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

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

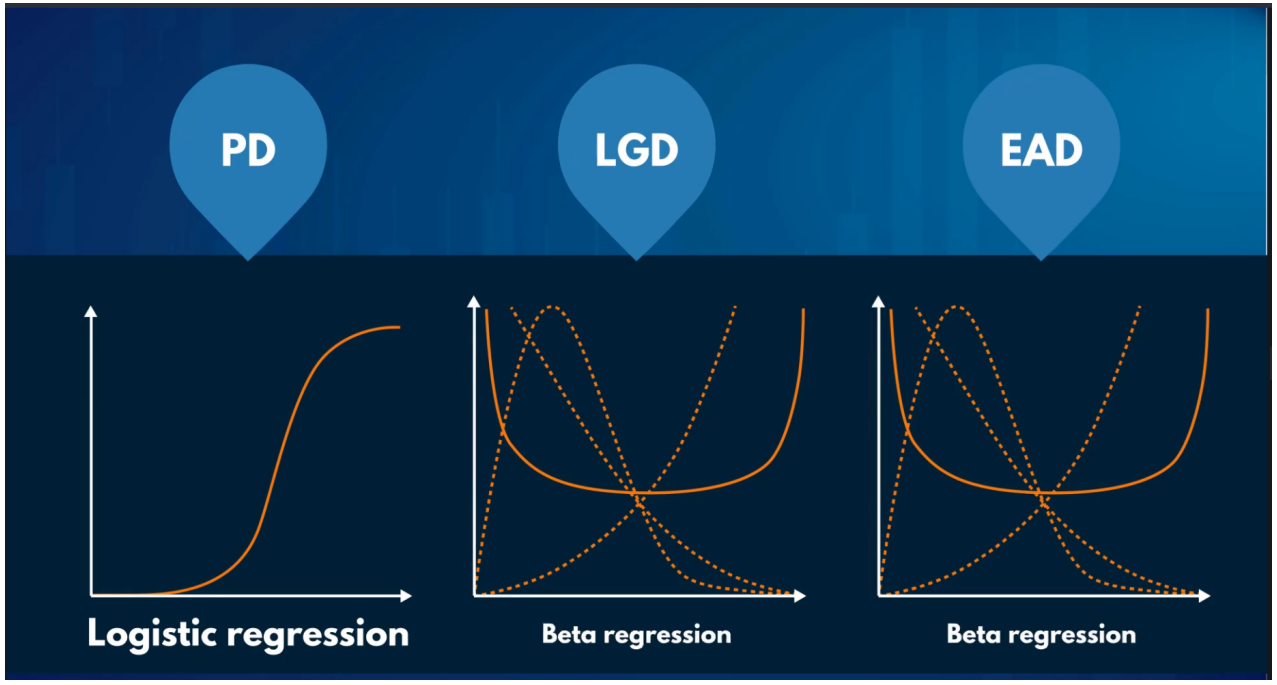
CL - Credit losses

PD - Probability of default

LGD - Losses given default

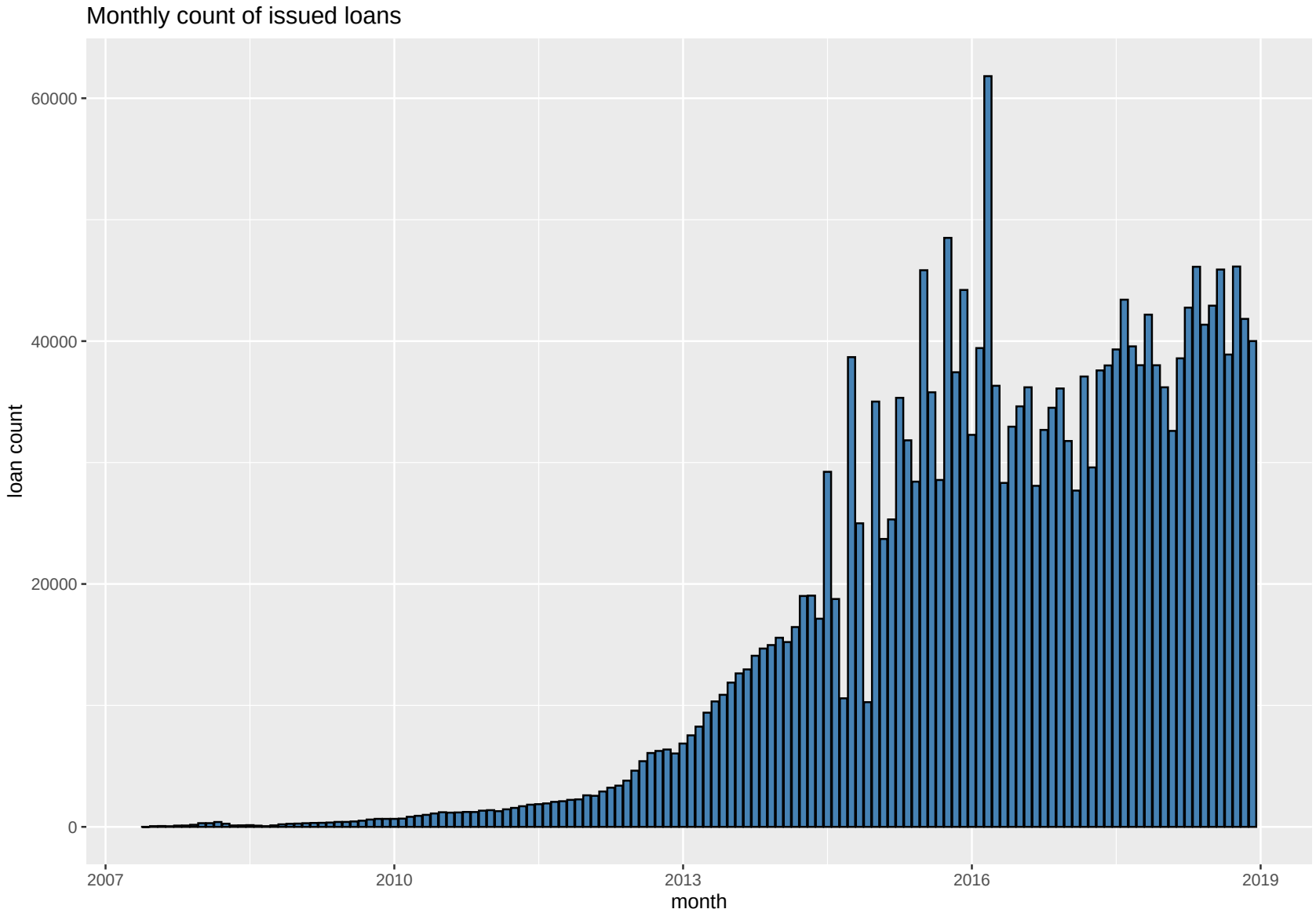
EAD - Exposure at default

Expected credit loss modeling in a nutshell



Exploring the data

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



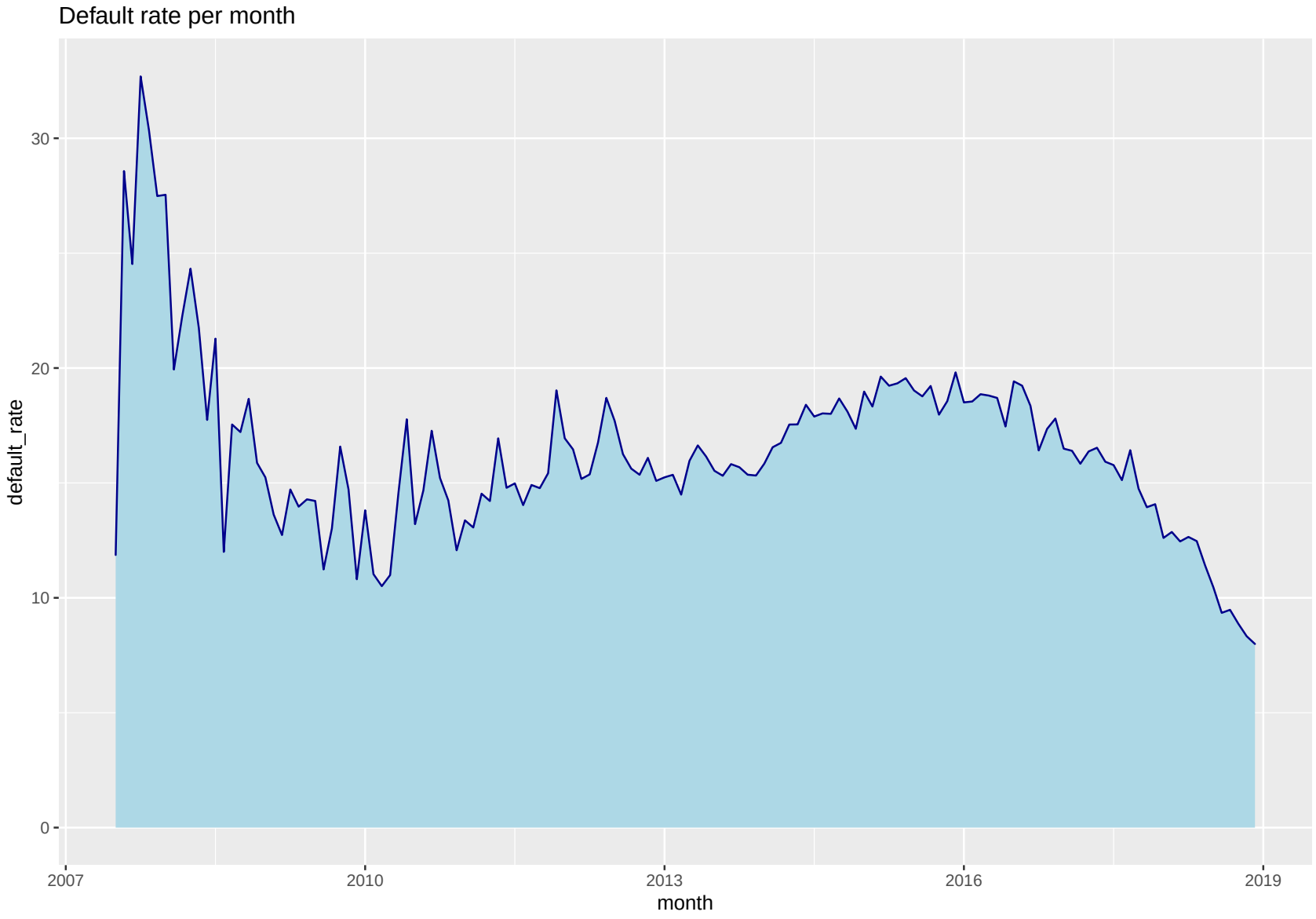
Loan status

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

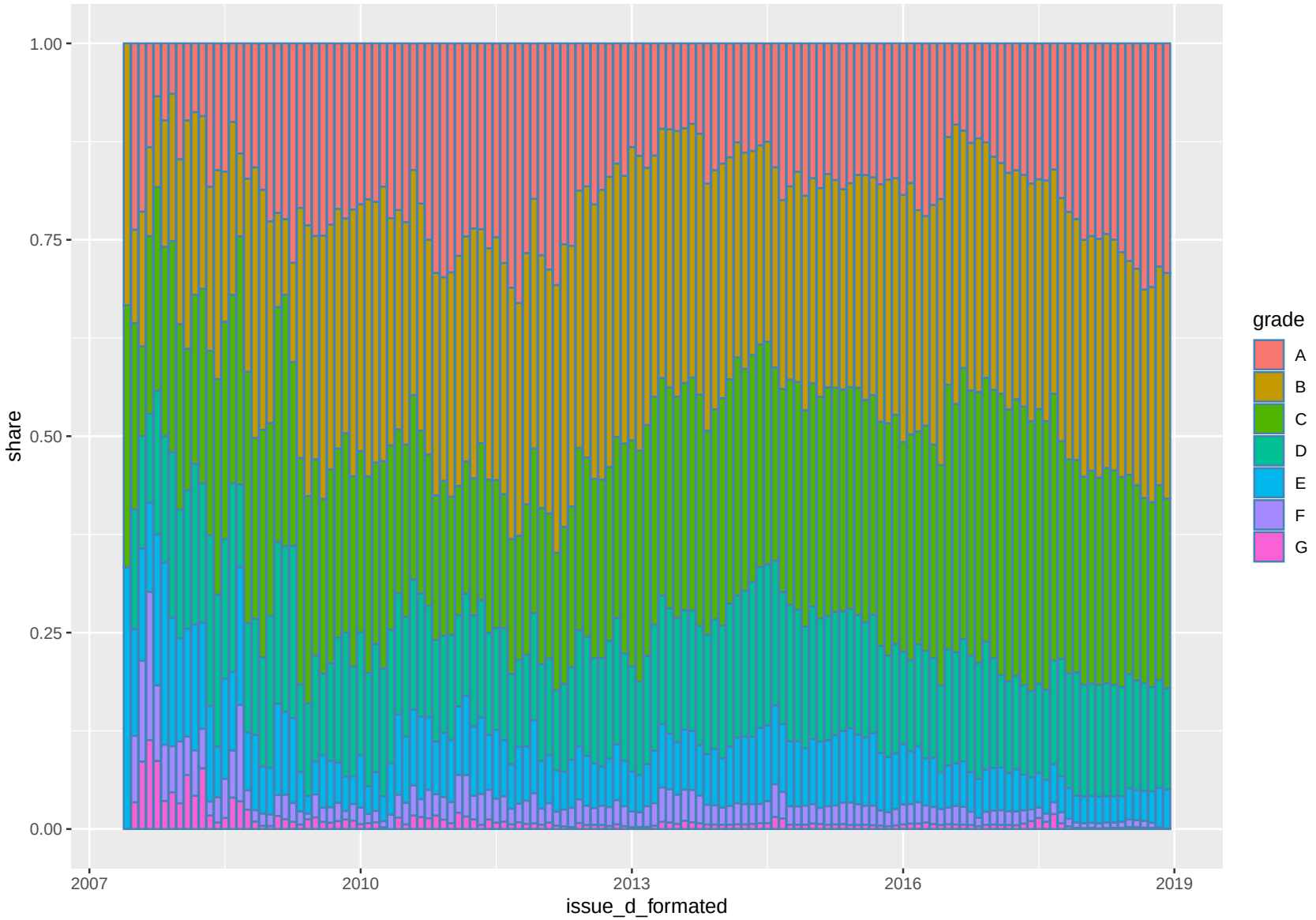
loan_status	N
Default	351
