INSY 5339 - Data mining

Project Final Report

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.

Variable Name	**Type**	**Description**
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>
```

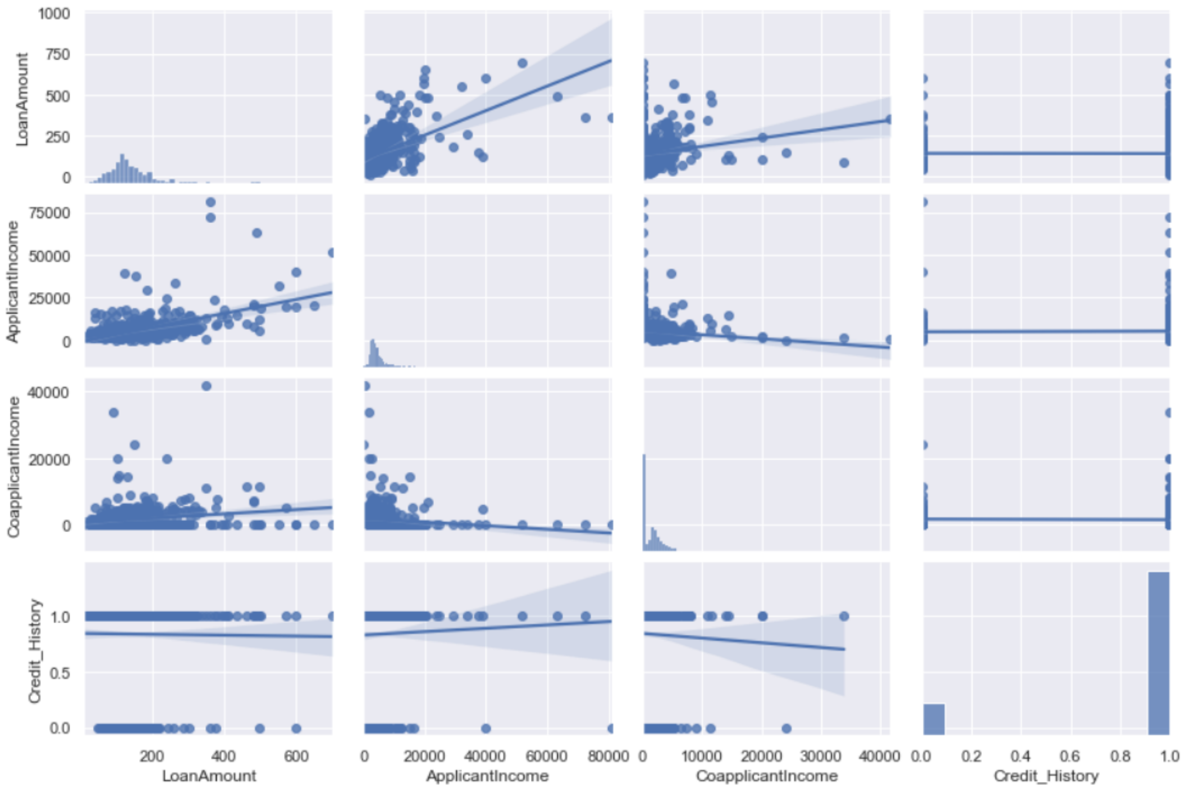


