Deliverable 1 - Loan Approval Predictor

For the final project, I plan to design a machine learning model that will predict loan approval by a bank institution based on a dataset containing the conditions of a load and the whether the loan was approved or not.

1- Dataset: https://www.kaggle.com/vikasukani/loan-eligibility-prediction-machine-learning/
2- Data Manipulation:

- a) Some preprocessing of the data will need to be completed in order to use it with the model, converting all of the categorical string fields into numbers, preferable 0/1. The fields that will need to be converted include Gender, Education Status, Property Area, Self Employed, Married Status and load status.
  Moreover, I will have to remove some of the rows that have any missing data values in them (after checking which data fields are critical to the correct approval of the loan).
  The id data can be preserved as a way to keep track of the different loans without being used as a feature of the model.
  Some analysis would have to be completed in order to determine whether the number of dependents is important or we just need to know if the client has a dependent or not independent of the number of them. (Would need to be processed in different ways).

- b) Machine Learning Model: I plan to use logistic regression to predict Yes/No for the loan approval. Since we are basically predicting the possible outcome which is composed of two values based on a bunch of parameters, the logistic regression model is the most appropriate model in my opinion since it does exactly this. More specifically the model is a binomial logistic regression model since the target variable can only be two values 0/1 indicating No/Yes in regards to the approval. I choose this model because it is a model that can be trained fast and it is a model that is also relatively easy to implement.
  The advantages of using this model is that it is a very accurate model and is less prone to overfitting. We have a high observations/data ratio.
  The cons of this model is that it cannot be used to predict all relationships and mostly complex relationships and is only limited to two target variables. In addition, we need to have enough data to prevent overfitting.

- c) I plan to use a classification matrix as a way to display the four possible prediction categories: correct positive/negative and false positive/negative. This would help assess the accuracy of the model. Moreover using the two variables of accuracy and precision recall we can see how good the model is at predicting on both the data it was trained with and other test data.

- d) I plan to showcase the results of my model training using a django web app. On the web app, I plan to display the data along with the results of quantifying the accuracy of the model on the train and test set to show how the model behaves with data it was not trained with.

3- Visualization/Application:

- I plan to build a django web app that would allow a user to predict whether they will get the loan or not by inputting the corresponding fields and quickly getting the results.