**Case Study: Analyze data for Home loan approval prediction**

**Introduction**

Dream Housing Finance company is a growing company that provides Home loans in all over US Areas. They first check the eligibility of the customers before customers apply for the application. The company is interested in automating the pre-approval eligibility process by evaluating the customer's data provided by them on the application. This is important, as this provides the customer their likelihood of getting approved before they apply for the loan. The data is provided by the company to determine the eligibility of the customer so they can target these customers. There are many factors that can determine the eligibility of the customer, such as education level, Income, Credit History, etc.

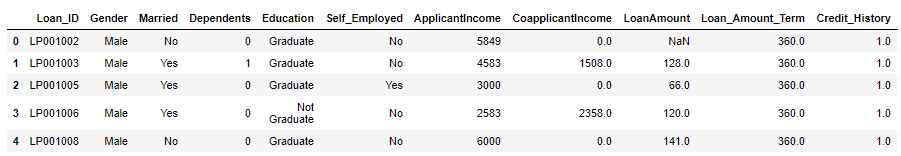
For this project, I will attempt to analyze what factors are significant in approving the loan, to predict who will get approved for the loan.

**Problem Statement:** Predict which customer is eligible for loan based on their background information.

**Data**

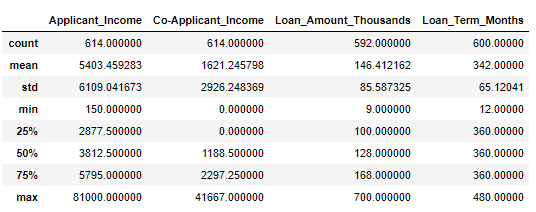
1. Data Source: [Loan Eligibility Data](https://www.kaggle.com/vikasukani/loan-eligible-dataset?select=loan-train.csv) from kaggle.com.
2. Data variables in the file are: Key Name Description

* Gender - Male/ Female
* Married - Applicant married (Y/N)
* Dependents - Number of dependents Education - Applicant Education (Graduate/ Undergraduate)
* Self\_Employed - Self-employed (Y/N)
* ApplicantIncome - Applicant income
* CoapplicantIncome - Coapplicant income
* LoanAmount - Loan amount in thousands
* Loan\_Amount\_Term - Term of a loan in months
* Credit\_History - credit history meets guidelines
* Property\_Area - Urban/ Semi-Urban/ Rural
* Loan\_Status - Loan approved (Y/N) – Target variable

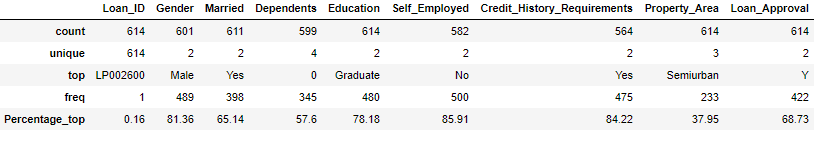


**Summary Statistics**

**Numerical**



**Categorical**



The minimum and maximum of applicant income is 150 and 81,000. So, Income can be an important feature. Co-applicant income can also be an important variable. About 84.22% of applicants met the credit history requirements. 85.91% of the applicants were self-employed. 81.36% of applicants were male.

**Key Observations**

1. There 81.36% male applicants and had 69.33% chance of approval. Slightly higher than females

2. There were 65.14% percent of married applicants, with approval rate of 71.61%

3. Although, there were higher number of applicants with 0 dependents, their approval rate (68.99%) was slightly lower than the applicants with at least 2 dependents (75.25%).

4. There were 78.18% college graduate applicants with approval rate of 70.83%.

5. There were 85.91% applicants who were not Self\_Employed. Approval rate for self-employed (68.3%) and non-self-employed (68.6%) were almost the same.

6. Applicants from Semi-urban areas are likely to be approved with 76.82% approval rate. There were also more applicants from semiurban areas.

