Μοντέλα Πιστωτικού Κινδύνου

Δρ. Σωτήριος Δ. Νικολόπουλος

Big Data & Analytics

Πανεπιστήμιο Πελοποννήσου

Τμήμα Λογιστικής & Χρηματοοικονομικών

s.nikolopoulos@go.uop.gr

Μοντέλα Πιστωτικού Κινδύνου

Μοντέλα Πιστωτικού Κινδύνου

Μοντέλα Πιστωτικού Κινδύνου

Τα μοντέλα πιστωτικού κινδύνου διαδραματίζουν κεντρικό ρόλο στο χρηματοοικονομικό τοπίο, παρέχοντας στα ιδρύματα ένα συστηματικό πλαίσιο για την αξιολόγηση και τη διαχείριση των κινδύνων που σχετίζονται με τις δανειοδοτικές και επενδυτικές δραστηριότητες.

Ένα κρίσιμο στοιχείο αυτών των μοντέλων είναι ο υπολογισμός της πιθανότητας πτώχευσης, η οποία ποσοτικοποιεί την πιθανότητα ένας δανειολήπτης να μην μπορέσει να ανταποκριθεί στις οφειλές του.

Η σημασία των μοντέλων πιστωτικού κινδύνου έγκειται στην ικανότητά τους να βελτιώνουν τις διαδικασίες λήψης αποφάσεων, προσφέροντας μια ολοκληρωμένη κατανόηση της πιθανής πιστοληπτικής ικανότητας ατόμων ή εταιρειών.

Μοντέλα Πιστωτικού Κινδύνου

Χρησιμοποιώντας στατιστικές μεθόδους, ιστορικά δεδομένα και διάφορους χρηματο-οικονομικούς δείκτες, αυτά τα μοντέλα επιτρέπουν στα χρηματοοικονομικά ιδρύματα να λαμβάνουν κατάλληλες αποφάσεις σχετικά με τη δανειοδότηση, την τιμολόγηση και τη διαχείριση χαρτοφυλακίου.

Ο υπολογισμός της πιθανότητας πτώχευσης δεν βοηθά μόνο στην αξιολόγηση κινδύνου αλλά υποστηρίζει επίσης τη συμμόρφωση με τους κανονισμούς, επιτρέποντας στα ιδρύματα να διατηρούν μια υγιή ισορροπία μεταξύ κινδύνου και απόδοσης στα χαρτοφυλάκια τους.

Σε ένα συνεχώς εξελισσόμενο χρηματοοικονομικό τοπίο, η συνεχής βελτίωση και εφαρμογή μοντέλων πιστωτικού κινδύνου είναι απαραίτητη για την ενίσχυση της σταθερότητας και της ανθεκτικότητας του χρηματοπιστωτικού συστήματος.

Ανάλυση Δανείου

```
# Logistic models
library(gmodels) # CrossTable()
library(ggplot2)
library(tidyr) # gather()
library(dplyr)
library(pROC) # roc
library(vembedr)
```

1.1 Εξερεύνηση της βάσης δεδομένων.

Ας φορτώσουμε τα δεδομένα που ονομάζονται **loan_data_ARF.rds** και στη συνέχεια ας κατανοήσουμε τη δομή τους πριν προχωρήσουμε σε οποιαδήποτε περαιτέρω ανάλυση.

```
dat <- readRDS("loan_data_ARF.rds")
str(dat)
```

Ανάλυση Δανείου

```
glimpse(dat)
Rows: 29,092
Columns: 10
$ loan_status     <int> 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, ...1
$ loan_amnt       <int> 5000, 2400, 10000, 5000, 3000, 12000, 9000, 3000, 10000, ...1
$ int_rate        <dbl> 10.65, 10.99, 13.49, 10.99, 10.99, 12.69, 13.49, 9.91, ...10.
$ grade           <fct> B, C, C, A, E, B, C, B, B, D, C, A, B, A, B, B, B, B, ...C
$ emp_length      <int> 10, 25, 13, 3, 9, 11, 0, 3, 3, 0, 4, 13, 1, 6, 17, 13, 5, ...
$ home_ownership  <fct> RENT, RENT, RENT, RENT, RENT, OWN, RENT, RENT, RENT, RENT...,
$ annual_inc      <dbl> 24000.00, 12252.00, 49200.00, 36000.00, 48000.00, ...75000.00
$ age             <int> 33, 31, 24, 39, 24, 28, 22, 22, 28, 22, 23, 27, 30, 24, ...29
$ sex             <fct> 0, 0, 0, 0, 1, 1, 1, 1, 0, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, ...1
$ region          <fct> E, E, S, S, N, N, N, N, W, E, E, S, E, E, E, E, N, N, E, ...E
```

Αυτά τα δεδομένα είναι μια τυπική βάση δεδομένων χρηματοοικονομικό ίδρυμα, όπως μια τράπεζα ή μια εταιρεία που χρησιμοποιεί πιστωτικά κανάλια για να πουλήσει τα προϊόντα ή τις υπηρεσίες της.

Ανάλυση Δανείου

Έχουμε 29.092 παρατηρήσεις 10 μεταβλητών.

Κάθε παρατήρηση αντιστοιχεί στα προσωπικά χαρακτηριστικά και τα χαρακτηριστικά δανείου ενός δανείου.

Μια σημαντική μεταβλητή, η **εξαρτημένη μεταβλητή**, είναι η **loan_status**.
Η τιμή 0 είναι ”χωρίς καθυστέρηση” και η τιμή 1 είναι ”καθυστέρηση αποπληρωμής”.

Μια καθυστέρηση αποπληρωμής συμβαίνει όταν ένας δανειολήπτης δεν μπορεί να κάνει έγκαιρες πληρωμές, χάνει πληρωμές ή αποφεύγει ή σταματά να κάνει πληρωμές για τους τόκους ή το κεφάλαιο που οφείλει.

Στη συνέχεια, ο ορισμός της καθυστέρησης εξαρτάται από τους στόχους της ανάλυσης, εδώ απλά διακρίνουμε μεταξύ καθυστέρησης ή μη καθυστέρησης. **Η μεταβλητή loan_status είναι δυαδική ή κατηγορική.**

Μας ενδιαφέρει να προβλέψουμε εάν μια νέα αίτηση θα έχει καθυστέρηση ή όχι.

Ανάλυση Δανείου

Είναι σαφές ότι το αρχείο **loan_data_ARF.rds** περιέχει ιστορικά δεδομένα, καθώς γνωρίζουμε με βεβαιότητα εάν το άτομο καθυστέρησε (1) ή όχι (0) το δάνειο.

Τα ιστορικά δεδομένα είναι χρήσιμα για να κατανοήσουμε καλύτερα πόσο πιθανό είναι να καθυστερήσει ένα άτομο σύμφωνα με τα προσωπικά του χαρακτηριστικά και τα χαρακτηριστικά του δανείου.

Τα ιστορικά δεδομένα είναι χρήσιμα για την εκπαίδευση των ποσοτικών μοντέλων για να κάνουμε προβλέψεις για νέους αιτούντες δάνειο, και ακόμη να αξιολογήσουμε την απόδοση των μοντέλων.

Άν το όνομα μιας μεταβλητής είναι πολύ μεγάλο μπορούμε να το αλλάζουμε με ένα μικρότερο όνομα που εξηγεί εξίσου καλά τον σκοπό της μεταβλητής.

