

# Who would be approved for home Loan?

## Ramsey Abdelrahman Metro College Final Capstone

### Introduction

Metro Financial Services as fictitious company which is in the area of providing home loans wants to lower its risk of default payments. Financial organization wants to know, which customers should be given the loans for buying the home. The home loans data set has the data of 990 loans with their loan status (approved vs. not). Certain features have been defined and could have an impact of the loan status. The aim is to build up a predictive model to find out which features of the applicants could affect the loan status and in which sense.

### Hypotheses

#### Service Level Hypothesis

**Marital Status:** Single applicants are more able to have approvals as they request less loan amounts.

**Education:** Applicants who are more educated are more able to get approval as they have higher incomes.

**Self-employment:** Applicants who are self-employed are more able to get approval as they get higher income.

**Co-application:** Applicants whose co-applicants are working and having income are more able to have loan approval than those who are sole-applicants.

**Property Area:** Applicants who are living in urban areas are more able to get approval as their salaries are higher.

**Credit History:** Applicants who have available credit history are more able to get approval than those who don't.

**The Loan Amount:** Applicants who have high income apply for higher loan amounts.

#### Demographic Level Hypothesis

**Dependents:** Applicants who live in urban areas have fewer dependents than who live in semi-urban or rural areas.

**Co-application:** Applicants who live in urban areas are more probable to have co-applicants than applicants in semi-urban or rural areas.

**Credit History:** Applicants who live in rural areas are less probable to have available credit history than the ones in urban and semi-urban areas.

**Marital Status:** Single applicants are more probably living in urban and semi-urban areas than rural areas.

**Dependents:** Singles applicants have fewer dependents than married applicants.

**Gender:** Males tend to be more the main applicants while the females are the co-applicants due to cultural reasons.

**Gender:** Males are more able to have approved loans as they have higher income.

**The Loan Amount:** Applicants who live in urban areas tend to apply for larger amounts than those living in semi-urban or rural.

### Description of Data

• Two Datasets: one with Loan-Status, and one without.

• Total observations: 981 (Training: 614 vs. Test: 367)

• Number of original features: 13

• Number of total features after engineering: 28

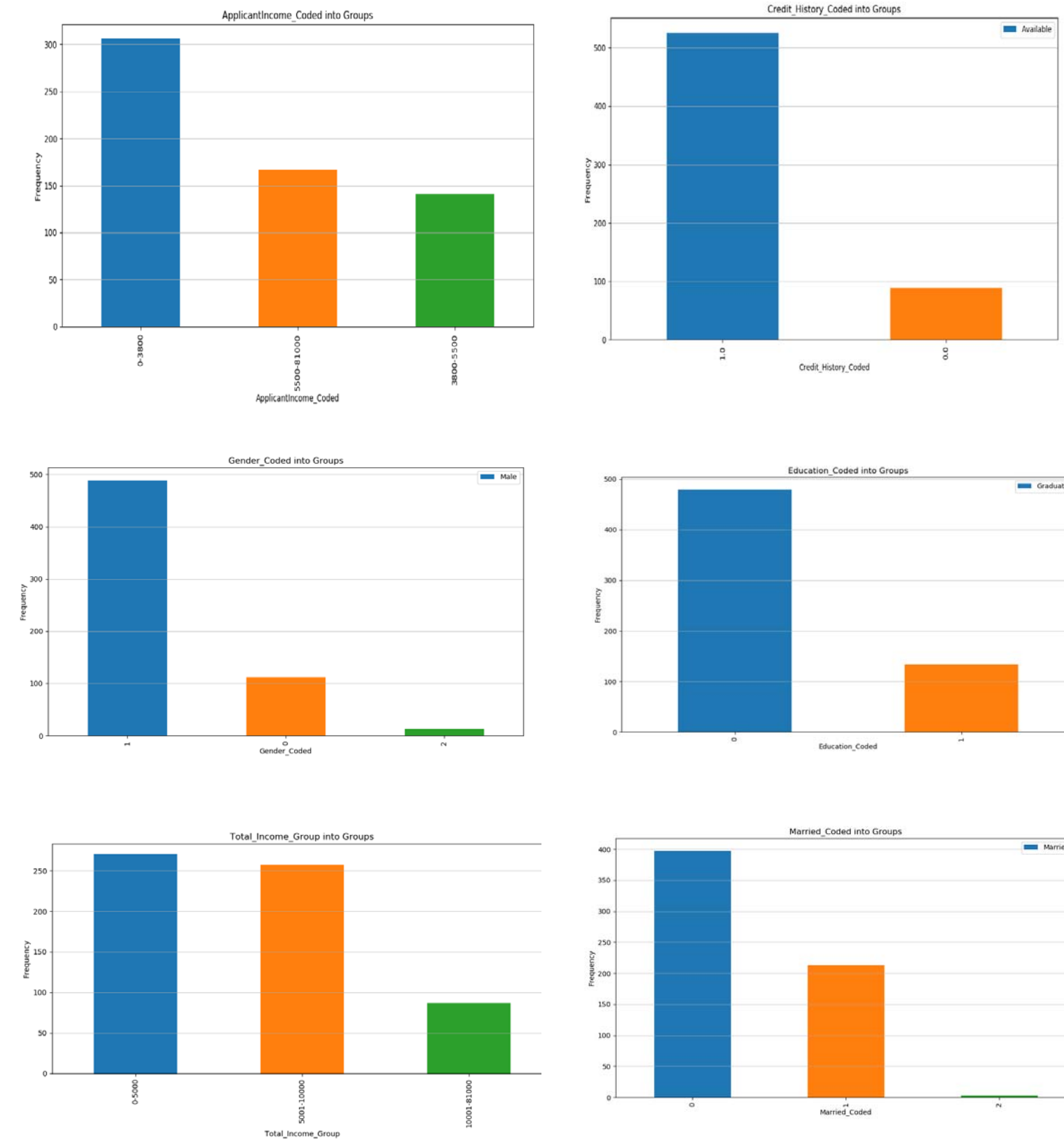
• Missing Values: All features have missing values except: ApplicantIncome, CoapplicantIncome, Education, Loan\_ID, and Property\_Area.

