Jiayu Charlene Wei

Weij22@wfu.edu

12/09/2022

## Executive Summary

A significant financial institution has been tasked with doing loan research. When a borrower skips payments for a predetermined amount of time, a loan default occurs. They are interested in what variables will affect loan default. Based on the crucial factors, they will strengthen their business plan to handle loan default scenarios. The project will make use of the techniques for logistic regression, XGB, and random forest to develop its predictive models. We found that the XGB model had the best performance after comparing those products.

## Problem

This financial organization is interested in learning what crucial elements will raise the rate of loan default. If the corporation wishes to manage its business well in the future, they have to deal with those crucial aspects.

## Key Findings

there is a positive relationship between installment and default

there is a positive relationship between interest rate and default

there is a positive relationship between annual rate and default

## Recommendations

- All of the actionable suggestions were developed based on the crucial elements of XGBoosting. Finding defaulted loans is more likely if you search for loans with high borrower monthly payments, loan subgrades of B4 and C1, a high number of negative public records, and addresses in California. Finding the default loan can also be done by looking for loans that lack indicators of LC-verified income, grades of C.
- If the annual income of these customers is below a certain amount, stop lending money to them.

# Detailed Analysis & Steps

## File Summary

| File Name | Record count | Column count | Numeric columns | Character columns |
|---|---|---|---|---|
| loan_train.csv | 29777 | 27 | 19 | 8 |
| loan_holdout.csv | 5000 | 27 | 20 | 7 |

## Field Summary

### Loan_train.csv

| | skim_variable <chr> | n_missing <int> | complete_rate <dbl> | ordered <lgl> | n_unique <int> | top_counts <chr> |
|---|---|---|---|---|---|---|
| 1 | grade | 3 | 0.9998993 | FALSE | 7 | B: 8620, A: 7142, C: 6068, D: 4268 |
| 2 | sub_grade | 3 | 0.9998993 | FALSE | 35 | B3: 2088, A4: 2044, A5: 1957, B5: 1932 |
| 3 | home_ownership | 3 | 0.9998993 | FALSE | 5 | REN: 14064, MOR: 13340, OWN: 2275, OTH: 91 |
| 4 | verification_status | 3 | 0.9998993 | FALSE | 3 | Not: 13128, Ver: 9460, Sou: 7186 |
| 5 | loan_status | 0 | 1.0000000 | FALSE | 2 | cur: 25300, def: 4477 |
| 6 | pymnt_plan | 3 | 0.9998993 | FALSE | 2 | n: 29773, y: 1 |
| 7 | purpose | 3 | 0.9998993 | FALSE | 14 | deb: 13816, cre: 3850, oth: 3108, hom: 2226 |
| 8 | addr_state | 3 | 0.9998993 | FALSE | 50 | CA: 5188, NY: 2836, FL: 2200, TX: 2067 |

| | skim_variable <chr> | n_missing <int> | complete_rate <dbl> | mean <dbl> | sd <dbl> | p0 <dbl> | p25 <dbl> | p50 <dbl> | p75 <dbl> |
|---|---|---|---|---|---|---|---|---|---|
| 1 | loan_amnt | 3 | 0.9998993 | 11109.434406 | 7404.652253 | 500.00 | 5225.000 | 9800.000 | 15000.00 |
| 2 | funded_amnt | 3 | 0.9998993 | 10843.637066 | 7147.052231 | 500.00 | 5100.000 | 9600.000 | 15000.00 |
| 3 | funded_amnt_inv | 3 | 0.9998993 | 10149.655315 | 7130.855966 | 0.00 | 4950.000 | 8500.000 | 14000.00 |
| 4 | term | 3 | 0.9998993 | 42.137435 | 10.470626 | 36.00 | 36.000 | 36.000 | 60.00 |
| 5 | int_rate | 3 | 0.9998993 | 12.166454 | 3.716096 | 5.42 | 9.630 | 11.990 | 14.72 |
| 6 | installment | 3 | 0.9998993 | 323.808152 | 209.771603 | 15.67 | 165.845 | 278.940 | 429.86 |
| 7 | annual_inc | 4 | 0.9998657 | 69201.232288 | 66566.415292 | 2000.00 | 40000.000 | 59000.000 | 82500.00 |
| 8 | dti | 3 | 0.9998993 | 13.384026 | 6.738964 | 0.00 | 8.190 | 13.490 | 18.70 |
| 9 | fico_range_low | 3 | 0.9998993 | 713.053167 | 36.310150 | 610.00 | 685.000 | 710.000 | 740.00 |
| 10 | fico_range_high | 3 | 0.9998993 | 717.053167 | 36.310150 | 614.00 | 689.000 | 714.000 | 744.00 |

1-10 of 19 rows | 1-10 of 11 columns

Previous 1 2 Next

### Loan_holdout.csv

| | skim_variable <chr> | n_missing <int> | complete_rate <dbl> | min <int> | max <int> | empty <int> | n_unique <int> | whitespace <int> |
|---|---|---|---|---|---|---|---|---|
| 1 | grade | 0 | 1 | 1 | 1 | 0 | 7 | 0 |
| 2 | sub_grade | 0 | 1 | 2 | 2 | 0 | 35 | 0 |
| 3 | home_ownership | 0 | 1 | 3 | 8 | 0 | 5 | 0 |
| 4 | verification_status | 0 | 1 | 8 | 15 | 0 | 3 | 0 |
| 5 | pymnt_plan | 0 | 1 | 1 | 1 | 0 | 1 | 0 |
| 6 | purpose | 0 | 1 | 3 | 18 | 0 | 14 | 0 |
| 7 | addr_state | 0 | 1 | 2 | 2 | 0 | 50 | 0 |

7 rows

| | skim_variable <chr> | n_missing <int> | complete_rate <dbl> | mean <dbl> | sd <dbl> | p0 <dbl> | p25 <dbl> | p50 <dbl> | p75 <dbl> |
|---|---|---|---|---|---|---|---|---|---|
| 1 | id | 0 | 1.0000000 | 6.682515e+05 | 2.179090e+05 | 55521.00 | 502532.000 | 651269.00 | 826140.00 |
| 2 | loan_amnt | 0 | 1.0000000 | 1.104373e+04 | 7.425672e+03 | 500.00 | 5000.000 | 9600.00 | 15000.00 |
| 3 | funded_amnt | 0 | 1.0000000 | 1.077014e+04 | 7.146609e+03 | 500.00 | 5000.000 | 9500.00 | 15000.00 |
| 4 | funded_amnt_inv | 0 | 1.0000000 | 1.011691e+04 | 7.133851e+03 | 0.00 | 4903.993 | 8400.00 | 14000.00 |
| 5 | term | 0 | 1.0000000 | 4.237003e+01 | 1.059775e+01 | 36.00 | 36.000 | 36.00 | 60.00 |
| 6 | int_rate | 0 | 1.0000000 | 1.216166e+01 | 3.688968e+00 | 5.42 | 9.620 | 11.99 | 14.72 |
| 7 | installment | 0 | 1.0000000 | 3.198580e+02 | 2.069255e+02 | 16.85 | 164.460 | 274.63 | 424.60 |
| 8 | annual_inc | 3 | 0.9997649 | 6.898562e+04 | 5.792595e+04 | 1896.00 | 40000.000 | 59000.00 | 82435.00 |
| 9 | dti | 0 | 1.0000000 | 1.334742e+01 | 6.696902e+00 | 0.00 | 8.230 | 13.43 | 18.63 |
| 10 | fico_range_low | 0 | 1.0000000 | 7.130511e+02 | 3.590427e+01 | 620.00 | 685.000 | 710.00 | 740.00 |

