The purpose of this assignment is to create a model that predicts whether or not a loan will default using historical data. For loan companies being able to predict if a loan will default or not is very important so that risk is managed appropriately.

The data that I have to analyse is from 2007 to 2015. The model will need to be able to predict categorical variables of not fully paid or fully paid.

The first step is to perform EDA (exploratory data analysis). By building a count plot I noticed that the data is highly imbalanced and when this happens if the data is not oversampled or under-sampled then the model will have a high bias towards the category that has most data. In the given dataset I was able to identify that 83% of the data was for loans that have been paid and 16% for loans that have not been paid.

Proceeding further with EDA I wanted to verify if there is a direct correlation between any feature and the ability of the borrower to pay back the loan.

I decided to build a bar plot to check what are the different purposes for which the loan has been taken, from the count plot I identified that the higher purpose for not paid debt is debt consolidation.

Followed by a histogram that compares the credit policy to their FICO score, from which I noticed that the histogram is slightly right skewed but but it looks as expected, a high FICO score will have a low rate of not paid loans, and the people with low FICO score are more likely to default on their debts, so it can be concluded that the FICO score is a good indicator of successful repayment, as I was able to observe by the next graph where I compared the successful repayment of the loan against the FICO score of the borrower.

Then I wanted to check if a high interest rate will cause the borrower to default on their payment. I first plotted in a line plot the FICO score against the interest rate, which shows that the higher the FICO score the lower the interest rate, and then I used a histogram to check the distribution of default of the borrower agains the interest rate and here I observed that the data is normally distributed, which means that the interest rate is not correlated to the successful loan repayment.

Following EDA I then proceeded with transforming the categorical features into numerical ones so that I could proceed with oversampling the not paid data. When oversampling the dataset becomes balanced and the model does not risk to be overfitted, for this process I used sample method from pandas.

Then I wanted to verify if there are highly correlated features that I could remove which will make the model simpler and faster. For this I used a heat map and also created a function that helps me identify highly correlated features, I set the threshold for this function to be around 65%, from this we can see that

interest rate and FICO have an inverse corration of 68% against the target variable, because of that I decided to remove interest rate from the dataset.

Finally I removed the target variable from the dataset and proceeded with splitting the data into train and test split. I then I applied a normalisation technique called MinMaxScaler, and then I proceeded with building the deep learning model.

The first layer has 19 neurones because the data has 19 features after transforming the categorical variables, for each layer I used a drop out of 0.2 to lower the cut-off line in binary prediction to reduce type 2 error at the cost of increasing type 1 error, for this problem type 2 error is more serious because a type 2 error will label borrowers as being able to pay back when they are not. For the loss function I used binary cross entropy which measures how well a classification model is performing. Also to avoid overfitting of the model without compromising on model accuracy I used early stopping which is a form of regularization.

The final model build has an accuracy of 65%.