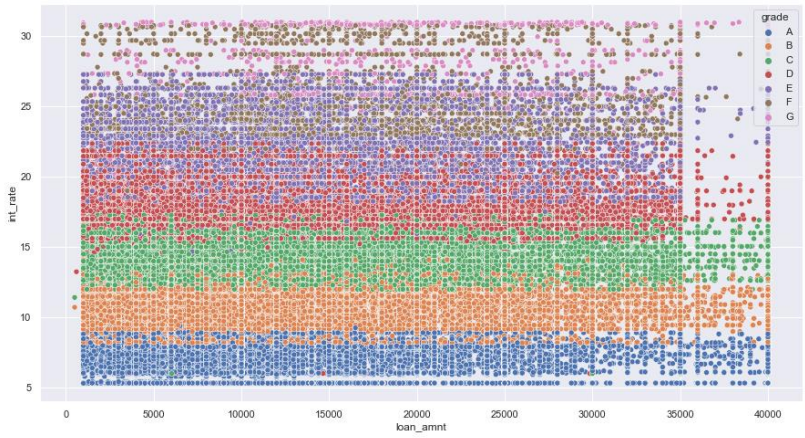
Prediction Report

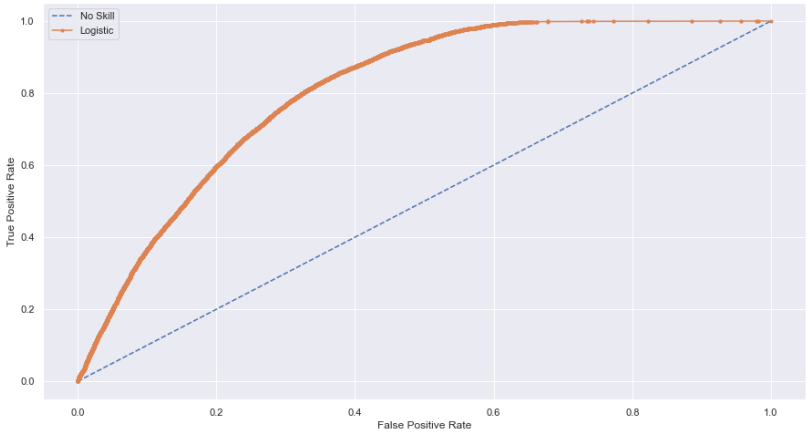
210229330 Yun-Wen Ku

To predict the default probability for borrowers on the LendingClub platform, we need to recognise our data first. Looking at the distribution of loan grades assigned by the LendingClub and their interest rate on loans, and the listed amount of the loan applied, we observed that the loan amount is irrelevant to the grade. We saw a significant relationship between the grade and the interest rate: the higher the interest rate, the lower the grade. It is possible that due to the higher default possibility (lower grade), we demand a higher interest rate in compensation.



Next, as we are interested in predicting whether a loan would be charged off, we build a logistic regression to predict whether a loan would default. The predictors we used were loan amount, the interest rate on loans, annual income, total mortgage payment, mortgage instalment (monthly payment owed), total paid in interest rates, and last payment amount. We estimate a logistic regression without shrinking the parameters using 70% of our data as the training set. Then, we use this model to produce the forecast of our testing set, and the confusion matrix of our prediction is below:

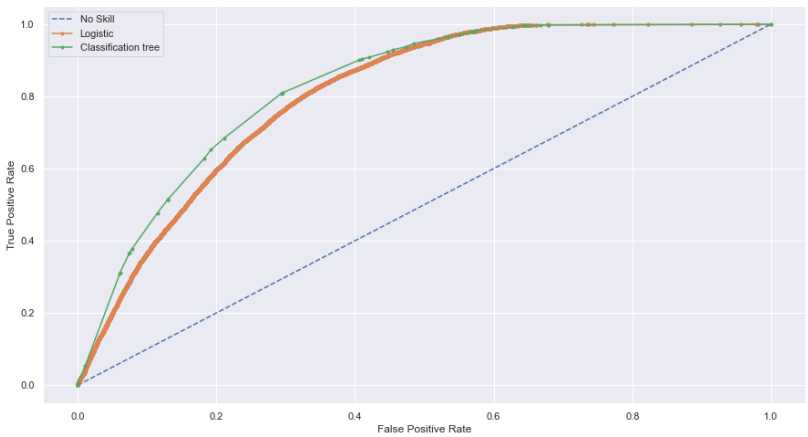
		Predicted Class	
Actual Class		Negative	Positive
	Negative	59465	505
	Positive	7604	245



Out of the total 850 predicted positives, i.e. loans charged off, only about 30% are true positives, that is, 505 false positives. There are 67,069 expected non-charged-off loans; among these, 11.3% were predicted wrong. The predicting ability of this regression is relatively weak but still better than a "no-skill" prediction, which we could also observe in the figure of the ROC curve below. As a result, we suggest that this model does not perform robustly in forecasting the loan default probability.

