Customer Default Identification Report
Prepared for Credit One
Prepared by Alberto Muniz

Objective:
Over the past year Credit One has taken notice of an increase in the number of customers who have defaulted on loans. This is despite the current classification and customer selection models. The objective of this urgent project is to examine current customer demographics to better understand traits that relate to whether or not a customer is likely to default on their credit obligations.

This brief project report includes 1. visualization of factors that may affect customer account status, 2. a description of the machine learning models that were trained using hand selected features, 3. a description of the machine learning models that were trained with the application of recursive feature elimination for feature selection, 4. a list customer attributes that are significantly related to customer loan default status. Using these identified attributes, a predictive model that Credit One can use to better identify existing customers as 'at-risk' was developed and customized for improved accuracy. Further, we suggest additional data that may be useful in future tuning of the model.

Available Data:
The collected data was made available by Tamkang University, Taiwan. This data contains 30,000 records containing customer loan information.
The loan information is composed of : Credit Limit provided to customer, Gender, Education, Marital status, customer Age, 6 month payment history, 6 month bill statement amounts, 6 month amount of payment made and Default status.

Interesting Factors that may be affecting customer account status:

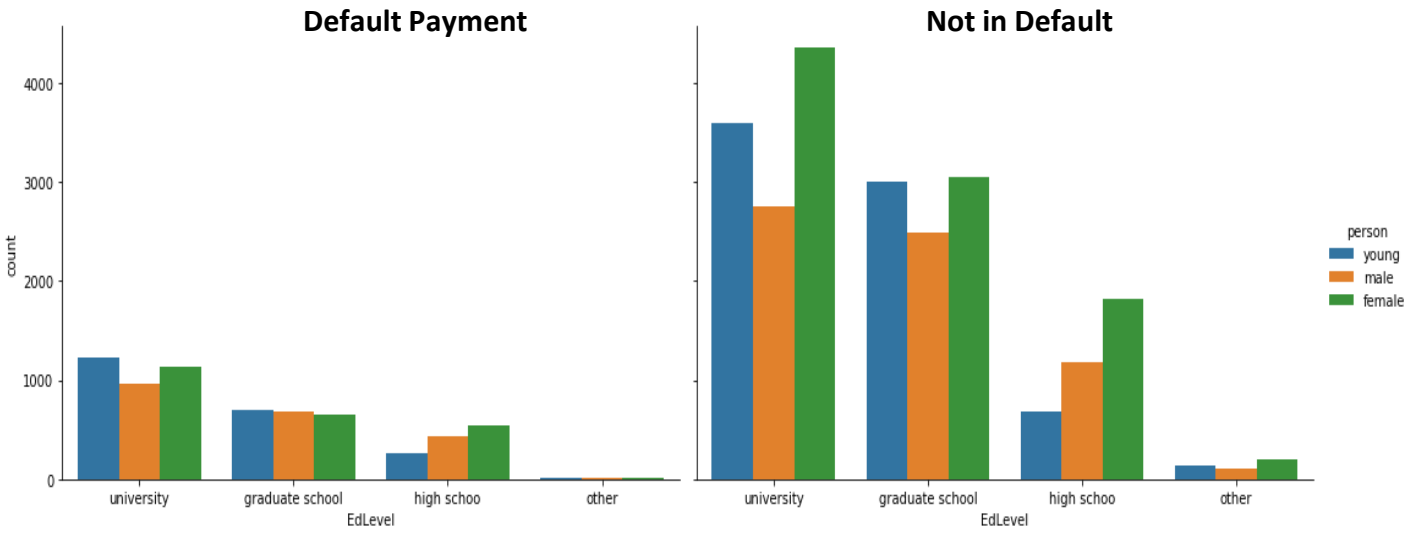

Figure1: Customers in default in Oct 2005. Plot by education, age and gender. Young= person is < 30, males and females are over 30. There are more customers in the university category that default on their Oct. 2005 payment than in grad school and high school categories. From the university education category there are more young adults that default on their Oct 2005 payment that mature males and females.