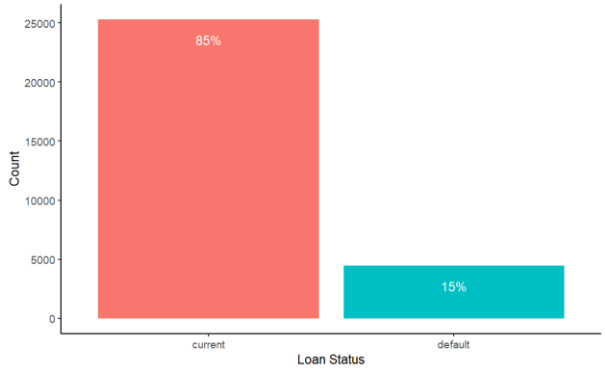
1-10 of 20 rows | 1-10 of 11 columns

Previous 1 2 Next

## Target Summary

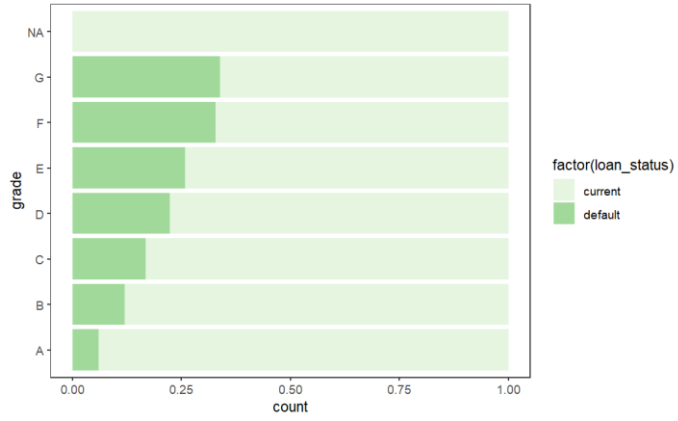
# Explorations relative to the target

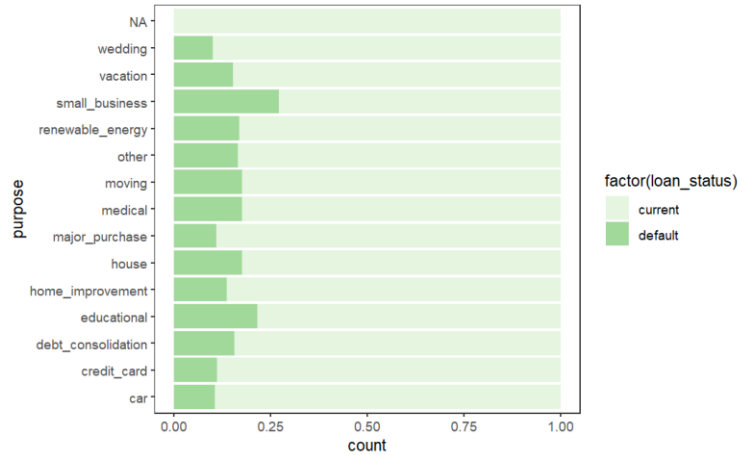
## Target



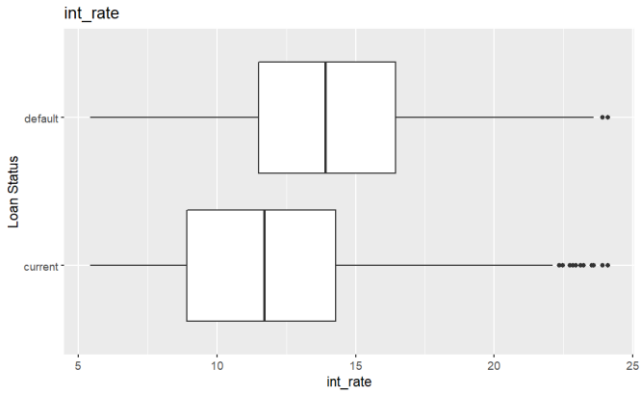

15% of data rows are

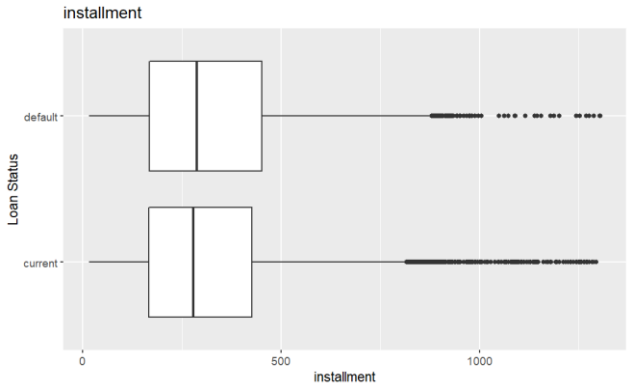
## Categorical

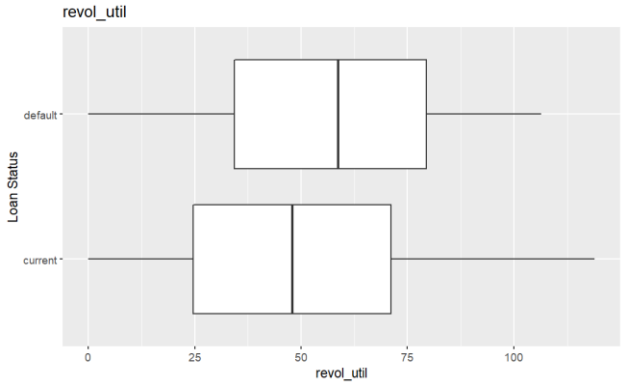


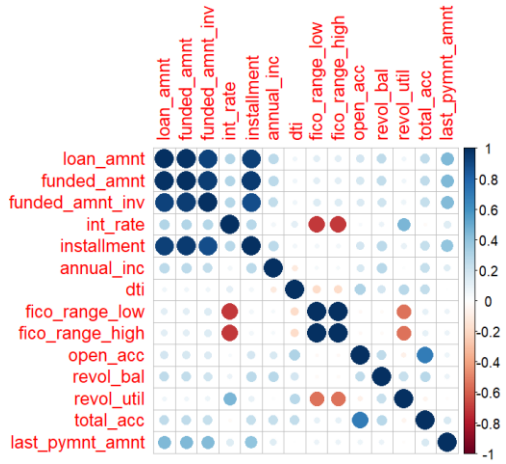
## Numeric


int_rate
default
current
Loan Status
5
10
15
20
25
int_rate


installment
default
current
Loan Status
0
500
1000
installment


revol_util
default
current
Loan Status
0
25
50
75
100
revol_util

## Correlations




# Transformations

## Missing values

Imputation with means or modes

```
# Recipe
```{r}
# deal w. categoricals
ir_recipe <- recipe(~.,loan_model) %>%
  step_unknown(all_nominal_predictors()) %>%
  step_novel(all_nominal_predictors()) %>%
  step_impute_median(all_numeric_predictors()) %>%
  step_dummy(all_nominal_predictors()) %>%
  step_impute_mode(all_nominal_predictors()) %>%
  prep()

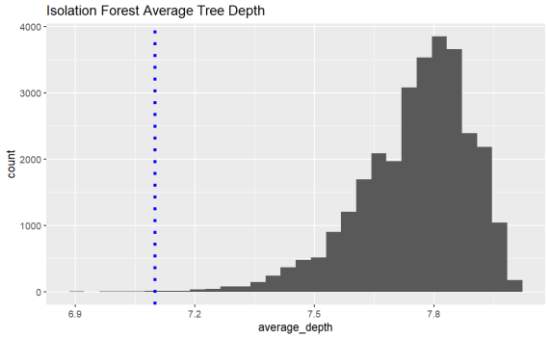

bake_loan <- bake(ir_recipe, loan_model)
skim(bake_loan)
```

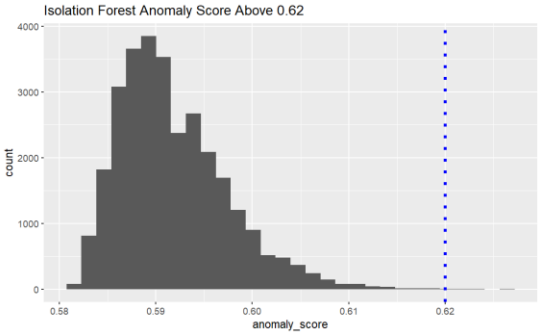
## Transformation

Transformation some strings as numbers

```
#transform some variables
loan_model = loan_del %>%
  mutate(term = as.numeric(str_replace(term, " months", "")),
         int_rate = as.numeric(str_replace(int_rate, "%", "")),
         revol_util = as.numeric(str_replace(revol_util, "%", "")))
```
```