Ας μετονομάσουμε μερικές μεταβλητές.

Ανάλυση Δανείου

```
old_names <- colnames(dat)
colnames(dat) <- c("loan_st", "l_amnt", "int", "grade", "emp_len",
"home", "income", "age", "sex", "region")
data.frame(old_names, "new_names" = colnames(dat))
```

```
          old_names new_names
1       loan_status   loan_st
2         loan_amnt    l_amnt
3          int_rate       int
4             grade     grade
5        emp_length   emp_len
6   home_ownership      home
7        annual_inc    income
8               age       age
9               sex       sex
10           region    region
```

Ανάλυση Δανείου

Μπορούμε να εξετάσουμε τις πληροφορίες με διάφορους τρόπους.

Για παράδειγμα, να δούμε τις πρώτες 10 γραμμές από τις 29.092.

```
head(dat, 10)
```

	loan_st	l_amnt	int	grade	emp_len	home	income	age	sex	region
1	0	5000	10.65	B	10	RENT	24000	33	0	E
2	0	2400	10.99	C	25	RENT	12252	31	0	E
3	0	10000	13.49	C	13	RENT	49200	24	0	S
4	0	5000	10.99	A	3	RENT	36000	39	0	S
5	0	3000	10.99	E	9	RENT	48000	24	1	N
6	0	12000	12.69	B	11	OWN	75000	28	1	N
7	1	9000	13.49	C	0	RENT	30000	22	1	N
8	0	3000	9.91	B	3	RENT	15000	22	1	N
9	1	10000	10.65	B	3	RENT	100000	28	0	W
10	0	1000	16.29	D	0	RENT	28000	22	1	E

Ανάλυση Δανείου

Αντί να δούμε τις λεπτομέρειες των πρώτων 10 γραμμών, μπορούμε να συνοψίσουμε τα δεδομένα ανά κατοικία χρησιμοποιώντας τη συνάρτηση CrossTable(). Με άλλα λόγια, μπορούμε να μετρήσουμε τον αριθμό των γυναικών και των αντρών

CrossTable(dat$home)

MORTGAGE	OTHER	OWN	RENT
12002	97	2301	14692
0.413	0.003	0.079	0.505

Ανάλυση Δανείου

Επίσης, μπορούμε να χρησιμοποιήσουμε τη συνάρτηση CrossTable() για να συνοψίσουμε δύο μεταβλητές.

Πιο συγκεκριμένα, αντί να μετράμε την κατοικία, μπορούμε να προσθέσουμε μια δεύτερη διάσταση όπως το loan_st (κατάσταση δανείου).

Αυτό μας επιτρέπει να δημιουργήσουμε καλύτερους (από την πλευρά της πληροφορίας) πίνακες.

Για παράδειγμα, μπορούμε να μετρήσουμε τον αριθμό των γυναικών και των αντρών σε κάθε κατοικία, διαχωρίζοντας τους ταυτόχρονα ανάλογα με την κατάσταση δανείου τους (έχουν δάνειο ή όχι).

Ανάλυση Δανείου

```
CrossTable(dat$home, dat$loan_st, prop.r = TRUE,
prop.c = FALSE, prop.t = FALSE, prop.chisq = FALSE)
```

Loan Status by Home Ownership			
	dat$loan_st		
dat$home	**0**	**1**	**Row Total**
MORTGAGE	10821	1181	12002
	0.902	0.098	0.413
OTHER	80	17	97
	0.825	0.175	0.003
OWN	2049	252	2301
	0.890	0.110	0.079
RENT	12915	1777	14692
	0.879	0.121	0.505
Column Total	25865	3227	29092

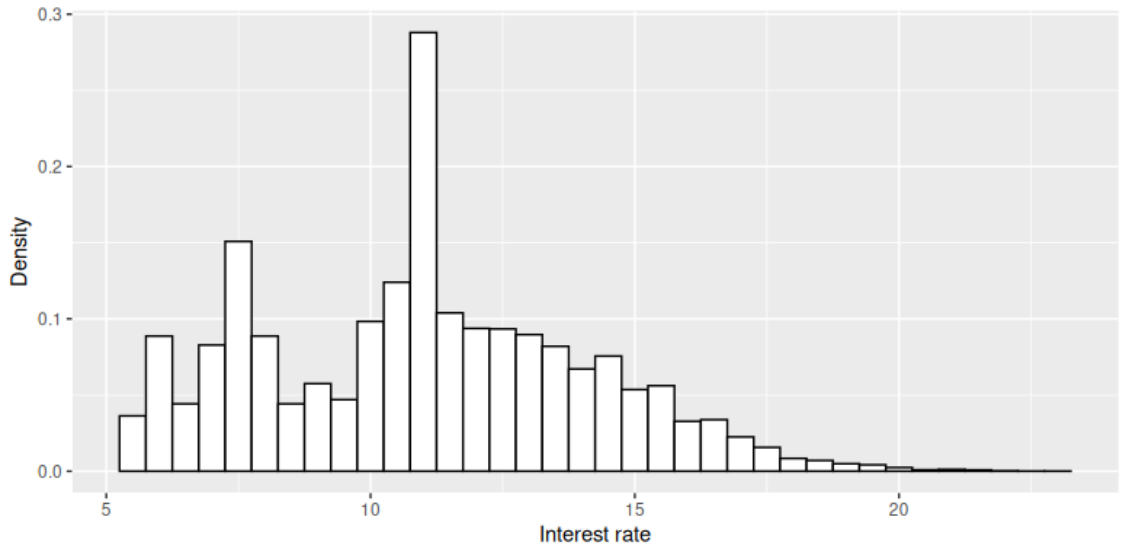
Ανάλυση Δανείου

Πίνακας αυτός αποκαλύπτει τις πτωχεύσεις ανά ιδιοκτησία κατοικίας.

Μπορούμε να χρησιμοποιήσουμε ιστογράμματα για να δούμε την κατανομή μίας μεταβλητής.

Σε αυτή την περίπτωση, έχουμε την κατανομή του επιτοκίου.