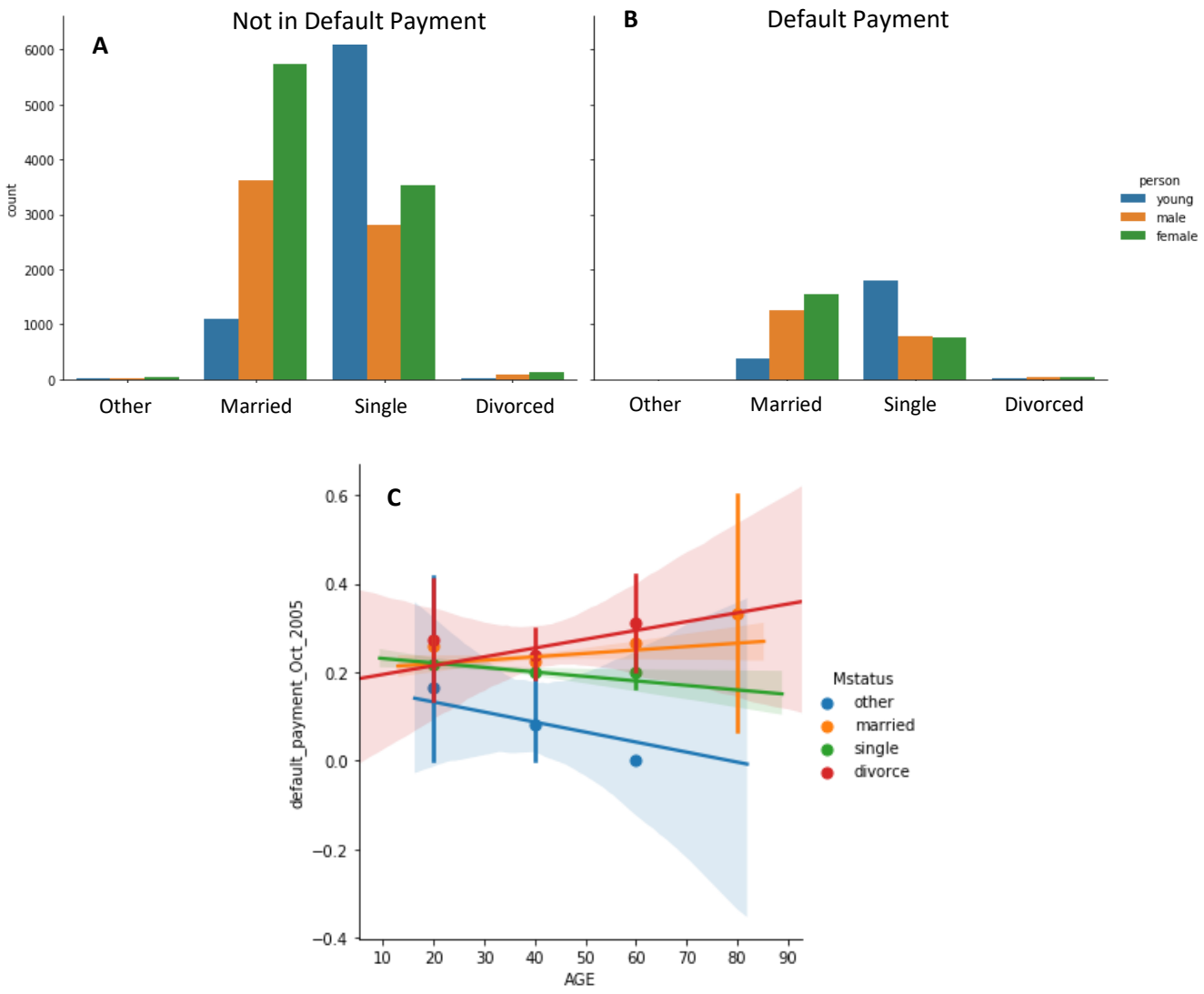

Figure 2: Marital status and Age plots. A. shows that the highest number of customers not in default fall with in the categories of married and single. Similarly, in **B.** those who do default seem to mostly fall with in the married and single categories. This may be explained by the ratio of these categories to the total customer population, in which these two groups are the majority of the population. However, when marital status and age are plotted **C**. The data suggests that those within the divorced category have a higher propensity to default as age increases..