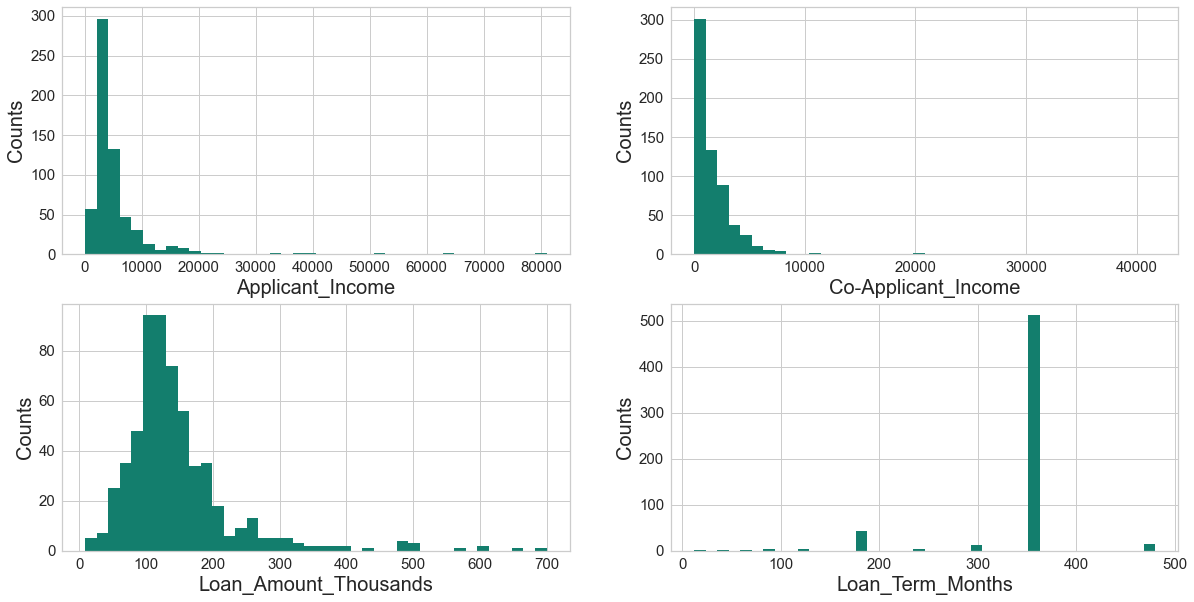
7. Applicants who meet the credit history requirements has 79.58% chance of getting approved, applicants who did not meet credit history requirements had 92.13% chance of getting denied.

8. About 68.73% of the Loans are approved out of total applications.

9. 85.3% loans were for 360 months, with 70.12% approval rate. Less than or equal 60 months loan term had 100% approval.

**Graph Analysis**

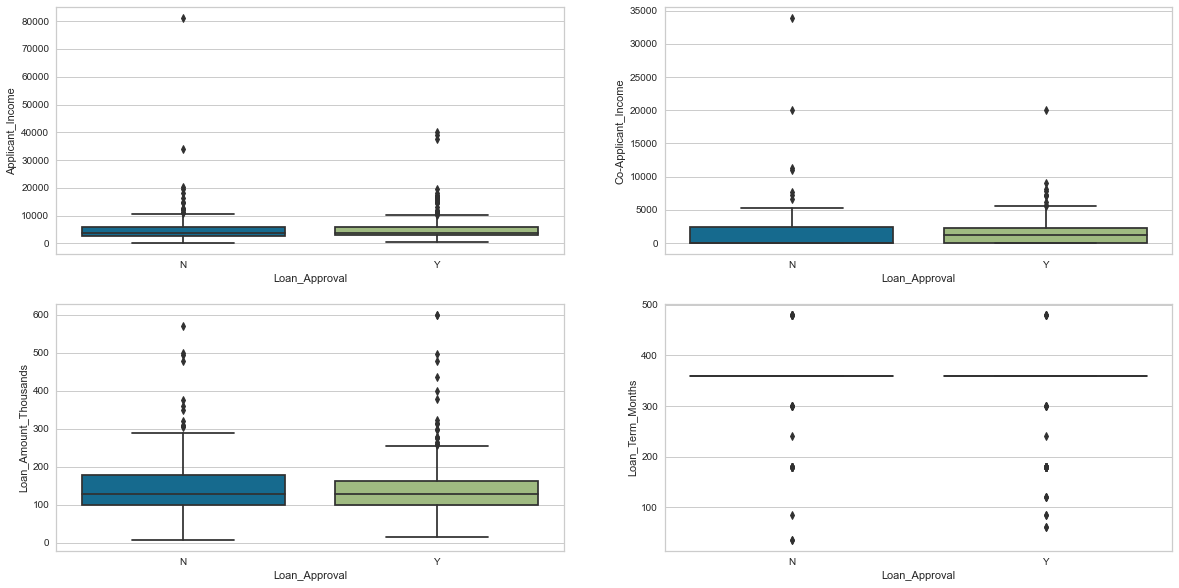
**Histograms for Numerical Data**



**Observations**

1. Applicant Income has a median 3812 with extreme value of 81,000.
2. Most Applicants had no Co-applicants with mean co-applicant income is 1621.
3. Median loan amount is ~14146,412.
4. Loan term seems like a categorical data type.