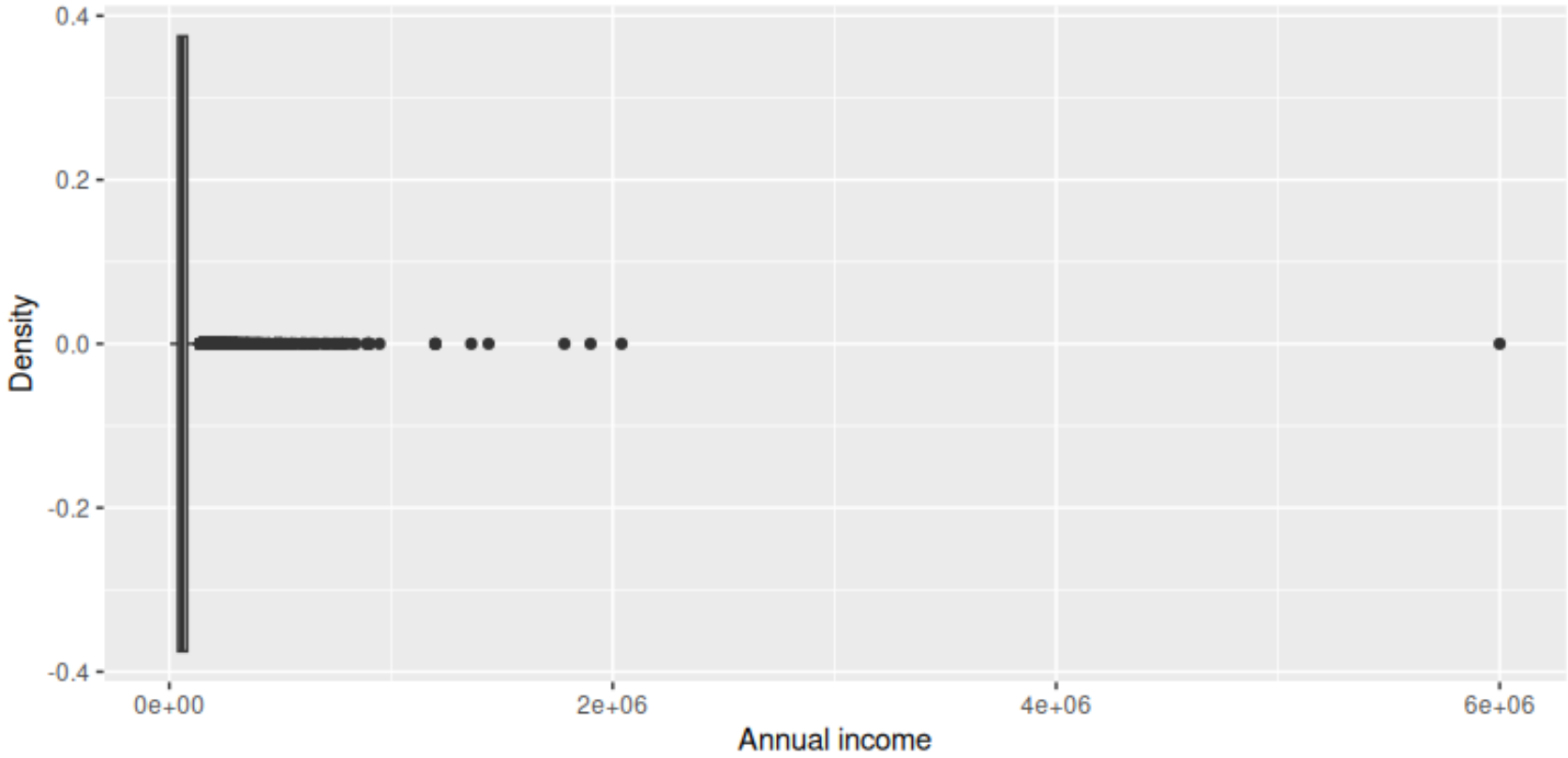
```
ggplot(dat, aes(x = int)) +
geom_histogram(aes(y=..density..), binwidth = 0.5, colour = "black",
fill = "white") +
labs(y = "Density", x = "Interest rate") +
theme(legend.position = "bottom", legend.title = element_blank())
```



Ανάλυση Δανείου

Η παρακάτω εικόνα είναι ένα διάγραμμα κουτιού (boxplot) για το ετήσιο εισόδημα.

```
ggplot(dat, aes(income)) +
geom_boxplot() +
labs(y = "Density",
x = "Annual income") +
theme(legend.position = "bottom", legend.title = element_blank())
```



Ανάλυση Δανείου

Το παραπάνω διάγραμμα κουτιού (boxplot) δείχνει κάποια προβλήματα στις τιμές.

Έχουμε μια πολύ μεγάλη τιμή ετήσιου εισοδήματος στον οριζόντιο άξονα (6.000.000).

Επίσης, υπάρχουν λίγα άτομα με πολύ υψηλό εισόδημα.

Θα πρέπει να ερευνήσουμε περαιτέρω και να διαπιστώσουμε εάν πρόκειται για έγκυρες παρατηρήσεις ή απλά για λάθος στην αρχική βάση δεδομένων.

Ανάλυση Δανείου

```
high_income <- dat[(dat$income > 1000000), ]
high_income
```

loan_st	l_amnt	int	grade	emp_len	home	income	age	sex	region
0	12025	14.27	C	13	RENT	1782000	63	0	E
0	10000	6.54	A	16	OWN	1200000	36	0	W
0	1500	10.99	A	5	MORTGAGE	1900000	60	1	N
0	12000	7.51	A	1	MORTGAGE	1200000	32	0	E
0	5000	12.73	C	12	MORTGAGE	6000000	**144**	1	E
0	10000	10.99	A	1	MORTGAGE	1200000	40	1	E
0	6400	7.40	A	7	MORTGAGE	1440000	44	1	E
0	6600	7.74	A	9	MORTGAGE	1362000	47	0	E
0	8450	12.29	C	0	RENT	2039784	42	0	E

Ένας άνδρας δεν είναι μόνο πλούσιος, είναι και 144 ετών. Έτσι, η απόφασή είναι να διαγράψω αυτά τα 9 άτομα. Ο καθαρισμός δεδομένων είναι μια συνηθισμένη εργασία όταν χειρίζεστε μεγάλες βάσεις δεδομένων. Αυτό δεν αποτελεί πρόβλημα, εφόσον δεν αλλοιώσουμε τη φύση των δεδομένων.