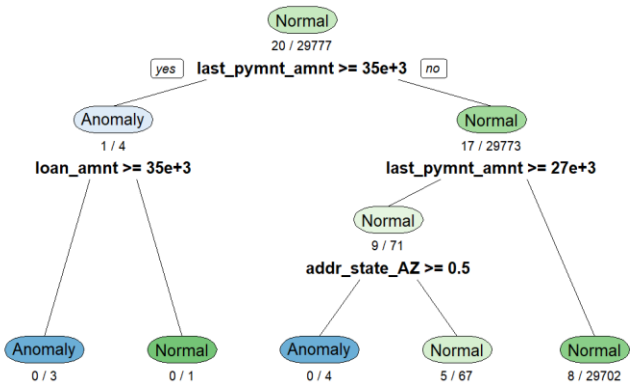
## Anomaly detection



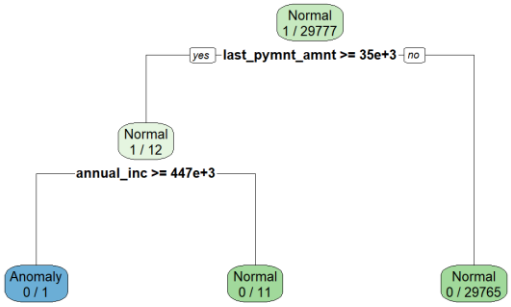




| anomaly <fctr> | n <int> |
|---|---|
| Anomaly | 20 |
| Normal | 29757 |

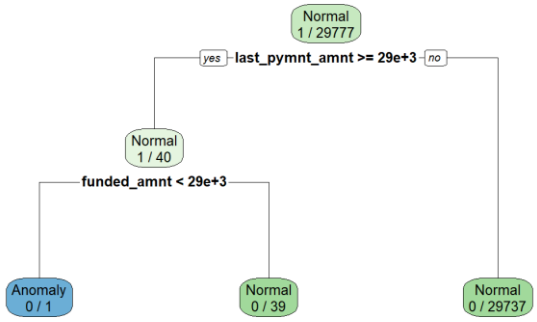
## Global Anomaly Rules



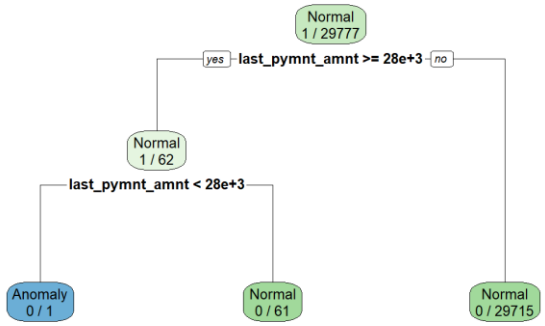

## Local Anomaly Rules



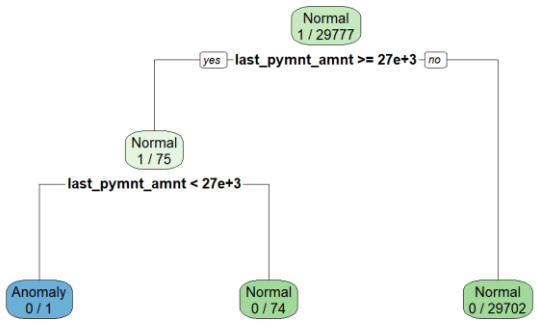

The last payment amount >= 35e+3 and annual inc > 447e+3 will be regarded as an anomaly.



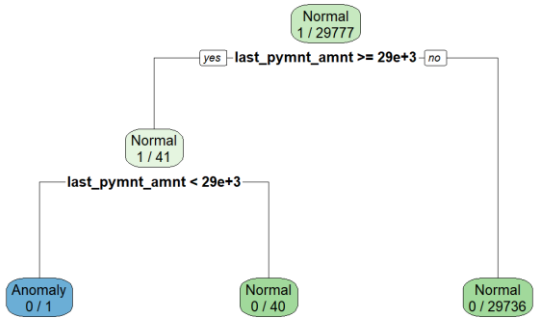

The last payment amount >= 29e+3 and the funded amount < 29e+3 will be regarded as an anomaly.




The last payment amount <28e+3 will be regarded as an anomaly.



The last payment amount <= 27e+3 will be regarded as an anomaly.




The last payment amount <= 29e+3 will be regarded as an anomaly.

# Model analysis

## Model building

Partitioning the data in some form either a Train / Test split and using K-Fold cross-validation

```
# -- XGB model & workflow
xgb_model <- boost_tree(
  trees=40, learn_rate = 0.1,tree_depth = 20) %>%
  set_engine("xgboost") %>%
  set_mode("classification")

xgb_workflow_fit <- workflow() %>%
  add_recipe(loan_recipe) %>%
  add_model(xgb_model) %>%
  fit(train)

xgb_workflow_fit2 <- workflow() %>%
  add_recipe(loan_recipe) %>%
  add_model(xgb_model) %>%
  fit_resamples(kfold_splits)

collect_metrics(xgb_workflow_fit2)
```

XGBoosting with k-fold

| .metric <chr> | .estimator <chr> | mean <dbl> | n <int> | std_err <dbl> | .config <chr> |
|---|---|---|---|---|---|
| accuracy | binary | 0.8796716 | 5 | 0.003565112 | Preprocessor1_Model1 |
| roc_auc | binary | 0.9055783 | 5 | 0.001612468 | Preprocessor1_Model1 |

Random Forest with k-fold

| .metric <chr> | .estimator <chr> | mean <dbl> | n <int> | std_err <dbl> | .config <chr> |
|---|---|---|---|---|---|
| accuracy | binary | 0.8685890 | 5 | 0.001582388 | Preprocessor1_Model1 |
| roc_auc | binary | 0.8818714 | 5 | 0.003008897 | Preprocessor1_Model1 |

Logistic Regression with k-fold

| .metric <chr> | .estimator <chr> | mean <dbl> | n <int> | std_err <dbl> | .config <chr> |
|---|---|---|---|---|---|
| accuracy | binary | 0.7319974 | 5 | 0.004459947 | Preprocessor1_Model1 |
| roc_auc | binary | 0.8606148 | 5 | 0.001834721 | Preprocessor1_Model1 |

## Models' evaluation metrics

| part <chr> | accuracy <dbl> | kap <dbl> | mn_log_loss <dbl> | roc_auc <dbl> | model_name <chr> |
|---|---|---|---|---|---|
| test | 0.8828073 | 0.4588482 | 0.25457340 | 0.9125882 | XGB model |
| train | 0.9964017 | 0.9857768 | 0.07802239 | 0.9978215 | XGB model |

| part <chr> | accuracy <dbl> | kap <dbl> | mn_log_loss <dbl> | roc_auc <dbl> | model_name <chr> |
|---|---|---|---|---|---|
| test | 0.8755317 | 0.3395795 | 0.2812104 | 0.8888103 | RF model |
| train | 0.9480401 | 0.7642212 | 0.1682611 | 0.9954718 | RF model |

| part <chr> | accuracy <dbl> | kap <dbl> | mn_log_loss <dbl> | roc_auc <dbl> | model_name <chr> |
|---|---|---|---|---|---|
| test | 0.8755317 | 0.3395795 | 0.2812104 | 0.8888103 | RF model |
| train | 0.9480401 | 0.7642212 | 0.1682611 | 0.9954718 | RF model |

Therefore, XGBoosting is selected as the best model because it has the highest ROC_AUC (imbalanced data).

## Operating table

XGBoosting

| fpr <dbl> | threshold <dbl> | tpr <dbl> |
|---|---|---|
| 0.25 | 0.09 | 0.9200000 |
| 0.26 | 0.08 | 0.9241176 |
| 0.27 | 0.08 | 0.9304706 |
| 0.28 | 0.07 | 0.9400000 |
| 0.29 | 0.07 | 0.9401205 |
| 0.30 | 0.07 | 0.9500000 |
| 0.31 | 0.06 | 0.9500000 |
| 0.32 | 0.06 | 0.9500000 |
| 0.33 | 0.06 | 0.9545000 |
| 0.34 | 0.05 | 0.9600000 |

FPR means that the percentage of 'current' is regarded as 'default' among actual 'current' and TPR means that the percentage of predicting 'default' correctly. In this situation, we would like to have a higher TPR without much attention to FPR, even though we expect a low FPR. So, we select threshold 0.05 as the line when FPR is 0.34 and TPR is 0.96. it means that we can predict 96% 'default' correctly and only 34% 'current' is predicted as default.

Random Forest

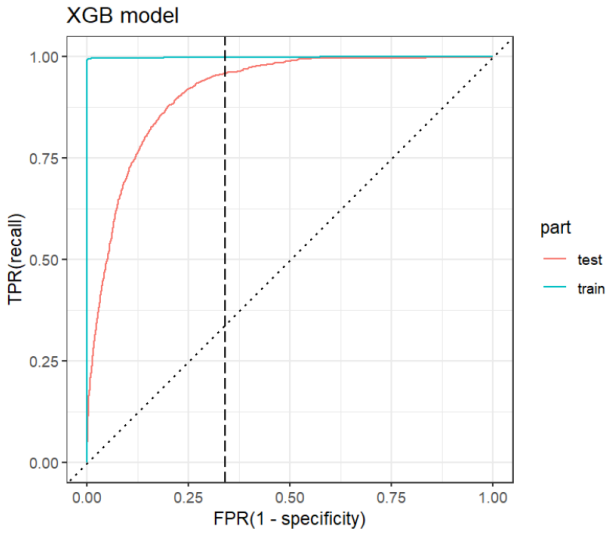
| fpr <dbl> | threshold <dbl> | tpr <dbl> |
|---|---|---|
| 0.31 | 0.14 | 0.9100000 |
| 0.32 | 0.13 | 0.9138372 |
| 0.33 | 0.13 | 0.9220930 |
| 0.34 | 0.12 | 0.9300000 |
| 0.35 | 0.12 | 0.9381928 |
| 0.36 | 0.12 | 0.9400000 |
| 0.37 | 0.11 | 0.9482278 |
| 0.38 | 0.11 | 0.9500000 |
| 0.39 | 0.10 | 0.9539506 |
| 0.40 | 0.10 | 0.9600000 |

