Expected credit loss

Eligijus Bujokas

01/07/2021

Aim of this presentation

The aim of this presentation is to model and present the various techniques of working with expected credit losses using real life data.

The subject in this presentation will be one of the biggest short term loan lenderer in US: **Lending Club**.

Acknowledgment

All the logic and models are presented in the course
https://www.udemy.com/course/credit-risk-modeling-in-python

The presenter of the material is Nikolay G. Georgiev, PhD from the Norwegian Business School.

The data is taken from kaggle
https://www.kaggle.com/ethon0426/lending-club-20072020q1

Lending club website https://www.lendingclub.com/

Expected credit loss

$$\mathbb{E}[CL] = \mathbb{E}[PD]\mathbb{E}[LGD]\mathbb{E}[EAD]$$

CL - Credit losses

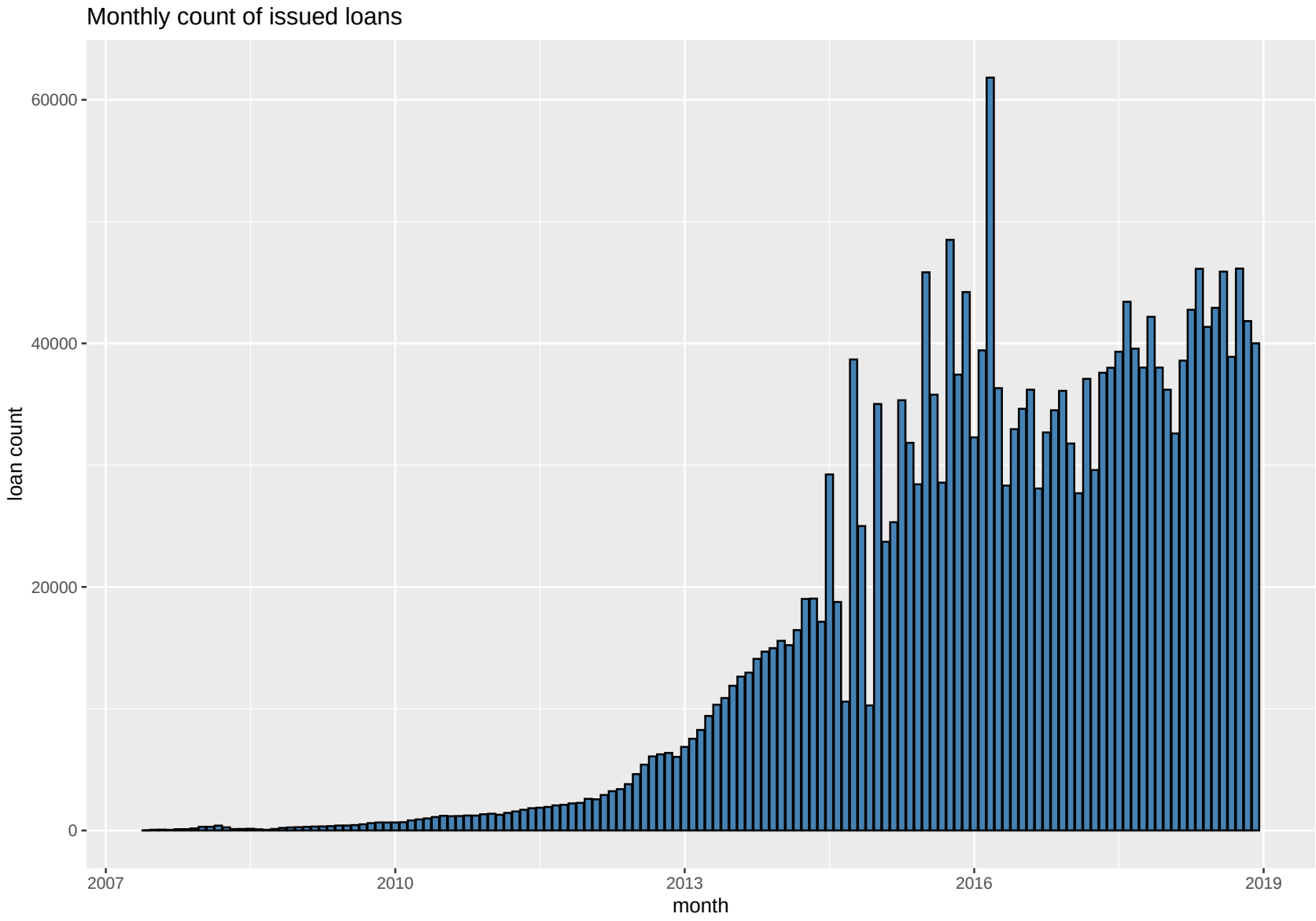
PD - Probability of default

LGD - Losses given default

EAD - Exposure at default

Exploring the data

The data is monthly spanning from 2008 January up until 2018 December (included).



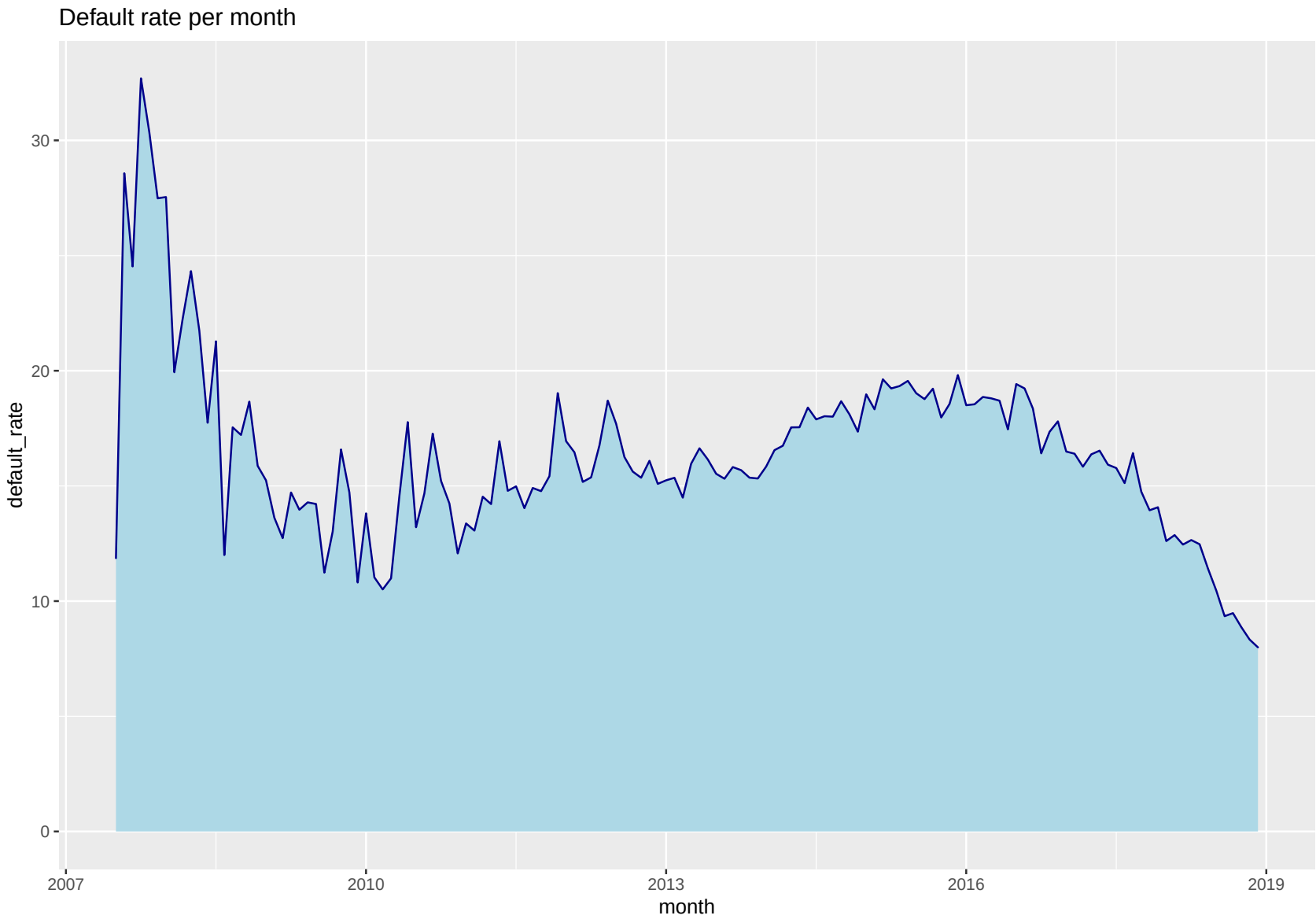
Loan status

There is a feature in the data called **loan_status**. The values can be:

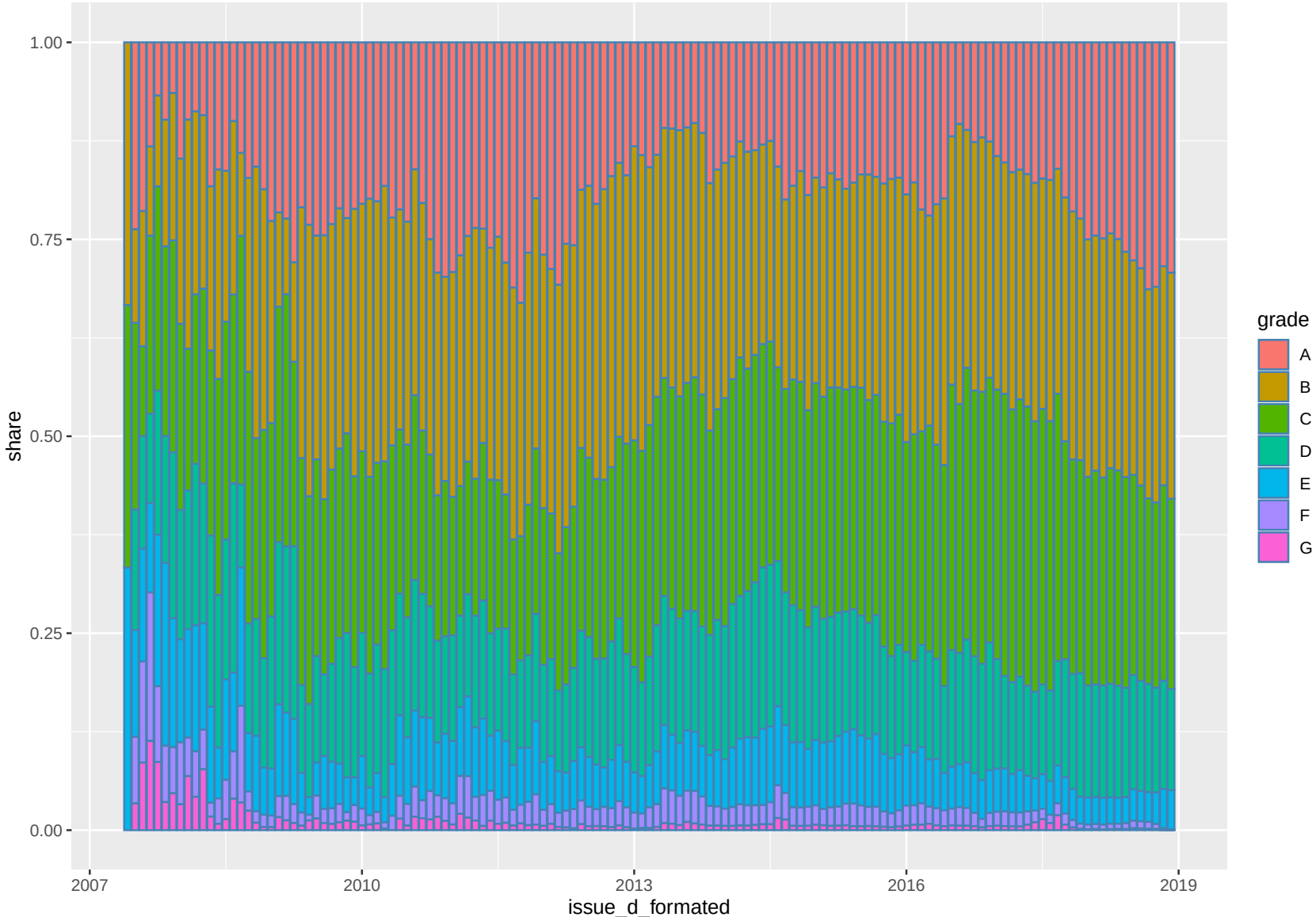
```
##                                                loan_status       N
## 1:                                                 Default     351
## 2: Does not meet the credit policy. Status:Charged Off     757
## 3:                                          Late (16-30 days)    1518
## 4:  Does not meet the credit policy. Status:Fully Paid    1957
## 5:                                            In Grace Period    5627
## 6:                                         Late (31-120 days)    9499
## 7:                                                Charged Off  347810
## 8:                                                    Current  458491
## 9:                                                 Fully Paid 1427932
```

Bad loans

```
## Warning: Removed 1 rows containing missing values (position_stack).
```



Distribution by grade



The Basel Accords

The Basel II accord, which was signed in 2004, defined three strict guidelines:

- How much capital banks need to have
- How capital is defined
- How capital is compared against risk-weighted assets

One of the main takeouts from both the basel II and subsequent basel III accords is that

The greater the risk a bank is exposed to, the greater the amount of capital it needs to hold

Probability of default (PD)

This is the most strict part of the three components of ECL and must follow certain rules in modeling. Every feature, both categorical and numeric, needs to be transformed into dummy variables.

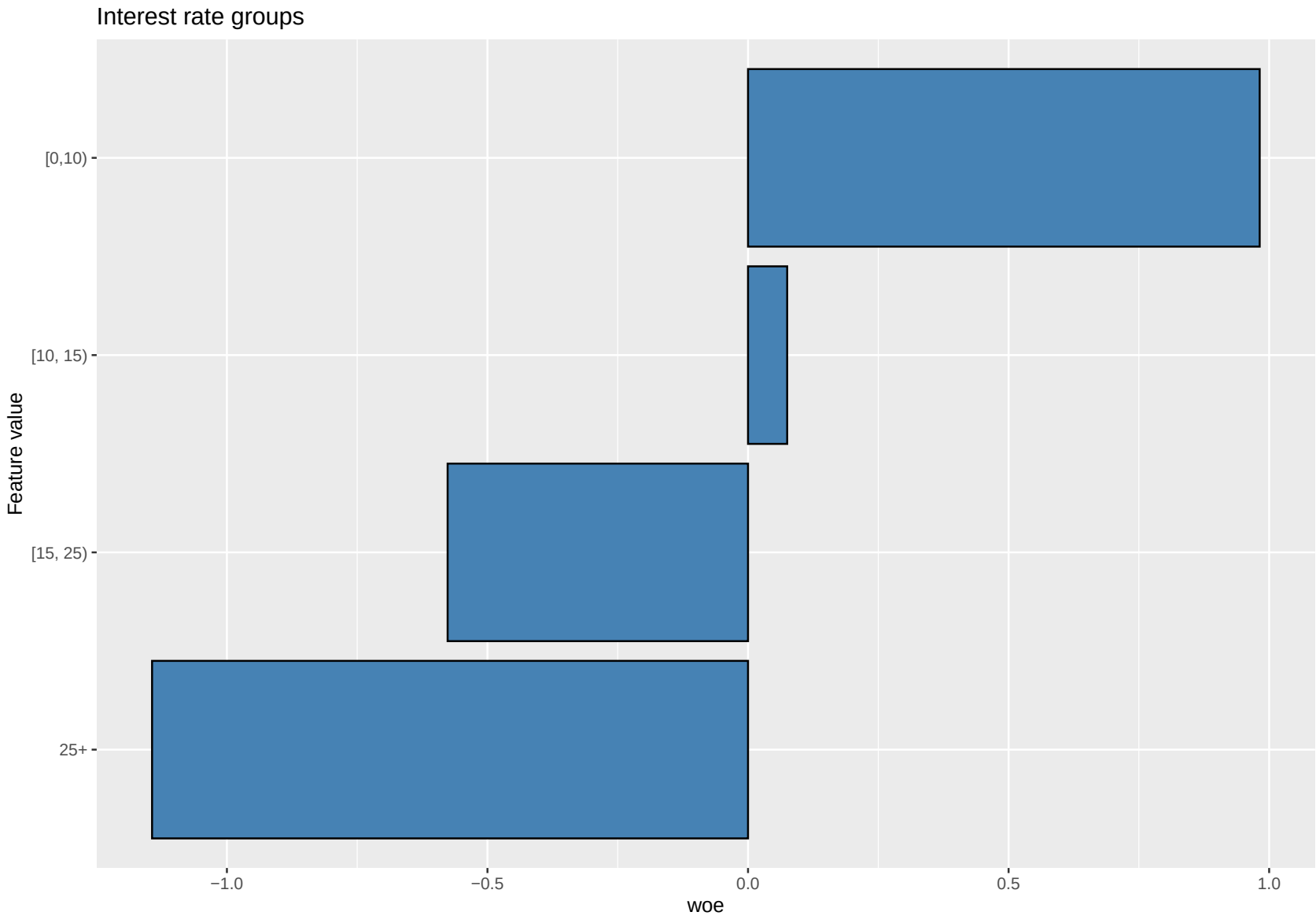
To infer what categorical feature values are the most influential in determining bad loan from a good loan we can use the weight of evidence (WOE for short) criteria. For a feature i and the feature level j the $WOE_{i,j}$ is calculated with the following formula:

$$WOE_{i,j} = log\left(\frac{P(X_i = j|Y = 1)}{P(X_i = j|Y = 0)}\right)$$

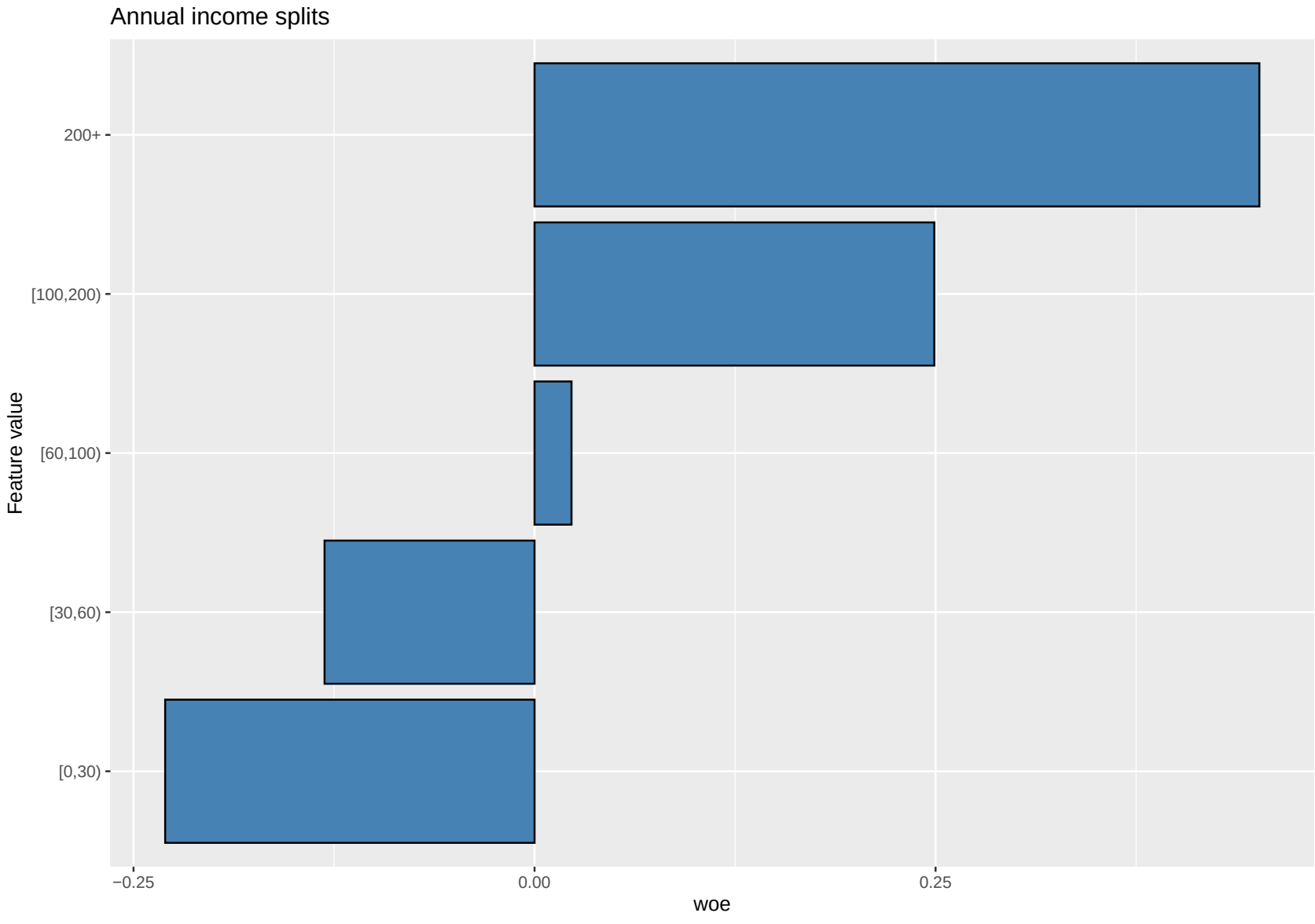
WOE example

feature	bad	good	prop_good	prop_bad	woe
36 months	214070	1390942	0.733803	0.5972652	0.2058794
60 months	144347	504583	0.266197	0.4027348	-0.4140418