Ανάλυση Δανείου

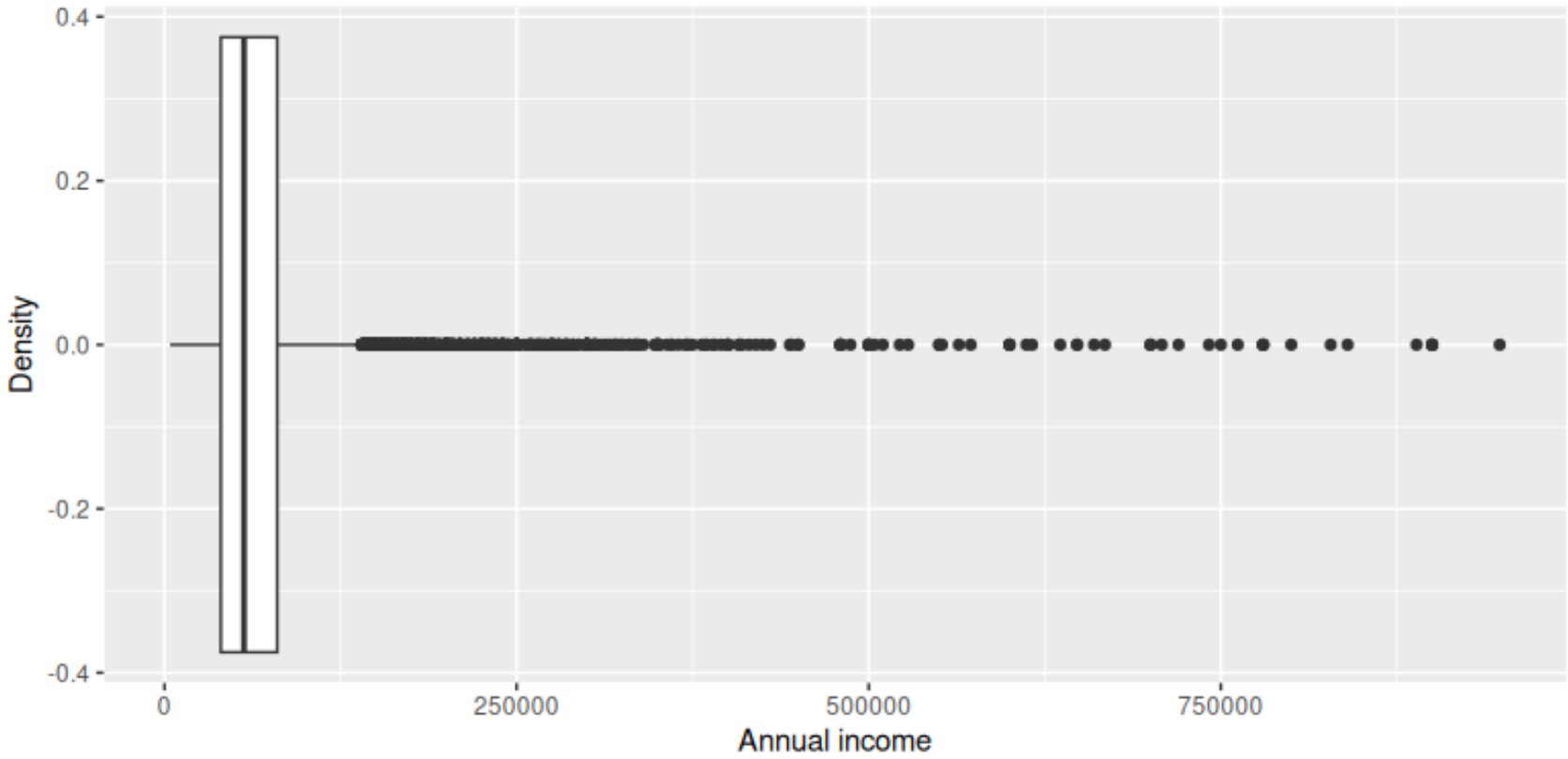
```
high_income_index <- data.frame(value = as.integer(rownames(high_income)))
dat <- dat[-high_income_index$value,]
```

```
> dim(dat)
[1] 29083    10
```

Αρχικά η βάση δεδομένων είχε 29.092 γραμμές και τώρα αφαιρούμε 9 οπότε καταλήγουμε με 29.083. Είδαμε το αποτέλεσμα.

Ανάλυση Δανείου

```
ggplot(dat, aes(income)) +
geom_boxplot() +
labs(y = "Density", x = "Annual income") +
theme(legend.position = "bottom", legend.title = element_blank())
```



Ανάλυση Δανείου

Τα **box plots** είναι ένα τυποποιημένος τρόπος εμφάνισης της κατανομής των δεδομένων βασισμένος σε μια πενταψήφια περίληψη: το ελάχιστο,
το πρώτο τεταρτημόριο (Q1),
η μεσαία τιμή (δεύτερο τεταρτημόριο ή Q2),
το τρίτο τεταρτημόριο (Q3) και το μέγιστο.

Λογιστική Παλινδρόμηση

Τα μοντέλα λογιστικής παλινδρόμησης μας επιτρέπουν να κάνουμε προβλέψεις για επισφαλείς δανειστές.

Η λογιστική παλινδρόμηση είναι ένα στατιστικό μοντέλο που στη βασική του μορφή χρησιμοποιεί μια λογιστική συνάρτηση για να μοντελοποιήσει μια δυαδική εξαρτημένη μεταβλητή όπως το loan_st.

Σε αυτή την περίπτωση, η δυαδική εξαρτημένη μεταβλητή είναι η επιλογή (1) ή η μη επιλογή (0).

Λογιστική Παλινδρόμηση

Πρώτα, φορτώνουμε τα δεδομένα και τα χωρίζουμε σε δύο σύνολα:
(1) εκπαίδευσης και
(2) δοκιμής.

Το σύνολο εκπαίδευσης χρησιμοποιείται για τη δημιουργία και την εκτίμηση μοντέλων, ενώ το σύνολο δοκιμής χρησιμοποιείται για την αξιολόγηση των προβλέψεων του μοντέλου μας με νέα δεδομένα.

Όταν εκτιμούμε μοντέλα, είναι συνηθισμένη πρακτική να διαχωρίζουμε τα διαθέσιμα δεδομένα σε δύο μέρη, δεδομένα εκπαίδευσης και δοκιμής, όπου τα δεδομένα εκπαίδευσης χρησιμοποιούνται για την εκτίμηση των παραμέτρων (εντός δείγματος) και τα δεδομένα δοκιμής χρησιμοποιούνται για την αξιολόγηση της ακρίβειας του (εκτός δείγματος).

Επειδή τα δεδομένα δοκιμής δεν χρησιμοποιούνται στον καθορισμό της εκτίμησης, θα πρέπει να παρέχουν μια αξιόπιστη ένδειξη για το πόσο καλά το μοντέλο είναι πιθανό να εκτιμήσει ή να προβλέψει νέα δεδομένα.

Λογιστική Παλινδρόμηση

Συνοψίζοντας, εκπαιδεύουμε το μοντέλο, το δοκιμάζουμε και μόλις είμαστε ικανοποιημένοι με την απόδοση του μοντέλου σε νέα δεδομένα, είμαστε έτοιμοι να το χρησιμοποιήσουμε σε πραγματικές εφαρμογές.

Αν αγνοήσουμε αυτόν τον διαχωρισμό και χρησιμοποιήσουμε ολόκληρη τη βάση δεδομένων για να εκτιμήσουμε τα μοντέλα μας, μπορεί να πετύχουμε να εξηγήσουμε τις προεπιλογές στη βάση δεδομένων μας, αλλά μπορεί να αποτύχουμε να εξηγήσουμε τις προεπιλογές για νέες αιτήσεις δανείου.

Λογιστική Παλινδρόμηση

```
# It is convenient to set the loan status as factor.
dat$loan_st <- as.factor(dat$loan_st)
set.seed(567)
index_train <- cbind(runif(1 : nrow(dat), 0 , 1), c(1 : nrow(dat)))
index_train <- order(index_train[, 1])
index_train <- index_train[1: (2/3 * nrow(dat))]
# Create training set
train <- dat[index_train , ]
# Create test set
test <- dat[-index_train , ]
```

Έχουμε 29.083 παρατηρήσεις στο dat.

Ο παραπάνω κώδικας επιλέγει τυχαία **29083*(2/3) = 19388** γραμμές για να σχηματίσει το σύνολο εκπαίδευσης (train).

Το σύνολο δοκιμής (test) αποτελείται από τις υπόλοιπες γραμμές.

Λογιστική Παλινδρόμηση

Ο τυχαίος διαχωρισμός πρέπει να γίνει καθώς το dat μπορεί να έχει κάποια δομή ή ταξινόμηση που θα μπορούσε να προκαλέσει προκατάληψη στην εκτίμηση του μοντέλου μας και να επηρεάσει αρνητικά τη δοκιμή του μοντέλου.

Για παράδειγμα, φανταστείτε ότι για κάποιο περίεργο λόγο η βάση δεδομένων είναι ταξινομημένη με τέτοιο τρόπο ώστε οι πρώτες παρατηρήσεις να είναι όλες περιπτώσεις χωρίς καθυστέρηση.