• Unique Values: All categorical features have unique features range from 2 – 5 except Loan\_Amount\_Term which had 13 and recoded into 2 features (other vs. 360). Continuous features have been recoded into ranges

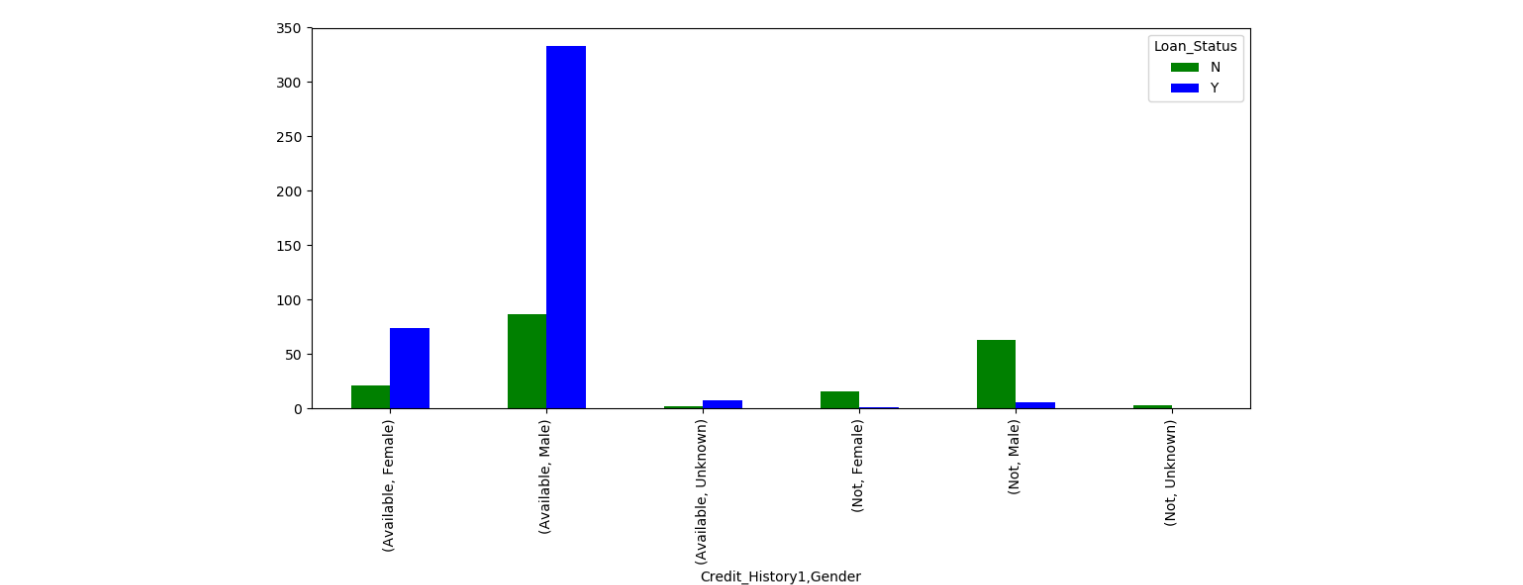
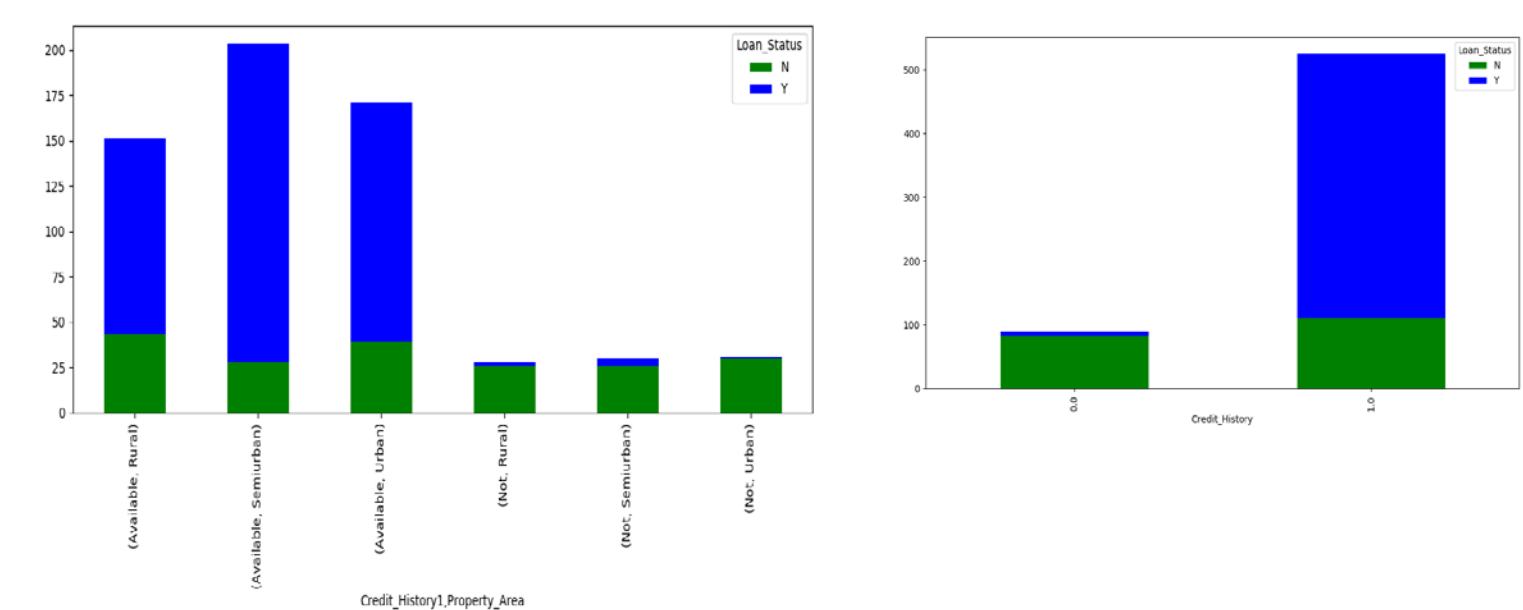
• All the missing values of the categorical features have been replaced by creating new category “unknown” – which believed to serve a better understanding and representation of population rather than using the mode or the median (clear biasness/ e.g. gender).

### Data Insights & Visualization

- Binning & Recoded Features

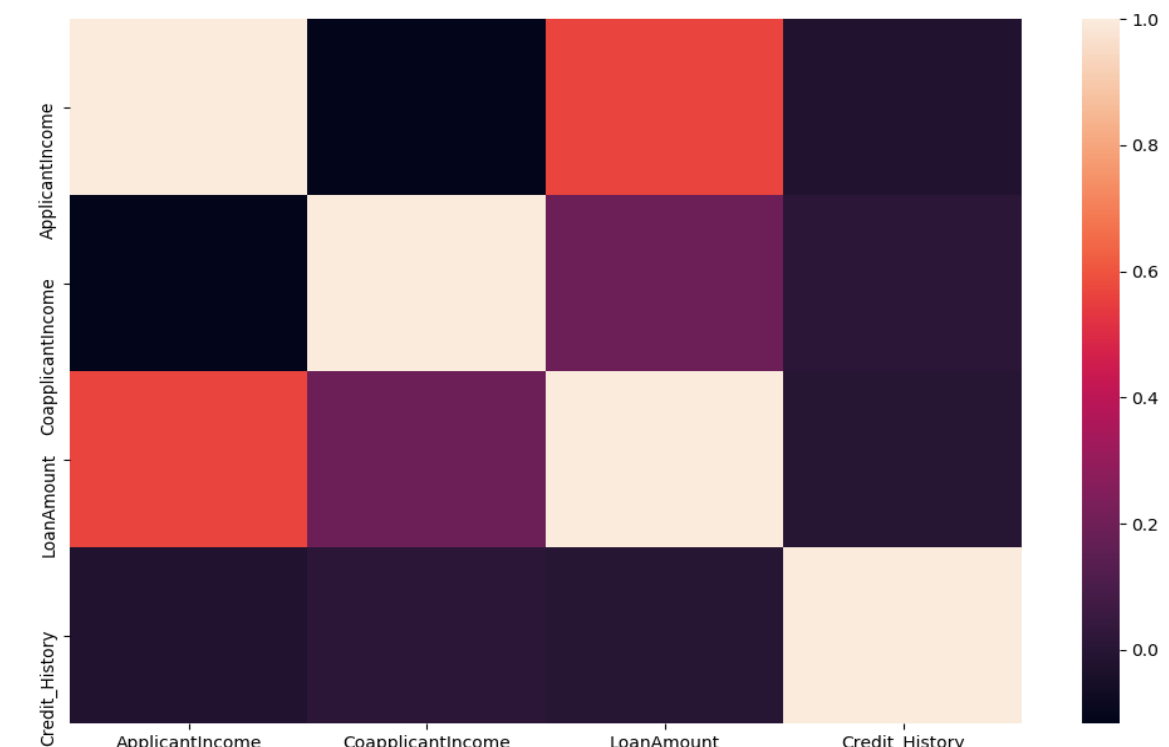


- Cross tab visualization

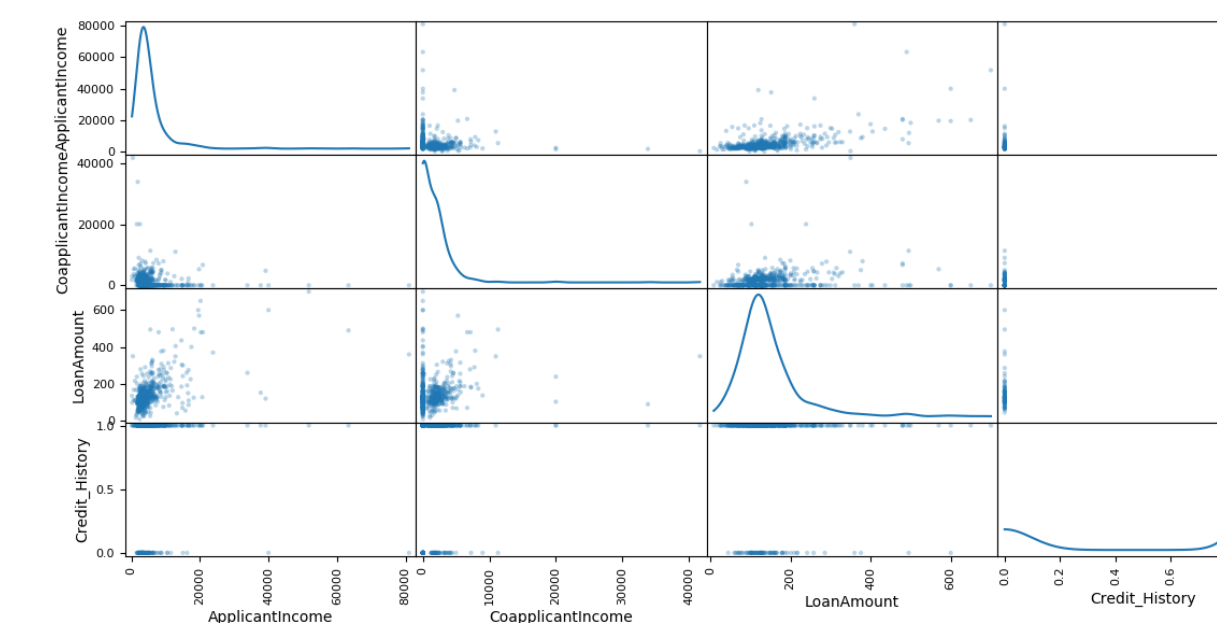


### Statistical Insights

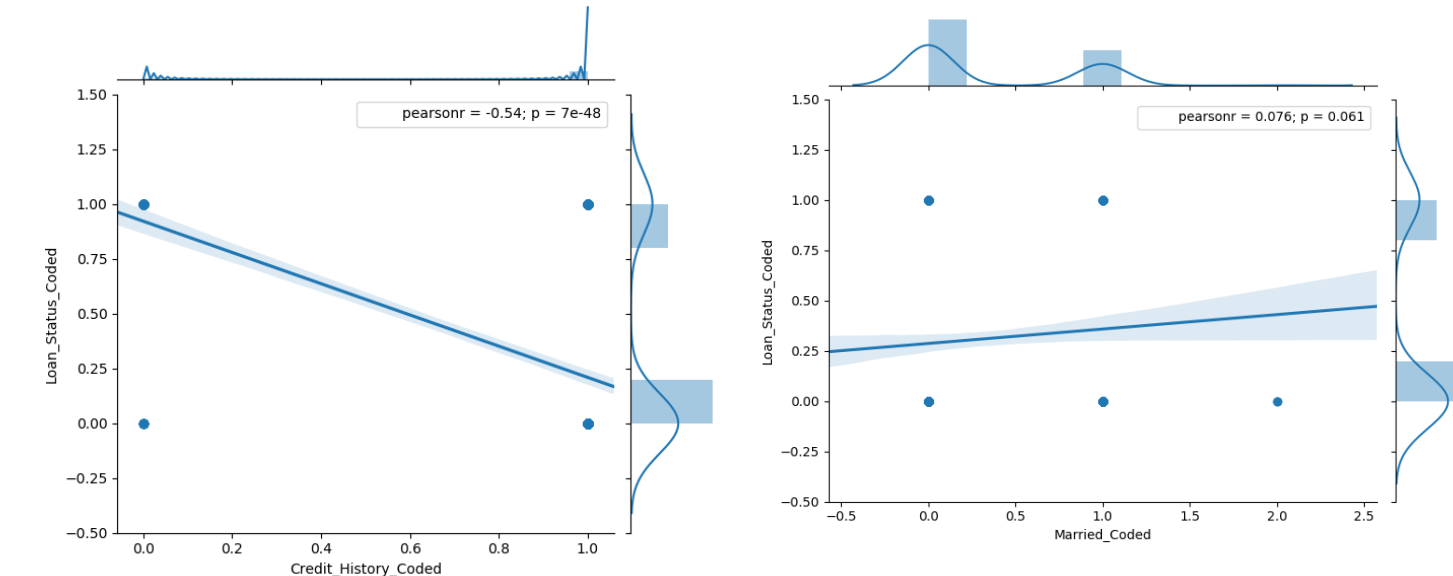
- Heat Matrix



- Distribution Matrix



- Joint Plots



- Baseline: Initial logit Model for estimating coefficient & signigance

Logit Regression Results					
Dep. Variable:	Loan_Status_Coded	No. Observations:	614		
Model:	Logit	Df Residuals:	591		
Method:	MLE	Df Model:	22		
Date:	Fri, 25 Jan 2019	Pseudo R-squ.:	0.2723		
Time:	11:55:35	Log-Likelihood:	-277.59		
converged:	False	LL-Null:	-381.45		
		LLR p-value:	3.495e-32		
	coef	std err	z	P> z	[0.025 0.975]
Intercept	2.3864	0.572	4.174	0.000	1.266 3.507
C((LoanAmount_Coded))[T.1]	0.3603	0.278	1.298	0.194	-0.184 0.904
C((LoanAmount_Coded))[T.2]	0.9771	0.489	1.999	0.046	-0.019 1.935
C((Total_Income_Coded))[T.1]	-0.6016	0.357	-1.687	0.092	-1.308 0.097
C((Total_Income_Coded))[T.2]	-0.7306	0.569	-1.285	0.199	-1.845 0.384
C((ApplicantIncome_Coded))[T.1]	0.2318	0.303	0.764	0.445	-0.363 0.826
C((ApplicantIncome_Coded))[T.2]	0.3580	0.453	0.791	0.429	-0.529 1.245
C((CoapplicantIncome_Coded))[T.1]	-0.2894	0.309	-0.935	0.350	-0.896 0.317
C((CoapplicantIncome_Coded))[T.2]	0.2099	0.358	0.586	0.558	-0.492 0.912
C((Education_Coded))[T.1]	0.4061	0.265	1.535	0.125	-0.112 0.925
C((Gender_Coded))[T.1]	0.0864	0.306	0.283	0.777	-0.512 0.685
C((Gender_Coded))[T.2]	0.3011	0.766	0.393	0.694	-1.200 1.802