Logistic Regression

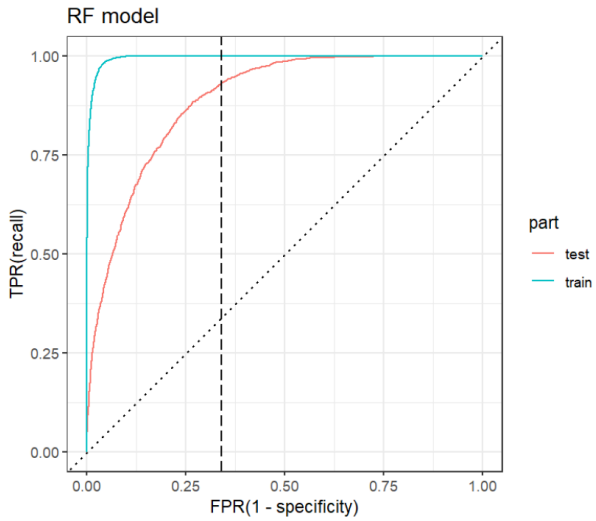
| fpr <dbl> | threshold <dbl> | tpr <dbl> |
|---|---|---|
| 0.44 | 0.09 | 0.9600000 |
| 0.45 | 0.09 | 0.9613750 |
| 0.46 | 0.09 | 0.9700000 |
| 0.47 | 0.08 | 0.9700000 |
| 0.48 | 0.08 | 0.9700000 |
| 0.49 | 0.07 | 0.9788750 |
| 0.50 | 0.07 | 0.9800000 |
| 0.51 | 0.06 | 0.9864198 |
| 0.52 | 0.06 | 0.9900000 |
| 0.53 | 0.05 | 0.9900000 |