Αν συμβαίνει αυτό, τότε το σύνολο εκπαίδευσης και δοκιμής δεν θα έχουν τμήματα περιπτώσεων καθυστέρησης και μη καθυστέρησης και μπορεί να διαστρεβλώσουμε ολόκληρη την ανάλυση.

Ο τυχαίος διαχωρισμός μας επιτρέπει να αναπαράγουμε μια πραγματική κατάσταση στην οποία η βάση δεδομένων μας δεν είναι ταξινομημένη, με διαφορετικά χαρακτηριστικά.

Λογιστική Παλινδρόμηση

```
dat_prop <- table(dat$loan_st)/sum(table(dat$loan_st))
train_prop <- table(train$loan_st)/sum(table(train$loan_st))
test_prop <- table(test$loan_st)/sum(table(test$loan_st))
prop <- data.frame(rbind(dat_prop, train_prop, test_prop))
colnames(prop) <- c("no defaults", "defaults")
prop
```

```
           no defaults   defaults
dat_prop     0.8890417 0.1109583
train_prop   0.8882298 0.1117702
test_prop    0.8906653 0.1093347
```

Λογιστική Παλινδρόμηση

Δείτε το σύνολο εκπαίδευσης (train).

```
# See the data structure.
head(train)
```

	loan_st	l_amnt	int	grade	emp_len	home	income	age	sex	region
21547	0	4000	10.25	B	6	RENT	26000	27	1	N
22625	0	5000	14.26	C	3	RENT	280000	36	1	N
20340	0	18000	18.30	F	10	RENT	121000	24	1	N
1911	0	5600	7.90	A	3	RENT	32000	22	0	W
4021	0	2700	14.27	C	5	MORTGAGE	88500	32	1	N
16887	0	15000	10.38	B	10	MORTGAGE	75000	23	0	W

Οι μεταβλητές ως παράγοντες (factors) είναι χρήσιμες για την εκτίμηση μοντέλων και την απεικόνιση δεδομένων στην γλώσσα R.

Οι παράγοντες είναι μεταβλητές στην R που λαμβάνουν έναν περιορισμένο αριθμό διαφορετικών τιμών.

Τέτοιες μεταβλητές αναφέρονται συχνά ως κατηγορηματικές μεταβλητές.

Λογιστική Παλινδρόμηση

Ας υποθέσουμε ότι πιστεύουμε ότι η κατάσταση δανείου (loan_st) εξαρτάται από την ηλικία του ατόμου.

Μπορούμε να εκτιμήσουμε ένα απλό λογιστικό μοντέλο για να μάθουμε τη σχέση μεταξύ της ηλικίας και της κατάστασης δανείου.

```
# Fitting a simple logistic model.
logi_age <- glm(loan_st ~ age, family = "binomial", data = train)
logi_age

Call:  glm(formula = loan_st ~ age, family = "binomial", data = train)

Coefficients:
(Intercept)          age
-1.90097     -0.00623

Degrees of Freedom: 19387 Total (i.e. Null);  19386 Residual
Null Deviance:      13580
Residual Deviance: 13580     AIC: 13580
```

Λογιστική Παλινδρόμηση

Φαίνεται πως υπάρχει μια αρνητική σχέση μεταξύ της ηλικίας και της κατάστασης δανείου.

Η τιμή AIC (13.580) είναι χρήσιμη για τη σύγκριση μοντέλων.
Το κριτήριο πληροφορίας Akaike (AIC) είναι μια μαθηματική μέθοδος για την αξιολόγηση του πόσο καλά ένα μοντέλο ταιριάζει στα δεδομένα από τα οποία δημιουργήθηκε.

Στην στατιστική, το AIC χρησιμοποιείται για τη σύγκριση διαφορετικών πιθανών μοντέλων και τον προσδιορισμό του ποιο είναι το πιο κατάλληλο για τα δεδομένα.

Προς το παρόν, το AIC δεν είναι χρήσιμη επειδή έχουμε μόνο ένα μοντέλο και δεν μπορούμε να το συγκρίνουμε με κάποιο άλλο.

Ας εκτιμήσουμε ένα άλλο απλό μοντέλο όπου η κατηγορία επιτοκίου χρησιμοποιείται ως παράγοντας πρόβλεψης της κατάστασης δανείου (**loan_st**).

Θυμηθείτε ότι δεν πραγματοποιούμε καμία πρόβλεψη αυτή τη στιγμή, απλώς εκτιμούμε μοντέλα χρησιμοποιώντας το σύνολο εκπαίδευσης (**train**).

Λογιστική Παλινδρόμηση

```
# Build a glm model with variable interest rate as a predictor.
logi_int <- glm(formula = loan_st ~ int, family = "binomial", data = train)
# Print the parameter estimates.
logi_int

Call:  glm(formula = loan_st ~ int, family = "binomial", data = train)

Coefficients:
(Intercept)          int
-3.710        0.142

Degrees of Freedom: 19387 Total (i.e. Null);  19386 Residual
Null Deviance:      13580
Residual Deviance: 13210    AIC: 13220
```

Το AIC είναι χαμηλότερο (13.220 έναντι 13.580), οπότε τώρα έχουμε ένα καλύτερο μοντέλο.

Λογιστική Παλινδρόμηση

Η χρήση ενός μόνο παράγοντα πρόβλεψης όπως η ηλικία ή το επιτόκιο είναι σαφώς μια περιορισμένη προσέγγιση.

Ας προσθέσουμε μερικούς ακόμη παράγοντες πρόβλεψης.

Επίσης, ας εισάγουμε τη συνάρτηση summary() για να εξαγάγουμε περισσότερες πληροφορίες σχετικά με τα αποτελέσματα εκτίμησης του μοντέλου.

Το **logi_multi** παρακάτω υποθέτει ότι η κατάσταση δανείου εξαρτάται από την ηλικία, το επιτόκιο, την κατηγορία πιστοληπτικής διαβάθμισης (grade), το ποσό του δανείου και το ετήσιο εισόδημα.

```
# Multiple variables in a logistic regression model.
logi_multi <- glm(loan_st ~ age + int + grade + log(l_amnt) +
log(income) , family = "binomial", data = train)
# Obtain significance levels using summary().
summary(logi_multi)

Call:
glm(formula = loan_st ~ age + int + grade + log(l_amnt) + log(income),
family = "binomial", data = train)

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.996240   0.477911   4.177 2.95e-05 ***
age         -0.002302   0.003825  -0.602   0.5473
int          0.038767   0.017249   2.247   0.0246 *
gradeB       0.503409   0.087435   5.758 8.54e-09 ***
gradeC       0.748229   0.117765   6.354 2.10e-10 ***
gradeD       0.964343   0.147283   6.548 5.85e-11 ***
gradeE       1.033442   0.190817   5.416 6.10e-08 ***
gradeF       1.619470   0.257900   6.279 3.40e-10 ***
gradeG       1.867494   0.440232   4.242 2.21e-05 ***
log(l_amnt)  0.015718   0.036341   0.433   0.6654
log(income) -0.470748   0.046423 -10.140  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
```

Λογιστική Παλινδρόμηση

```
Null deviance: 13579  on 19387  degrees of freedom
Residual deviance: 13028  on 19377  degrees of freedom
AIC: 13050

Number of Fisher Scoring iterations: 5
```

Το μοντέλο μας με πολλαπλούς παράγοντες λειτουργεί καλά.