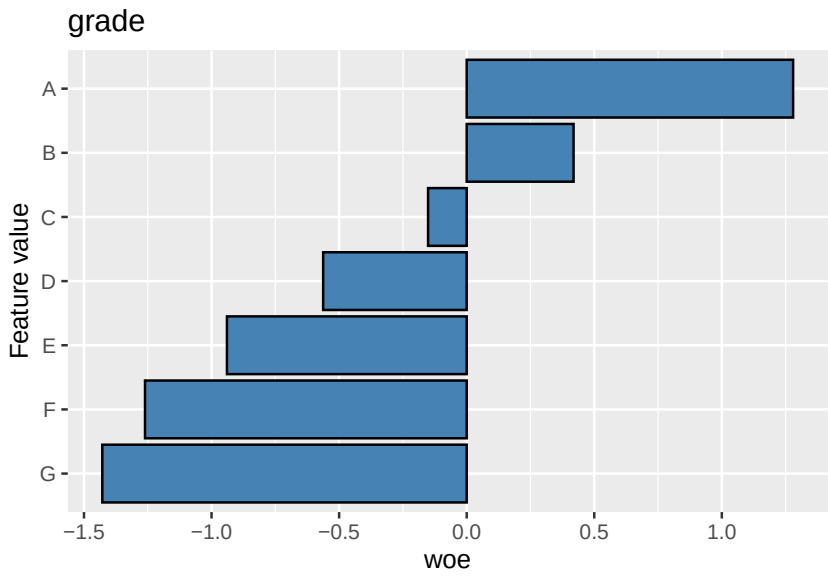
WOE for interest rates

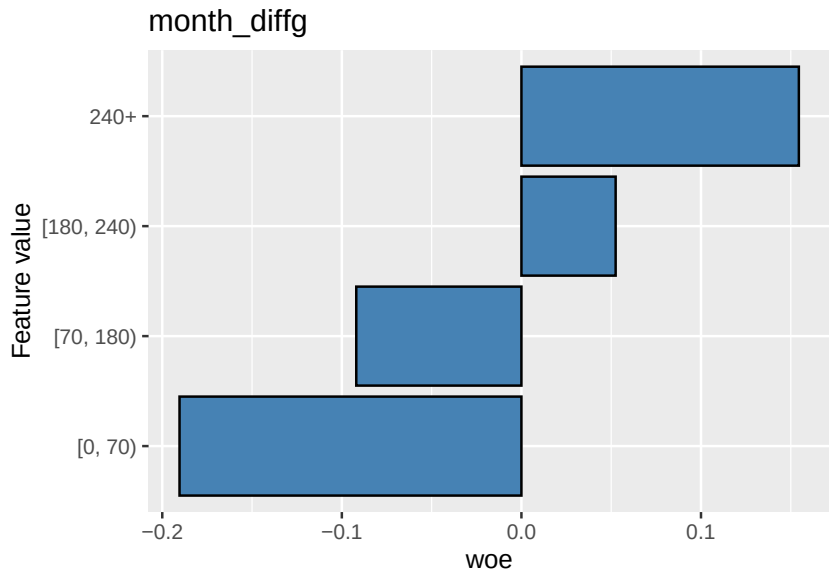


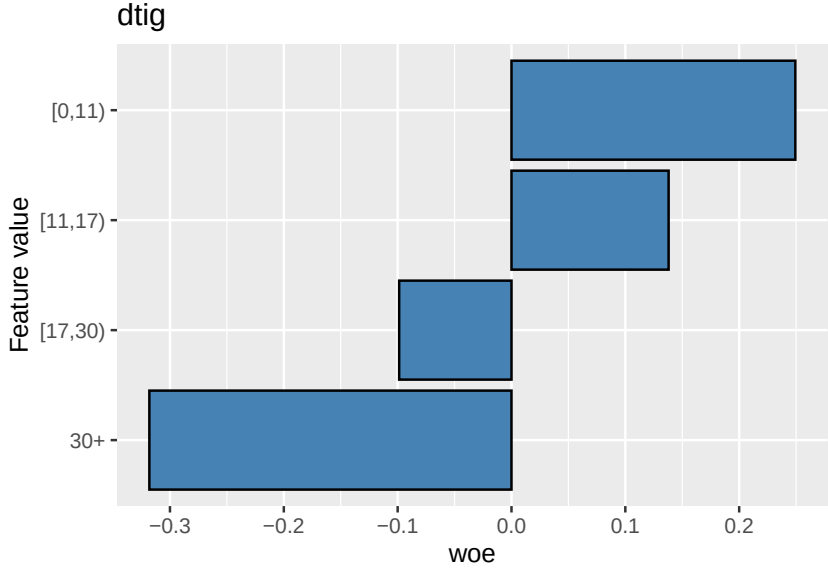
Annual income

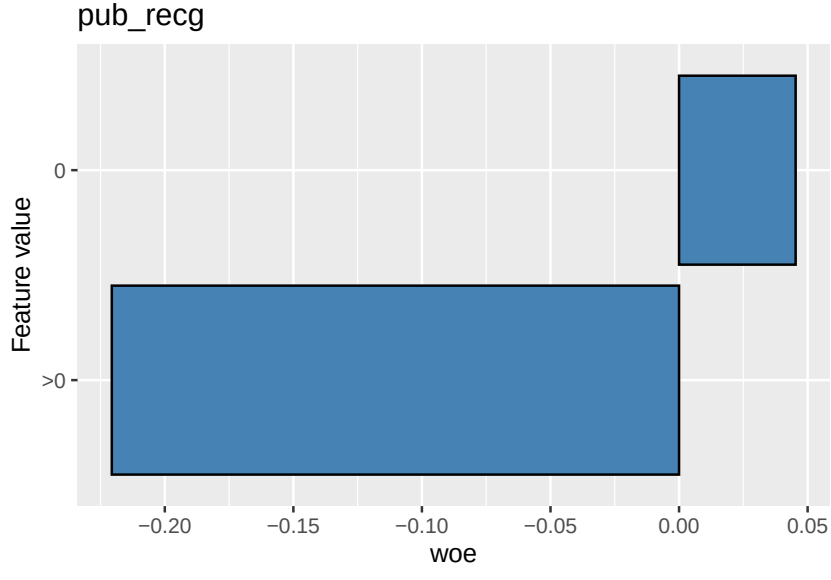


Some more features









Final variable list used for PD

term

pub_rec

annual_inc

dti

month_diff

emp_length

int_rate

loan_amnt

grade

Machine learning method

Now that we have our X matrix and our Y matrix, we need a method to model the relationship between them.

A popular choice is the **logistic regression** model for binary classification.

We want to estimate the following conditional probability:

$$P(Y = 1|X)$$

In our case:

$$P(Y = good\_loan|data)$$

Logistic regression - regression part

Regress - "coming back to" (liet. - grįžimas).

The term is accredited to Francis Galton in the 19th century in his biological work.

Regression $\approx$ "coming back to the mean"

Regression models try to model the expected value (average) of the dependent variable with the independent ones.

In general terms:

$$\mathbb{E}[Y|X] = \mu = g^{-1}(\beta_0 + \Sigma_{i=1}^{k}(\beta_i X_i))$$

$g(.)$ is called the link function.

Logistic regression - logistic part

The standard logistic function is:

$$logistic(x) = \frac{1}{1 + e^{-x}}$$

$$logistic : (-\infty, +\infty) \rightarrow (0, 1)$$

The logit (log-odds) function is:

$$logit(x) = log(\frac{x}{1 - x})$$

$$logit : (0, 1) \rightarrow (-\infty, +\infty)$$

$$logit^{-1}(x) = logistic(x)$$

Logistic regression equation

Putting "logistic" and "regression" together:

Lets define:

$$z := \beta_0 + \Sigma_{i=1}^{k} (\beta_i X_i))$$

Logistic regression is form of general linear models (GLM) where the link function is the logit function.

$$\mathbb{E}[Y|X] = \mathit{logit}^{-1}(z) = \mathit{logistic}(z) = \frac{1}{1 + e^{-z}}$$

Estimating the coefficients

The way we obtain the coefficients β is by using a procedure called maximum likelihood (ML).

$$\Pi_{i=1}^{n} \left[P(Y = 1|X)^{y_i} (P(Y = 0|X)^{1-y_i}) \right] \rightarrow max$$

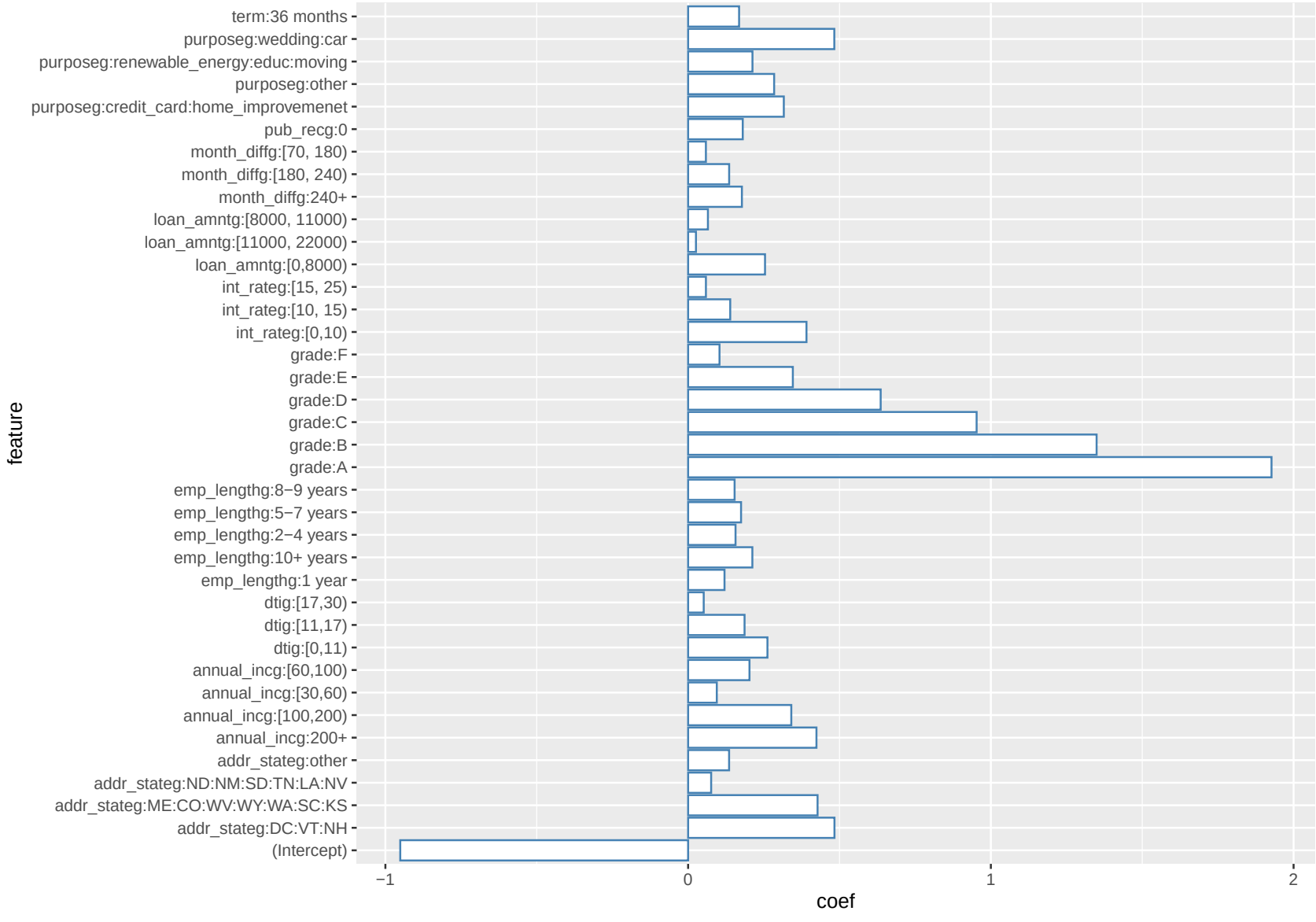
$$L(\beta|Y, X) = \Pi_{i=1}^{n} \left[logistic(z)^{y_i} ((1 - logistic(z))^{1-y_i}) \right]$$

$$l(\beta|Y, X) = log(L(\beta|Y, X))$$

$$l(\beta) = \Sigma_{i=1}^{n} [y_i log(logisitc(z)) + (1 - y_i) log(1 - logistic(z))]$$

The computer tries to find the "best" β values such that the probability is as big as possible for witnessing the Y in our sample from the given X.