## ROC-AUC

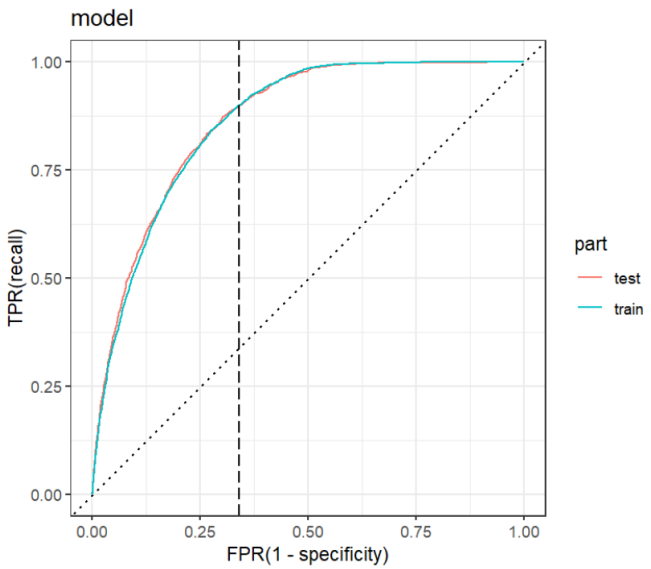
XGBoosting




Random Forest

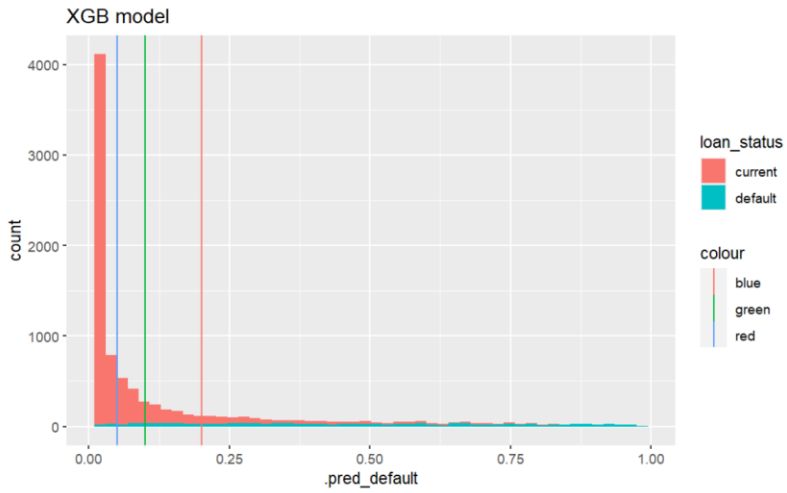



Logistic Regression

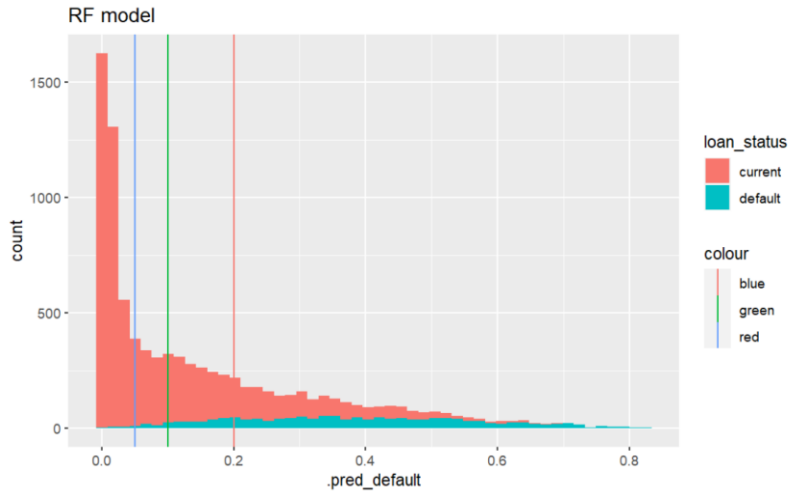



## Threshold graphics

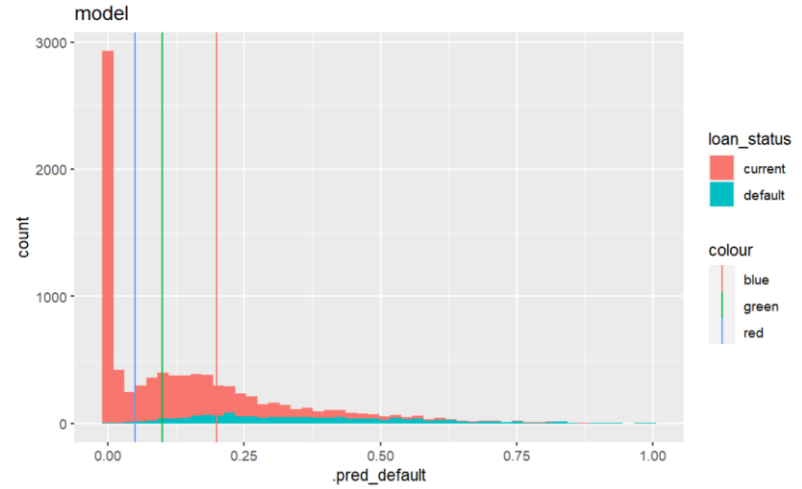
XGBoosting




Random Forest

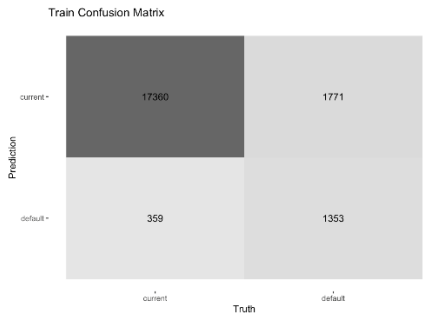


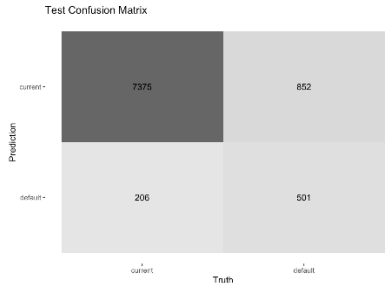

Logistics Regression



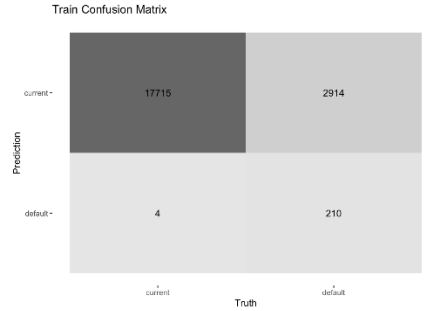

## Confusion matrix

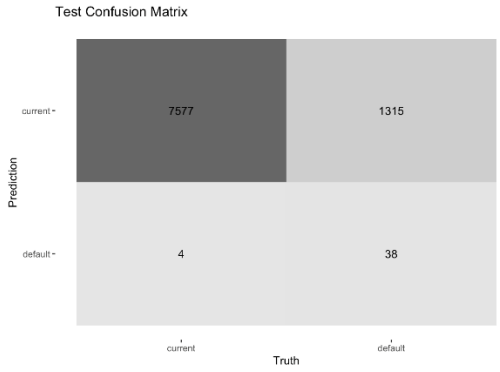
XGBoosting



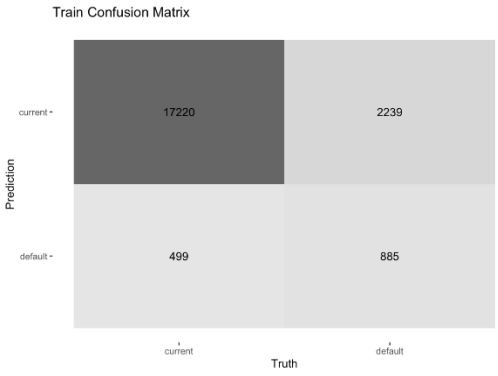


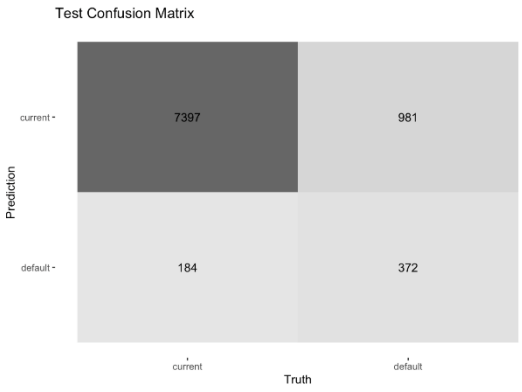