Does not meet the credit policy. Status:Charged Off	757
Late (16-30 days)	1518
Does not meet the credit policy. Status:Fully Paid	1957
In Grace Period	5627
Late (31-120 days)	9499
Charged Off	347810
Current	458491
Fully Paid	1427932

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.

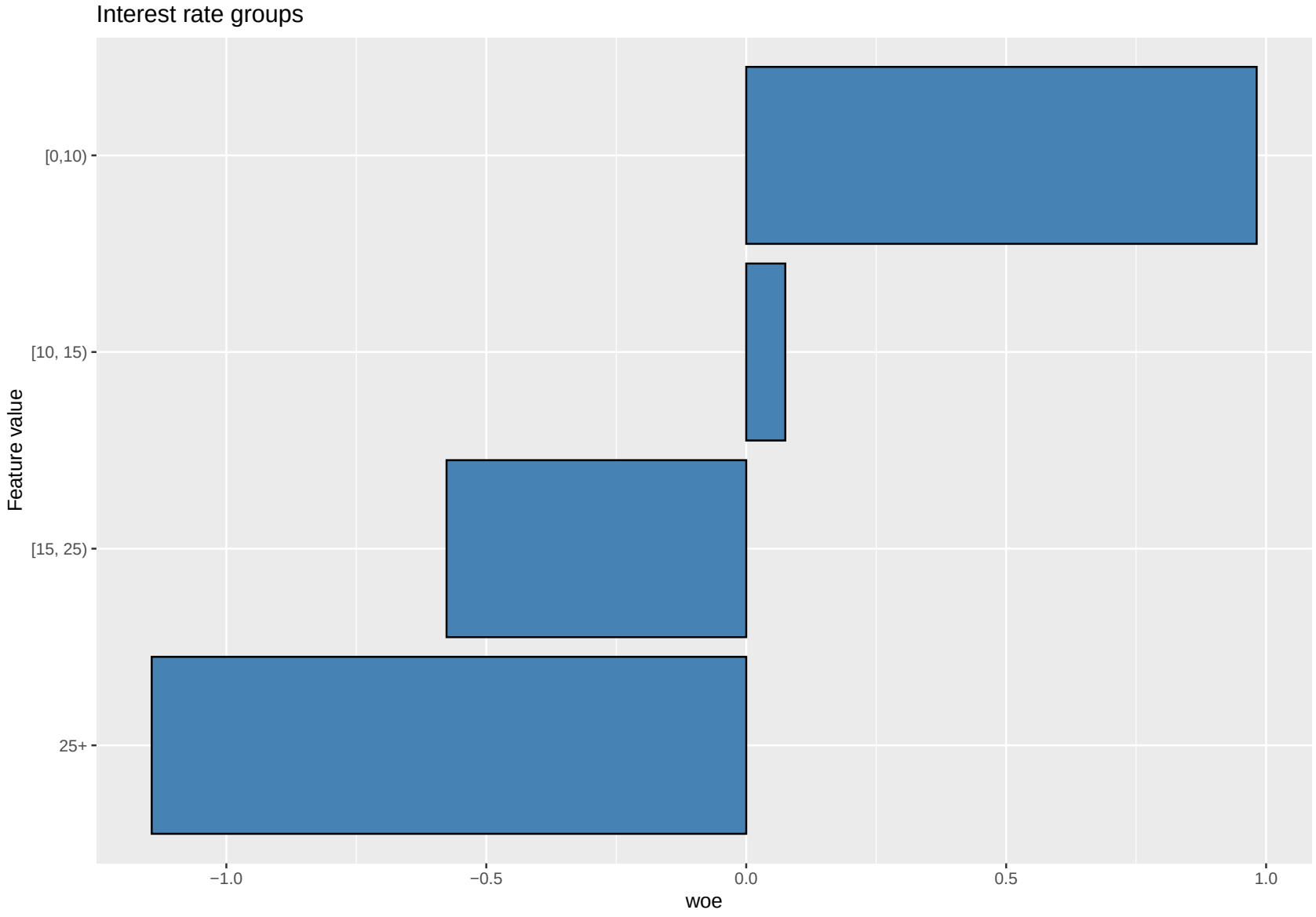
To infer what categorical feature values are the most influential in determining bad loan from a good loan we can use the weight of evidence (WOE for short) criteria. For a feature i and the feature level j the $WOE_{i,j}$ is calculated with the following formula:

