Description of preliminary data pre-processing:

Below are the first steps for data pre-processing:

1. Identify the null rows in the data set.
2. Fill in the null rows in one of the features “self employed” as “Other”
3. Drop the remining null rows.
4. Drop the Loan application column as it is not a feature that could impact the loan approval decision.
5. Export the csv of the clean data to be loaded to Machine leaning model.

Description of preliminary feature engineering and preliminary feature selection, including their decision-making process

1. First, looking at the statistics to identify if some the features had a large spread that required standardizing and scaling.
2. Look at the EDA to identify outliers and correlation amongst the features.
3. Remove the outliers (Tried removing the features but that did not result in improvement in model results to let it be there.
4. Also looked at the Class imbalance with the Counter formula.
5. Create a new column of Total Income from summing Applicant Income and Co Applicant Income
6. Encode the various categorical features using get dummies
7. Encode the Loan\_Status feature i.e. our outcome column using Label Encode since the output needs to be encoded in one single column

Description of how data was split into training and testing sets

Data was split into Train and Test usingthe *train\_test\_split()* function with

1. random state=1 so the results remain same each time the model is run and and
2. stratify=y  to ensure that both the train and test sets have the proportion of examples in each class and ensure that the Imbalance is not worsened.

Explanation of model choice, including limitations and benefits

This is a classification problem with binary output so Logistic regression, Decision trees and Random Forest classier can be utilized to analyse the model (supervised learning) can be used. If none of the models gives the right results then, Gradient Boost can be used to improvise.

**Logistic regression** is the base line model used in classification models. It evaluates the probability of an occurrence and draws a strict cut-off line at 50% that divides one classification from the other.

Advantage:Advantage of this model is that it is easier to implement, interpret, and very efficient to train. Advantage is that Continuous data work better with Logistic Regression.

Disadvantages:

1. The disadvantage of using the model is that Categorical data does not work well with Logistic Regression.
2. Also, Logistic regression cannot handle skewed data well and will push the decision boundary towards the outlier.

**Decision Tree**

Decision Tree (DT) handles the disadvantages of the Logistic Regression well. Some benefits of decision tree are:

1. Categorical data works well with Decision Trees, while continuous data work well with Logistic Regression.
2. Decision Trees handle skewed classes nicely if we let it grow fully.
3. Won’t be affected by outliers.
4. Decision tree will handle missing values. It will keep all missing values in one node during the split.

Limitation of Decision tree is that

1. Complex tree: Complexity to increase as the input increases.
2. Overfit: Decision Tree will overfit if we allow to grow
3. Train/Test time: Comparing to logistics regression decision Tree takes more time for training and it is slower for testing.

**Random forest:**

We also decided to use random forest model as decision trees sometimes are prone to overfitting and single tree may become complex and prone to errors, but many of them can be combined to form a stronger model. One nice byproduct of the random forest algorithm is to rank the features by their importance, which allows us to see which features have the most impact on the decision. The main limitation of random forest is that a large number of trees can make the algorithm too slow and ineffective for real-time predictions.

**Gradient Boost:**

Gradient boosting, is an ensemble method that works sequentially to minimize similar errors. It works by adding a small tree (called a stump) sequentially to the model.

Advantage of the model is that it often provides predictive accuracy that cannot be trumped. No data pre-processing required - often works great with categorical and numerical values as is.

Disadvantage is that it can be computationally expensive and is prone to overfitting.