Fig.1. Matrix plot of independent variables

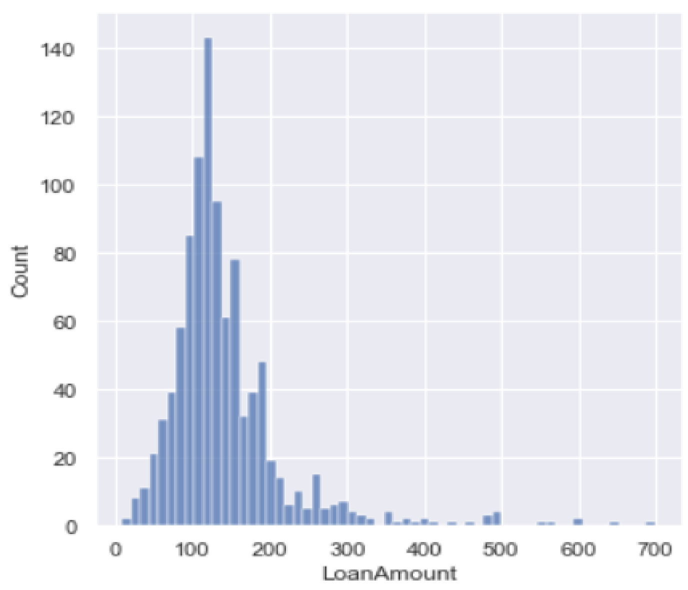


Fig.2. Probability plot of loan amount

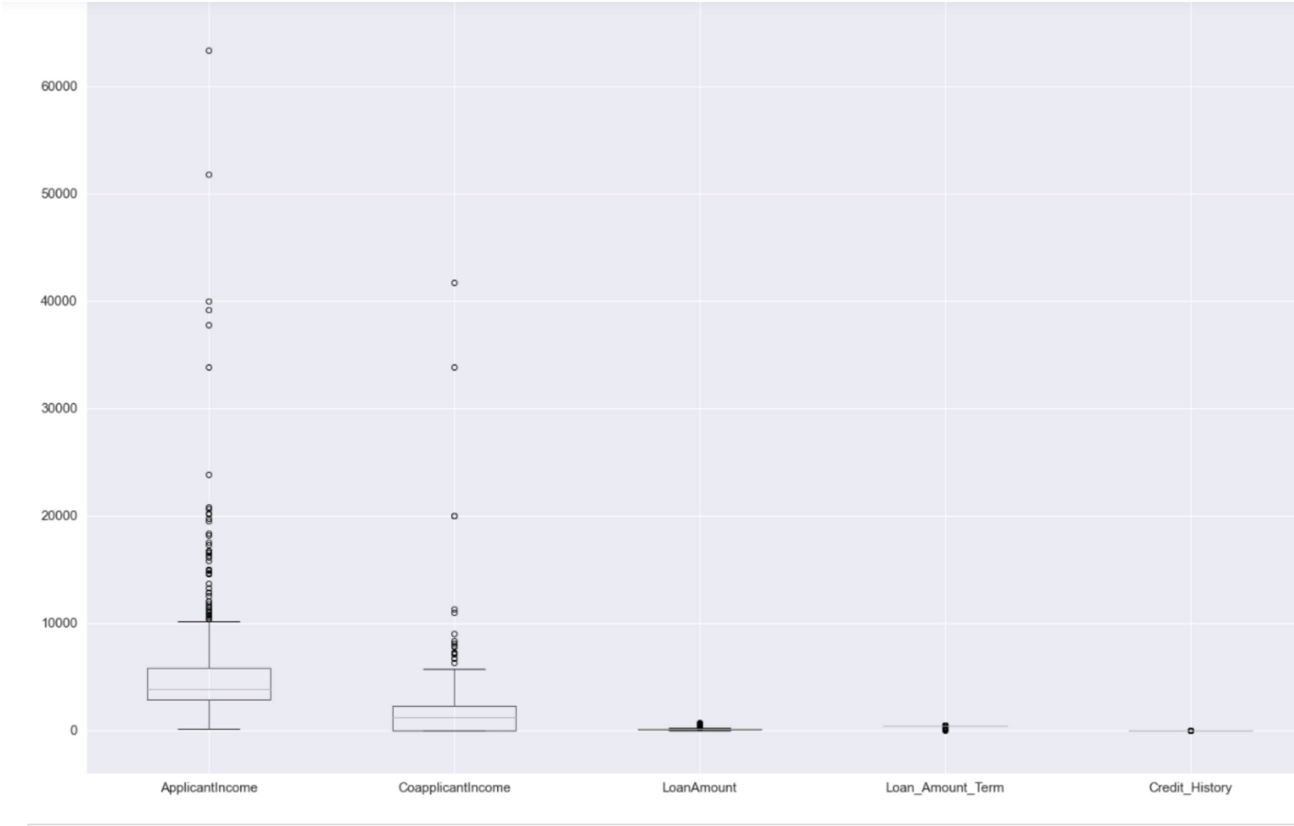


Fig.3. Box Plots for continuous independent variables

- 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**.

