous variables?	3
What is the outcome / target variable?	Loan Status

If binary or categorical: What percentage of the variables belong to each class.

**Loan Approved (1) = 422**  
**Loan Approval rate = 69%**  
**Loan Denied (0) = 192**  
**Loan Denial rate = 31%**  
**male population = 80%**  
**Female population = 18%**  
**Other population = 2%**  
**Graduates = 78%**  
**Non-graduates = 22%**  
**Self-Employed = 13%**  
**Non-Self-Employed = 81%**  
**Others = 6%**  
**Property Area Urban = 33%**  
**Property Area semi-urban = 38%**  
**Property Area Rural = 29%**  
**“0” Dependents = 56%**  
**“1” Dependent = 17%**  
**“2” Dependents = 16%**  
**“3+” Dependents = 8%**

**Table 2. Summary of Data set and Report**