## Random Forest



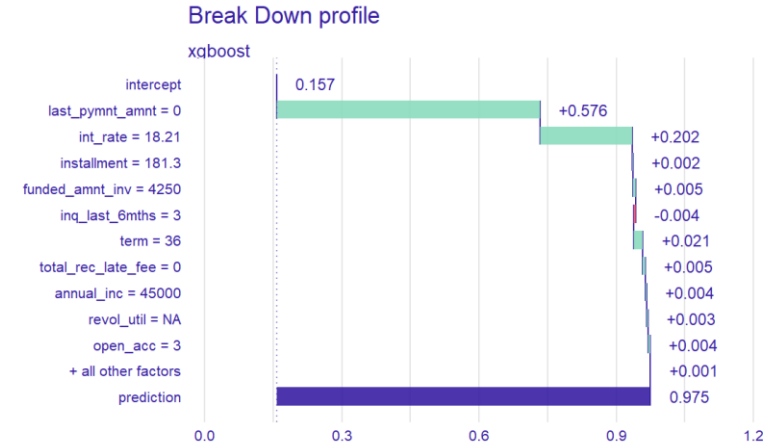

## Logistic regression

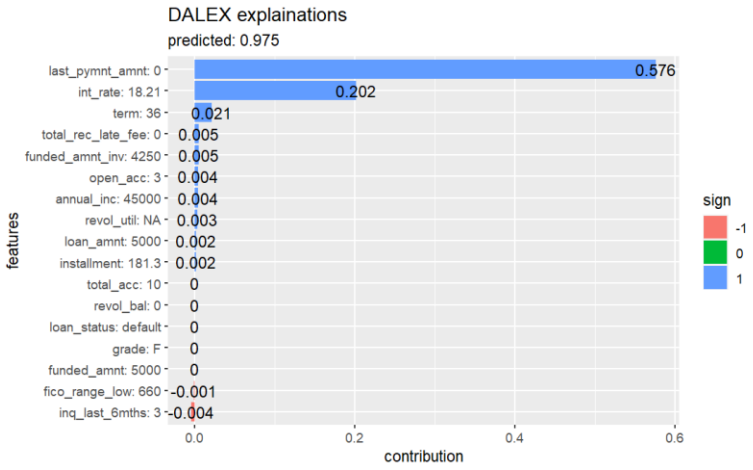


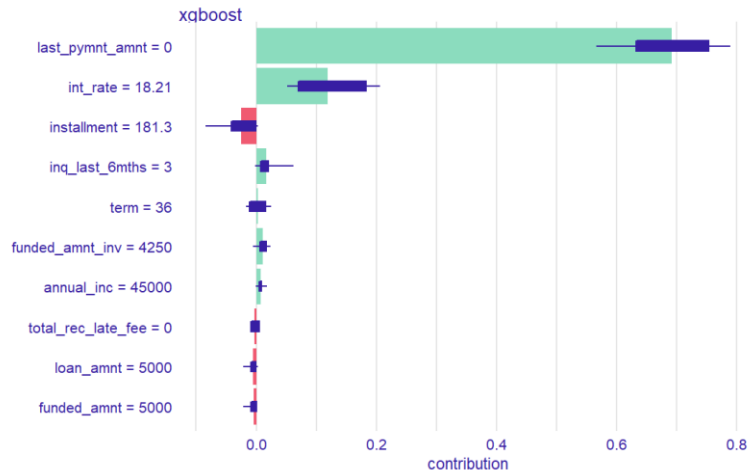


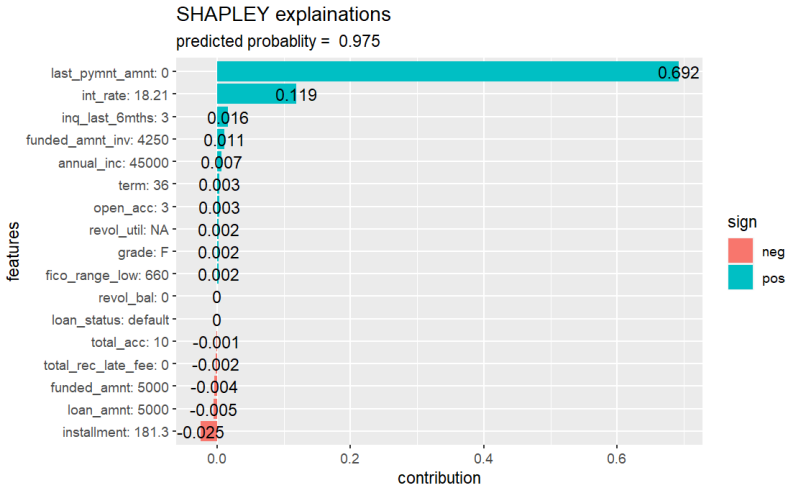

# Methodology

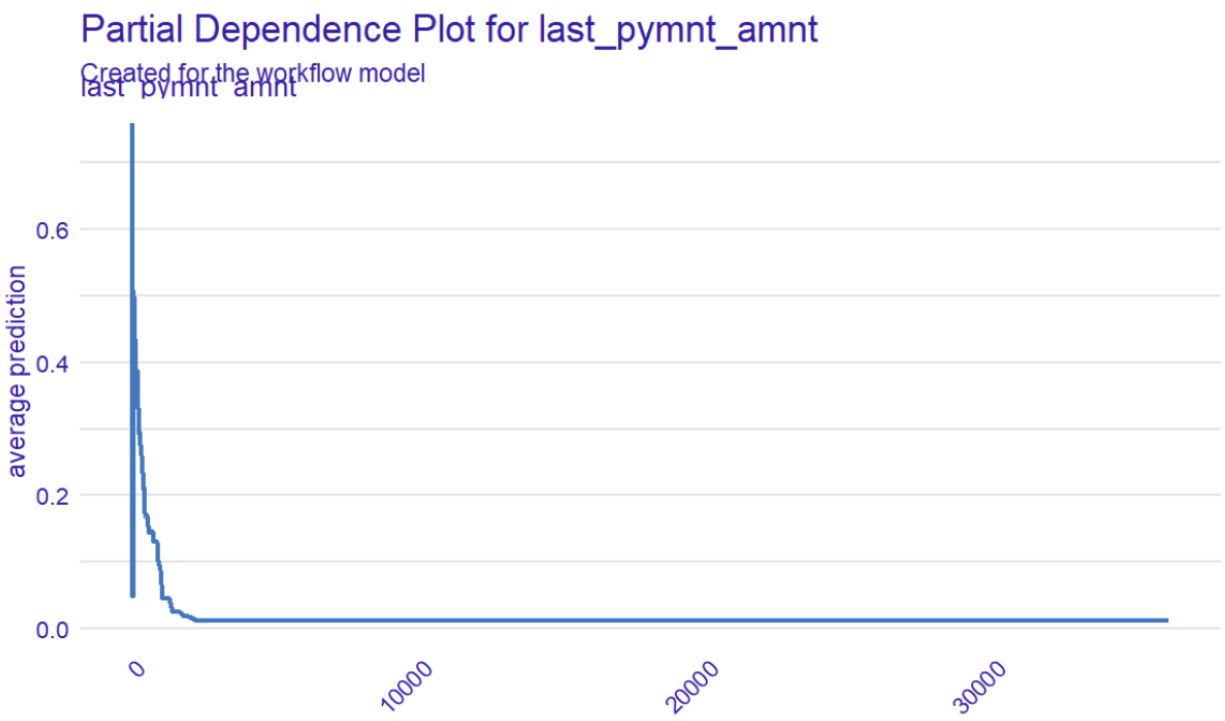
## Global explanations

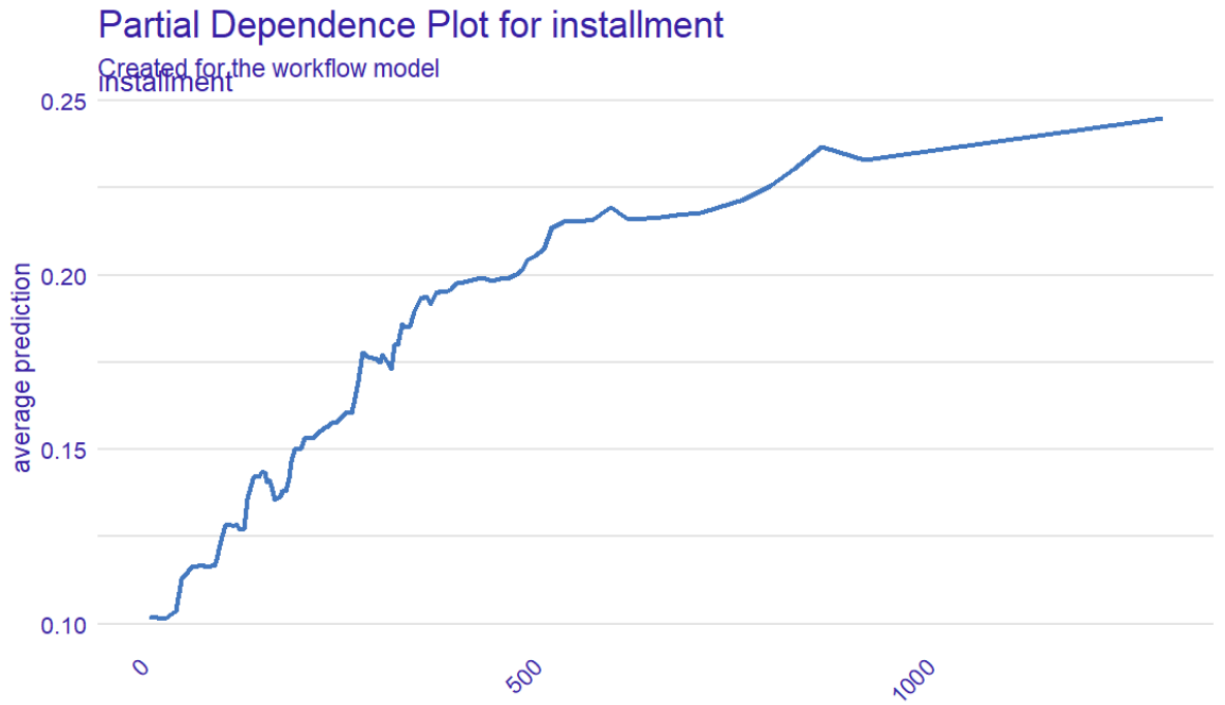


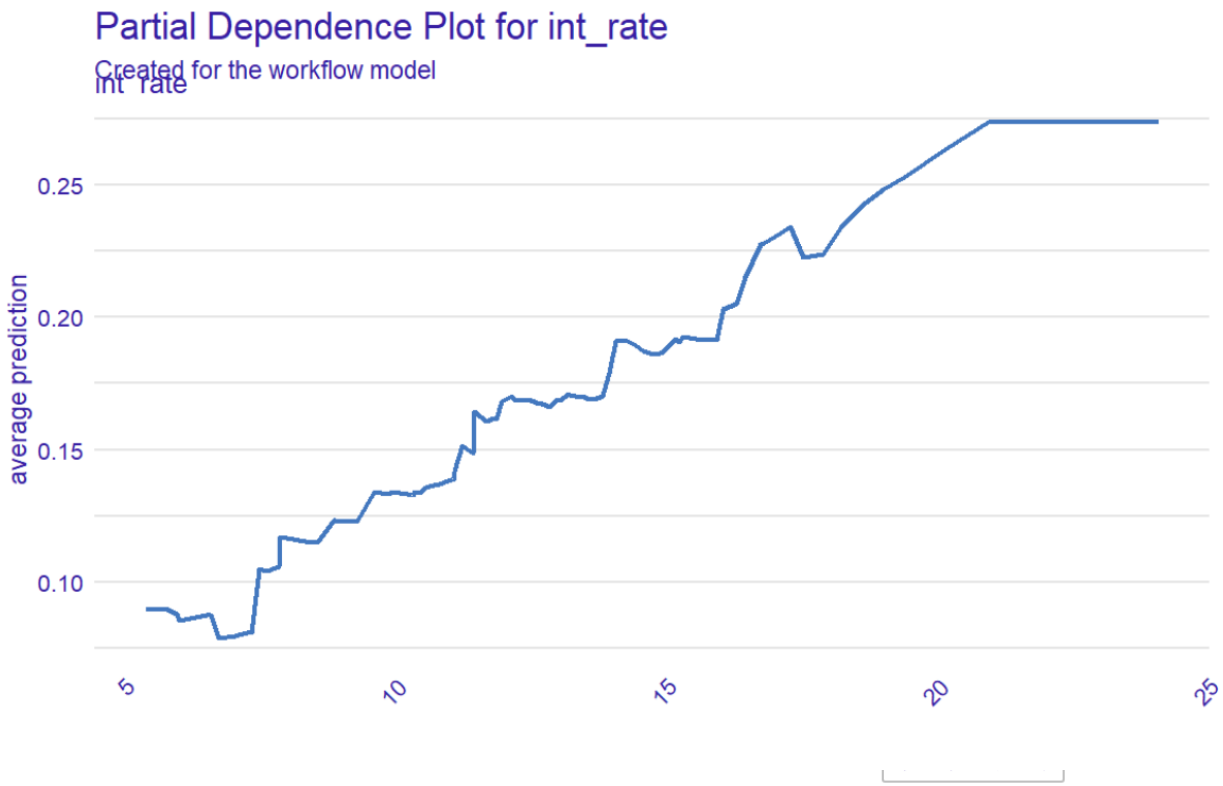



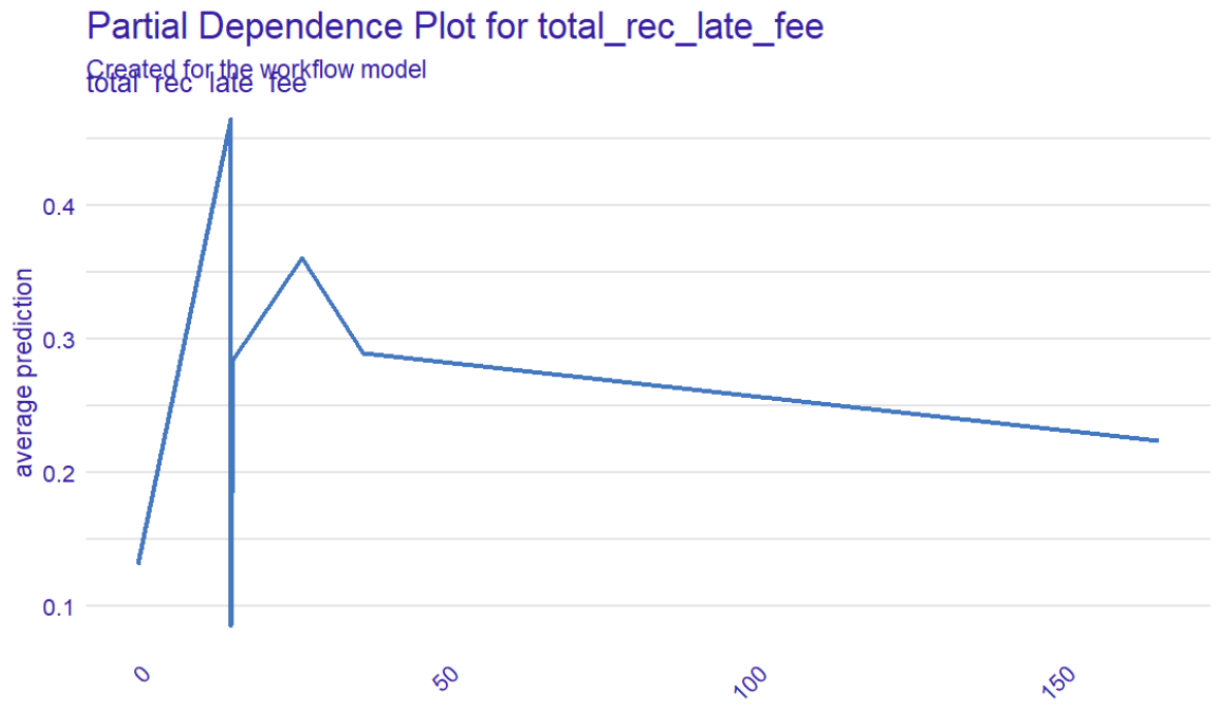

xgboost
last_pymnt_amnt = 0
int_rate = 18.21
installment = 181.3
inq_last_6mths = 3
term = 36
funded_amnt_inv = 4250
annual_inc = 45000
total_rec_late_fee = 0
loan_amnt = 5000
funded_amnt = 5000
0.0
0.2
0.4
0.6
0.8
contribution