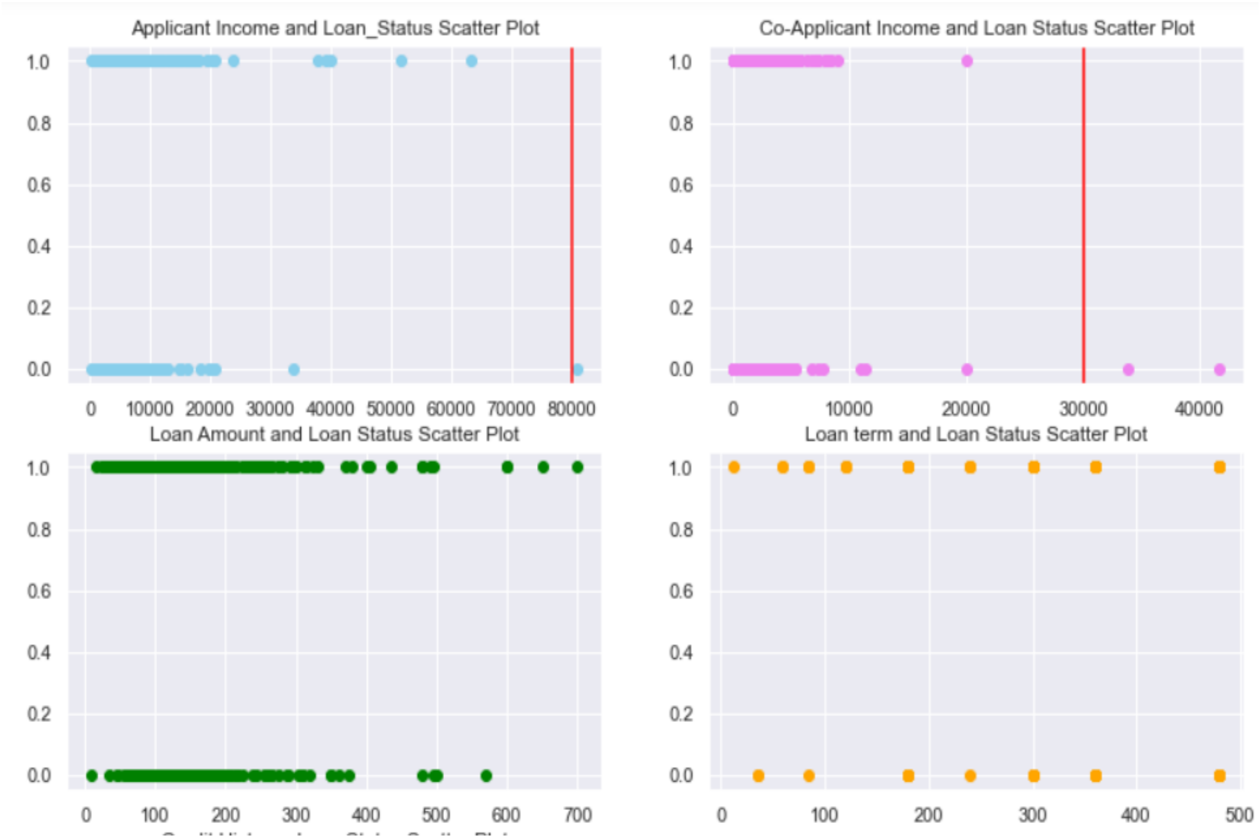


Fig.4. Potential Outliers

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%.

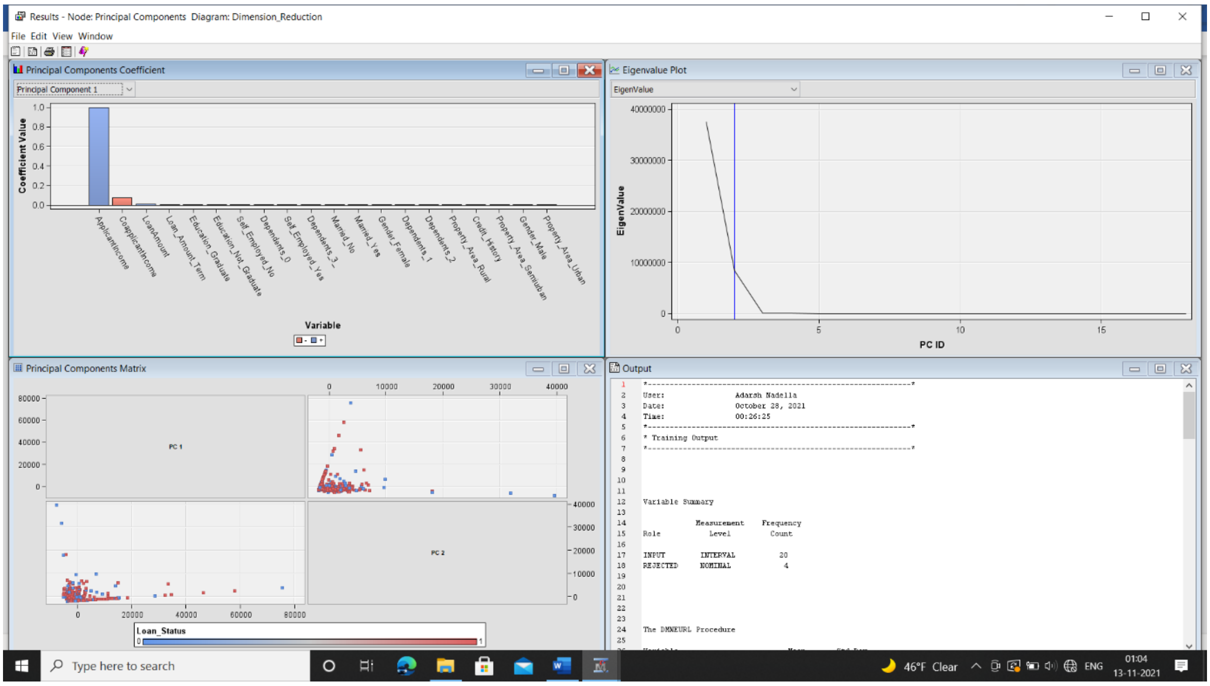


Fig.5. Principal Component Analysis

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