C((Loan_Amount_Term_Coded))[T.1]	0.0997	0.299	0.334	0.738	-0.485 0.685
C((Married_Coded))[T.1]	0.5516	0.260	2.122	0.034	0.042 1.061
C((Married_Coded))[T.2]	-25.4560	4.29e+05	-5.93e-05	1.000	-8.41e+05 8.41e+05
C((Property_Area_Coded))[T.1]	0.6900	0.272	-2.562	0.010	-1.232 -0.164
C((Property_Area_Coded))[T.2]	0.1809	0.267	0.678	0.498	-0.342 0.704
C((Self_Employed_Coded))[T.1]	0.0005	0.331	0.001	0.999	-0.649 0.650
C((Self_Employed_Coded))[T.2]	0.2905	0.442	0.657	0.511	-0.576 1.157
C((Credit_History_Coded))[T.1,0]	-3.9838	0.427	-9.321	0.000	-4.822 -3.146
C((Dependents_Coded))[T.1]	0.0739	0.259	0.285	0.776	-0.434 0.582
C((Dependents_Coded))[T.2]	-0.2065	0.429	-0.482	0.630	-1.047 0.634
C((Dependents_Coded))[T.3]	-0.1191	0.843	-0.141	0.888	-1.772 1.534

### Modeling & Findings

• As the test set is the target to predict, 4 main classification models have been operated to choose best accuracy and performance: KNN, Logit, Random Forest, & Perceptron.

• SVM & MLP have been also explored additionally.

#### Procedure:

• The two datasets (train & test) were marked initial for later stages. (e.g. train[Trainsource]='train')

• The two datasets were concatenated into one large dataset for preprocessing and feature engineering.

• Later, the two datasets were separated.

• Four objects were created based on the required features (predictors) to predict the target: X\_train, X\_test, y\_train, y\_test.

• As the target is (y\_test), each model was run twice. The first time is to predict the y\_test and use it as true y\_test in the second time for performance measuring.

