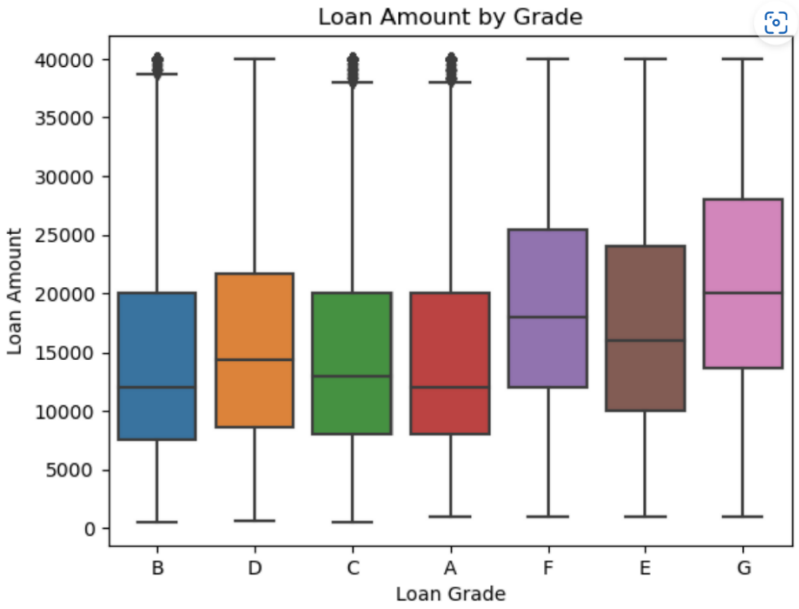
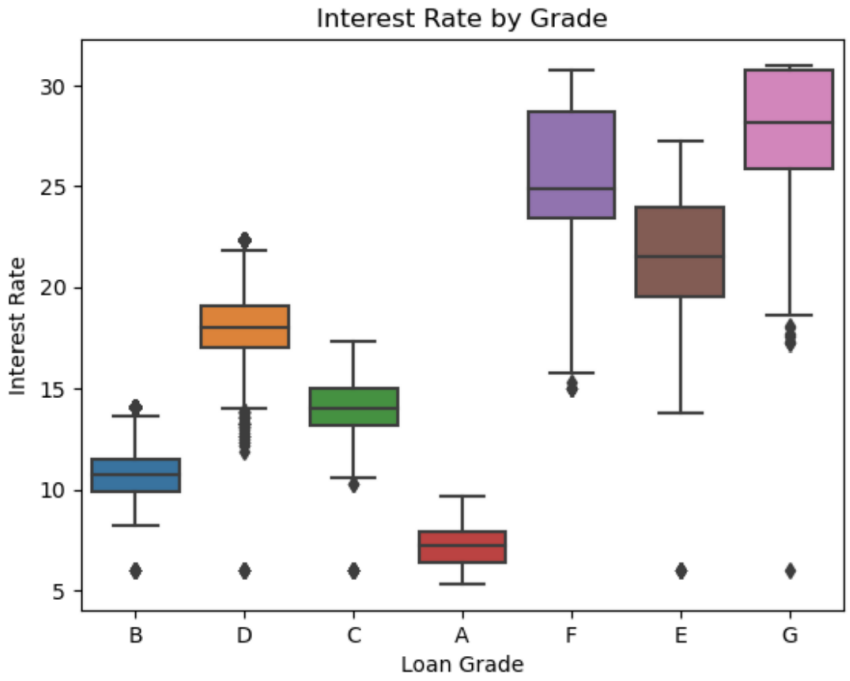
Prediction Report of the Big Data Assignment

After cleaning the data, we tried to find a relationship between the Loan Grade and Loan Amount, Loan Grade and Interest Rate. We have found the following relationship:



In this illustration, we can see that there is a clear pattern of the distribution of the loan amount. Even though the range of the A,B and C graded loans are same, the median is lower for A and it is slowly increasing as we are moving from A to C. From D, the range of the loan amount is increasing as well as the median level. As the loan grade is changing from A, the median is also increasing. That's the pattern we are clearly observing.