Machine learning description and selection:

Based on the available data four classification models were trained for predicting customers who would default. See table 1 and figure 3. The four models are Logistic Regression, Random Forest, Adaboost, and Gradient Boosting.

	LogisticReg	Random Forest	Adaboost	Gradient Boosting
accuracy	77.33	81.24	81.02	81.82
precision	100	66.39	69..04	69.86
recall	0.049	35.76	30.24	35.51

Table1: Metrics describing the models' for predicting customer loan default. The models were trained using hand selected features.

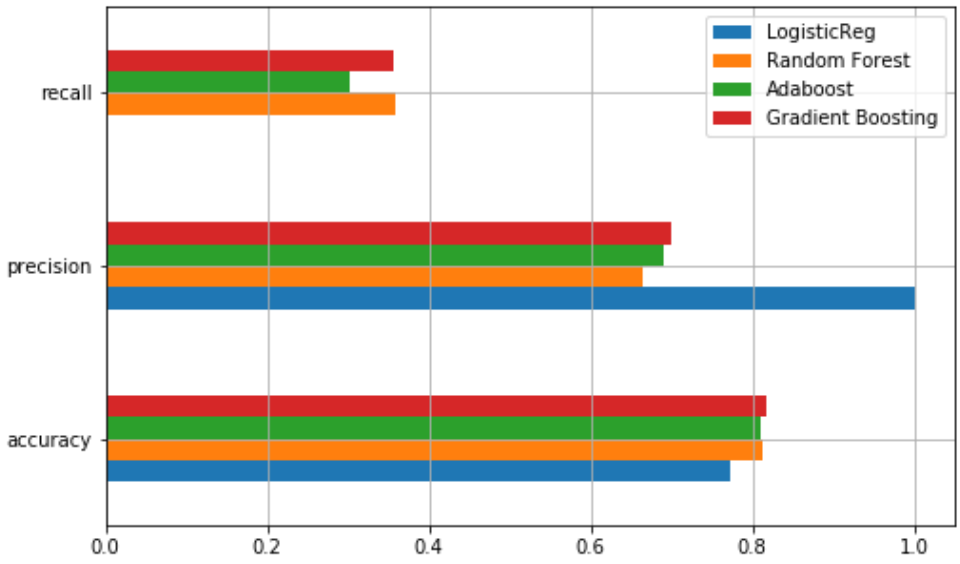

Figure 3: Plot of accuracy, precision and recall metrics for all four trained models. We can see that Random Forest and Gradient boosting have the highest precision and recall metrics while all four models are similar in accuracy. Logistic regression shows the highest precision at 100. This may be a sing of overfitting.

Additionally, the models were trained by applying recursive feature elimination for model improvement. See Table 2 and figure 4.

	RFECV_LogisticReg	RFECV_Random Forest	RFECV_Adaboost	RFECV_Gradient Boosting
accuracy	81.62	80.99	81.50	81.77
precision	71.76	66.31	71.04	69.72
recall	32.85	33.61	31.71	35.27

Table2: Metrics describing the models' for predicting customer loan default. The models were trained using features selected by the recursive feature elimination process.

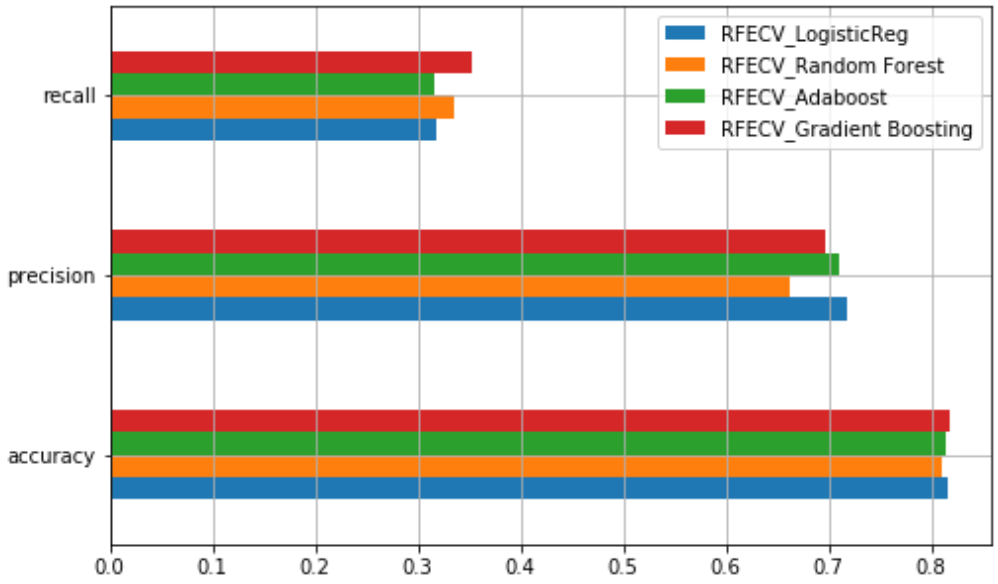

Figure 4: Plot of accuracy, precision and recall metrics for all four trained models with recursive feature elimination. We can see that Logistic regression and Adaboost have the highest precision score while Gradient boosting and Aboost have the highest recall metric. Further, all four models are similar in accuracy.

To determine between the rfecv_Adaboosting and the rfecv_Gradient Boosting models a plot was generated (figure 5) to compare the precision and recall scores . Based on this plot it is evident that the rfecv_Gradient Boosting model has a higher recall score thru the majority of the precision score range, see figure 5.

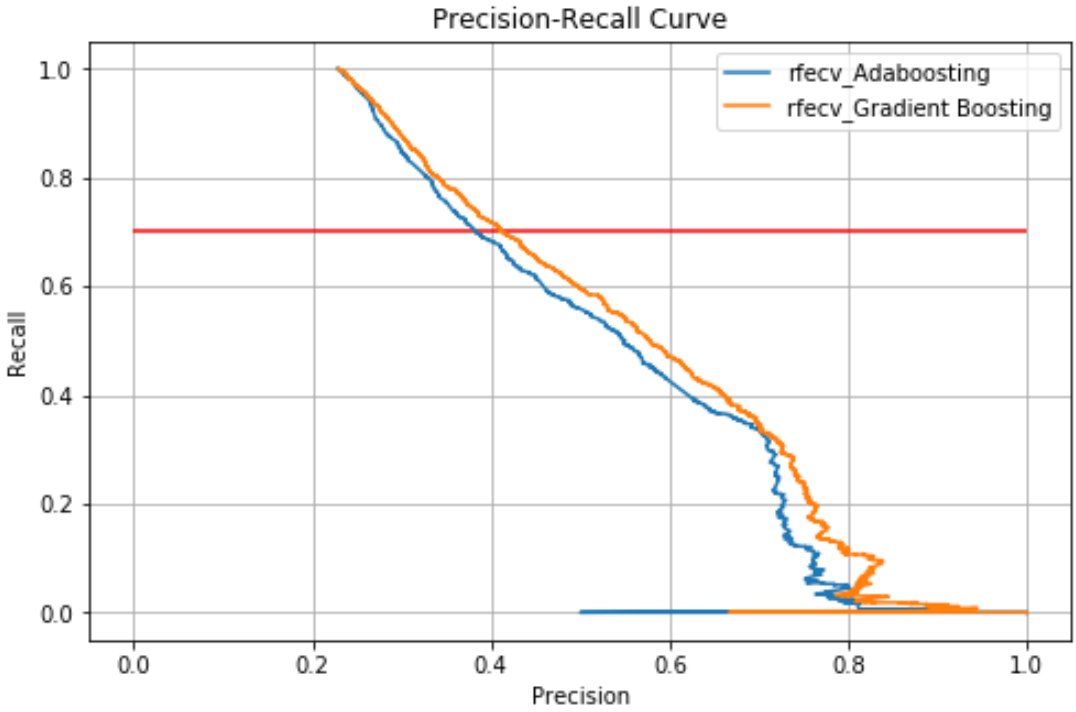

Figure 5: Plot of accuracy, precision and recall metrics for rfecv_Adaboosting and the rfecv_Gradient Boosting models. Based on the higher recall score thru out the precision score range the rfecv_Gradient Boosting model was selected.

The classification threshold for the selected model was then analyzed and changed to improve model prediction results. See figure 6.

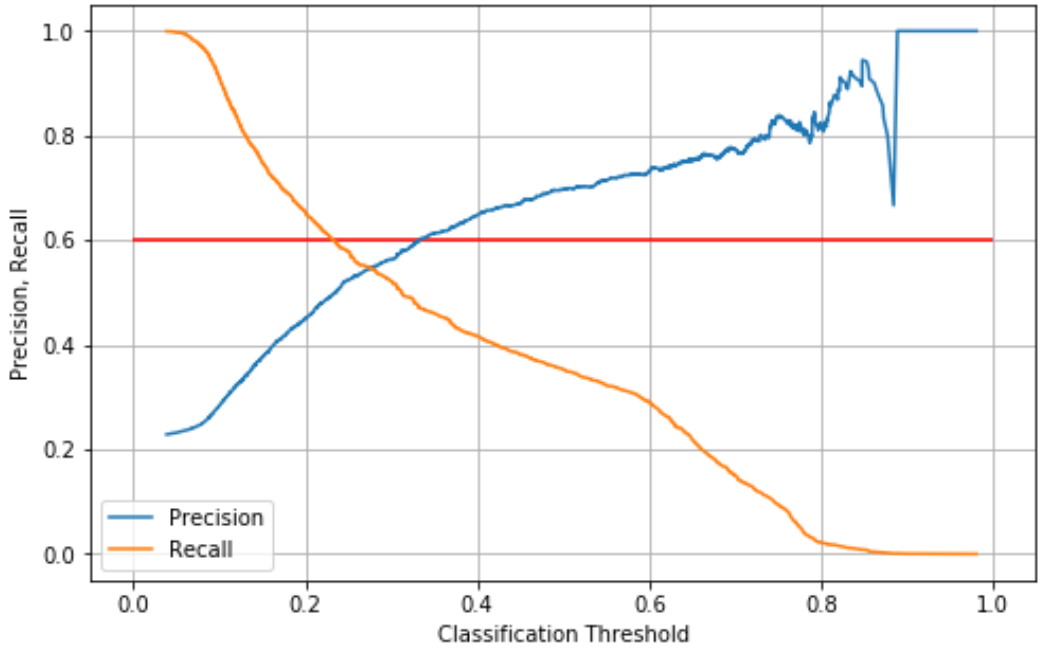

Figure 6: Plot of accuracy, precision and classification threshold metrics for the selected model, rfecv_Gradient Boosting. This allows us to see that a change to the classification threshold to 0.2 will provide a better model for classification predictions.

The selected model for predicting weather customers would default payment was RFECV_Gradient Boosting. By applying Recursive Feature Elimination, the most effective features for this task were selected.
These were determined to be:
Education levels
Age
Payment history for the last six months
Bill amounts for the last six months
Payment amount for the last six months.

Additionally, the classification threshold was changed to 0.2 to finalize the model.

Rfecv_Gradient Boosting Model with improved classification threshold of 0.2 was then tested on a fraction of the data not used for training. The results of the test are:

A Recall score of 64
And Precision score of : 46

Rfecv_ Gradient Boosting Classifier with a modified classification threshold to 0.2 was selected for predictions in this project. This model yields a precision score of 46 % and a recall score of 64%. This means that 64% of the predicted to default actually do default.

Potential issue and Reminder:
The available data represents information on loans already made and accepted and no information regarding the features or scores used to not approve a loan is available. Thus, using this Machine learning model provides a method to scrutinize already existing potentially low risk loans.

Additional information that may be helpful in determining future loan default:
Up to date Household income
Debt to income ratio
Total credit available to customer