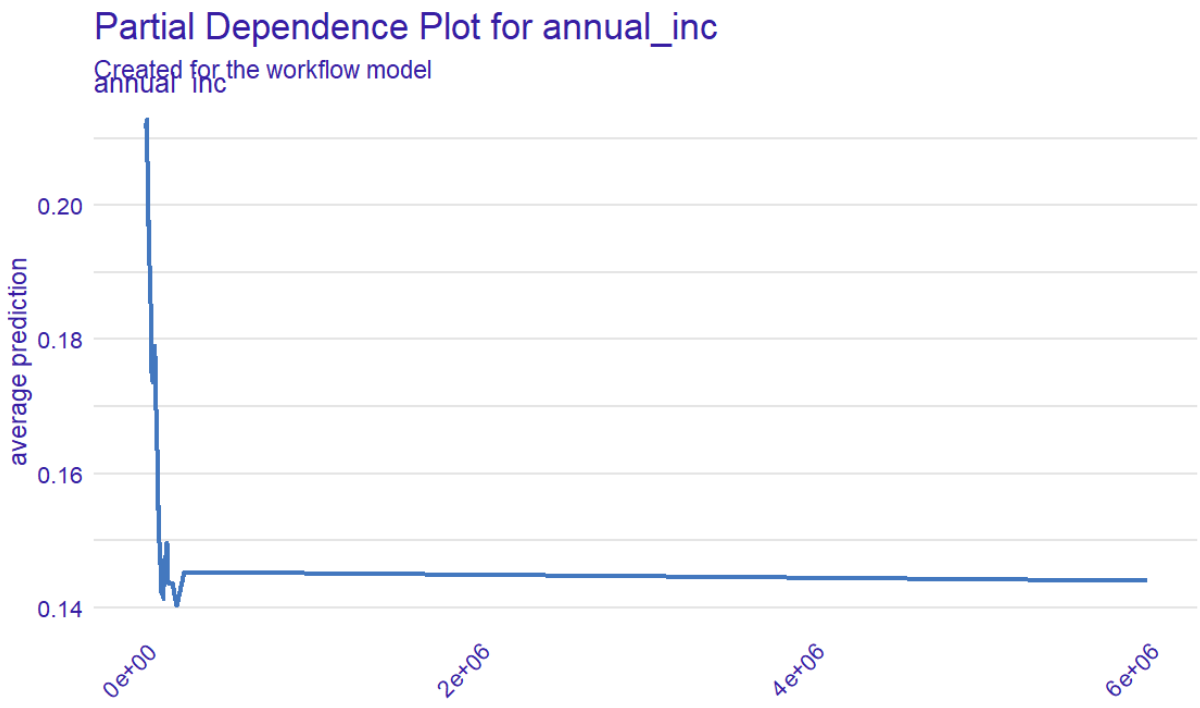

SHAPLEY explainations
predicted probablity = 0.975
last_pymnt_amnt: 0
int_rate: 18.21
inq_last_6mths: 3
funded_amnt_inv: 4250
annual_inc: 45000
term: 36
open_acc: 3
revol_util: NA
grade: F
fico_range_low: 660
revol_bal: 0
loan_status: default
total_acc: 10
total_rec_late_fee: 0
funded_amnt: 5000
loan_amnt: 5000
installment: 181.3
0.692
0.119
0.016
0.011
0.007
0.003
0.003
0.002
0.002
0.002
0
0
-0.001
-0.002
-0.004
-0.005
-0.025
features
0.0
0.2
0.4
0.6
contribution
sign
neg
pos


Partial Dependence Plot for last_pymnt_amnt
Created for the workflow model
last_pymnt_amnt
average prediction
0.6
0.4
0.2
0.0
0
10000
20000
30000


Partial Dependence Plot for installment
Created for the workflow model
installment
average prediction
0.25
0.20
0.15
0.10
0
500
1000


Partial Dependence Plot for int_rate
Created for the workflow model
int_rate
average prediction
0.25
0.20
0.15
0.10
5
10
15
20
25


Partial Dependence Plot for total_rec_late_fee
Created for the workflow model
total_rec_late_fee
average prediction
0.4
0.3
0.2
0.1
0
50
100
150