```
The DMNEURL Procedure

              Eigenvalues of Covariance Matrix

        Eigenvalue    Difference    Proportion    Cumulative

   1    37472895.1    29059716.3        0.8165        0.8165
   2     8413178.8     8408493.7        0.1833        0.9998
   3        4685.1         896.7        0.0001        0.9999
   4        3788.4        3787.9        0.0001        1.0000
   5           0.6           0.2        0.0000        1.0000
   6           0.4           0.0        0.0000        1.0000
   7           0.3           0.0        0.0000        1.0000
   8           0.3           0.0        0.0000        1.0000
   9           0.3           0.0        0.0000        1.0000
  10           0.2           0.0        0.0000        1.0000
  11           0.2           0.0        0.0000        1.0000
  12           0.2           0.0        0.0000        1.0000
  13           0.1           0.0        0.0000        1.0000
  14           0.1           0.1        0.0000        1.0000
  15           0.0           0.0        0.0000        1.0000
  16           0.0           0.0        0.0000        1.0000
  17           0.0           0.0        0.0000        1.0000
  18           0.0           0.0        0.0000        1.0000
  19           0.0           0.0        0.0000        1.0000
  20           0.0                      0.0000        1.0000









*-------------------------------------------------------------*
Summary of Exported Principal Components
*-------------------------------------------------------------*

Total number of input variables: 20
Maximum number cutoff of principal components: 20
Cumulative proportional eigenvalue cutoff: 0.95
Proportional eigenvalue increment cutoff: 0.001
Number of the selected principal components: 2
Total variation explained by the selected principal components: 0.9998153104
```

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.

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Loan Status		AIC	Akaike's Informati...	-656.474	.	.
Loan Status		ASE	Average Squared ...	0.216247	0.220505	0.207576
Loan Status		AVERR	Average Error Fun...	0.216247	0.220505	0.207576
Loan Status		DFE	Degrees of Freed...	429	.	.
Loan Status		DFM	Model Degrees of ...	1	.	.
Loan Status		DFT	Total Degrees of ...	430	.	.
Loan Status		DIV	Divisor for ASE	430	61	123
Loan Status		ERR	Error Function	92.98605	13.45082	25.5319
Loan Status		FPE	Final Prediction Er...	0.217255	.	.
Loan Status		MAX	Maximum Absolut...	0.683721	0.683721	0.683721
Loan Status		MSE	Mean Square Error	0.216751	0.220505	0.207576
Loan Status		NOBS	Sum of Frequencies	430	61	123
Loan Status		NW	Number of Estima...	1	.	.
Loan Status		RASE	Root Average Su...	0.465023	0.46958	0.455606
Loan Status		RFPE	Root Final Predicti...	0.466106	.	.
Loan Status		RMSE	Root Mean Squar...	0.465565	0.46958	0.455606
Loan Status		SBC	Schwarz's Bayesi...	-652.411	.	.
Loan Status		SSE	Sum of Squared E...	92.98605	13.45082	25.5319
Loan Status		SUMW	Sum of Case Wei...	430	61	123

Output