Στο **logi_multi**, η τιμή AIC είναι η χαμηλότερη μέχρι στιγμής (13.050 έναντι 13.220), επομένως προς το παρόν θα πρέπει να θεωρηθεί ως το καλύτερο μοντέλο εντός δείγματος.

Η συνάρτηση summary() δείχνει τα επίπεδα σημαντικότητας των εκτιμητών, αλλά προς το παρόν μας ενδιαφέρει περισσότερο η καλή εφαρμογή των μοντέλων επειδή θέλουμε να κάνουμε προβλέψεις για την κατάσταση δανείου (**loan_st**).

Δηλαδή, μας ενδιαφέρει να χρησιμοποιήσουμε ένα μοντέλο για να μάθουμε εάν αναμένεται να καθυστερήσουν οι νέοι αιτούντες στο σύνολο δοκιμής (test) ή όχι, παρά στους παράγοντες κινδύνου πίστωσης των αιτούντων. Αυτός είναι ο λόγος που επικεντρωνόμαστε τώρα στο AIC.

Λογιστική Παλινδρόμηση

Όταν ένας πελάτης συμπληρώνει μια φόρμα αίτησης πιστώσεως, συλλέγουμε πληροφορίες αλλά δεν γνωρίζουμε με σιγουριά εάν θα καθυστερήσει τελικά την αποπληρωμή.

Ένα μοντέλο πιστωτικού κινδύνου μπορεί να μας βοηθήσει σε αυτό το έργο.

Λογιστική Παλινδρόμηση

Όταν ένας πελάτης συμπληρώνει μια φόρμα αίτησης πιστώσεως, συλλέγουμε πληροφορίες αλλά δεν γνωρίζουμε με σιγουριά εάν θα καθυστερήσει τελικά την αποπληρωμή.

Ένα μοντέλο πιστωτικού κινδύνου μπορεί να μας βοηθήσει σε αυτό το έργο.

1.3 Πρόβλεψη και Αξιολόγηση Μοντέλου

Ας πάρουμε τα τρία μας μοντέλα: logi_age, logi_int και logi_multi από το προηγούμενο υποτμήμα για να πραγματοποιήσουμε μια απλή άσκηση πρόβλεψης.

Ξεκινάμε εντοπίζοντας μια παρατήρηση στο σύνολο δοκιμής (test) και ζητάμε από τα μοντέλα να προβλέψουν την κατάσταση δανείου (loan_st).

Δηλαδή, παίρνουμε την ηλικία του πρώτου ατόμου, στη συνέχεια εφαρμόζουμε το μοντέλο logi_age και συγκρίνουμε την προβλεπόμενη κατάσταση δανείου (loan_st) με αυτό που συνέβη πραγματικά.

Θυμηθείτε ότι γνωρίζουμε τι συνέβη πραγματικά με αυτό το άτομο επειδή έχουμε τις πληροφορίες στο σύνολο δοκιμής.

Αναμένεται κάθε μοντέλο να παράγει διαφορετικές προβλέψεις για την κατάσταση δανείου. Εάν το μοντέλο είναι καλό, τότε η προβλεπόμενη κατάσταση δανείου θα πρέπει να ταιριάζει με αυτό που συνέβη πραγματικά.

1.3 Πρόβλεψη και Αξιολόγηση Μοντέλου

```
# Define one single observation in test_set.
John_Doe <- as.data.frame(test[1, ])
John_Doe
```

```
  loan_st l_amnt   int grade emp_len home income age sex region
1       0   5000 10.65     B      10 RENT  24000  33   0      E
```

Εντάξει, γνωρίζουμε εκ των προτέρων ότι η τιμή loan_st για αυτή την παρατήρηση που λάβαμε από το σύνολο δοκιμής (test) είναι 0.

Ωστόσο, τα μοντέλα δεν μπορούν να το γνωρίζουν αυτό απλά επειδή δεν χρησιμοποιήσαμε το σύνολο δοκιμής για να εκτιμήσουμε τα λογιστικά μοντέλα.

Τα μοντέλα μας εκτιμήθηκαν χρησιμοποιώντας το σύνολο εκπαίδευσης (train). Ένα καλό μοντέλο πιστωτικού κινδύνου θα πρέπει να προβλέπει μη καθυστέρηση για αυτόν τον νέο αιτούντα.

Οι τιμές της μεταβλητής loan_st στο σύνολο δοκιμής είναι είτε 0 είτε 1.

1.3 Πρόβλεψη και Αξιολόγηση Μοντέλου

Ωστόσο, τα λογιστικά μοντέλα εκτιμούν την loan_st ως τιμές στο εύρος από 0 έως 1.

Αυτό σημαίνει ότι θα περιμέναμε η εκτιμώμενη τιμή loan_st να είναι κοντά στο 0.

Αλλά πόσο κοντά; Θα ασχοληθούμε με αυτό το ζήτημα αργότερα.

```
# Predict the loan status.
logi_age_pred <- predict(logi_age, newdata = John_Doe, type = "response")
logi_int_pred <- predict(logi_int, newdata = John_Doe, type = "response")
logi_multi_pred <- predict(logi_multi, newdata = John_Doe, type = "response")
# Collect all.
pred_John <- rbind("logi_age" = logi_age_pred,
"logi_int" = logi_int_pred,
"logi_multi" = logi_multi_pred)
# Prepare a table.
colnames(pred_John) <- "Loan status predictions for John Doe."
pred_John
```

```
           Loan status predictions for John Doe.
logi_age                                0.10846236
logi_int                                0.09996702
logi_multi                              0.14461840
```

Αυτές οι τιμές είναι χαμηλές καθώς είναι κοντά στο 0.

Θα μπορούσαμε να το ερμηνεύσουμε αυτό ως μια συγκεκριμένη ικανότητα των μοντέλων να προβλέψουν αυτή τη μεμονωμένη περίπτωση από το σύνολο δοκιμής (test).

Ωστόσο, πολλά ερωτήματα παραμένουν αναπάντητα και απαιτούν περαιτέρω ανάλυση.

Για παράδειγμα:
Πώς μπορούμε να προσδιορίσουμε εάν η πρόβλεψη είναι αρκετά χαμηλή ώστε να θεωρηθεί μη καθυστέρηση;

Ίσως χρειαστούμε μια τιμή ορίου για να αποφασίσουμε.

Θα εξερευνήσουμε αυτό το ζήτημα αργότερα.

Μια άλλη πτυχή είναι: Τι γίνεται με τις υπόλοιπες περιπτώσεις στο σύνολο δοκιμής;

Έχουμε 9.695 παρατηρήσεις στο σύνολο δοκιμής και στο παραπάνω παράδειγμα δοκιμάζουμε μόνο για την πρώτη.

Μας ενδιαφέρει ολόκληρο το σύνολο δοκιμής και όχι μόνο ο John Doe.

Ευτυχώς, αυτό το ζήτημα είναι εύκολο να αντιμετωπιστεί καθώς χρειάζεται μόνο να αλλάξουμε την παράμετρο newdata στη συνάρτηση predict().

Συγκεκριμένα, αντί για newdata = John_Doe, που είναι μια παρατήρηση, μπορούμε να την αλλάξουμε σε newdata = test, που είναι ολόκληρο το σύνολο δοκιμής των 9.695 παρατηρήσεων.