**Boxplot for Numerical Data**

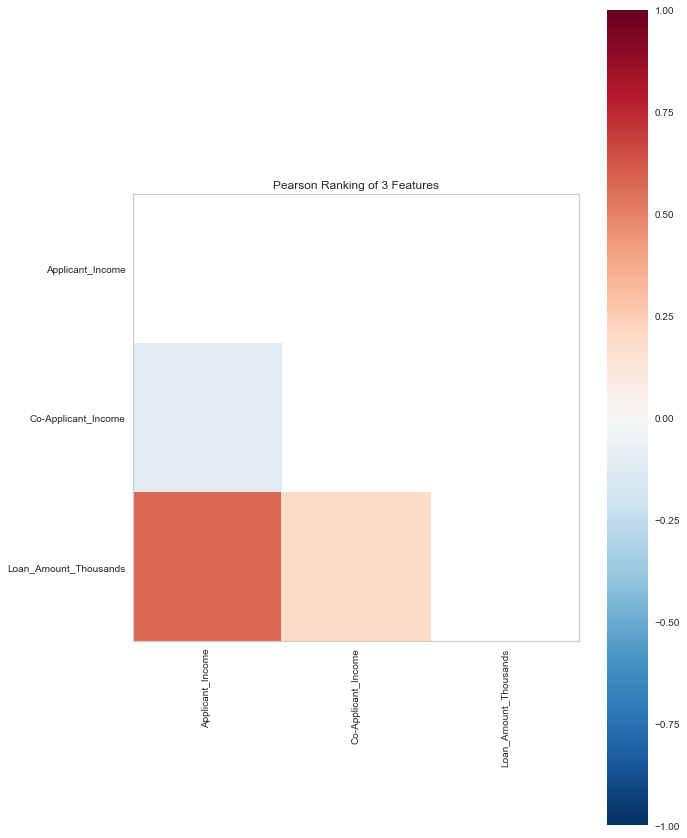
****

**Observation:** Numerical Data has no significant relevance to the Loan Approval

**Bar Charts for Categorical Data**

****

**Correlation Plot for Numerical Data**

****

**Observation:** There seem to have some positive correlation (~.50) between loan amount and applicant salary. There is a weak correlation between loan amount and co-applicant salary.

**Graph Analysis Summary**

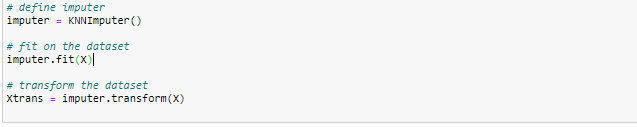
**Observations**

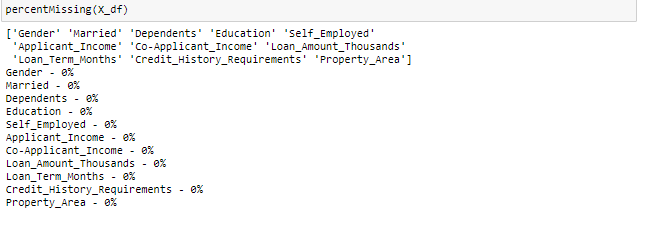
1. Married couple with 2 dependents is more likely to get approved.
2. It is possible for higher approval in males (need more analysis to confirm).
3. Applicants with graduate degree and applicants who are not self-employed has more chances for approval.
4. Applicant's residence area matter.

**Dimensionality Reduction**

**Handling Missing Values & Splitting the Data.**

I used KNNImputer() to predict and replace missing values. I then split the data in two sets of training and validation. Training set has 460 observations and 11 columns. Validation set has 154 observations and 11 columns.