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

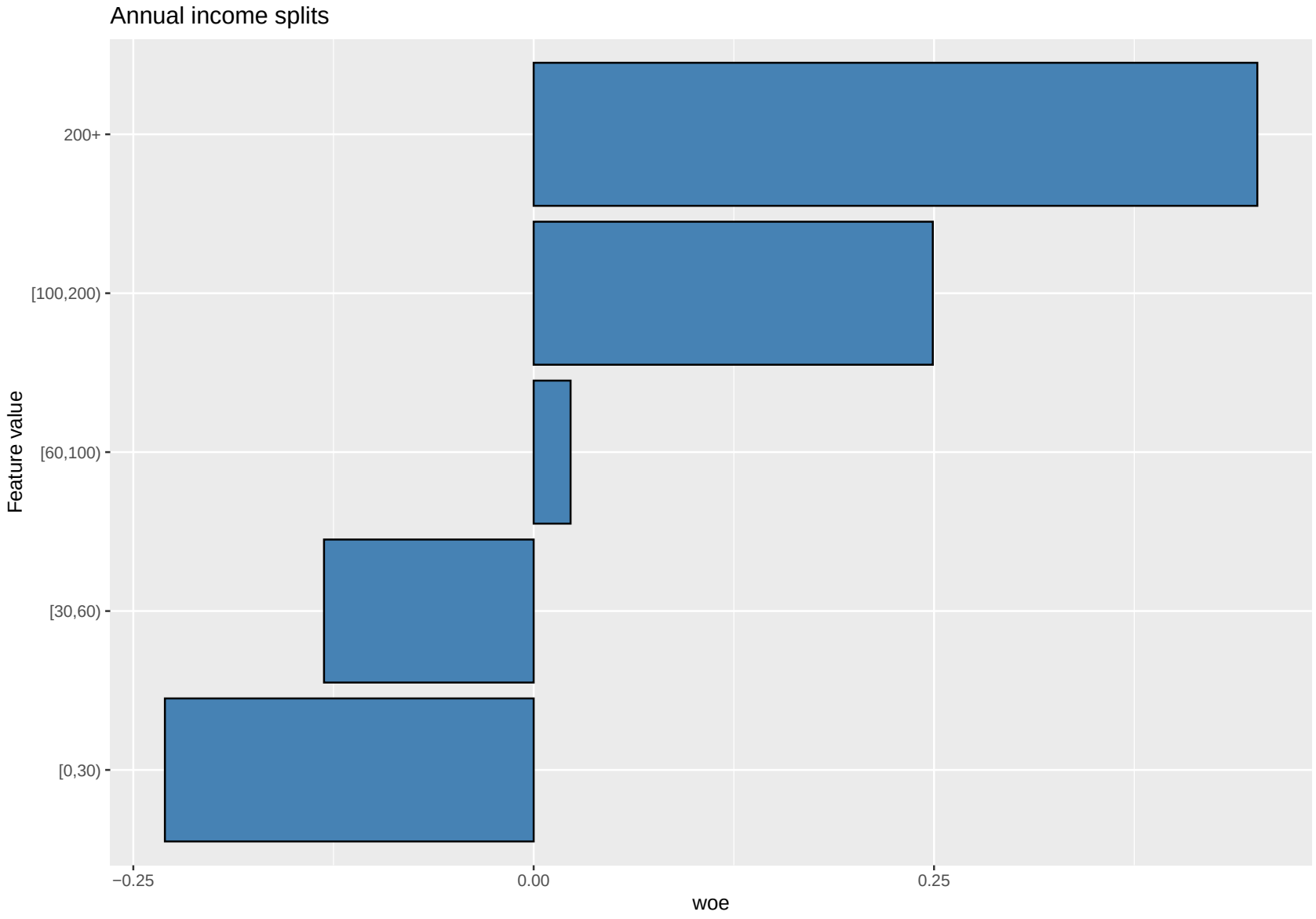
WOE example

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

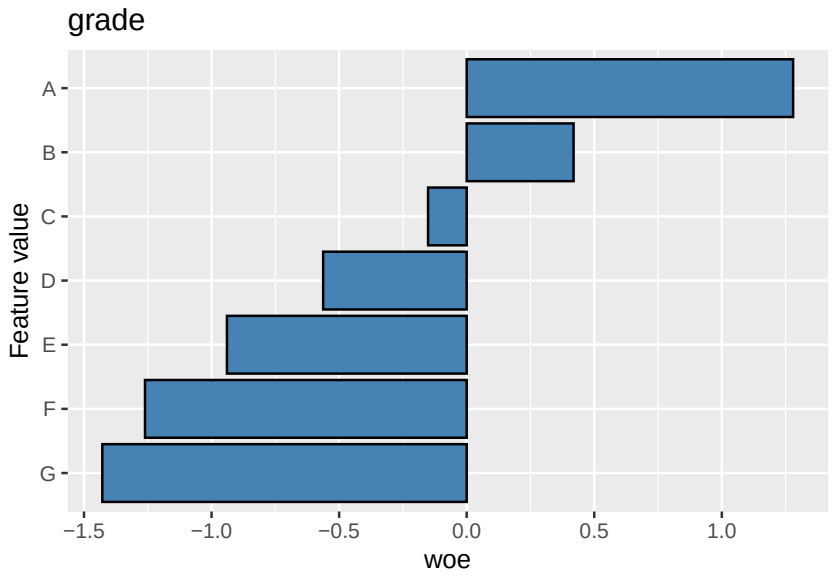
WOE for interest rates

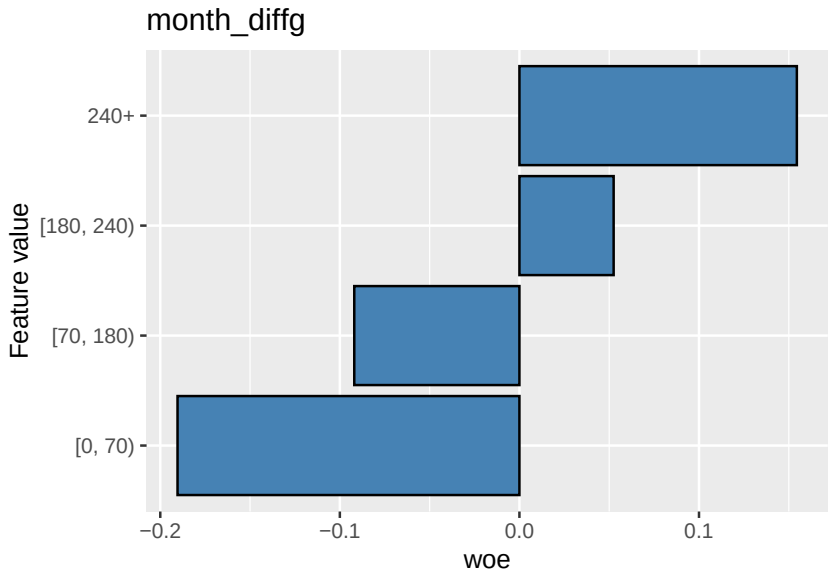


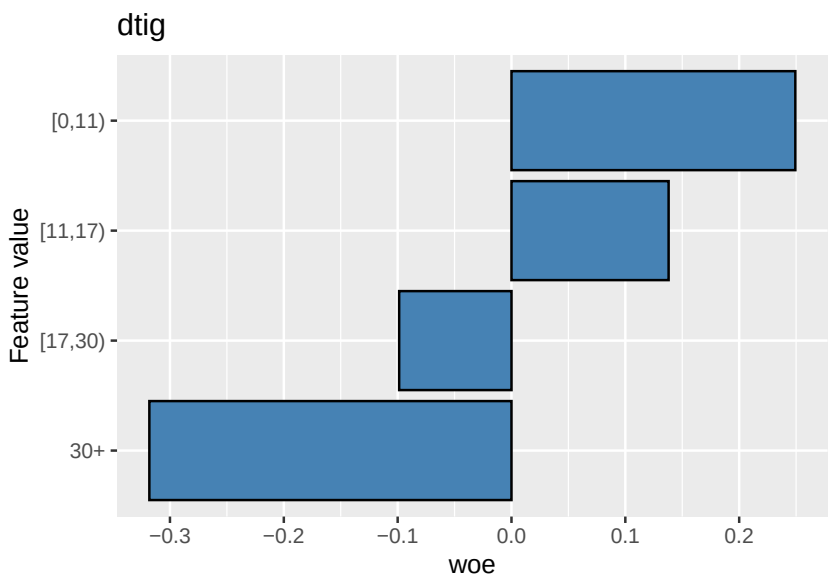
Annual income

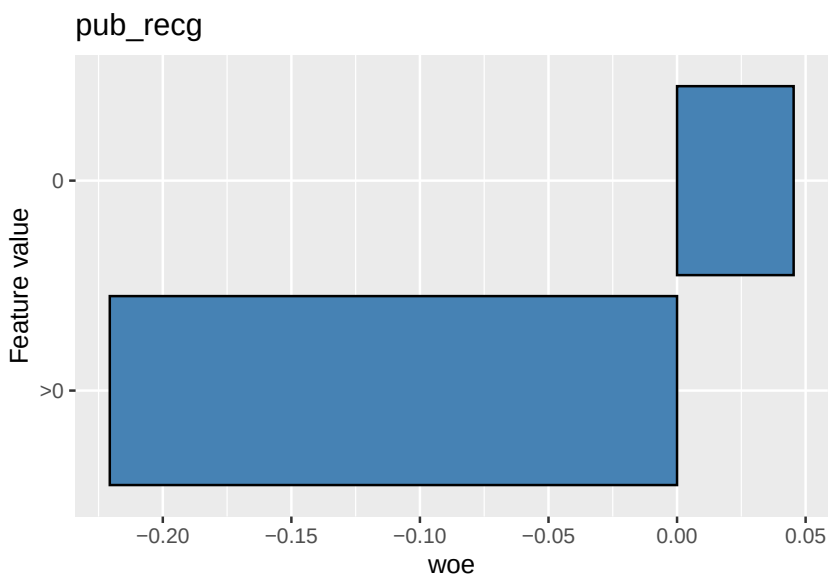


Some more features









Final variable list used for PD

x
term
pub_rec
annual_inc
dti
month_diff
emp_length
int_rate
loan_amnt
grade
purpose
addr_state

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.

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

Bernouli distribution

Because our Y variable is either 1 or 0 (good or bad loan) we should talk about the Bernoulli distribution.

$$Y \in \{0, 1\}$$

$$P(Y = 1) = p = 1 - P(Y = 0)$$

The distribution is:

$$f(y, p) = p^y (1 - p)^{1-y}$$