1.3 Πρόβλεψη και Αξιολόγηση Μοντέλου

```
# Predict the loan status with the three models.
pred_logi_age <- predict(logi_age, newdata = test, type = "response")
pred_logi_int <- predict(logi_int, newdata = test, type = "response")
pred_logi_multi <- predict(logi_multi, newdata = test,
type = "response")

pred_range <- rbind("logi_age" = range(pred_logi_age),
"logi_int" = range(pred_logi_int),
"logi_multi" = range(pred_logi_multi))
aic <- rbind(logi_age$aic, logi_int$aic, logi_multi$aic)
pred_range <- cbind(pred_range, aic)
colnames(pred_range) <- c("min(loan_st)", "max(loan_st)", "AIC")
pred_range
```

```
           min(loan_st) max(loan_st)      AIC
logi_age     0.08417892    0.1165453 13580.65
logi_int     0.05019756    0.3982977 13215.61
logi_multi   0.01739107    0.4668004 13050.30
```

1.3 Πρόβλεψη και Αξιολόγηση Μοντέλου

Τώρα, αντί για την πρόβλεψη ενός μόνο αιτούντα, πραγματοποιήσαμε μια πρόβλεψη για όλους τους 9.695 αιτούντες στο σύνολο δοκιμής (test).

Η στήλη της χαμηλότερης τιμής αντιστοιχεί στην χαμηλότερη προβλεπόμενη κατάσταση δανείου (loan_st) για κάθε μοντέλο.

Τα λογιστικά μοντέλα παράγουν τιμές στο εύρος από μηδέν έως ένα, και στην περίπτωση αυτή τα εύρη είναι αρκετά στενά.

Τα στενά εύρη (η διαφορά μεταξύ των προβλεπόμενων τιμών loan_st υψηλότερης και χαμηλότερης) μπορεί να αποτελέσουν πρόβλημα επειδή το μοντέλο δεν θα μπορούσε να διακρίνει μεταξύ καθυστερήσεων (προβλέψεις πιο κοντά στο 1) και μη καθυστερήσεων (προβλέψεις πιο κοντά στο 0).

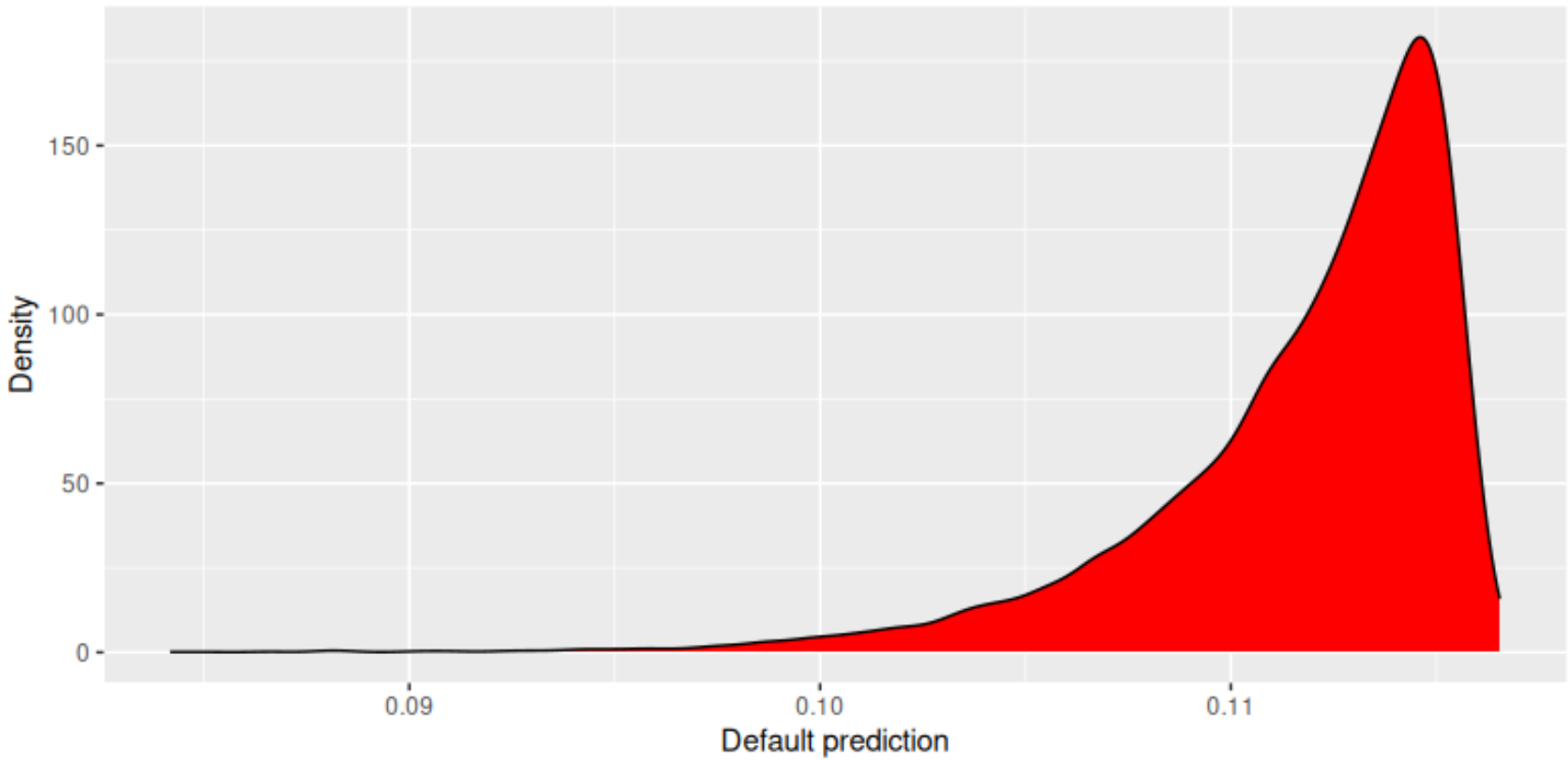
1.3 Πρόβλεψη και Αξιολόγηση Μοντέλου

Το υψηλότερο AIC αντιστοιχεί στο χειρότερο εντός δείγματος μοντέλο και το χαμηλότερο AIC στο καλύτερο εντός δείγματος μοντέλο.

Εδώ, μπορούμε να δούμε κάποια συνέπεια εντός και εκτός δείγματος επειδή το καλύτερο μοντέλο σύμφωνα με το AIC, αντιστοιχεί στο μοντέλο με το υψηλότερο εύρος πρόβλεψης. Ας διερευνήσουμε όλες τις προβλεπόμενες τιμές loan_st για το logi_age:

```
ggplot(data.frame(pred_logi_age), aes(x = pred_logi_age)) +
geom_density(fill = "red") +
labs(y = "Density", x = "Default prediction") +
theme(legend.position = "bottom", legend.title = element_blank())
```

1.3 Πρόβλεψη και Αξιολόγηση Μοντέλου



Σχήμα: ”Ιστόγραμμα πρόβλεψης μοντέλου ηλικίας.”

1.3 Πρόβλεψη και Αξιολόγηση Μοντέλου

Το μοντέλο **logi_age** δεν μπορεί να προβλέψει τιμές που κυμαίνονται από 0 έως 1.