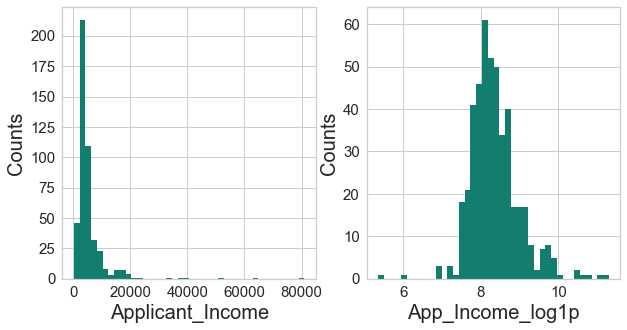
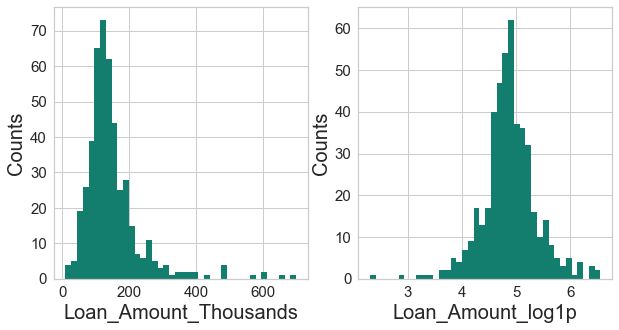
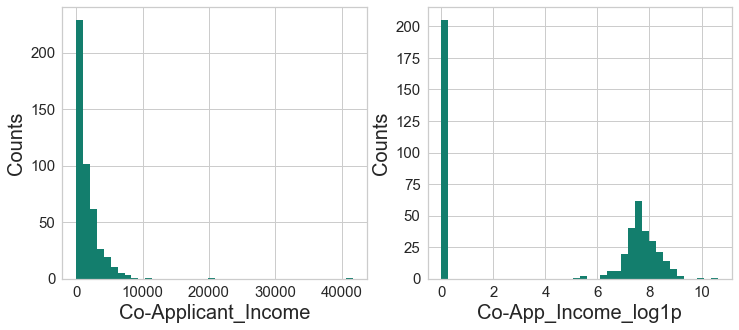




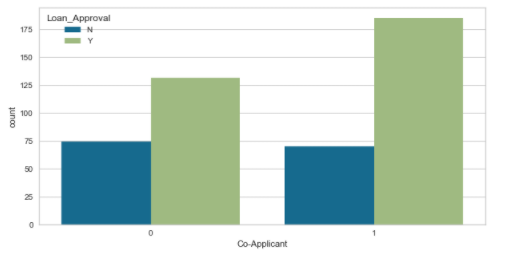
**Variable Transformation**

Log transforming variables (applicant's income, co-applicant's income, and loan amount because it is highly skewed data. Adding a Boolean variable of co-applicant and further extend the graph analysis for these three variables. I performed log transformation for both training and validation set.

**Plots for new variables**

****

**Co-Applicant**



Applicants with Co-applicant are more likely to get approved for the loan compared to applicants with no co-applicant. All the log transformed variables are approximately normal except co-applicant’s income, it has lot of 0’s indicating no applicant. Loan term with 360 months is more likely to get approved compared to other terms

**Dimensionality Reduction**

For Dimensionality reduction I will be creating 2 main data frames from original data frame: df\_cat with numerical variables turned into categorical and df\_num with all Boolean, Nominal numeric and continuous variables from training set.

* **Converting Numerical variables into categorical**

I converted numerical variables ‘App\_Income\_log1p’, ‘Co-App\_Income\_log1p’, ‘Co-App\_Income\_log1p’ by grouping them into their percentile bins (below 25th, 26-50th, 51-75th and 75th above). I did this so I can use chi-square test for feature selection. Created data frame **df\_cat** with 12 variables and data frame **subset\_cat** with 9 variables (excluding ‘**App\_Log\_Income\_Group’**, ‘**Co-App\_Log\_Income\_Group**’ and ‘**Loan\_Log\_Amount\_Group**’).