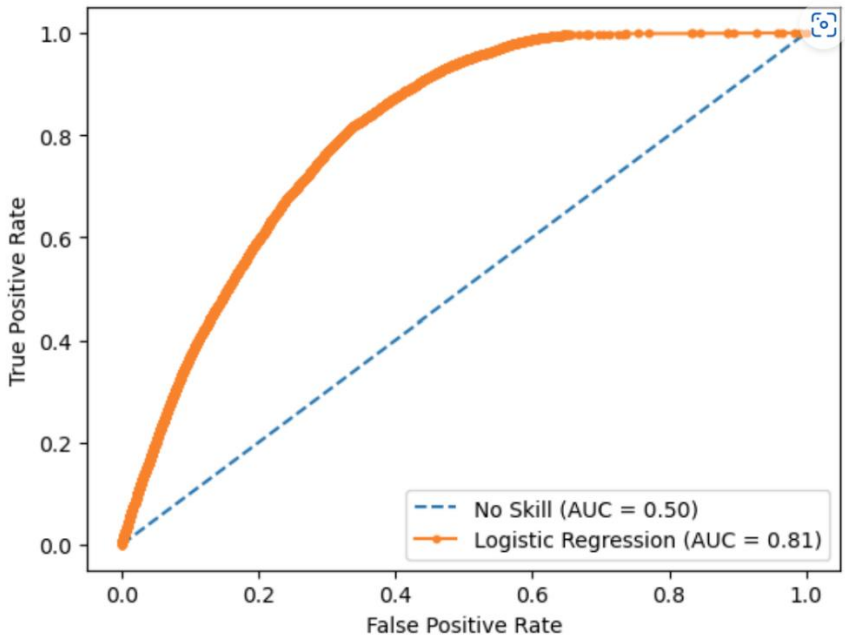
In this example, we can observe that the distribution of the loan amount follows a very apparent pattern. Even if the range of loans with grades of A, B, and C is the same, the median for grade A loans is lower and gradually rises as we move from A to C. The range of loan amounts and the median level both rise from D. The median is rising along with the change in loan grade from A. That is the pattern that is readily apparent.

The confusion matrix is a performance evaluation metric for classification models. It is a table that shows the true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions made by the model.

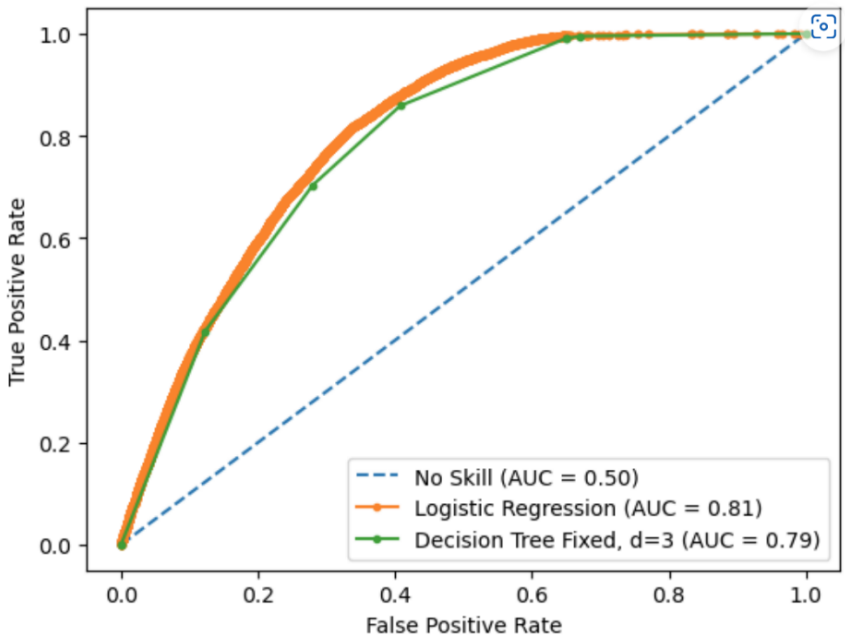
This is a 2 by 2 confusion matrix that illustrates the following:

- True positives (TP): 255 - This means that 255 positive cases (in this case, the class 'Charged Off') were correctly predicted by the model.
- True negatives (TN): 59554 - This means that 59554 negative cases (in this case, the class 'Fully Paid') were correctly predicted by the model.
- False positives (FP): 434 - This means that 434 negative cases were incorrectly predicted as positive cases. In other words, the model predicted that 434 loans would be 'Charged Off', but they were actually 'Fully Paid'.
- False negatives (FN): 7576 - This means that 7576 positive cases were incorrectly predicted as negative cases. In other words, the model predicted that 7576 loans would be 'Fully Paid', but they were actually 'Charged Off'.