• Gridsearch and hyperparameter were also ran eventually for performance improvement.

#### Key Findings

• The loan amount requested affects positively on the loan approval based on the coefficient & level of significance.

• Marriage & Credit history also affects positively on loan approvals and have significance.

• Based on coding, large amount of request loan is more probable to be approved. Also positive & significant.

• People who are living in urban and semi-urban areas are more to be approved. (Significant & positive)

• Unavailable credit history affects negatively on loan approval. (significant & negative)

• The best model performance was for Random Forest:

• Score: 0.94

Best Accuracy: 0.85

RF_y_pred - NumPy array		RF_y_pred_proba - NumPy array		RF_Score 0.9413680781758957
0	0	0	0	
1	0	0.95	0.05	
2	0	0.673889	0.326111	
3	1	0.9	0.1	
4	1	0.15	0.85	
5	0	0.0571429	0.942857	
6	0	0.75	0.25	
7	1	0.9	0.1	
8	0	0.3	0.7	
9	0	0.85	0.15	
10	1	1	0	
11	0	0.0571429	0.942857	
12	0	0.8	0.2	
13	1	0.3	0.7	
14	0	0.70881	0.29119	
15	0	0.825	0.175	
16	0	0.8	0.2	
17	1	1	0	
18	0	1	0	
RF_cm - NumPy array		Accuracies_scores - NumPy array		
0	0	0	0	
1	0	0.808917	0	
2	0	0.853503	0	
3	0	0.840764	0	
4	0	0.821656	0	
5	0	0.846154	0	

### Models Results

Comparison of the ML Models Results					
ML Algorithm	Accuracy %	Score	Approved	Not Approved	Total
KNN	79%	0.81	320	47	367
Logit	83%	0.83	308	59	367
Perceptron	81%	0.78	212	155	367
Random Forest (RF)	85%	0.94	274	93	367