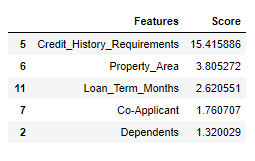
1. Gender: Male = 1, Female = 0
2. Married: Yes = 1, No = 0
3. Education: Graduate = 1, Not Graduate = 0,
4. Self\_Employed: Yes = 1, No = 0,
5. Credit\_History\_Requirements: Yes = 1, No = 0,
6. Co-Applicant: Yes = 1, No = 0
7. Dependents: 1 = 1, 0 = 0, 2 = 2 and 3+ = 3
8. Property\_Area: Semiurban = 1, Urban = 2, Rural = 3
9. App\_Log\_Income\_Group: below\_25th\_percentile = 1, 26th-50th\_percentile = 2, 51th-75th\_percentile = 3, above\_75th\_percentile = 4
10. Co-App\_Log\_Income\_Group: no\_applicants = 0, below\_25th\_percentile = 1, 26th-50th\_percentile = 2, 51st-75th\_percentile = 3, above\_75th\_percentile = 4
11. Loan\_Log\_Amount\_Group: below\_25th\_percentile = 1, 26th-50th\_percentile = 2, 51th-75th\_percentile = 3, above\_75th\_percentile = 4
12. Loan\_Term\_Months: 36, 60, 84, 120, 180, 240, 300, 360, and 480 are the categories

* **Converting categorical variables into Boolean variables, adding numerical variables, and identifying nominal variable**

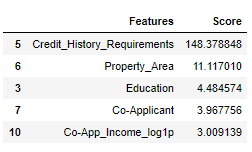
I created **subset\_num** data frame with only continuous variables 'App\_Income\_log1p','Co-App\_Income\_log1p', 'Loan\_Amount\_log1p' and created **df\_num** with all continuous variables and discrete variables.

* **Feature Selection: Chi-Square Method**

Top 5 Features:

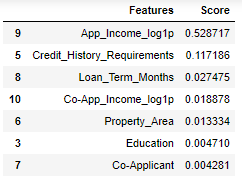


* **Feature Selection: ANOVA F-value statistic**

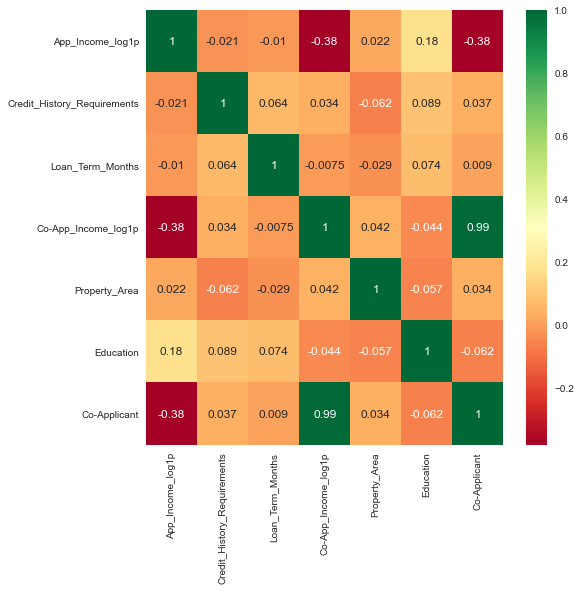


* **Feature Selection: Mutual information (MI) selector**

This method is for set of variables, that have continuous and discrete variables such as Boolean or Nominal. It uses score\_func=mutual\_info\_classif

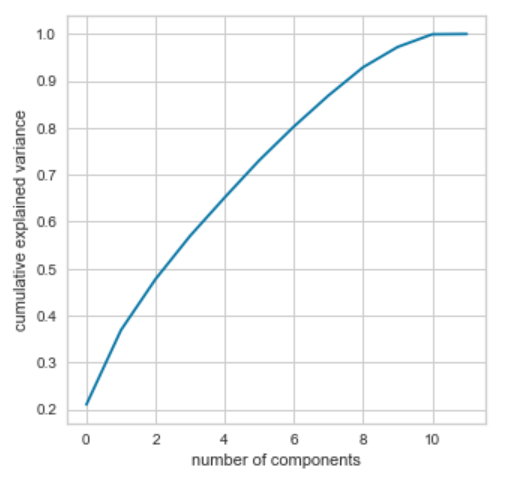


* **Correlation Check**



* **Feature Extraction: Principal Component Analysis**

Variance Plot for Components



I had 12 features to start with, and with PCA, it seems like 10 components is a good choice.

**Dimensionality Reduction Summary**

All Feature extraction or Selection methods has some limitations. Chi-Square works best when all features are categorical, and ANOVA works best when all features are numerical. I had mixture of both data types. For this reason, I attempted to convert continuous variables into category bins. Although, I had the option to do separate analysis for both data types, I do not believe separate analysis would give me any insights about patterns among all features. Mutual Information method works for both data types, but I must identify which ones are discrete. I believe this method gave me better insights on features. Further correlation analysis of these features showed some of the top features had some correlation.

I have decided to use feature selection method using Mutual Information classifier along with correlation chart instead of using PCA because I have too many categorical variables. Since feature selection methods use target variable to choose features, I have added another step of splitting the data into train and validation set before performing any data transformation or dimensionality reduction so this way I am not leaking any information from my validation set to training set.

I selected following Features based on Mutual Information and Correlation Chart:

1. Applicant Income log transformed

2. Credit History Requirements

3. Loan\_Term Months

4. Property Area

5. Education

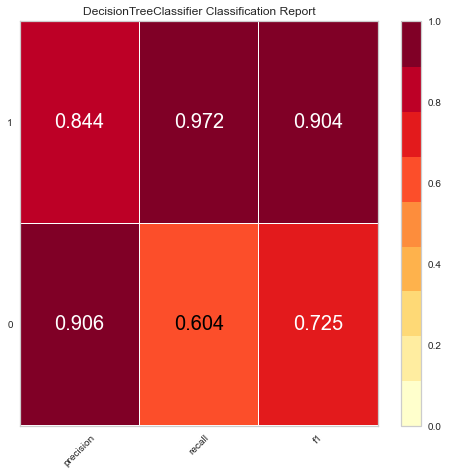
I decided to drop Co-App\_Income log transformed because it had some negative relationship with Applicant income log transformed (r = -.38).

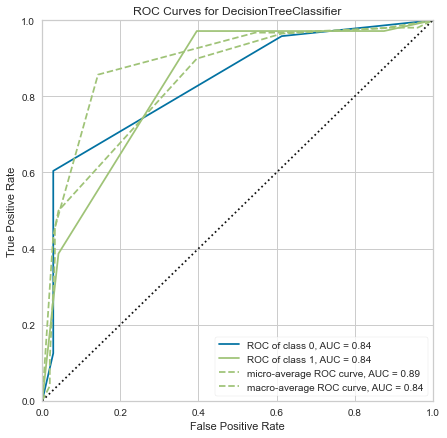
**Model Evaluation**

For this part I will be using Decision Tree, Random Forest and Logistic Regression.