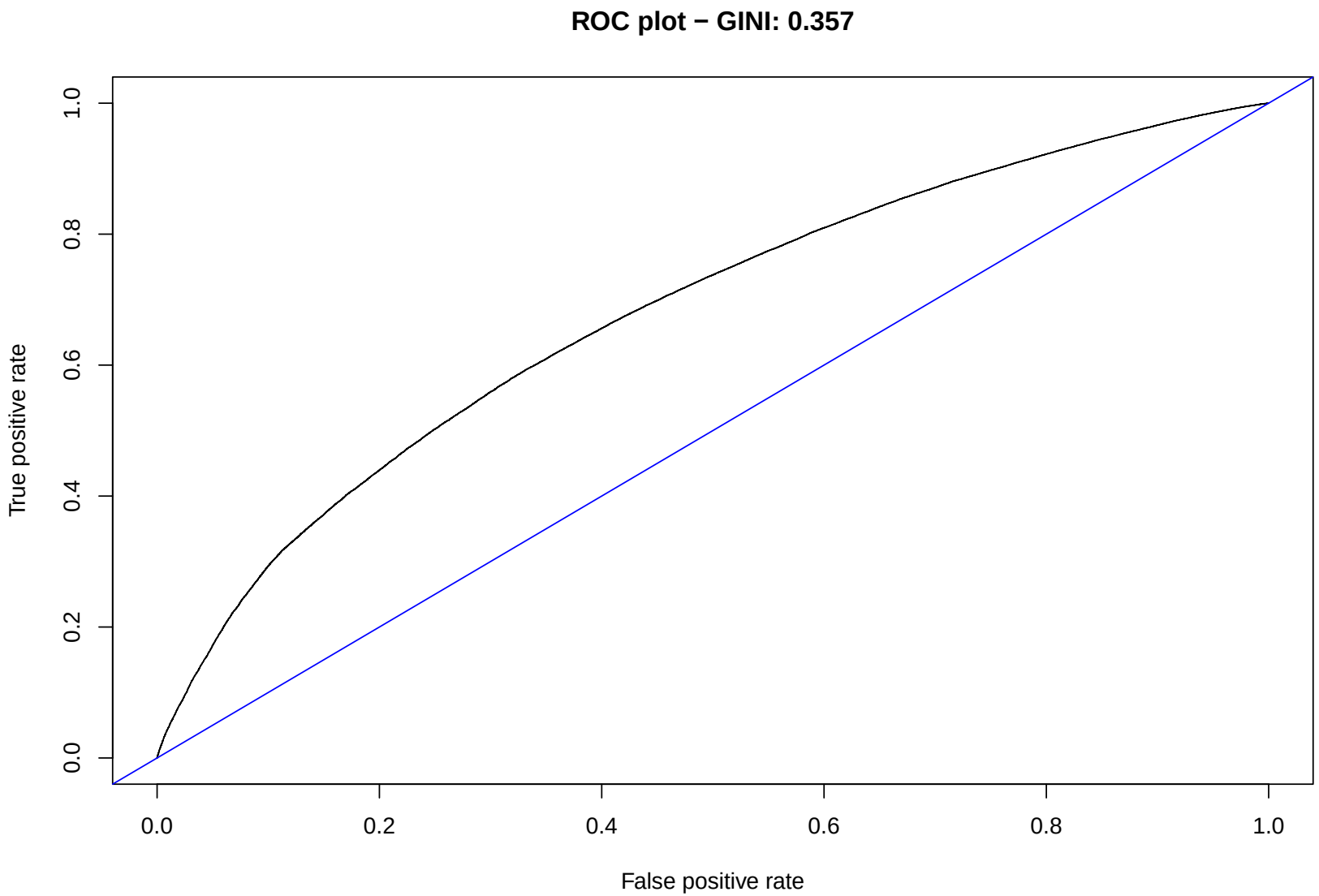
The model for PD



Results on the test set

```
FALSE Setting levels: control = bad, case = good

FALSE Setting direction: controls < cases
```



Converting to a scorecard

We will recalibrate the coefficients to be between 200 and 800.

```
##     origFeature   min_coef   max_coef
##  1: (Intercept) -0.9508579 -0.9508579
##  2: addr_stateg  0.0000000  0.4834281
##  3: annual_incg  0.0000000  0.4239534
##  4:        dtig  0.0000000  0.2620422
##  5: emp_lengthg  0.0000000  0.2122388
##  6:       grade  0.0000000  1.9270668
##  7:   int_rateg  0.0000000  0.3909776
##  8:  loan_amntg  0.0000000  0.2540543
##  9: month_diffg  0.0000000  0.1776638
## 10:    pub_recg  0.0000000  0.1804778
## 11:    purposeg  0.0000000  0.4829255
## 12:        term  0.0000000  0.1684319
```

Formulas for converting

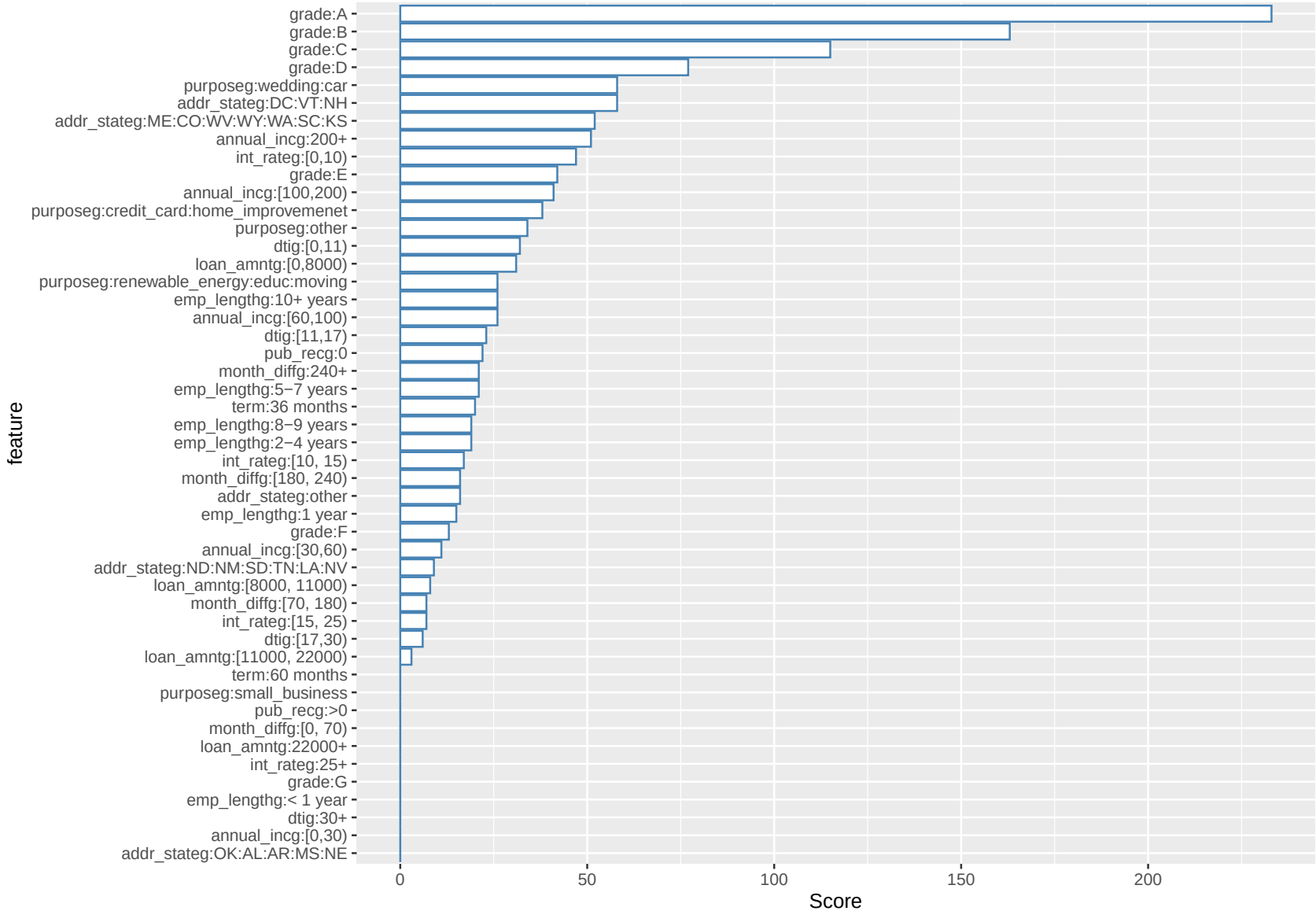
To convert the coefficient to a score, we need to follow the following formula:

$$score = coef \frac{(max\_score - min\_score)}{maxsum\_coef - minsum\_coef}$$

To adjust the coefficient for the intercept we will use the formula:

$$score\_intercept = \frac{(intercept\_coef - min\_sum\_coef)}{(max\_sum\_coef - min\_sum\_coef)}(max\_score - min\_score) + min\_score$$

Final scorecard



EAD modeling - Y variable

The dependent variable for the exposure at default is the amount of funds that a bank is at risk at a default event. They way we model it is using a variable called credit conversion factor (CCF):

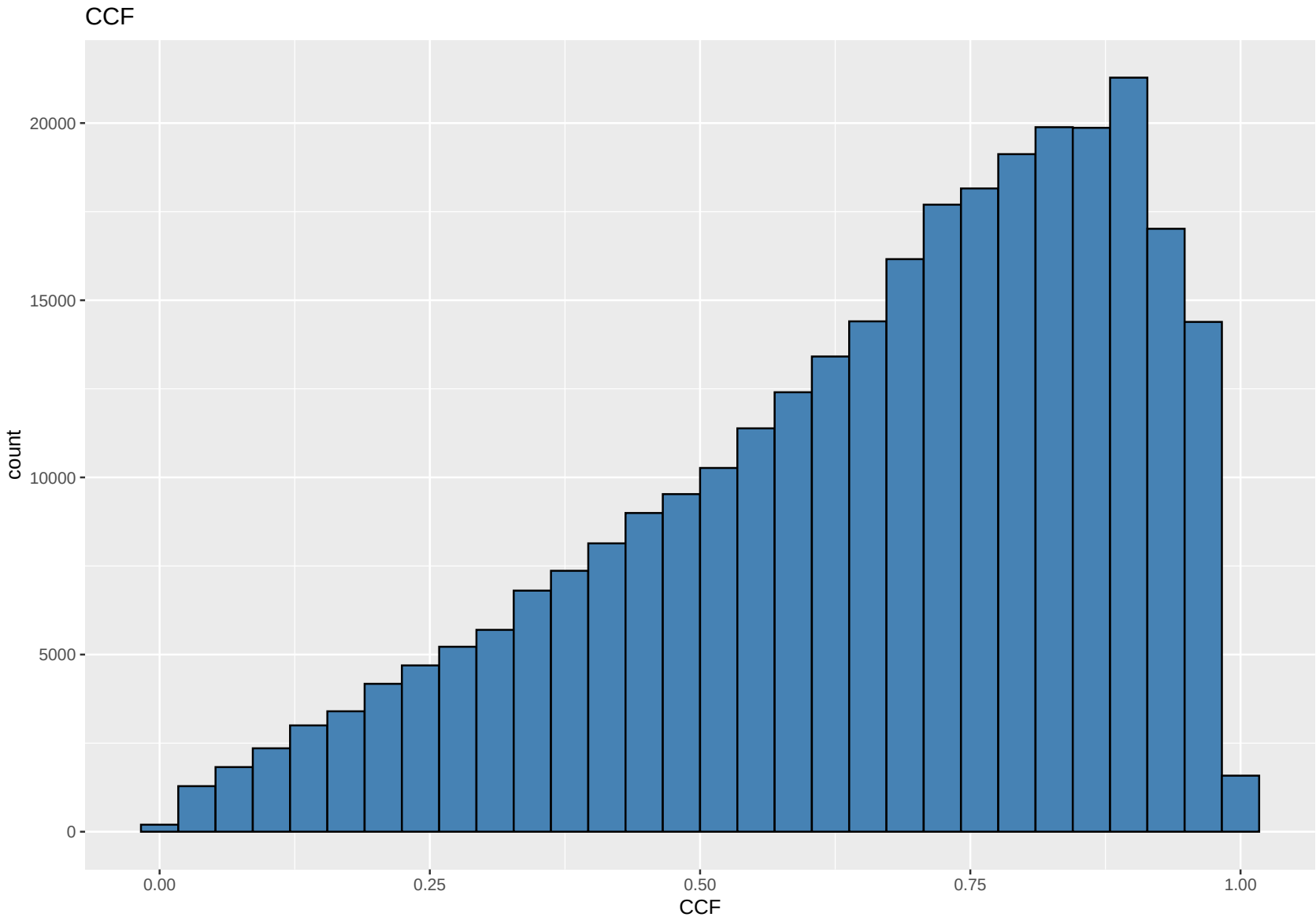
$$CCF_i = \frac{funded\_amount_i - received\_payments_i}{funded\_amount_i}$$

The higher the CCF for a given loan, the bigger is the EAD sum:

$$EAD_i = CCF_i * funded\_amount_i$$

EAD Y variable distribution

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

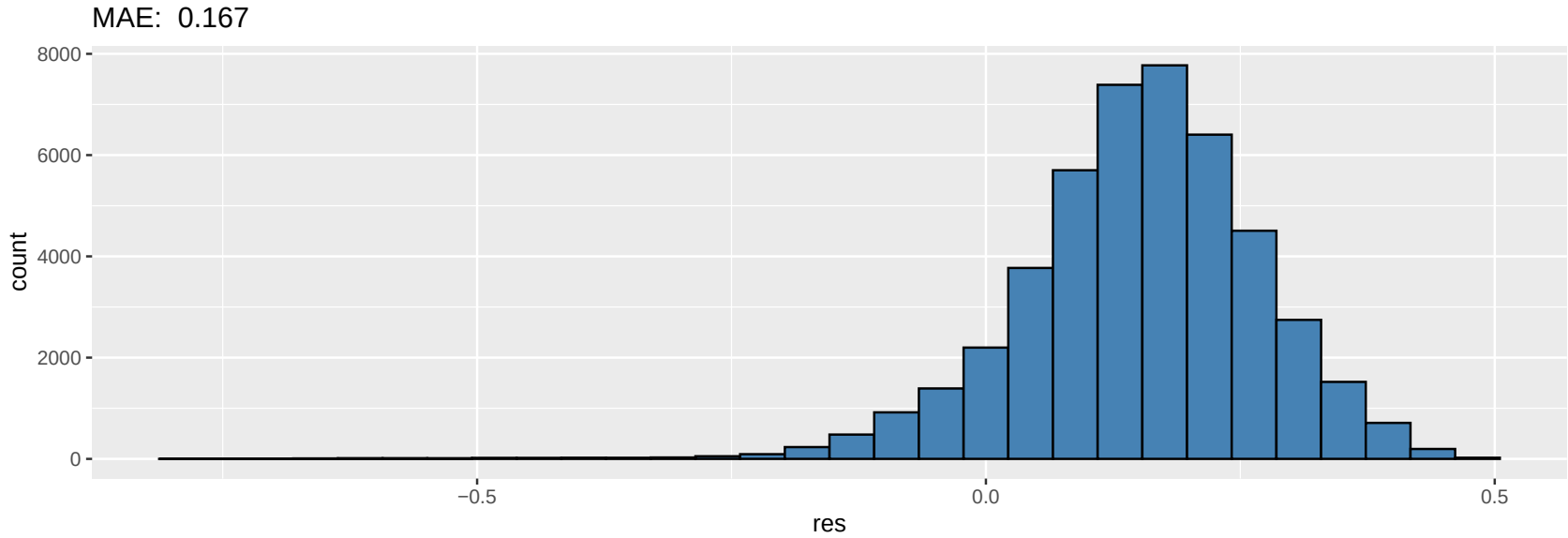


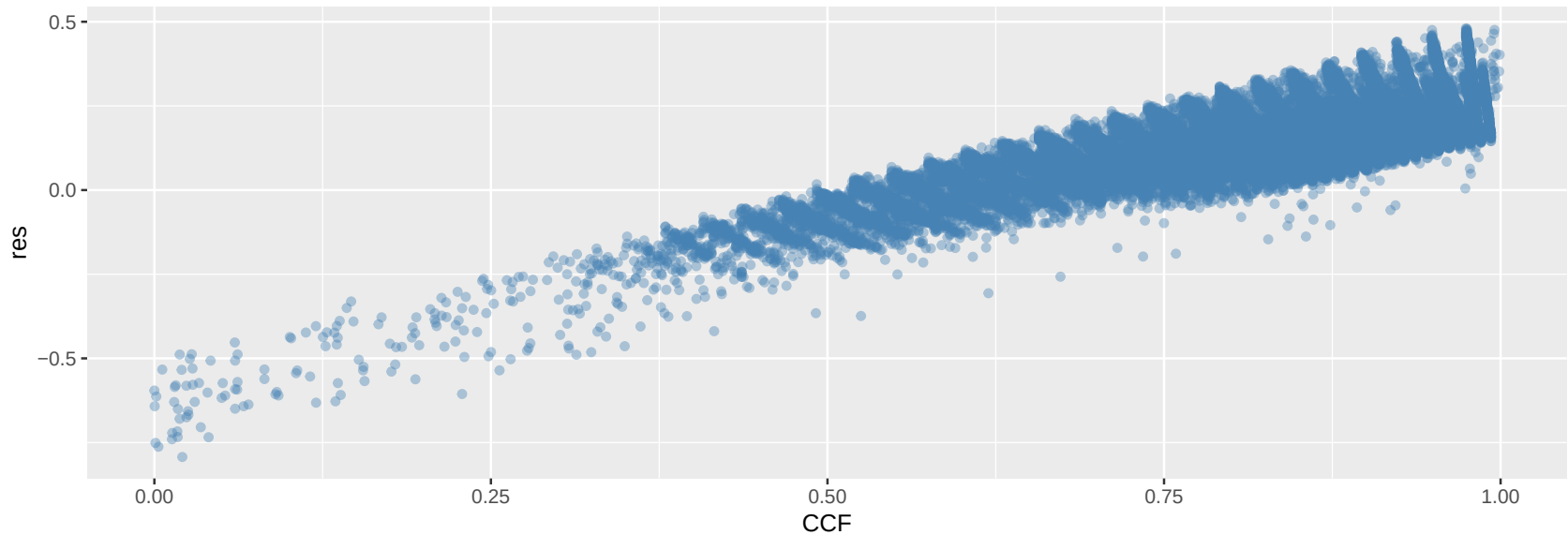
The model for EAD

V1	feature	coef	exp_beta
1	(Intercept)	-2.6088228	0.0736212
2	term60 months	0.1145067	1.1213201
3	pub_recg>0	0.0692237	1.0716759
4	annual_inc	0.0000019	1.0000019
5	dti	0.0075368	1.0075652
6	month_diff	0.0001042	1.0001042
7	int_rate	0.0007019	1.0007022
8	loan_amnt	-0.0000012	0.9999988
9	gradeB	0.0369728	1.0376648
10	gradeC	0.1302691	1.1391348
11	gradeD	0.1570964	1.1701084
12	gradeE	0.2032814	1.2254172
13	gradeF	0.2153040	1.2402389
14	gradeG	0.2705374	1.3106686
15	addr_stategME:CO:WV:WY:WA:SC:KS	0.0163051	1.0164387
16	addr_stategND:NM:SD:TN:LA:NV	0.1295943	1.1383665
17	addr_stategOK:AL:AR:MS:NE	0.1492610	1.1609760
18	addr_stategother	0.0970180	1.1018802
19	purposegother	-0.0042949	0.9957143
20	purposegrenewable_energy:educ:moving	0.0116414	1.0117094
21	purposegsmall_business	-0.1964943	0.8216060
22	purposegwedding:car	-0.1431135	0.8666557

EAD residuals in the test set

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```





LGD modeling - Y variable

LGD stands for losses given default. When dealing with loan data we can model recovery rate. Then, for each loan,

$$LGD = 1 - recovery\_rate$$

In this data set, we calculate the recovery rate using the following equation:

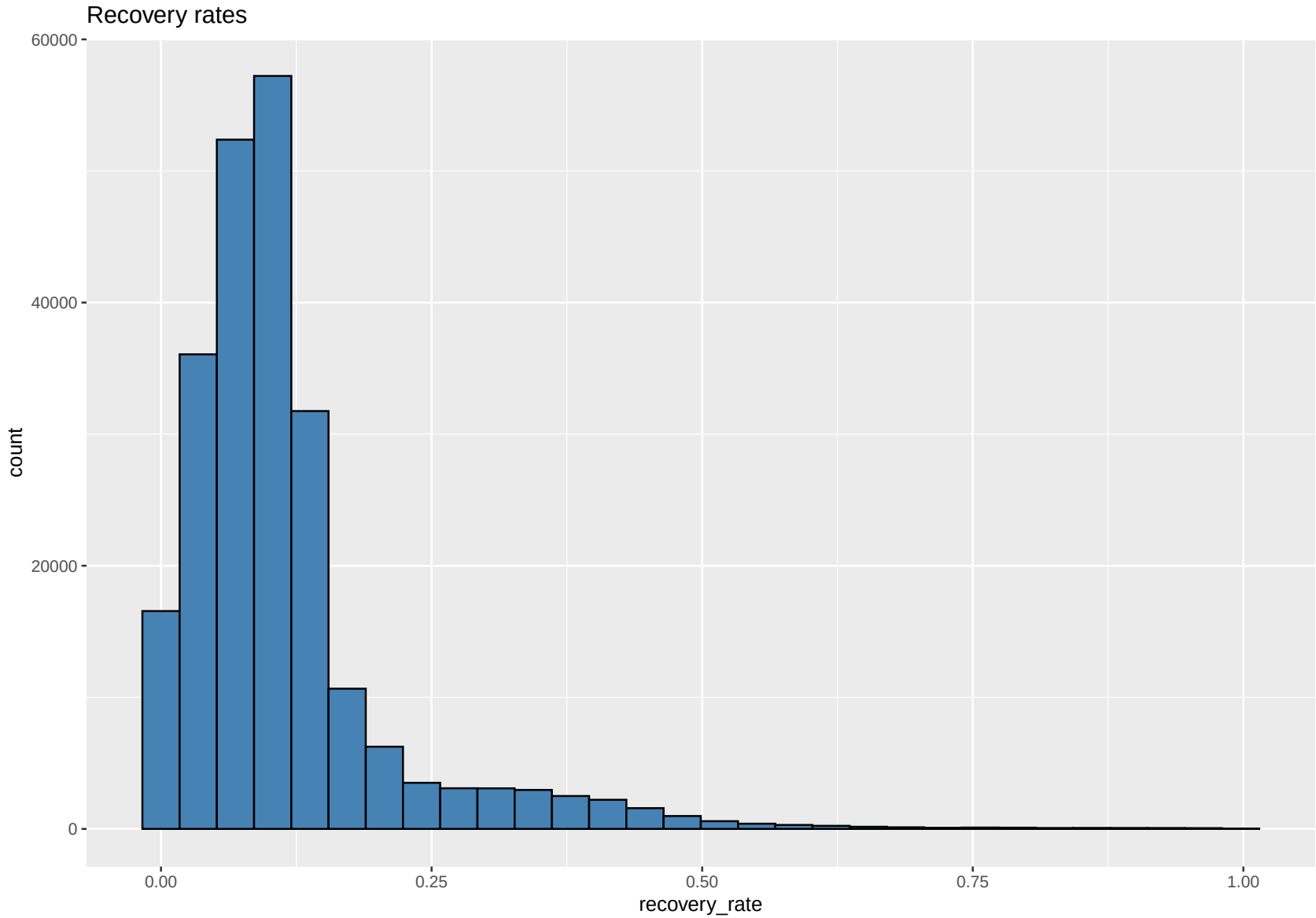
$$rt_i = \frac{recoveries_i}{funded\_amount_i}$$

i - loan i.

When modeling the rt in this dataset we need to take into account only those loans who were charged off.