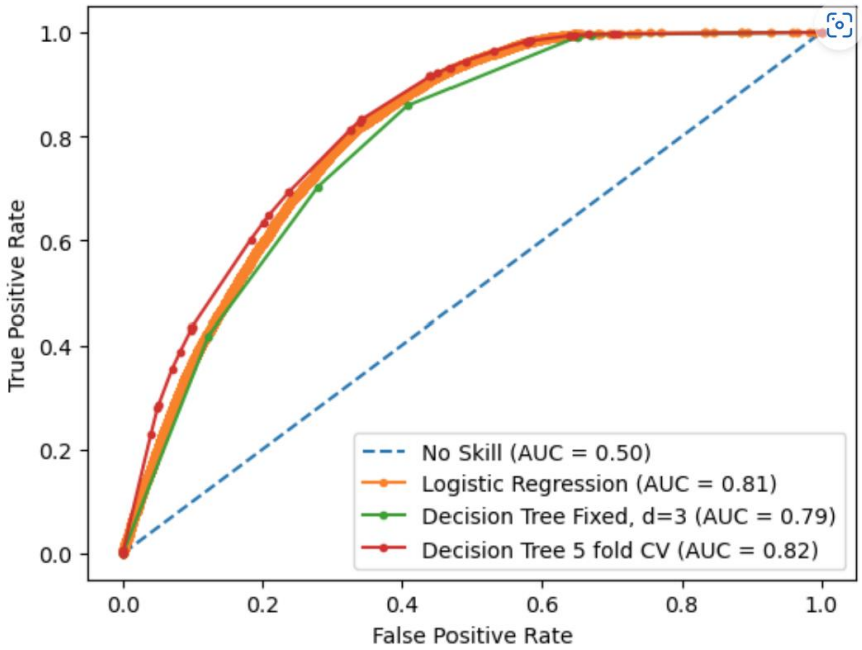
Then we have illustrated the ROC of No Skill predictors and the Logistic regression and we have found out that No skill is nothing but a random guess as AUC is .50, but logistic regression's AUC is .81 which is reasonably a good model. The picture is given below:



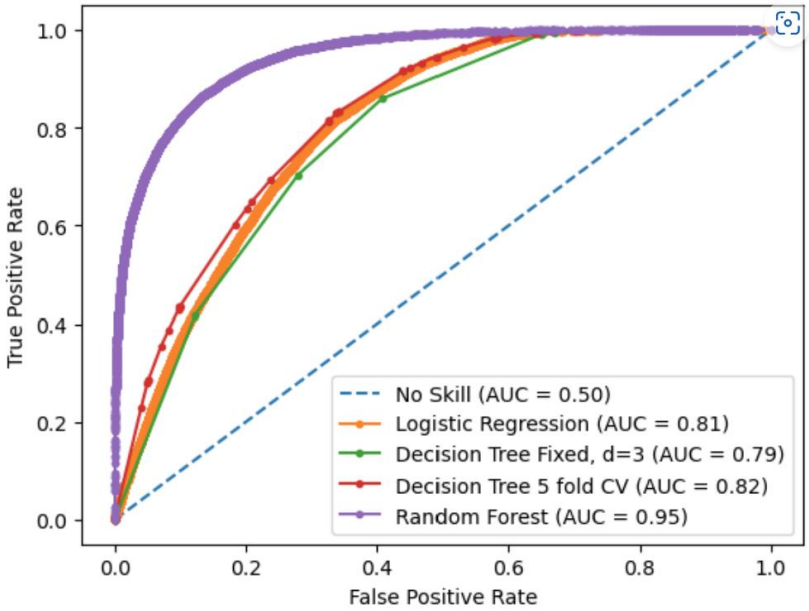
Next, we have tried to show a ROC of No skill, Logistic regression and Decision Tree without cross validation. We can see that even though this decision tree is not a random guess, but it is not better than logistic regression. The picture is as follows:



Now, we are plotted the ROC curve with decision tree with cross validation and it is so far the best among four models. We can see that the AUC of the decision tree with cross validation is .82 which is the highest among the four. The picture is given below:



This is the final ROC curve where we have shown all the five models and we can clearly see that Random Forest has outperformed all other models that we have considered. The AUC is .95 which is the highest than the rest of the four. Hence, our random forest is the best model and No skill is the worst one which nothing but a random guess.



In case of feature importance using permutation on full model, we can see that “Interest has the highest importance as it has the highest point. Total payment is the second and it continues till Annual income which has the last importance. In between, we have variables like Total_rec_int, Last payment amount, instalment and Loan amount. The picture is given below: