Credit Card Default

December 2005

Grace and Mahima

Introduction

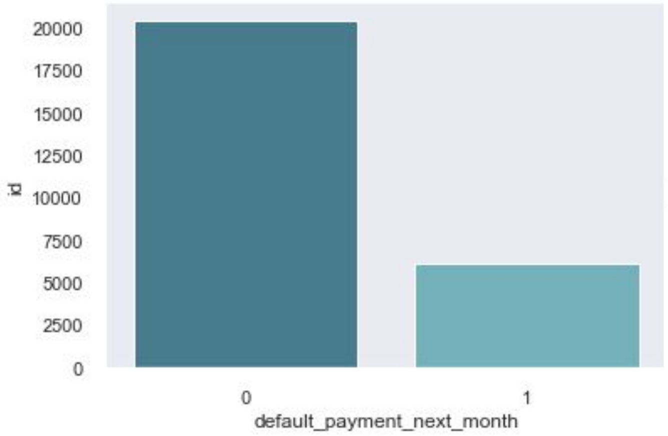
We are trying to predict whether a customer won't be able to pay their credit card bill the following month

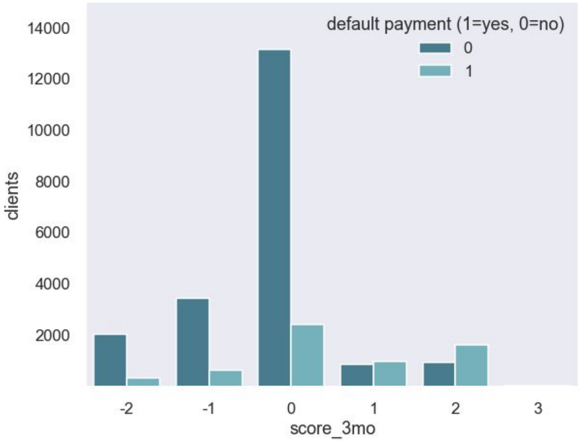
Our data consists of the following information on 30,000 customers from April to September 2005 :

1. Demographics - age, sex, marital status, education
2. Amount of credit given or the limit balance
3. Repayment status from April to September 2005
4. Bill statement as of April to September 2005
5. Amount of previous payment as in each month
6. Default next month (1 = yes, 0 = no)

Initial EDA

Non-defaulters: 78% of the dataset

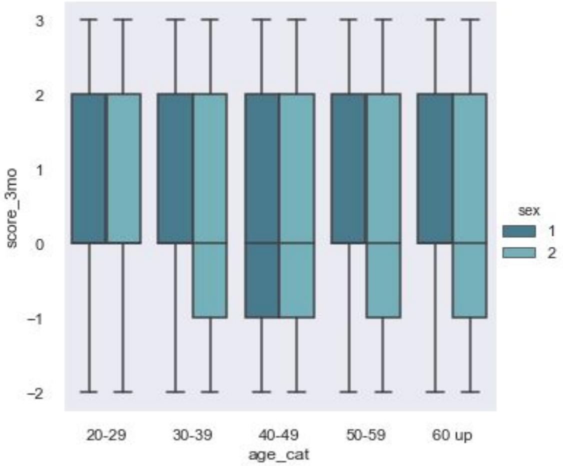


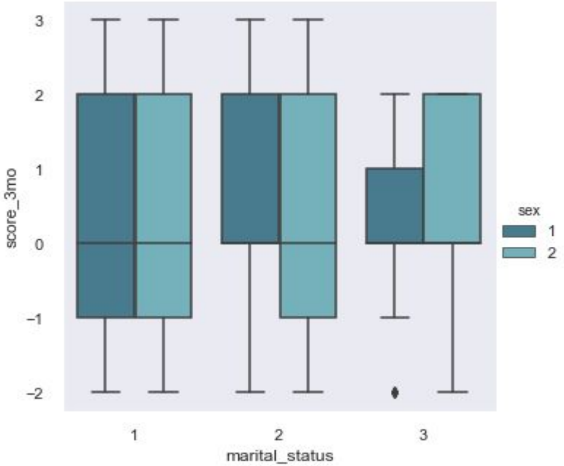


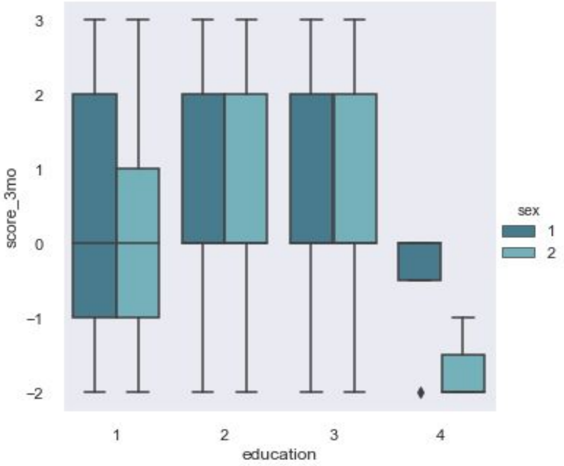
citibank

Initial EDA

Defaulters by demographics (past 3 months)

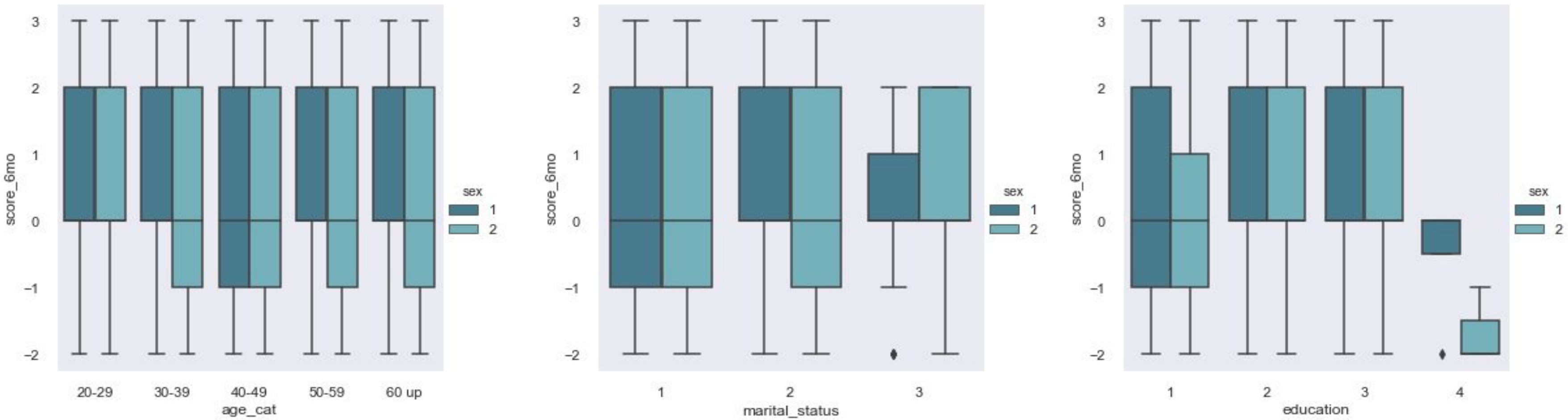






Initial EDA

Defaulters by demographics (past 6 months)



citibank

Initial EDA

“Best” predictors based on correlation matrix:

- pay_sep
- pay_aug
- score_3mo: weighted average of last July - September pay columns
- score_6mo: weighted average of April - September pay columns
- pay_jul
- pay_jun
- pay_may
- pay_apr
- limit_bal (negative correlation)

citibank

Model Selection

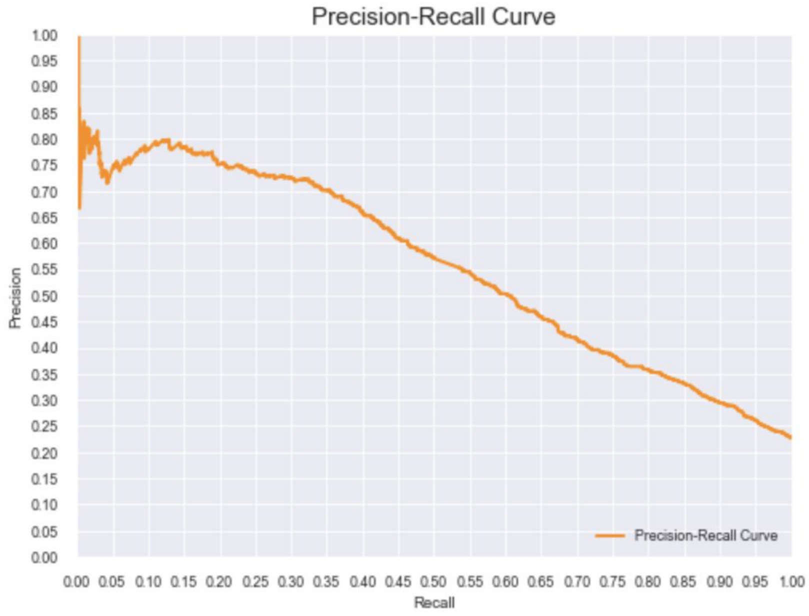
Logistic Regression

- Feature engineering:
  - weighted average of ‘pay’ columns for the past 3 and 6 months
  - bill amount as of September as percentage of the total credit limit
- Feature selection:
  - SelectFromModel: 32 features selected out of 85
- Class imbalance:
  - SMOTE
- Correlated features
  - PCA

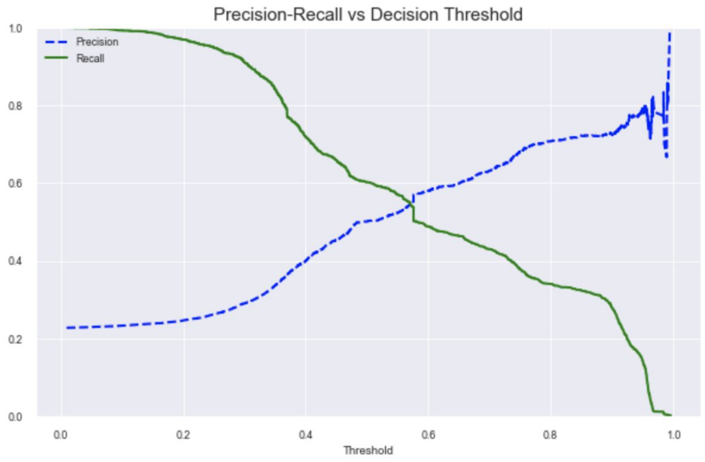
Model Selection

Logistic Regression (L2):

```
F1 score:  0.5467148884870403
Precision-Recall AUC: 0.5569426470458486
```




```
Classification Report:
              precision    recall  f1-score   support

           0       0.88      0.82      0.85      5123
           1       0.50      0.60      0.55      1507

    accuracy                           0.77      6630
   macro avg       0.69      0.71      0.70      6630
weighted avg       0.79      0.77      0.78      6630
```

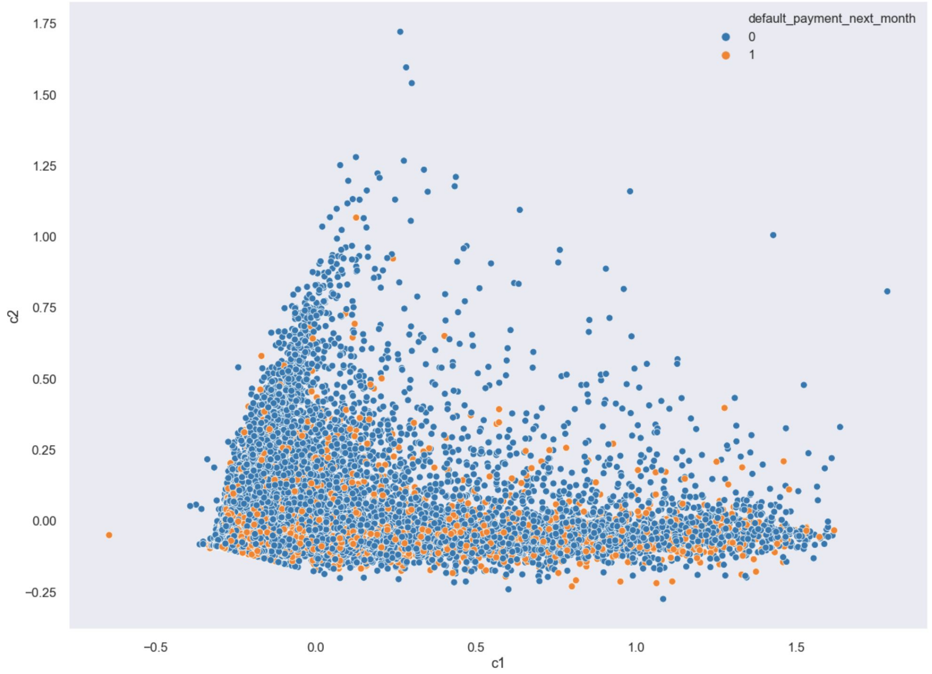


citibank

Model Selection

Principal Component Analysis + Logistic Regression (L2 Norm)

1	0.670239
2	0.102631
3	0.043163
4	0.039036
5	0.037631
6	0.035778
7	0.029958
8	0.027473
9	0.006832
10	0.003431
11	0.002140
12	0.001688

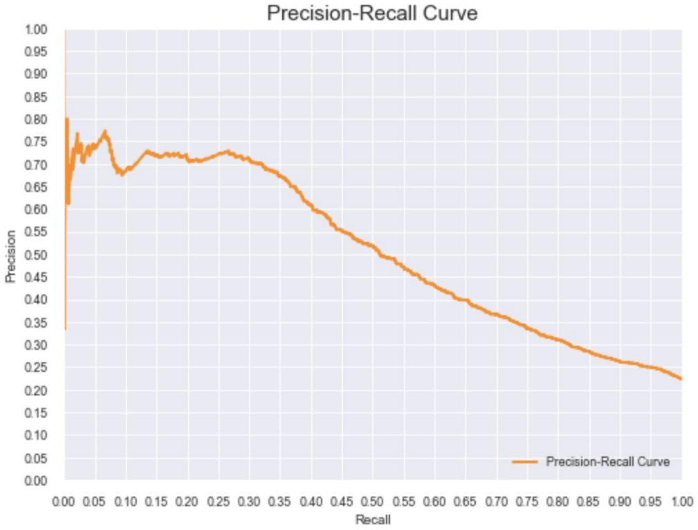


citibank

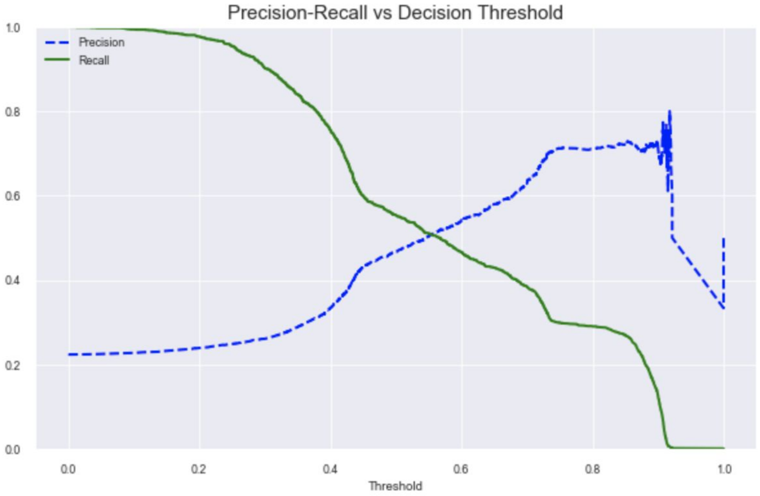
Model Selection

Principal Component Analysis + Logistic Regression

```
Best Parameters: {'C': 1000, 'penalty': 'l1'}
-----------------------------------------------------------------
F1 score:  0.5073049424930058
Precision-Recall AUC: 0.5107404220408498
```



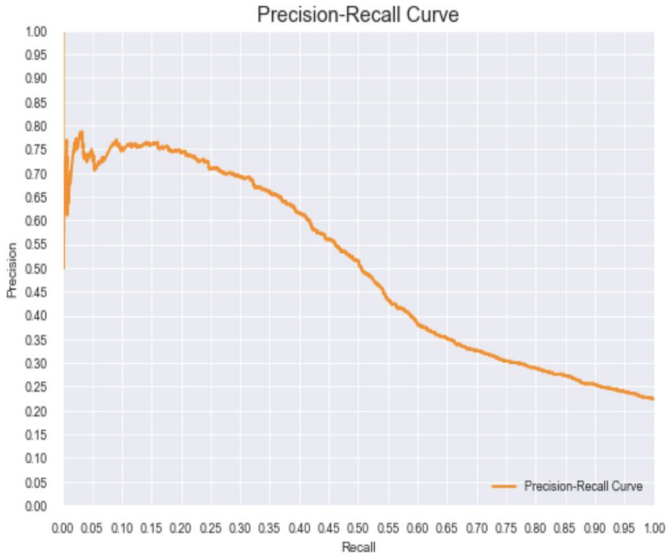

```
Classification Report:
              precision    recall  f1-score   support

           0       0.86      0.82      0.84      5113
           1       0.47      0.55      0.51      1472

    accuracy                           0.76      6585
   macro avg       0.67      0.69      0.67      6585
weighted avg       0.78      0.76      0.77      6585
```



Model Selection

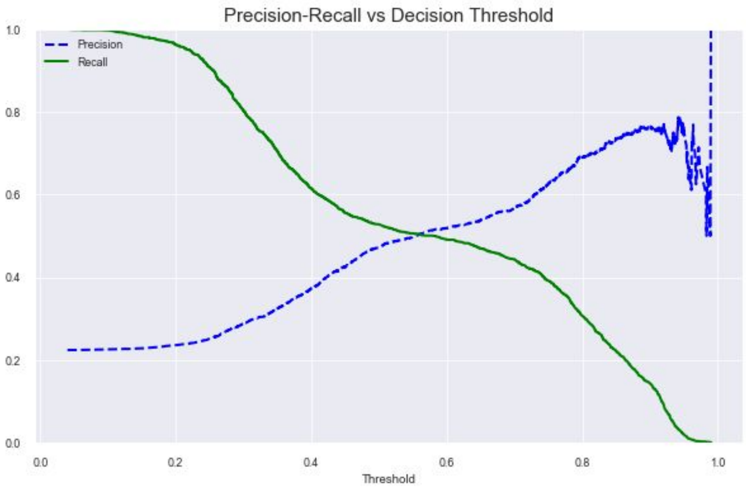
Support Vector Machine

F1 score: 0.5035629453681711
Precision-Recall AUC: 0.5035994024163514



Classification Report:

	precision	recall	f1-score	support
0	0.86	0.86	0.86	5113
1	0.50	0.50	0.50	1472
accuracy			0.78	6585
macro avg	0.68	0.68	0.68	6585
weighted avg	0.78	0.78	0.78	6585



citibank

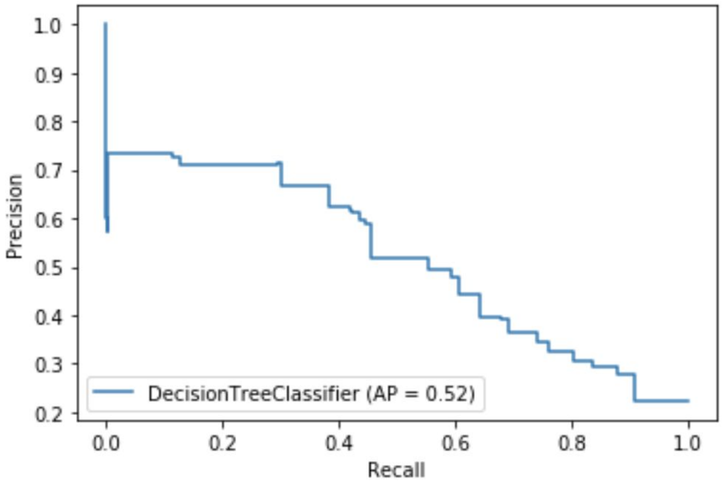
Model Selection

Random Forest

- Feature engineering:
  - weighted average of ‘pay’ columns for the past 3 and 6 months
  - bill amount as of September as percentage of the total credit limit
- Feature selection:
  - all features
- Class imbalance:
  - SMOTE

citibank

Random Forest



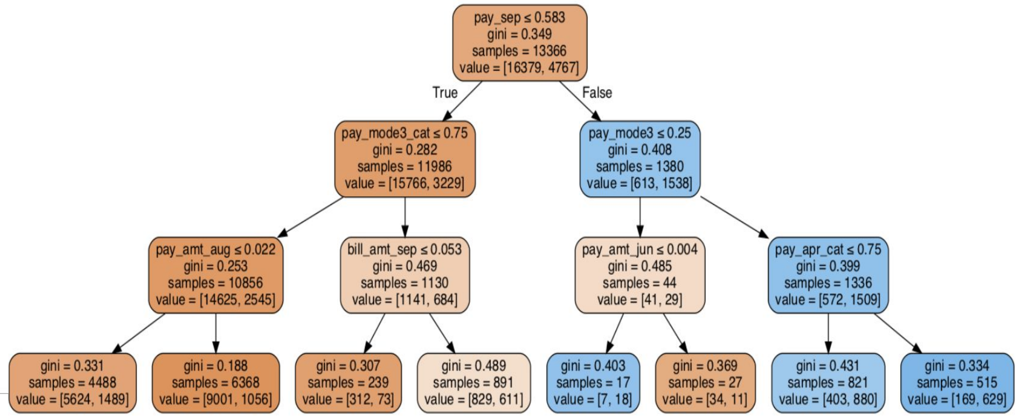

```
#calculating the auc
precision_recall_auc = auc(recall, precision)
precision_recall_auc
```

0.5940512167372005

	precision	recall	f1-score	support
0	0.84	0.95	0.89	4106
1	0.67	0.38	0.49	1181
accuracy			0.82	5287
macro avg	0.75	0.66	0.69	5287
weighted avg	0.80	0.82	0.80	5287

Random Forest

	rank_test_score	params	mean_test_score	std_test_score
0	7	{'max_depth': 3, 'min_samples_leaf': 5}	0.436625	0.020016
1	7	{'max_depth': 3, 'min_samples_leaf': 10}	0.436625	0.020016
2	7	{'max_depth': 3, 'min_samples_leaf': 15}	0.436625	0.020016
3	1	{'max_depth': 5, 'min_samples_leaf': 5}	0.468013	0.020908
4	2	{'max_depth': 5, 'min_samples_leaf': 10}	0.465494	0.020496
5	3	{'max_depth': 5, 'min_samples_leaf': 15}	0.464945	0.020169
6	4	{'max_depth': 7, 'min_samples_leaf': 5}	0.461309	0.025128
7	6	{'max_depth': 7, 'min_samples_leaf': 10}	0.458842	0.018647
8	5	{'max_depth': 7, 'min_samples_leaf': 15}	0.460636	0.017897



Conclusions

- Recall is more important than Precision
- None of the models are good in predicting defaulters
- Our best model: Logistic (recall = 0.6, f1= 0.55)
- We need to revisit and get more features and samples
    - Income, estimate monthly expense
    - Related accounts (savings, loans, other credit cards)
    - Up to 12 or 24 month history

Thank you

Appendix

True positive (TP): the model predicts default and the client is indeed defaulting

False positive (FP): the model predicts default but the client is not defaulting

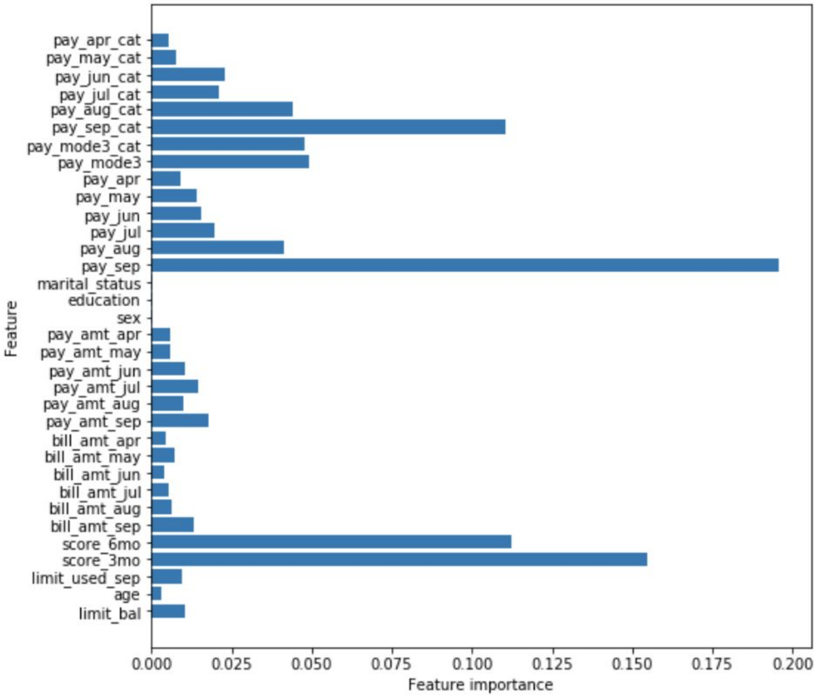
True negative (TN): the model predicts not defaulting and the client is not defaulting

False negative (FN): the model predicts not defaulting but the client is in fact defaulting

Precision = TP/TP+FP

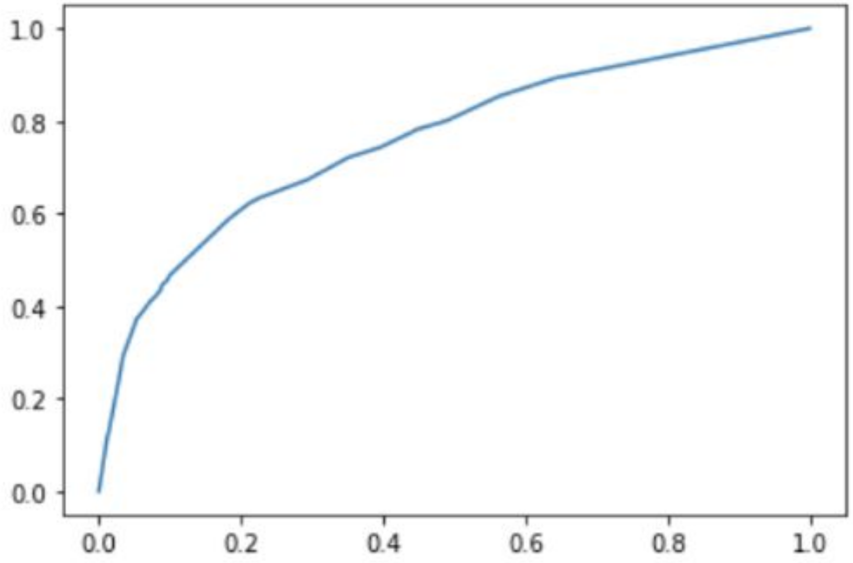
Recall / Sensitivity = TP/TP+FN = TPR (probability of detection)

Why recall matters to us is because of the False Negatives component i.e. the chance of knowing if someone has defaulted if they have actually defaulted. Not knowing this number will be a heavy cost to the bank and hence this is our key metric.


pay_apr_cat
pay_may_cat
pay_jun_cat
pay_jul_cat
pay_aug_cat
pay_sep_cat
pay_mode3_cat
pay_mode3
pay_apr
pay_may
pay_jun
pay_jul
pay_aug
pay_sep
marital_status
education
sex
pay_amt_apr
pay_amt_may
pay_amt_jun
pay_amt_jul
pay_amt_aug
pay_amt_sep
bill_amt_apr
bill_amt_may
bill_amt_jun
bill_amt_jul
bill_amt_aug
bill_amt_sep
score_6mo
score_3mo
limit_used_sep
age
limit_bal
Feature
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
Feature importance