```
              Model Fit Statistics

R-Square         0.0000     Adj R-Sq         0.0000
AIC           -656.4744     BIC           -654.4629
SBC           -652.4106     C(p)            -0.4646


           Analysis of Maximum Likelihood Estimates

                                  Standard
Parameter     DF    Estimate        Error    t Value    Pr > |t|

Intercept      1      0.6837       0.0225      30.45      <.0001


*------------------------------------------------------------*
* Score Output
*------------------------------------------------------------*


*------------------------------------------------------------*
* Report Output
*------------------------------------------------------------*
```

Fig.8. Linear Regression with PCA

III. Linear Regression without PCA

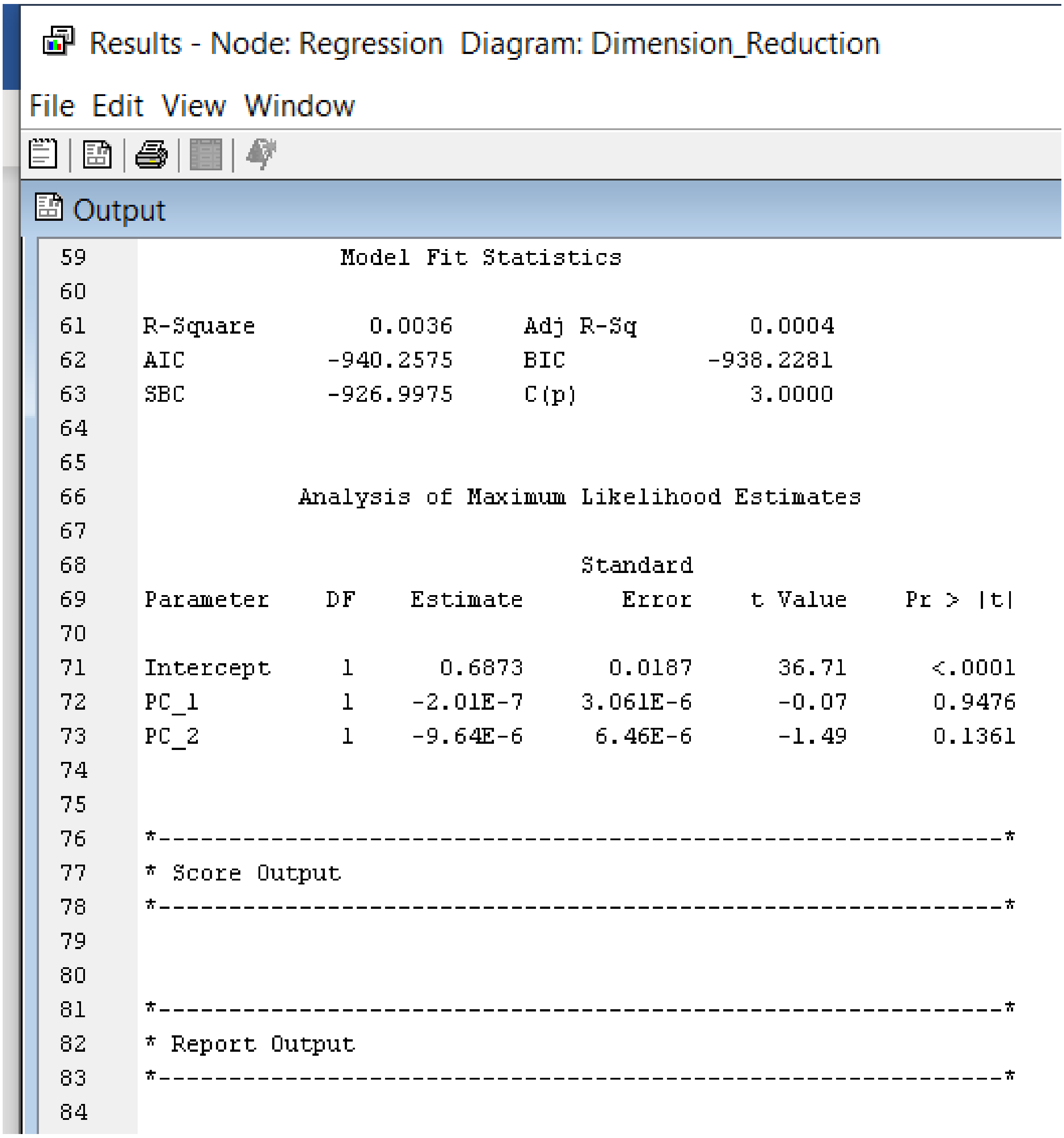
- Linear regression yielded more accurate results when with raw data with partitioning without PCA. Fig.8 shows the results and Hidden patterns.
- The root mean squared error (RMSE) is 37.91% which means the model classified 62.19% of loan status of customers accurately.
- Additionally, the hidden pattern 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

    1    Credit_History              1        1      176.61    <.0001              66.2879
    2    Property_Area_Semiurban     1        2        9.09    0.0027              65.1944
    3    Married_No                  1        3        9.67    0.0020              64.0772

The selected model, based on the cross-validation error rate, is the model trained in Step 3. It consists of the following effects:

Intercept  Credit_History  Married_No  Property_Area_Semiurban


                              Analysis of Variance

                                         Sum of
Source                   DF             Squares      Mean Square    F Value    Pr > F

Model                     3           29.963790         9.987930      67.51    <.0001
Error                   426           63.022256         0.147940
Corrected Total         429           92.986047