Bernouli and logistic regression maximum likelihood

The likelihood function (likelihood) measures the goodness of fit of a statistical model to a sample of data for **given values of the unknown parameters**.

$$l(\theta|x) = p_\theta(x) = P_\theta(X = x)$$

The maximum likelihood is a method to maximize the likelihood function in terms of the parameter θ such that the observed data is most probable given the assumption about the distribution of the data.

Bernoulli:

$$l(p) = \Pi_{i=1}^{n} p^{y_i}(1-p)^{1-y_i}$$

Logistic:

$$l(\beta) = \Pi_{i=1}^{n} \left[logistic(z)^{y_i}((1 - logistic(z))^{1-y_i}) \right]$$

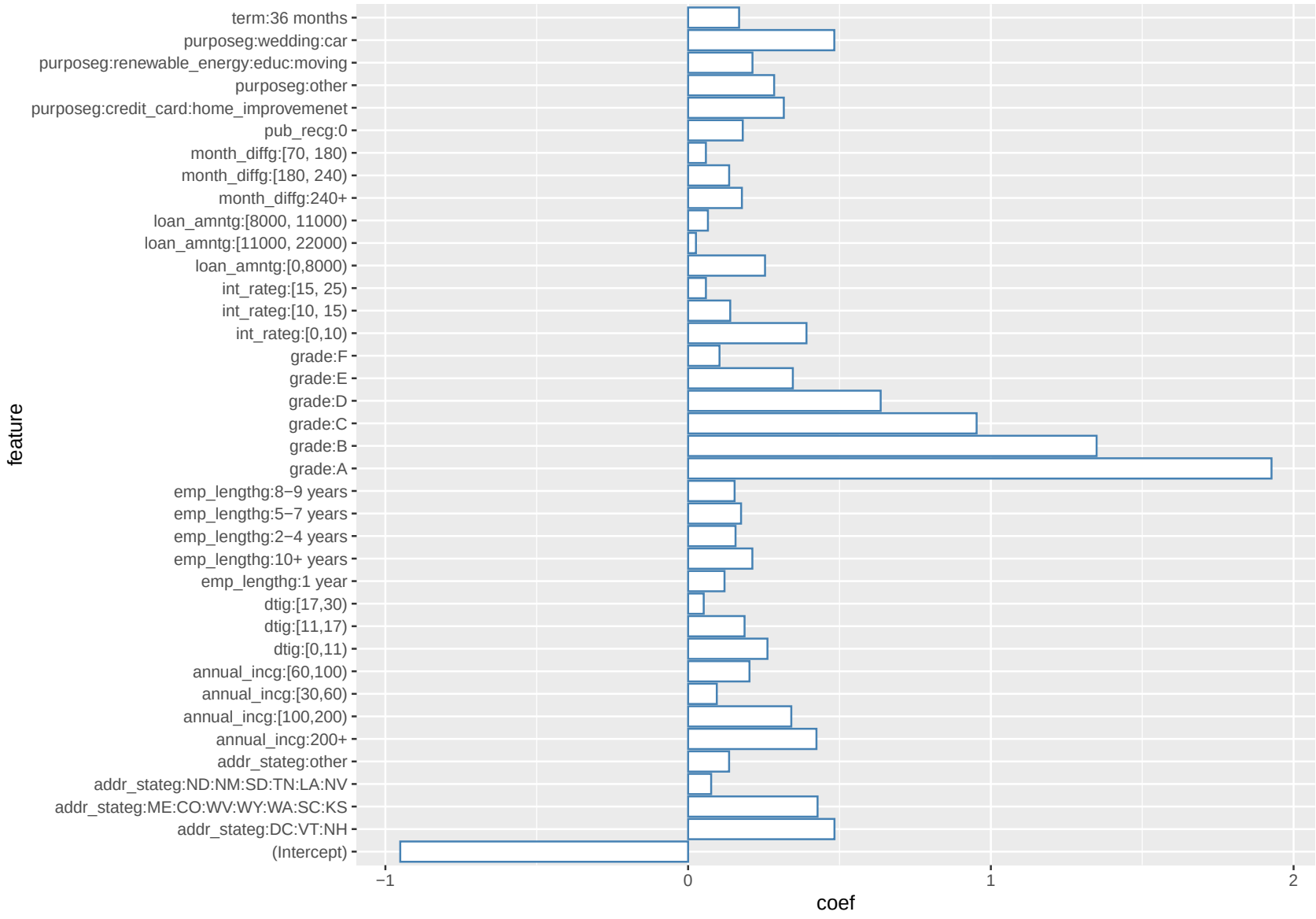
Estimating the coefficients

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

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

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.