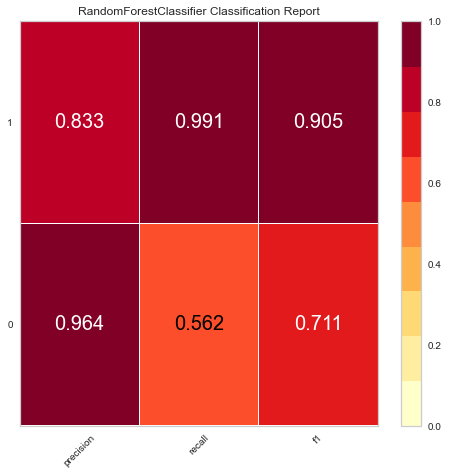
**Model Evaluation**

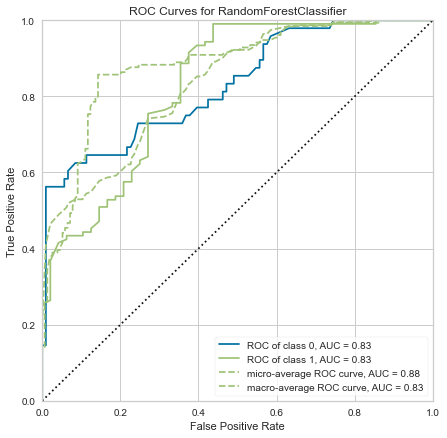
**Decision Tree Classifier**



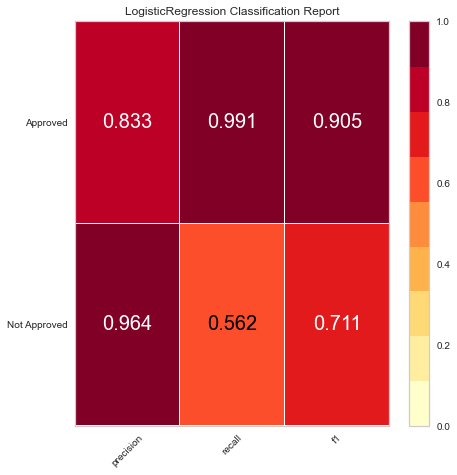


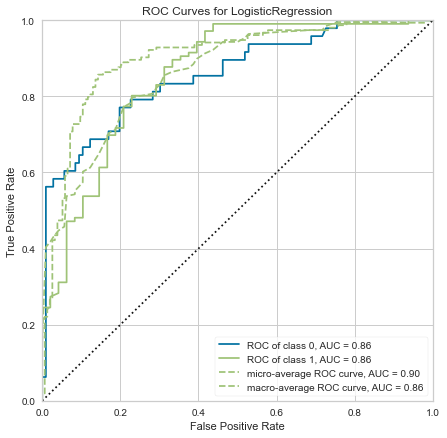
**Random Forest Classifier**

****

****

**Logistic Regression Classifier**

****

****

**Model Evaluation Summary**

**Key Steps before Model Evaluation**

To perform Model evaluation and selection carefully, I re-did part 1 & part 2. I moved my data cleaning step to part 2 and instead of simply dropping the variables to handle missing data, I added a step of imputing missing variables using KNNImputer(). This helped me keeping my data intact. I also split the data into training and validation set before I performed any data transformations and Dimensionality reduction. In part 1 I added another step of analyzing the summary statistics of the data.

**Key Insights**

Decision Tree

1. Cross validation Accuracy score: 0.817
2. Cross validation Precision score: 0.824
3. Cross validation Recall score: 0.943
4. Cross validation F1 score: 0.877
5. Decision Tree Score for test set 0. 8571428571428571

Random Forest

1. Cross validation Accuracy score: 0.850
2. Cross validation Precision score: 0.837
3. Cross validation Recall score: 0.981
4. Cross validation F1 score: 0.901
5. Random Forest Score for test set 0.8571428571428571

Logistic Regression

1. Cross validation Accuracy score: 0.857
2. Cross validation Precision score: 0.838
3. Cross validation Recall score: 0.990
4. Cross validation F1 score: 0.906
5. Logistic Regression Score for test set 0.8571428571428571

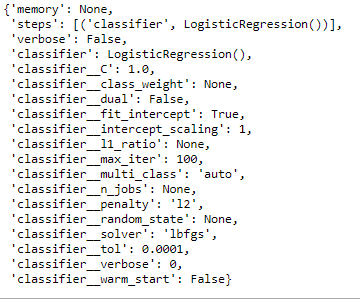
Accuracy score against test set for both models are exact same, however, when cross validated, logistic regression scores are slightly better than decision tree classifier and random forest classifier. The f1 score is higher by 1.26%. Although, Logistic regression looks like a better choice for this data, Further hyperparameter tuning will help make a better decision. Further analysis and cross validation,

**Model Selection**

I automated the model selection process by creating a dictionary of candidate learning algorithms (Random forest, decision tree and logistic regression) and their hyperparameters.

**Model Selection Summary**

Logistic regression with the following parameters seems be the best model out of the three algorithms:



**Conclusion**

All three models had not much difference in the accuracy score. Somehow, they had the same accuracy score. However, when dig deep and cross-validated. Logistic regression seems to outperform all the other model with accuracy score, precision score, recall score and f1 score slightly higher than the other two models. The accuracy score of logistic regression model is .857, meaning based on applicant’s income log transformed, credit history, loan term, property area and education, this model predicted 85.7% of the loan approval correctly. In real world though, this model will be hard to generalized because of imbalanced class. In my sample data as well, I had lot of imbalanced class such as 81% were male applicants. Although, I did split the data and stratified it to minimize the problem, in real world that will not be the case.

**References**

1. Brownlee, J. (2020, August 20). How to Choose a Feature Selection Method for Machine Learning. Retrieved February 02, 2021, from https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/
2. Wijaya, C. Y. (2020, July 07). 5 Must-Know Dimensionality Reduction Techniques via Prince. Retrieved February 02, 2021, from <https://towardsdatascience.com/5-must-know-dimensionality-reduction-techniques-via-prince-e6ffb27e55d1>

​