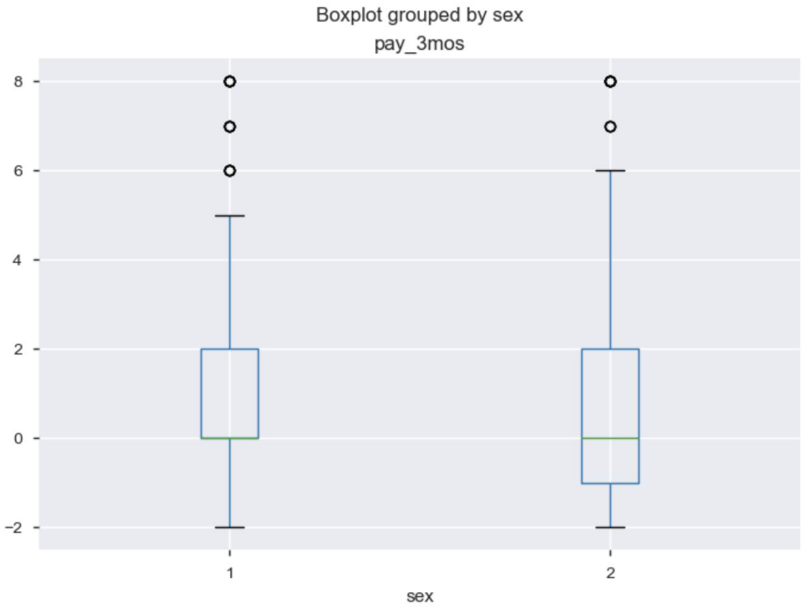
```
auc(fpr,tpr)
```

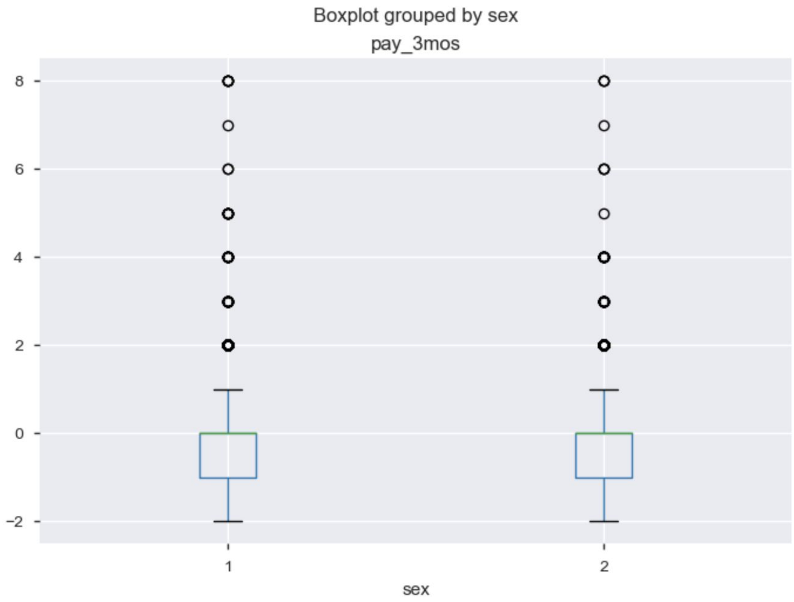
0.7592424790469989



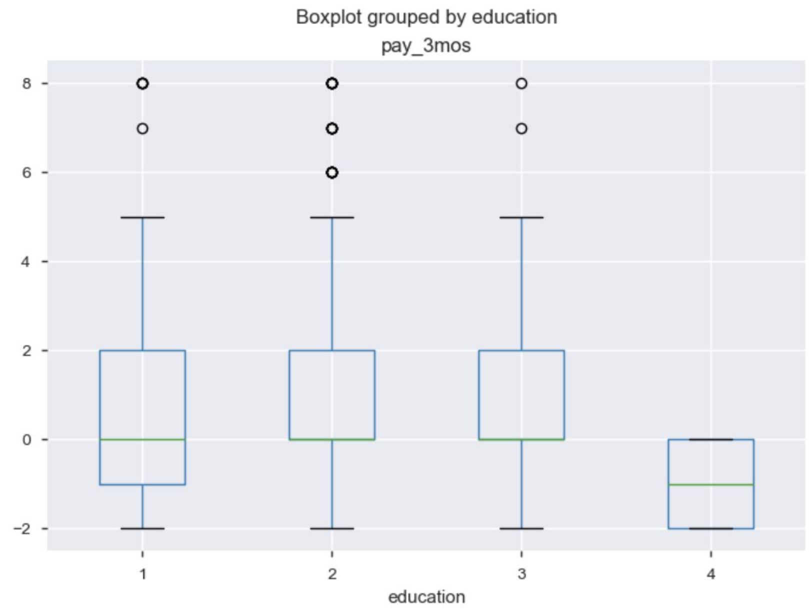
1. **Max depth - 3, 5, 7**
2. **Min samples leaf - 5,10,15**
3. **Best -** {'max_depth': 5, 'min_samples_leaf': 5}

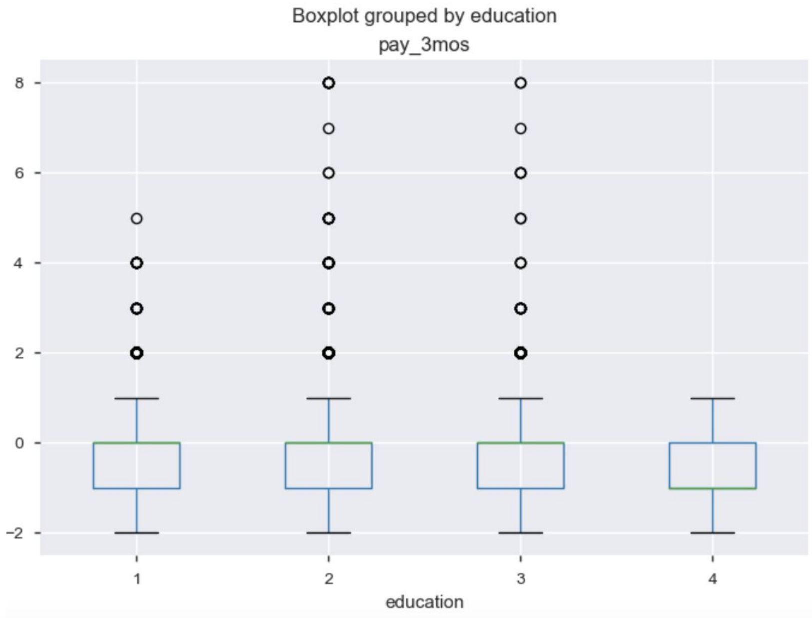
Class : default vs no default (by sex)



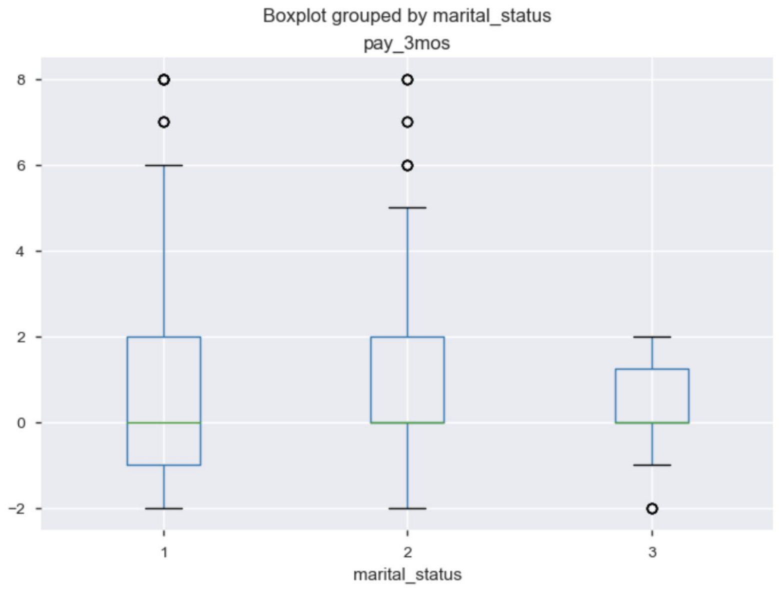


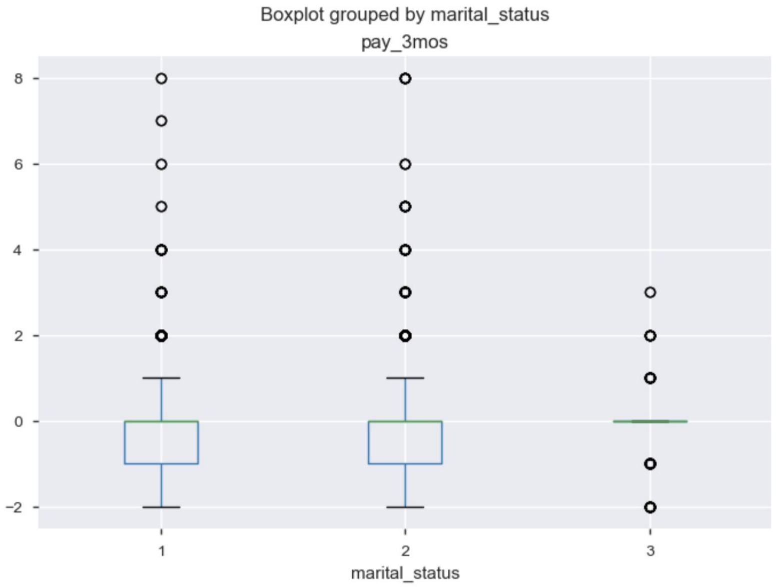
Class : default vs no default (by education)





Class : default vs no default (by marital status)





Class : default vs no default (by age)