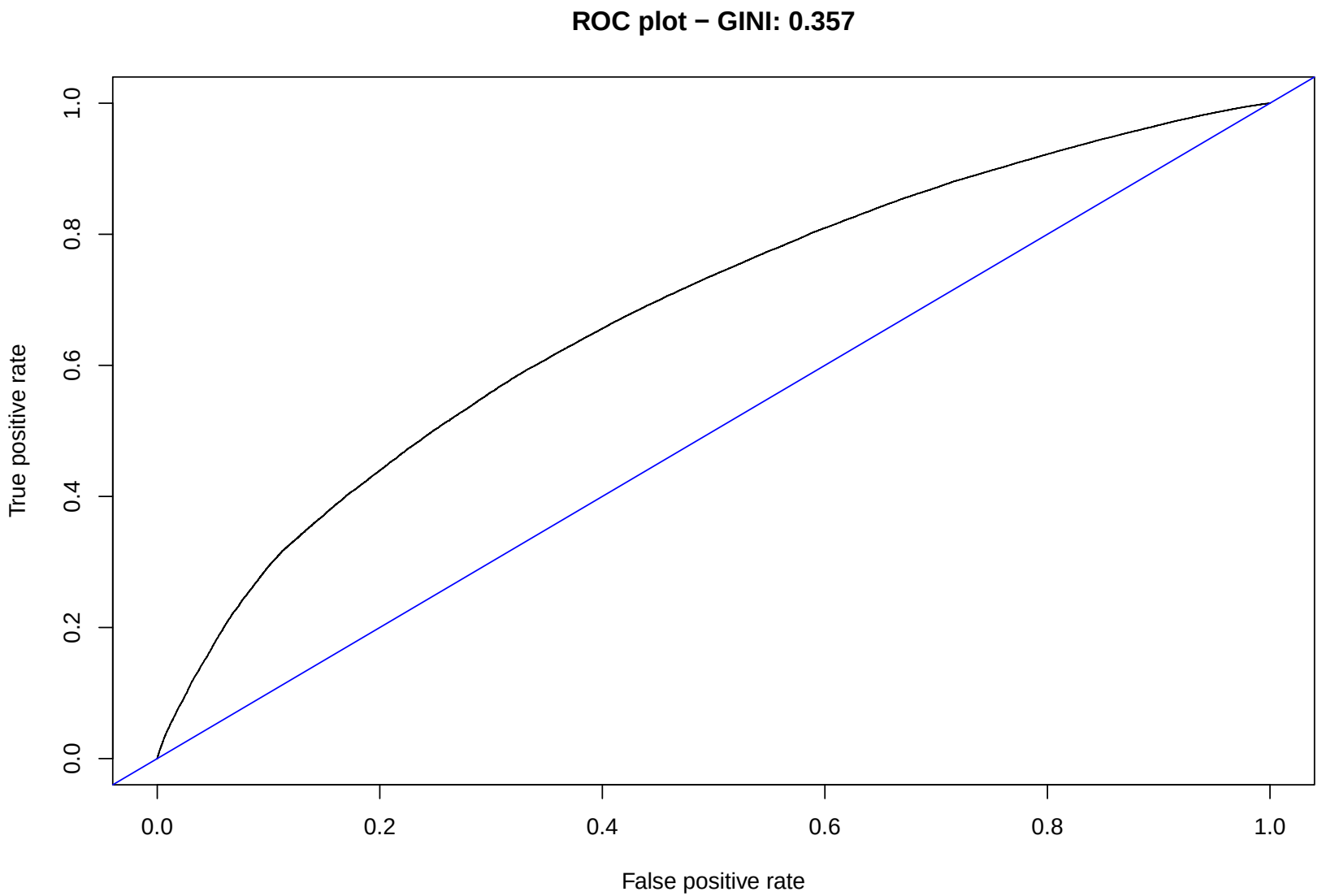
The model for PD



Results on the test set

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

Setting direction: controls < cases
```



Converting to a scorecard

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

origFeature	min_coef	max_coef
(Intercept)	-0.9508579	-0.9508579
addr_stateg	0.0000000	0.4834281
annual_incg	0.0000000	0.4239534
dtig	0.0000000	0.2620422
emp_lengthg	0.0000000	0.2122388
grade	0.0000000	1.9270668
int_rateg	0.0000000	0.3909776
loan_amntg	0.0000000	0.2540543
month_diffg	0.0000000	0.1776638
pub_recg	0.0000000	0.1804778
purposeg	0.0000000	0.4829255
term	0.0000000	0.1684319

Formulas for converting

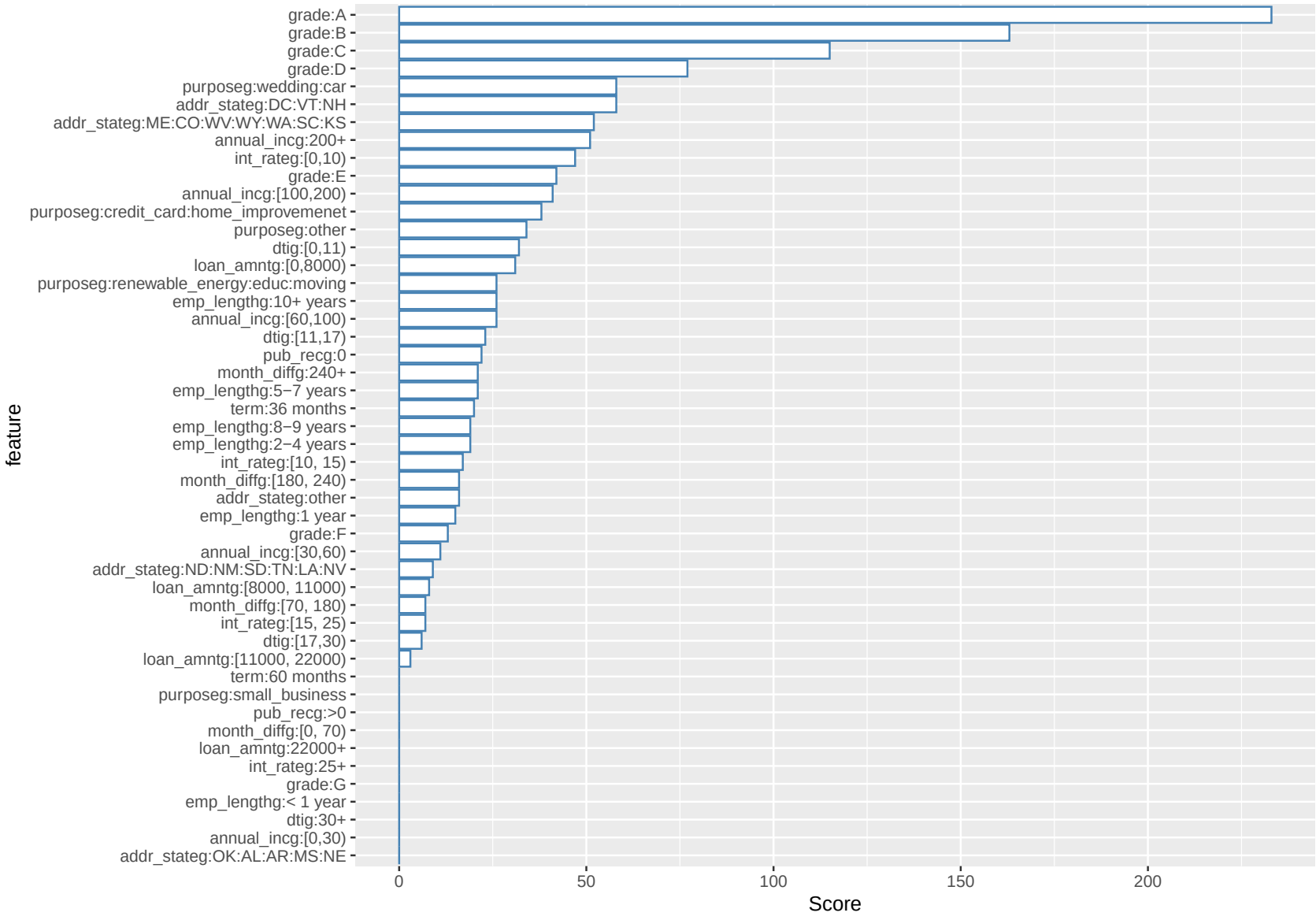
To convert the coefficient to a score, we need to follow the following formula:

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

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):

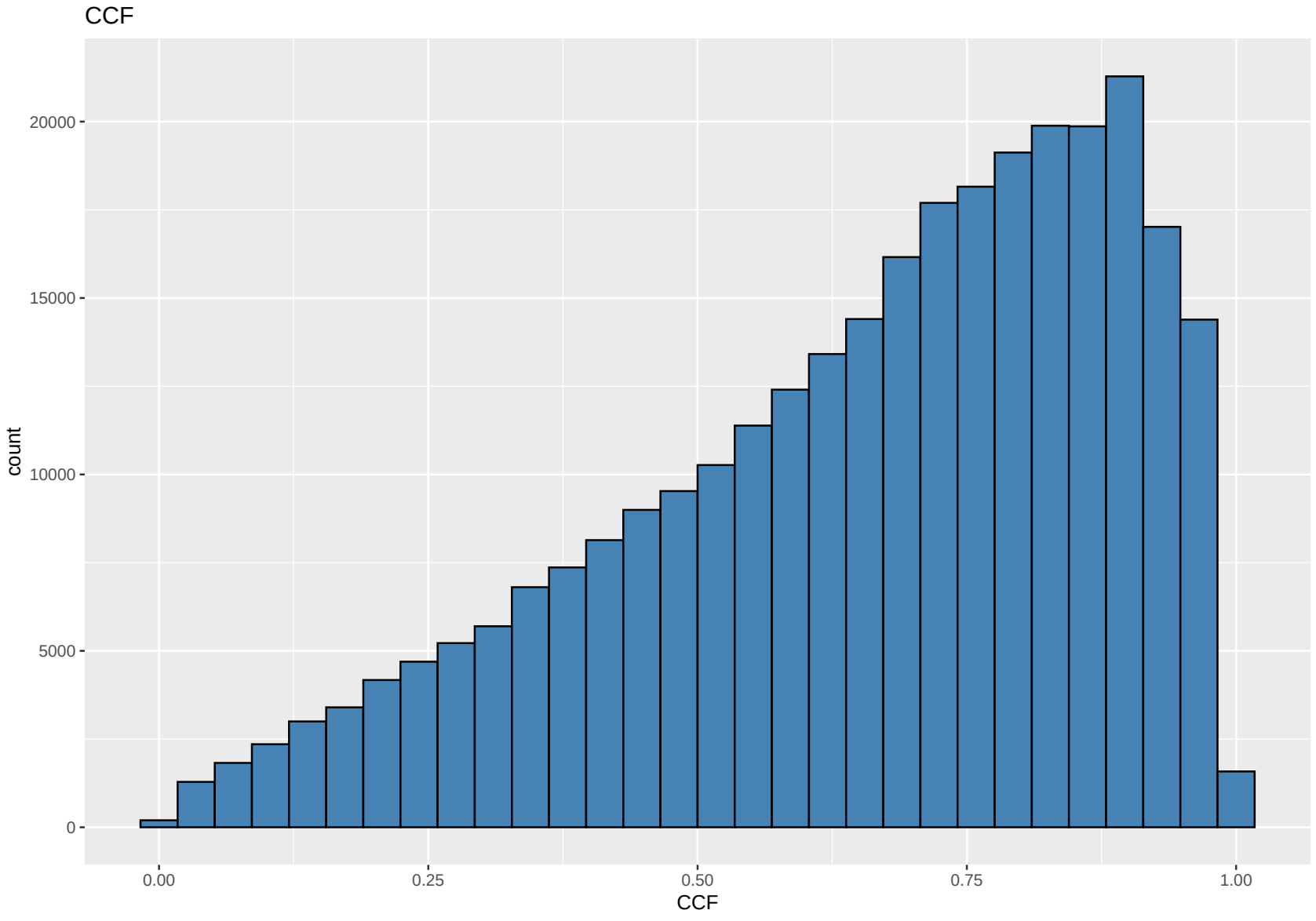
$$CCF_i = \frac{funded\_amount_i - received\_payments_i}{funded\_amount_i}$$

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`.
```