                   Model Fit Statistics

R-Square          0.3222    Adj R-Sq          0.3175
AIC            -817.7278    BIC            -815.6426
SBC            -801.4727    C(p)              3.4552


                       Analysis of Maximum Likelihood Estimates

                                              Standard
Parameter                  DF     Estimate       Error    t Value    Pr > |t|

Intercept                   1       0.0664      0.0515       1.29      0.1984
Credit_History              1       0.7129      0.0534      13.36      <.0001
Married_No                  1      -0.1237      0.0398      -3.11      0.0020
Property_Area_Semiurban     1       0.1208      0.0385       3.14      0.0018


                              Estimated Covariance Matrix

                                                Credit_       Married_       Property_Area_
Parameter                      Intercept        History             No            Semiurban

Intercept                       0.002656       -.002360       -.000428             -.000399
Credit_History                  -.002360       0.002848       -.000073             -.000163
Married_No                      -.000428       -.000073       0.001582             -.000048
Property_Area_Semiurban         -.000399       -.000163       -.000048             0.001480
```

Fig.9. Hidden Pattern in Data

IV. Logistic Regression with PCA

- Similar to linear regression with PCA logistic regression yielded un-satisfactory 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 t model underfitted the data.




```
                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

                                  Standard
Parameter    DF    Estimate          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


*------------------------------------------------------------*
* Score Output
*------------------------------------------------------------*


*------------------------------------------------------------*
* Report Output
*------------------------------------------------------------*
```