LGD Y variable distribution

```
FALSE `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

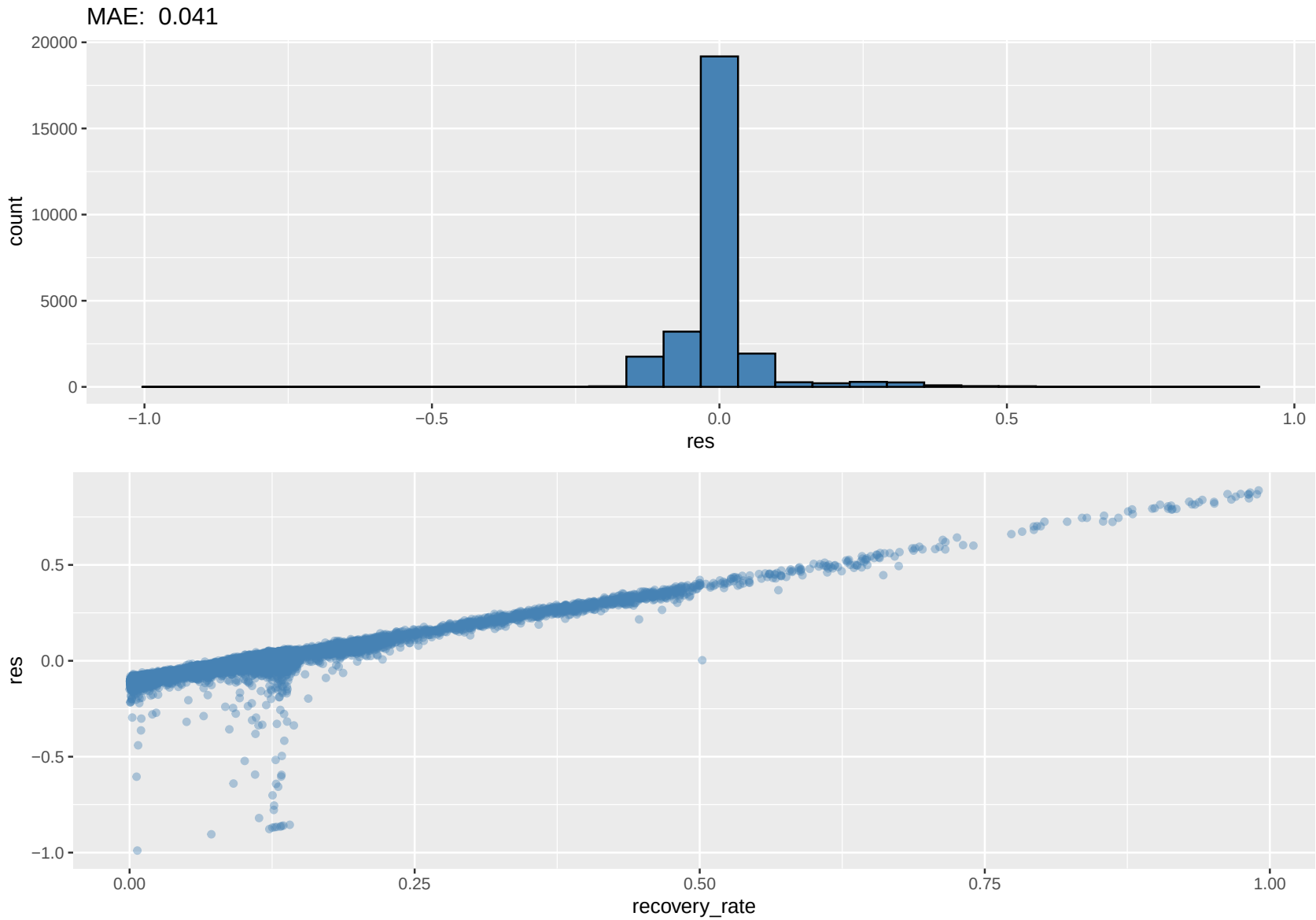


The model for LGD

feature	coef
(Intercept)	-2.6088228
term60 months	0.1145067
pub_recg>0	0.0692237
annual_inc	0.0000019
dti	0.0075368
month_diff	0.0001042
int_rate	0.0007019
loan_amnt	-0.0000012
gradeB	0.0369728
gradeC	0.1302691
gradeD	0.1570964
gradeE	0.2032814
gradeF	0.2153040
gradeG	0.2705374
addr_stategME:CO:WV:WY:WA:SC:KS	0.0163051
addr_stategND:NM:SD:TN:LA:NV	0.1295943
addr_stategOK:AL:AR:MS:NE	0.1492610
addr_stategother	0.0970180
purposegother	-0.0042949
purposegrenewable_energy:educ:moving	0.0116414
purposegsmall_business	-0.1964943
purposegwedding:car	-0.1431135

LGD residuals in the test set

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



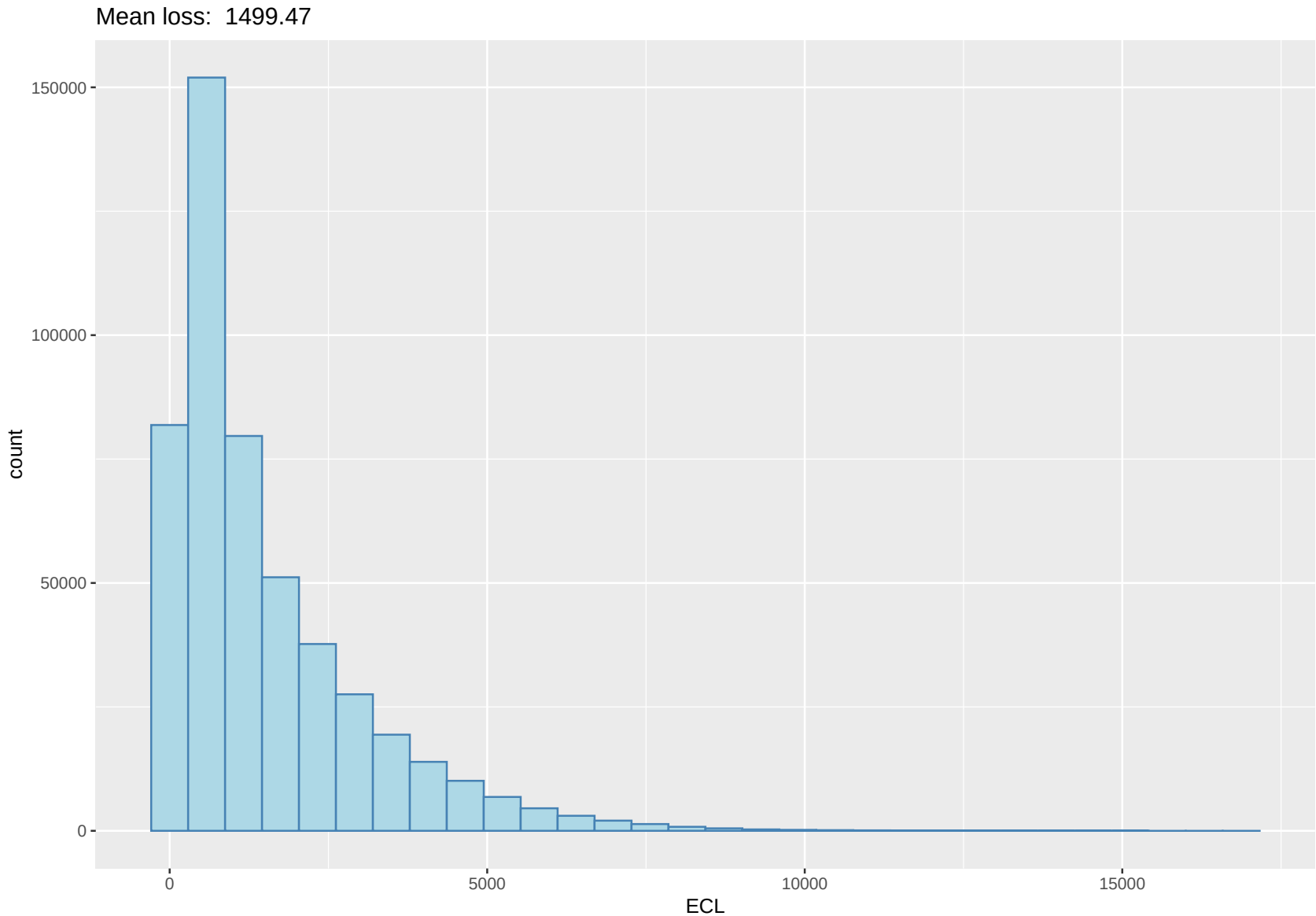
Combining all the parts together

The final ECL estimate using the dependent variables we estimated:

$$ECL = (1 - P(Y = 1)) * (1 - recovery\_rate) * CCF * loan\_amount$$

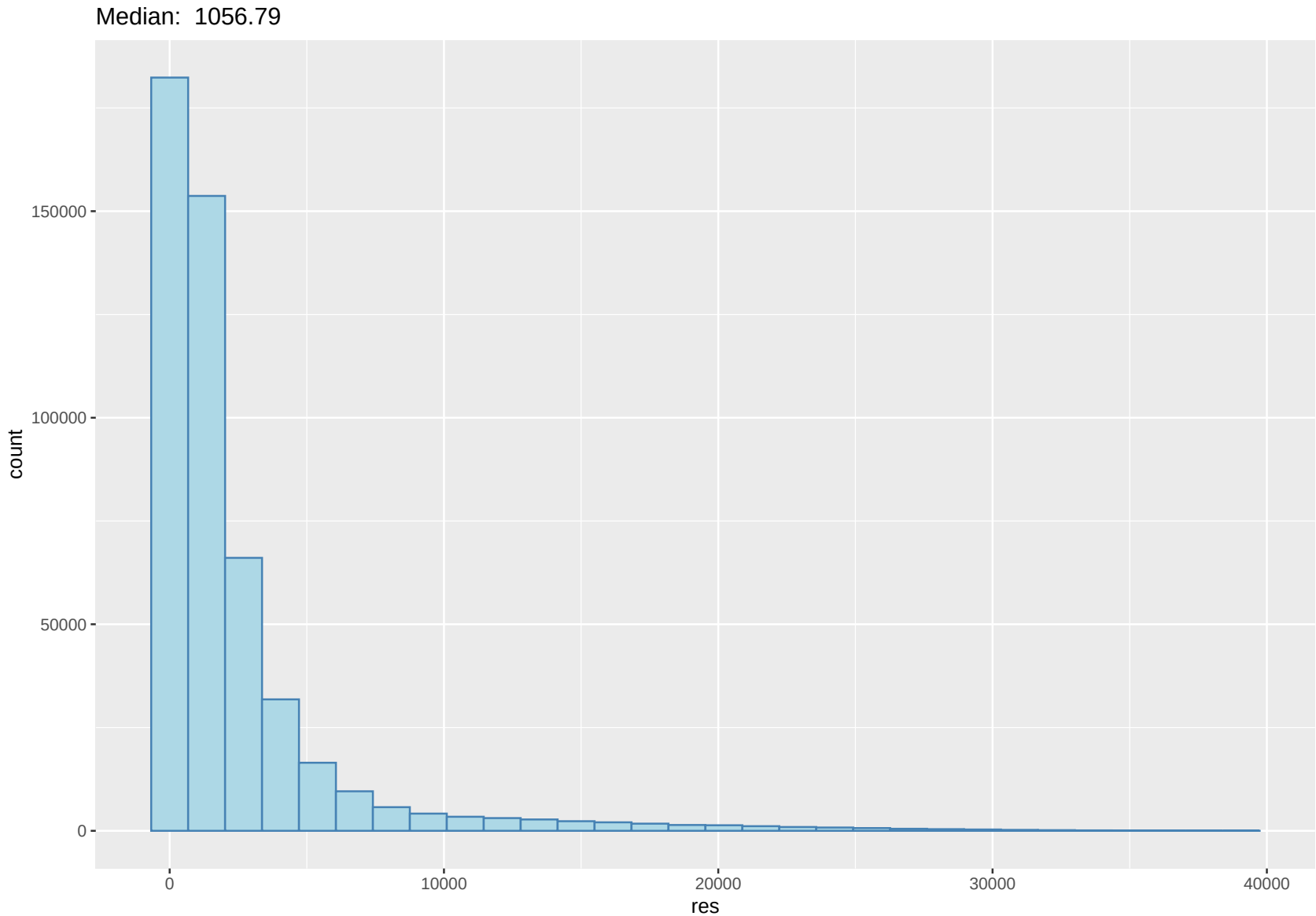
Distribution of expected credit loss

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Absolute residual distribution

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Final results

total_loss	ECL	total_funded	amnt_default	amnt_default_fc
676.0985	739.6699	7900.658	0.085575	0.0936213