Beta regression

When the dependant variable is a ratio and distributed similar to a beta distribution, we can use the beta regression method.

The density for a beta distributed variable can be defined as:

$$f(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}$$

$0 < y < 1$

$0 < \mu < 1$

$\phi > 0$

$$\mu = g^{-1}(X\beta)$$

Beta regression - estimation

Same as in logistic regression - maximizing likelihood:

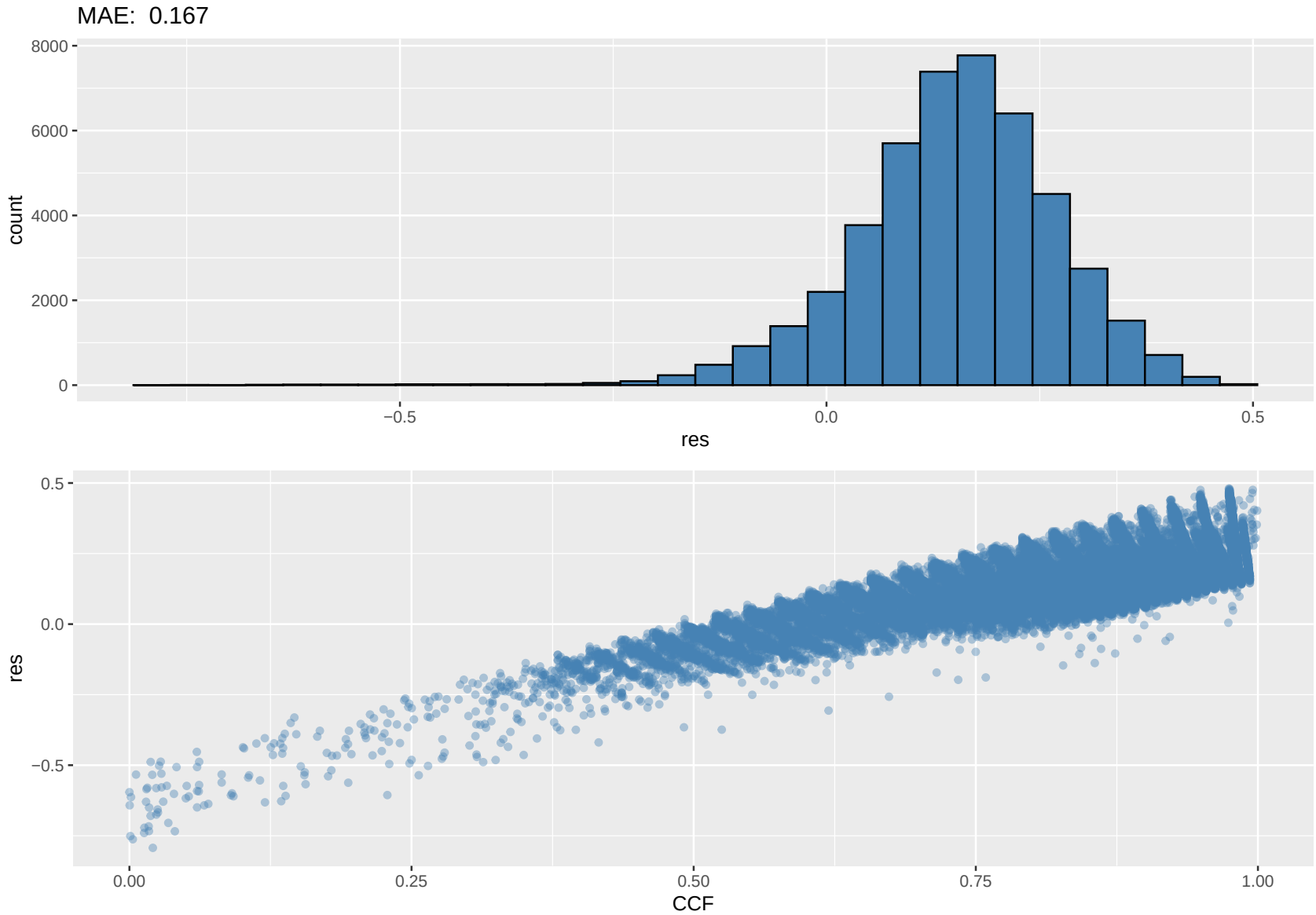
$$l(\mu, \phi) = log(\Gamma(\phi)) - log(\Gamma(\phi\mu)) - log(\Gamma((1-\mu)\phi)) + (\mu\phi - 1)log(y) + ((1-\mu)\phi - 1)log(1-y)$$

The model for EAD

V1	feature	coef	exp_beta
1	(Intercept)	-0.4558294	0.6339219
2	term60 months	0.3989245	1.4902211
3	pub_recg0	0.0276405	1.0280260
4	annual_inc	0.0000007	1.0000007
5	dti	0.0020766	1.0020788
6	month_diff	-0.0001695	0.9998306
7	int_rate	0.0531326	1.0545695
8	loan_amnt	0.0000011	1.0000011
9	gradeB	-0.0525258	0.9488298
10	gradeC	-0.0582123	0.9434496
11	gradeD	-0.1063639	0.8990974
12	gradeE	-0.1555601	0.8559356
13	gradeF	-0.1909953	0.8261365
14	gradeG	-0.1787634	0.8363037
15	addr_stategME:CO:WV:WY:WA:SC:KS	0.0493218	1.0505584
16	addr_stategND:NM:SD:TN:LA:NV	0.0911131	1.0953929
17	addr_stategOK:AL:AR:MS:NE	0.0575705	1.0592599
18	addr_stategother	0.0637639	1.0658408
19	purposegother	0.0244828	1.0247850
20	purposegrenewable_energy:educ:moving	0.1733795	1.1893174
21	purposegsmall_business	0.1115421	1.1180008
22	purposegwedding:car	0.0210354	1.0212582

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:

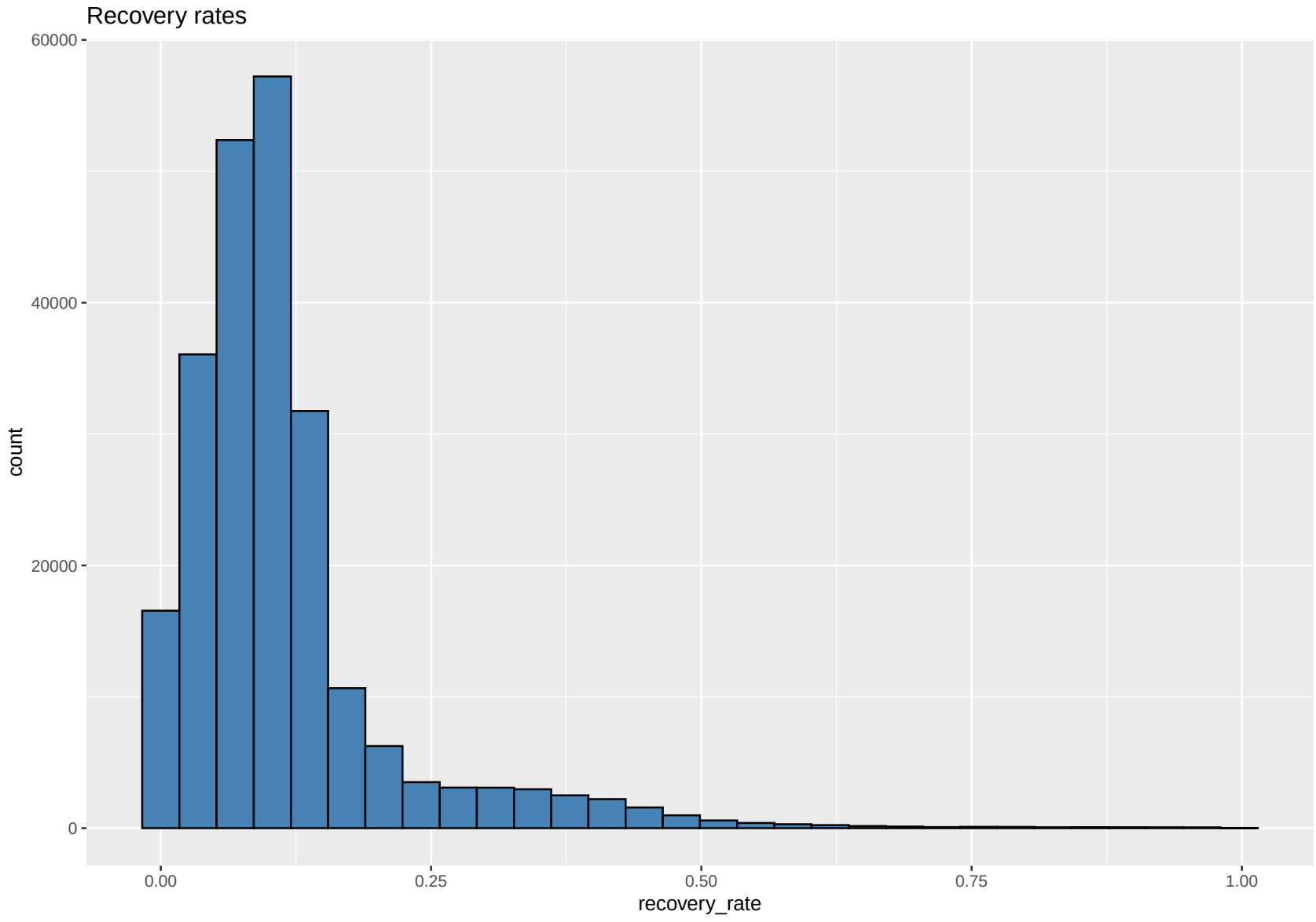
$$rt_i = \frac{recoveries_i}{funded\_amount_i}$$