Fig.10. Under fitted Logistic regression

V. Logistic Regression without PCA

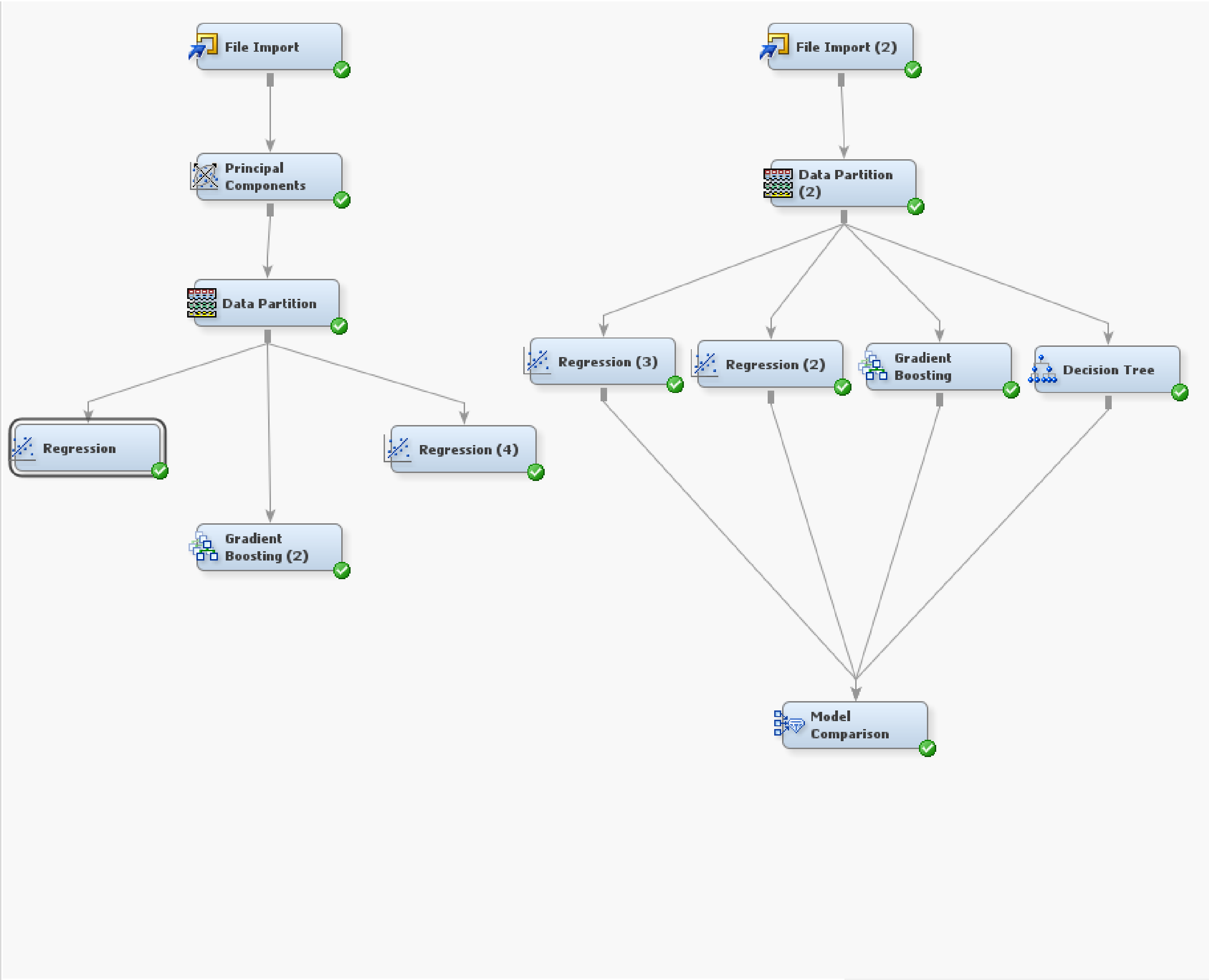
- Similar to linear regression without PCA, logit also gave same results with Adjusted R-squared of 31.66%.
- Logistic regression also gave the same hidden pattern criteria given by previous linear regression results further reinforcing our evidence that non married applicant with collateral property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