Next, we implement a classification tree with a max_depth of 6 to forecast mortgage defaults. The results of the confusion matrix and the ROC curve are shown below:

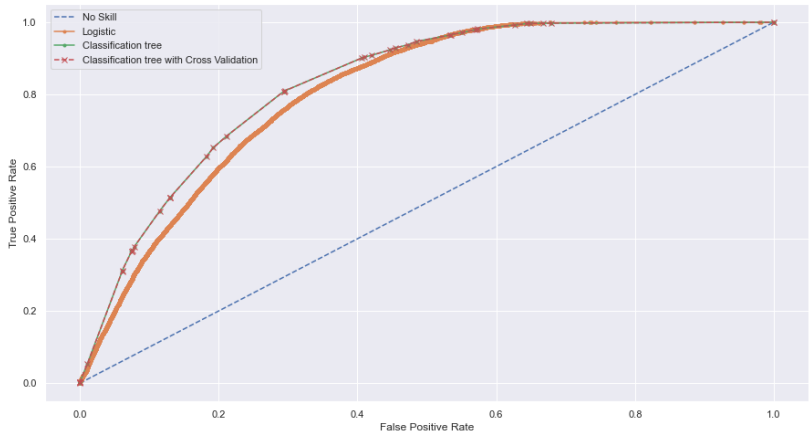
		Predicted Class	
		Negative	Positive
Actual Class	Negative	59954	16
	Positive	7819	30



This result is only slightly better than the predicting power of the previous logistic regression. The false-positive rate is still high, with one-third of all predicted-default loans.

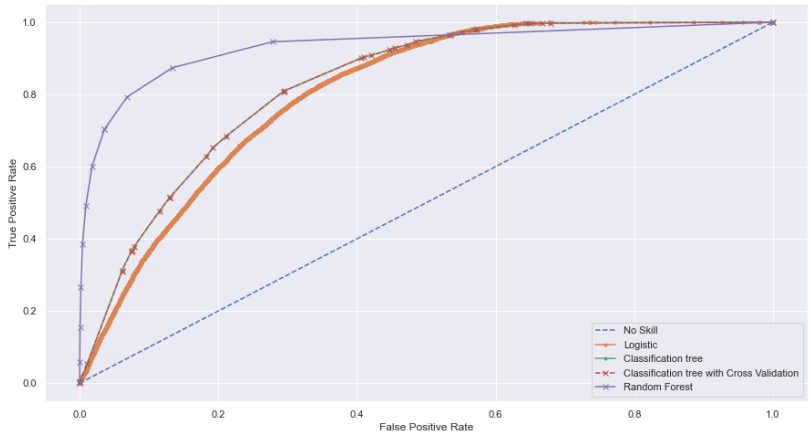
Next, we recalculate a classification tree with max_depth and min_samples_leaf parameters estimated by a 5-fold cross-validation method to find out the best parameters. The result turns out to be almost identical to the previous result with a max_depth of 6, which can be observed in both the confusion matrix and the ROC curve. This might indicate that the above parameters have given us almost the best result possible.

		Predicted Class	
Actual Class		Negative	Positive
	Negative	59956	14
	Positive	7819	30

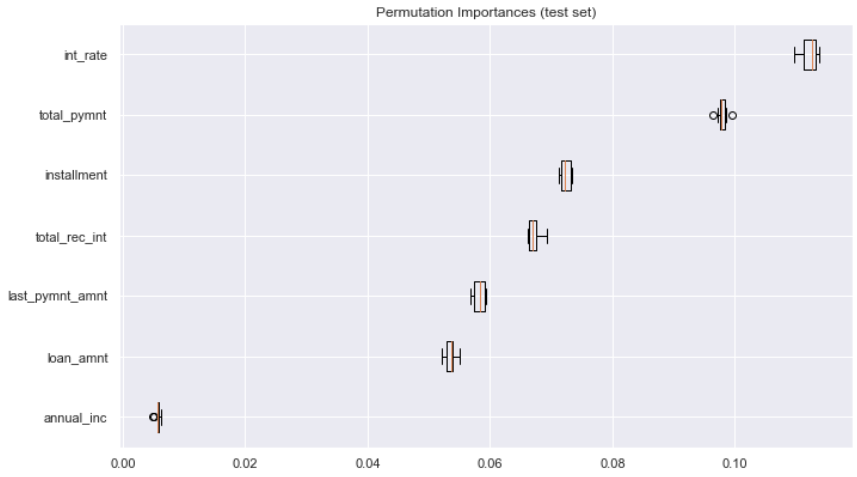


Then, we established a random forest model to see if a more robust framework allows us to have a more precise forecast of mortgage probability of default, assuming the number of estimators (n_estimators) is equal to 10. The confusion matrix and ROC curve results are below:

		Predicted Class	
Actual Class		Negative	Positive
	Negative	59455	515
	Positive	3978	3871



The false-negative and false-positive rates have significantly reduced. Less than 12% of predicted positives and about 6.3% of predicted negatives were predicted wrong. The ROC curve also shows that the random forest model has outperformed all other established models, indicating that this is the model with the best predicting power in loan default so far.



Lastly, we look at the relative importance of each predictor based on a permutation importance algorithm. From the figure above, we could state that the interest rate on loans has the highest importance in the forecasting model, with a weight of around 0.13. Payments received to date for the total amount funded (total_pymnt) has a slightly lower relative importance, with a weight of about 0.10. The monthly payment owed, interest received to date, last total payment amount received, and listed amount of the loan applied all have weights around 0.06. However, the self-reported annual income shows little importance. Therefore, we suggest that this predictor has little or no power in forecasting the probability of the mortgage loan defaulting.