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

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

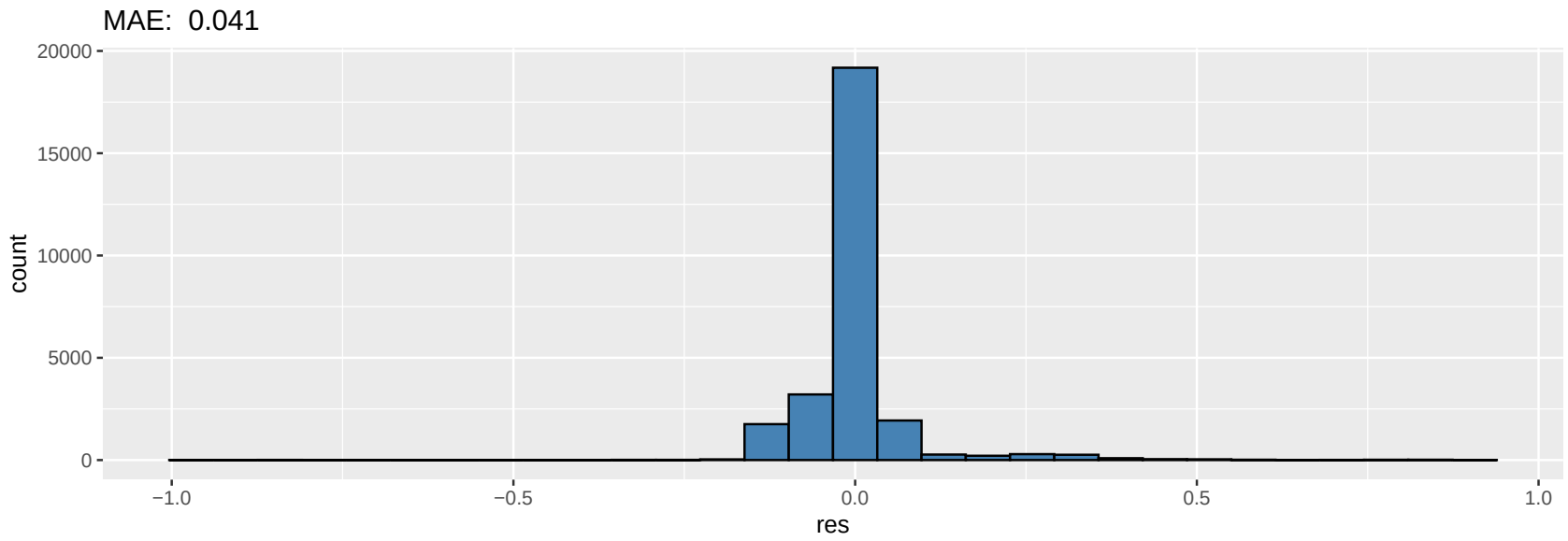


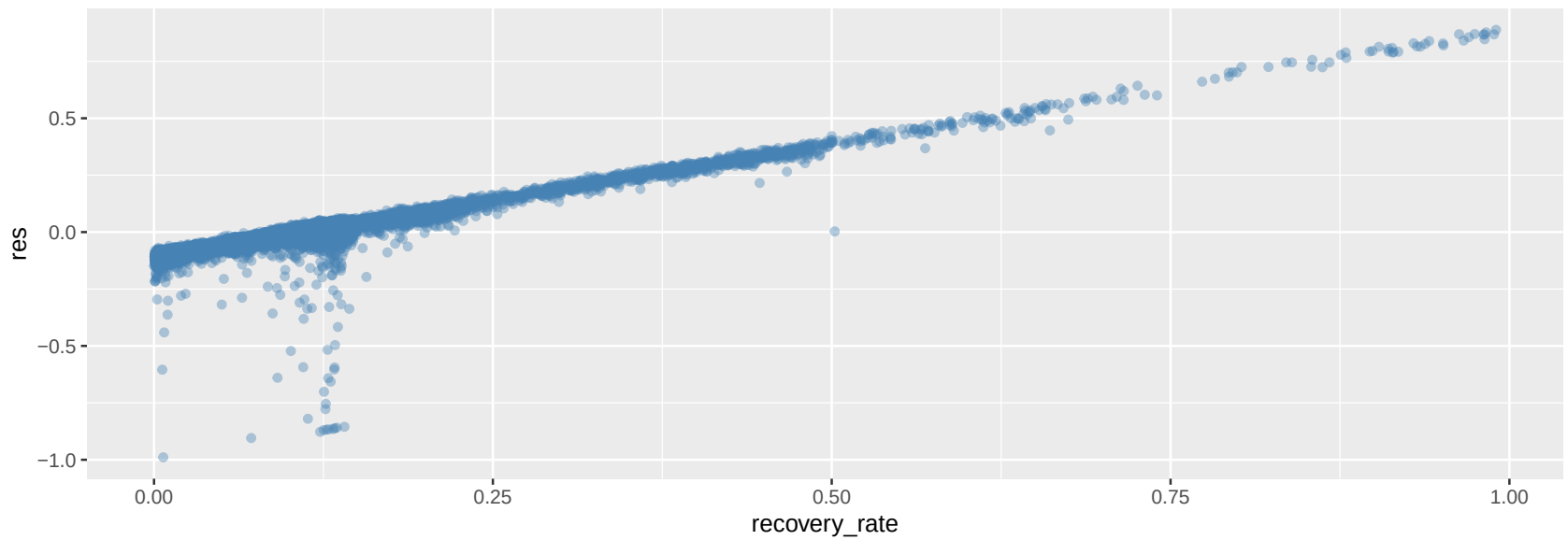
The model for LGD

feature	coef
(Intercept)	-2.5395991
term60 months	0.1145067
pub_recg0	-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`.
```





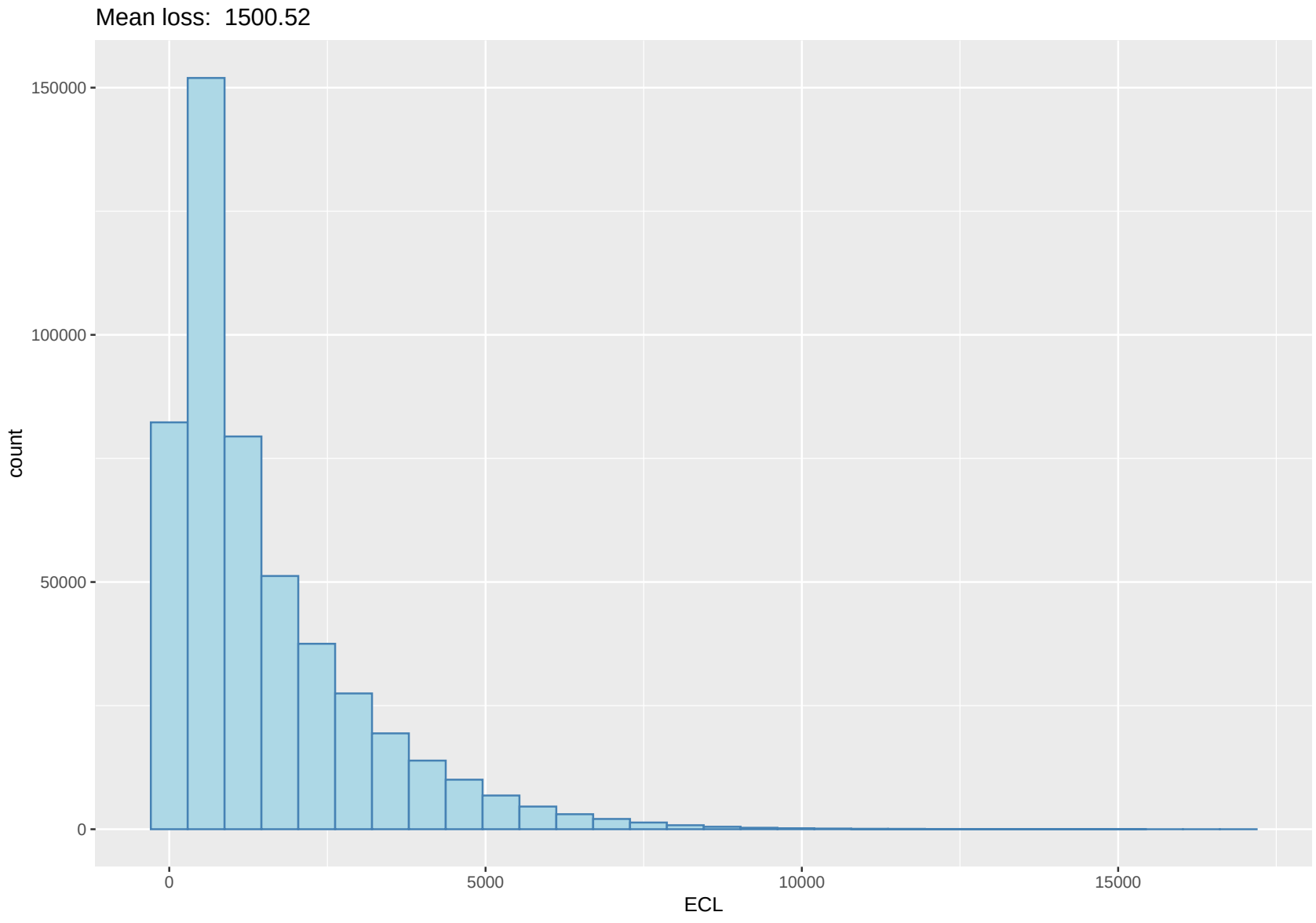
Combining all the parts together

The final ECL estimate using the dependent variables we estimated:

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

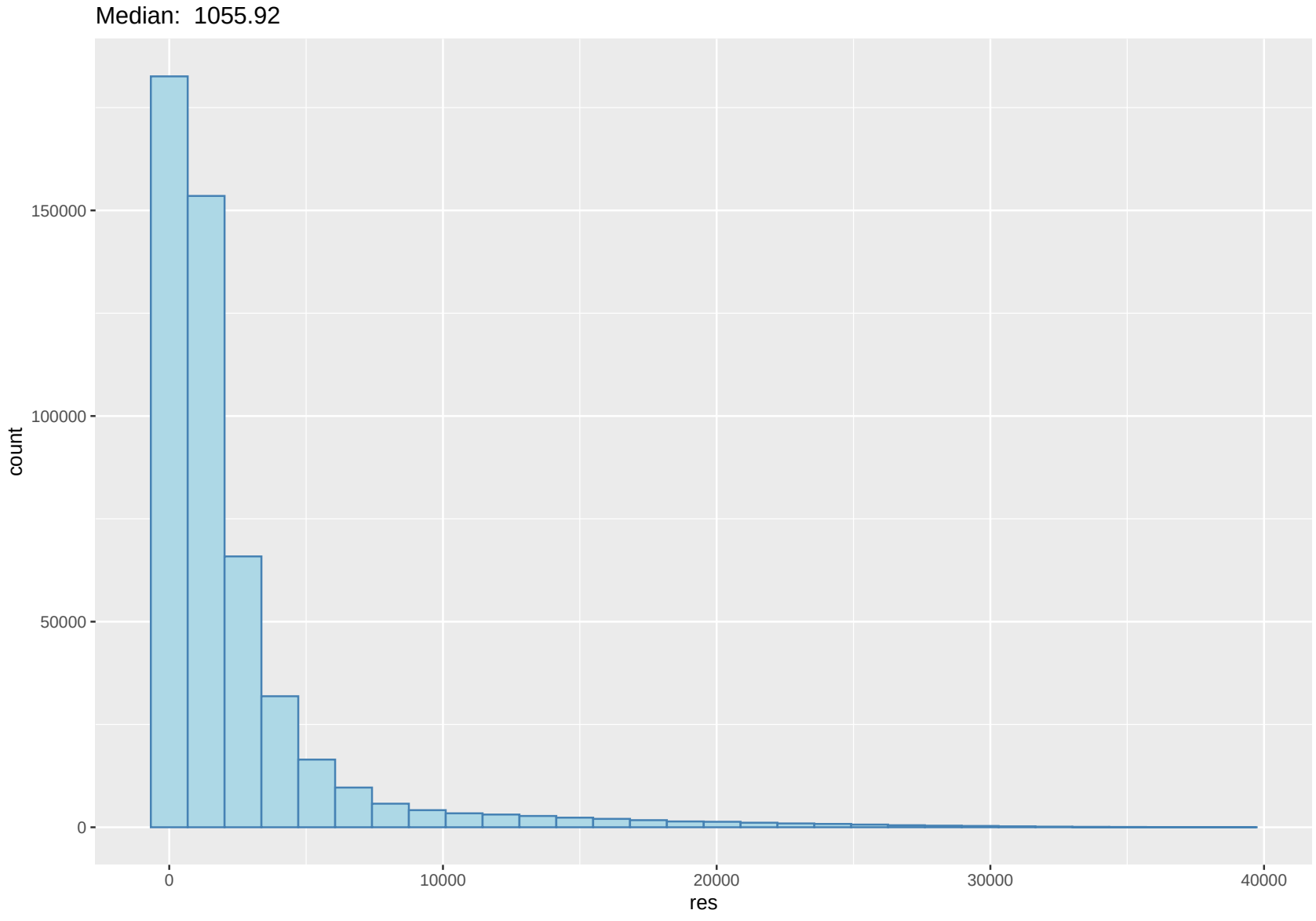
Distribution of expected credit loss

```
`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	740.192	7900.658	0.085575	0.0936874