Output

```
                Model Fit Statistics

R-Square        0.3437     Adj R-Sq        0.3166
AIC          -803.5465     BIC          -799.9775
SBC          -730.3983     C(p)           18.0000


                  Analysis of Maximum Likelihood Estimates

                                                Standard
Parameter                  DF    Estimate       Error    t Value    Pr > |t|

Intercept                   1      0.1952      0.2421       0.81      0.4204
ApplicantIncome             1    3.569E-6    3.855E-6       0.93      0.3551
CoapplicantIncome           1    0.000010    0.000011       0.98      0.3265
Credit_History              1      0.7191      0.0542      13.26      <.0001
Dependents_0                1     -0.0237      0.1204      -0.20      0.8442
Dependents_1                1     -0.0461      0.1259      -0.37      0.7143
Dependents_2                1      0.0584      0.1270       0.46      0.6458
Dependents_3_               1      0.0637      0.1333       0.48      0.6330
Education_Graduate          1      0.0527      0.0460       1.14      0.2532
Education_Not_Graduate      0           0           .          .           .
Gender_Female               1      0.0187      0.1582       0.12      0.9060
Gender_Male                 1     -0.0304      0.1503      -0.20      0.8399
LoanAmount                  1    -0.00048    0.000271      -1.79      0.0750
Loan_Amount_Term            1    -0.00039    0.000311      -1.24      0.2158
Married_No                  1     -0.1184      0.0457      -2.59      0.0100
Married_Yes                 0           0           .          .           .
Property_Area_Rural         1     -0.0700      0.0483      -1.45      0.1483
Property_Area_Semiurban     1      0.0833      0.0455       1.83      0.0678
Property_Area_Urban         0           0           .          .           .
Self_Employed_No            1      0.0574      0.0913       0.63      0.5295
Self_Employed_Yes           1      0.0670      0.1028       0.65      0.5152


*------------------------------------------------------------*
* Score Output
*------------------------------------------------------------*


*------------------------------------------------------------*
* Report Output
*------------------------------------------------------------*
```

Fig.11. Logistic regression summary output.

Tree Diagram:



Insights:

- For flexibility in extracting appropriate applicants' data with above conditions (Married_No, Good_Credit History and property_Type_Semiurban), python's pandas data frame is used
- It is found that **14 out of 614 applicants are satisfying above criteria and the loan was denied** as shown in Fig.12.
- These **14 applicants constitute 2.83%** of the bank's customer.
- A separate CSV file is generated through python which contains the 14 applicants' and their corresponding details.

	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	Loan_Amo	Credit_His	Gender_F	Gender_M	Married_N	Married_Y	Dependen	Dependen	Dependen	Dependen	Education	Education	Self_Empl	Self_Empl	Property_	Property_	Property_	Loan_Stat	Loan_ID
2	360	1	1	0	1	0	1	0	0	0	1	0	1	0	0	1	0	0	LP001164
3	360	1	0	1	1	0	1	0	0	0	1	0	1	0	0	1	0	0	LP001656
4	360	1	0	1	1	0	1	0	0	0	1	0	1	0	0	1	0	0	LP001702
5	360	1	1	0	1	0	1	0	0	0	1	0	1	0	0	1	0	0	LP001708
6	36	1	1	0	1	0	0	1	0	0	1	0	1	0	0	1	0	0	LP001870
7	300	1	1	0	1	0	1	0	0	0	1	0	0	1	0	1	0	0	LP001925
8	360	1	1	0	1	0	0	1	0	0	1	0	1	0	0	1	0	0	LP002161
9	480	1	0	1	1	0	1	0	0	0	1	0	1	0	0	1	0	0	LP002187
10	360	1	1	0	1	0	0	1	0	0	0	1	0	1	0	1	0	0	LP002318
11	360	1	0	1	1	0	0	1	0	0	0	1	0	1	0	1	0	0	LP002444
12	360	1	0	1	1	0	1	0	0	0	1	0	1	0	0	1	0	0	LP002683
13	360	1	0	1	1	0	1	0	0	0	1	0	1	0	0	1	0	0	LP002697
14	360	1	0	1	1	0	0	1	0	0	1	0	1	0	0	1	0	0	LP002729
15	36	1	0	1	1	0	1	0	0	0	0	1	1	0	0	1	0	0	LP002768

Fig.12. Eligible applicants with loan denials.

Policy Implications:

- Provide the loan offers to above customers to minimize churn rate and improve customer satisfaction.
- Improve customer relationship management (CRM) by understanding the customer needs for loan and estimating risk involved in offering loan. Doing so will uncover the hidden patterns which are visually not possible to capture.
- Add the following set of features to the existing loan eligibility criteria if not done so:
    1. **Un-married**
    2. **Good credit history**
    3. **Collateral Property in Semi urban Area (or) Developmental area zone.**

Summary:

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 = 290%** **“0” Dependents = 56%** **“1” Dependent = 17%** **“2” Dependents = 16%** **“3+” Dependents = 8%**

Table 2. Summary of Data set and Report