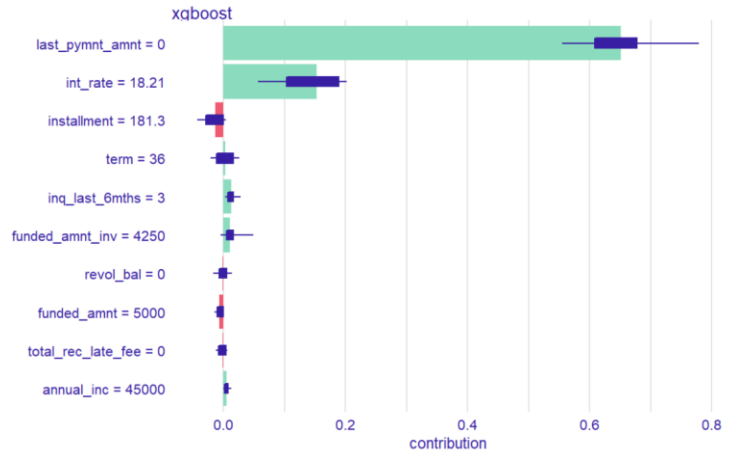


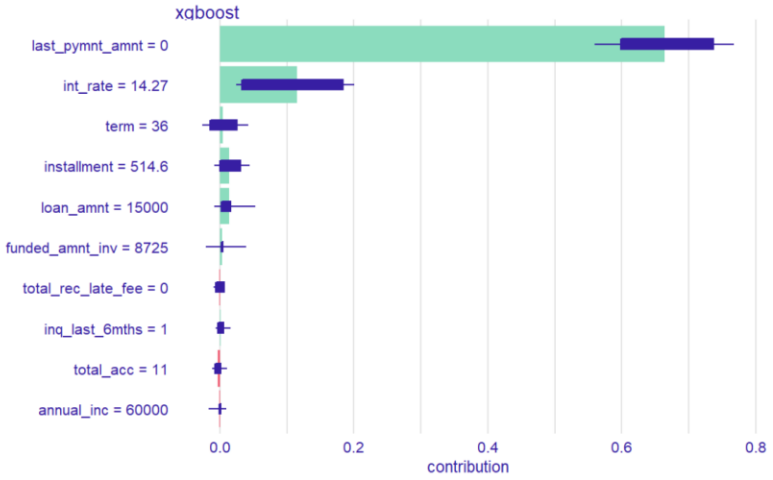

# Local explanations

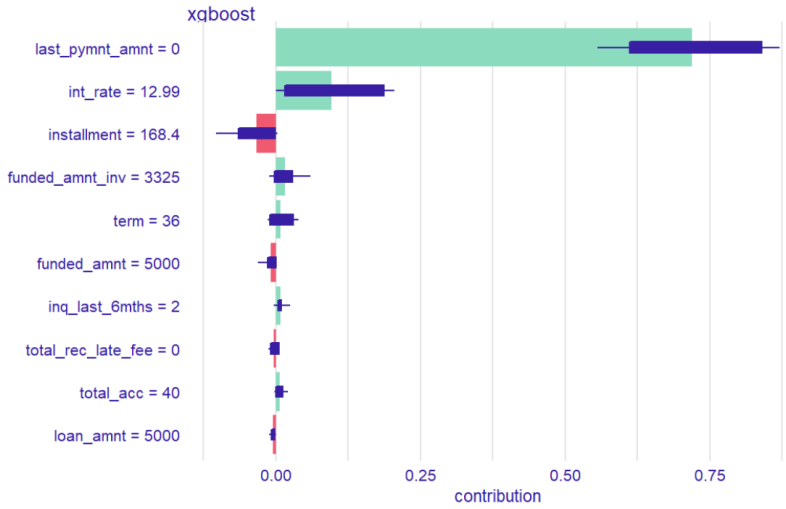
## TP-top 10

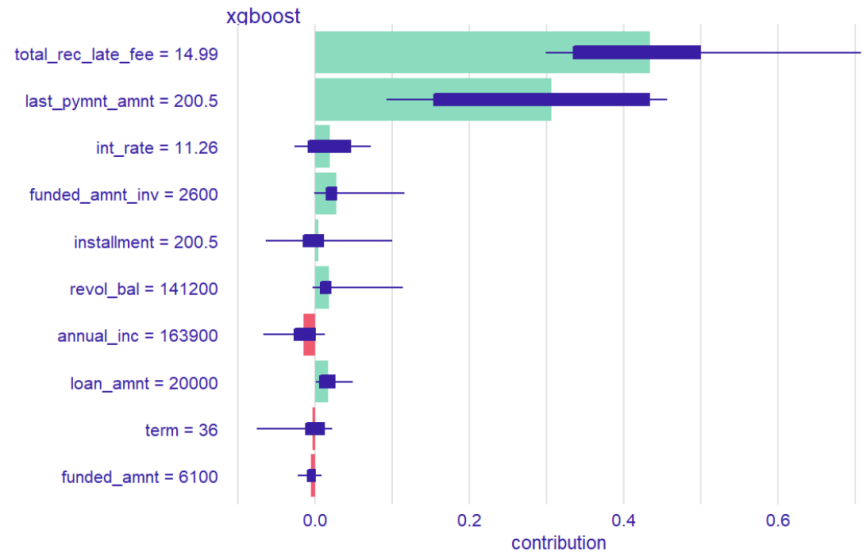
A tibble: 2,491 × 31

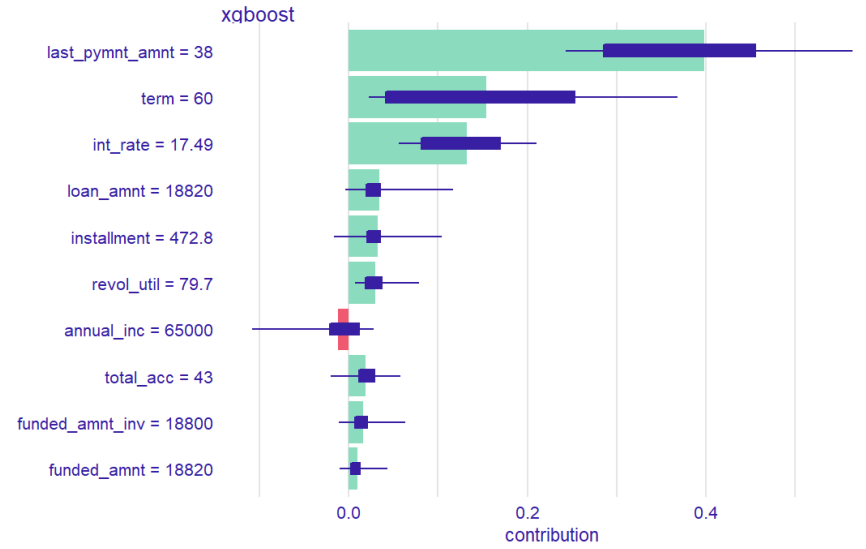
| .pred_current <dbl> | .pred_default <dbl> | .pred_class <fctr> | loan_amnt <int> | funded_amnt <int> | funded_amnt_inv <dbl> | term <dbl> | int_rate <dbl> | installment <dbl> | grade <fctr> |
|---|---|---|---|---|---|---|---|---|---|
| 0.9899091 | 0.01009095 | current | 15000 | 15000 | 15000.000 | 36 | 7.90 | 469.36 | A |
| 0.9899091 | 0.01009095 | current | 6000 | 6000 | 6000.000 | 36 | 12.42 | 200.50 | B |
| 0.9899091 | 0.01009095 | current | 16000 | 16000 | 15950.000 | 60 | 19.91 | 423.11 | E |
| 0.9899091 | 0.01009095 | current | 16000 | 16000 | 16000.000 | 60 | 17.58 | 402.65 | D |
| 0.9899091 | 0.01009095 | current | 9500 | 9500 | 9500.000 | 36 | 8.90 | 301.66 | A |
| 0.9899091 | 0.01009095 | current | 12000 | 12000 | 12000.000 | 36 | 11.71 | 396.92 | B |
| 0.9899091 | 0.01009095 | current | 10000 | 10000 | 10000.000 | 36 | 9.91 | 322.25 | B |
| 0.9899091 | 0.01009095 | current | 16000 | 16000 | 15975.000 | 60 | 19.91 | 423.11 | E |
| 0.9899091 | 0.01009095 | current | 20000 | 20000 | 18590.338 | 60 | 13.49 | 460.10 | C |
| 0.9899091 | 0.01009095 | current | 12000 | 12000 | 11975.000 | 36 | 14.27 | 411.71 | C |


xgboost
last_pymnt_amnt = 0
int_rate = 18.21
installment = 181.3
term = 36
inq_last_6mths = 3
funded_amnt_inv = 4250
revol_bal = 0
funded_amnt = 5000
total_rec_late_fee = 0
annual_inc = 45000
0.0
0.2
0.4
0.6
0.8
contribution


xgboost
last_pymnt_amnt = 0
int_rate = 14.27
term = 36
installment = 514.6
loan_amnt = 15000
funded_amnt_inv = 8725
total_rec_late_fee = 0
inq_last_6mths = 1
total_acc = 11
annual_inc = 60000
0.0
0.2
0.4
0.6
0.8
contribution


xgboost
last_pymnt_amnt = 0
int_rate = 12.99
installment = 168.4
funded_amnt_inv = 3325
term = 36
funded_amnt = 5000
inq_last_6mths = 2
total_rec_late_fee = 0
total_acc = 40
loan_amnt = 5000
0.00
0.25
0.50
0.75
contribution




## FP-top 10

A tibble: 10 × 31

| .pred_current <dbl> | .pred_default <dbl> | .pred_class <fctr> | loan_amnt <int> | funded_amnt <int> | funded_amnt_inv <dbl> | term <dbl> | int_rate <dbl> | installment <dbl> | grade <fctr> |
|---|---|---|---|---|---|---|---|---|---|
| 0.02504502 | 0.9749550 | default | 5000 | 5000 | 4250 | 36 | 18.21 | 181.30 | F |
| 0.02514792 | 0.9748521 | default | 15000 | 15000 | 8725 | 36 | 14.27 | 514.64 | C |
| 0.02791864 | 0.9720814 | default | 5000 | 5000 | 3325 | 36 | 12.99 | 168.45 | C |
| 0.02796551 | 0.9720345 | default | 20000 | 6100 | 2600 | 36 | 11.26 | 200.47 | B |
| 0.02930098 | 0.9706990 | default | 18825 | 18825 | 18800 | 60 | 17.49 | 472.83 | D |
| 0.03149085 | 0.9685092 | default | 8000 | 8000 | 8000 | 36 | 14.91 | 276.98 | D |
| 0.03211951 | 0.9678805 | default | 20000 | 20000 | 19150 | 36 | 13.85 | 682.08 | C |
| 0.03287901 | 0.9671210 | default | 2400 | 2400 | 2400 | 36 | 13.98 | 82.01 | C |
| 0.03326800 | 0.9667320 | default | 25000 | 25000 | 23225 | 36 | 13.92 | 853.43 | C |
| 0.03454849 | 0.9654515 | default | 12000 | 12000 | 11875 | 60 | 17.56 | 301.86 | E |

## FN-top 10

| .pred_current <dbl> | .pred_default <dbl> | .pred_class <fctr> | loan_amnt <int> | funded_amnt <int> | funded_amnt_inv <dbl> | term <dbl> | int_rate <dbl> | installment <dbl> | grade <fctr> |
|---|---|---|---|---|---|---|---|---|---|
| 0.06193544 | 0.9380646 | default | 35000 | 35000 | 34950.00 | 60 | 14.17 | 817.48 | C |
| 0.06662183 | 0.9333782 | default | 17200 | 17200 | 17200.00 | 60 | 20.89 | 464.26 | F |
| 0.06935586 | 0.9306441 | default | 18000 | 14525 | 14289.98 | 60 | 15.95 | 352.84 | D |
| 0.08473454 | 0.9152655 | default | 35000 | 35000 | 34852.55 | 60 | 22.06 | 967.86 | F |
| 0.09500479 | 0.9049952 | default | 35000 | 35000 | 32908.79 | 60 | 20.11 | 929.43 | G |
| 0.09598619 | 0.9040138 | default | 15000 | 15000 | 15000.00 | 60 | 15.58 | 361.44 | D |
| 0.10767475 | 0.8923253 | default | 12000 | 12000 | 12000.00 | 60 | 12.69 | 271.14 | B |
| 0.11423338 | 0.8857666 | default | 25000 | 25000 | 24975.00 | 60 | 21.36 | 681.41 | F |
| 0.12237268 | 0.8776273 | default | 5600 | 5600 | 5600.00 | 60 | 19.36 | 146.38 | F |
| 0.12534994 | 0.8746501 | default | 19075 | 19075 | 19050.00 | 60 | 20.30 | 508.57 | E |