Στην πραγματικότητα, αυτές οι τιμές είναι αρκετά συγκεντρωμένες σε ένα πολύ μικρό εύρος τιμών.

Ως συνέπεια, αυτό το μοντέλο δεν μπορεί να διαφοροποιήσει μεταξύ προβλέψεων καθυστέρησης και μη καθυστέρησης.

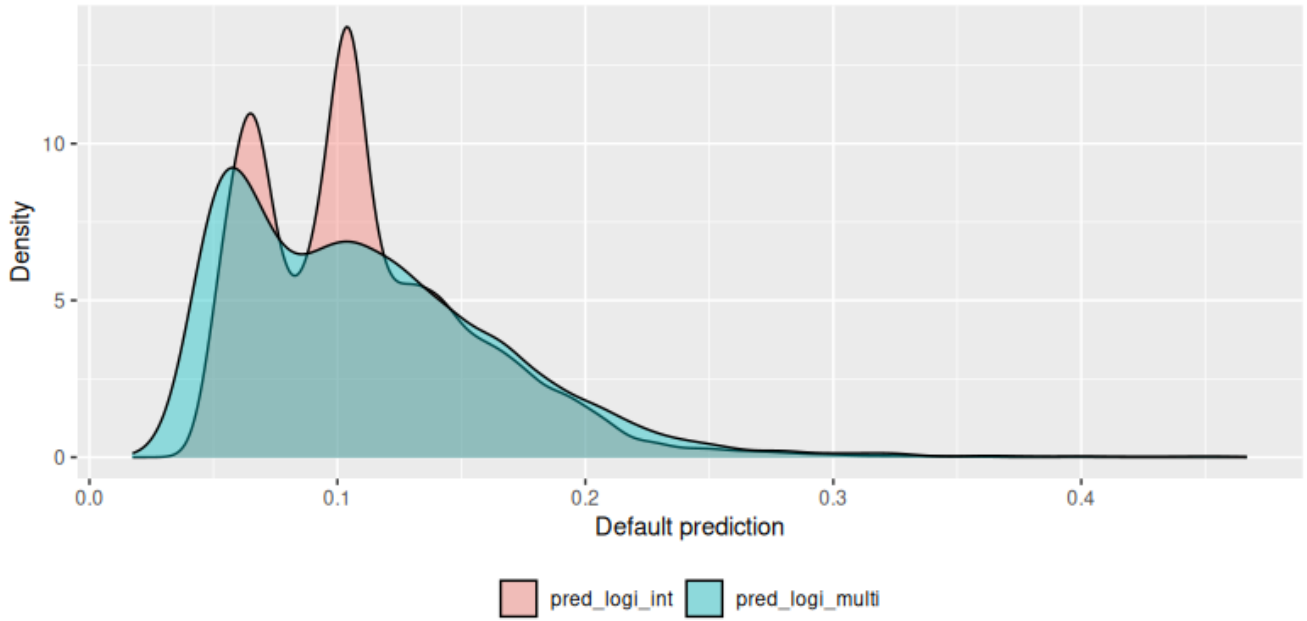
Ας οπτικοποιήσουμε τις προβλέψεις των logi_int και logi_multi.

Αρχικά, συγκεντρώνουμε όλες τις προβλέψεις σε ένα μόνο πλαίσιο δεδομένων για ευκολία.

```
pred_logi <- data.frame(cbind(pred_logi_age, pred_logi_int,
pred_logi_multi))
pred_logi <- gather(pred_logi, key = "model", value = "pred")
```

Τώρα ας απεικονίσουμε τα logi_int και logi_multi.

1.3 Πρόβλεψη και Αξιολόγηση Μοντέλου

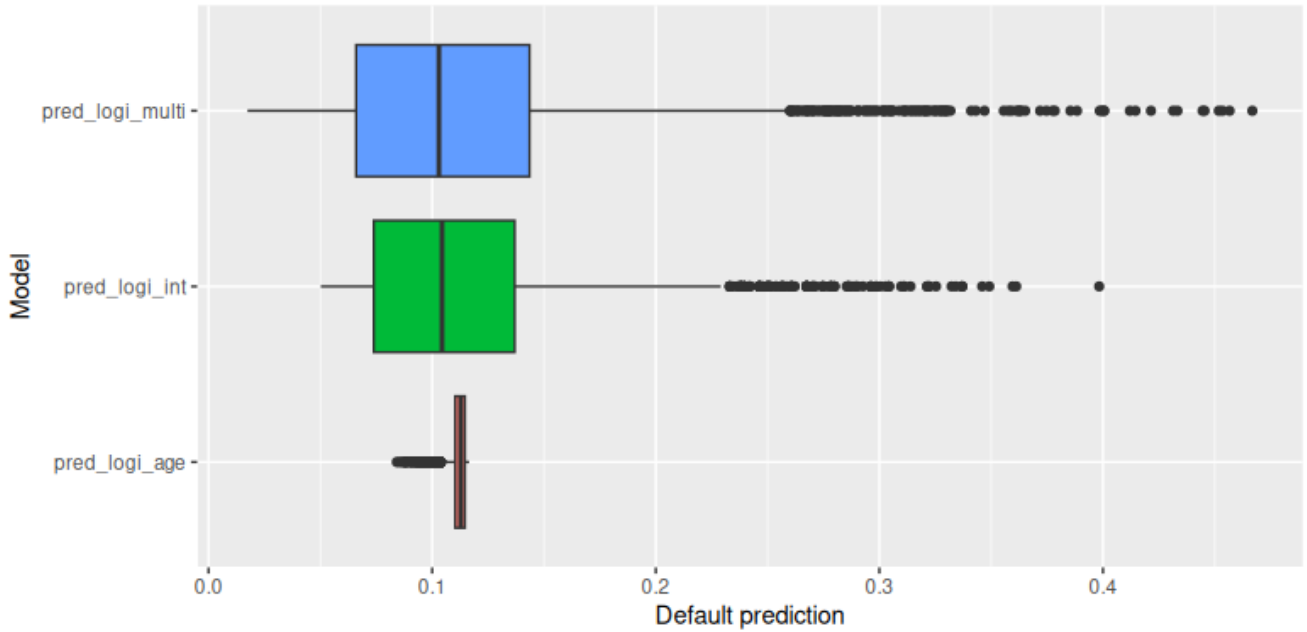


Σχήμα: Κατανομή προβλέψεων επιτοκίου από διάφορα μοντέλα

Ας προσθέσουμε και το μοντέλο ηλικίας.

1.3 Πρόβλεψη και Αξιολόγηση Μοντέλου

```
ggplot(pred_logi, aes(x = pred, y = model, fill = model)) +
geom_boxplot() +
labs(y = "Model", x = "Default prediction") +
theme(legend.position = "none", legend.title = element_blank())
```



Σχήμα: Ηλικίας, επιτοκίου και πολλαπλών μοντέλων.

1.3 Πρόβλεψη και Αξιολόγηση Μοντέλου

Πιθανολογούμε ότι ένα μοντέλο που λαμβάνει υπόψη όλες τις διαθέσιμες μεταβλητές θα μπορούσε να είναι καλύτερο για την πρόβλεψη της μεταβλητής loan_st.

```
# Logistic regression model using all available predictors in the data set.
logi_full <- glm(loan_st ~ age + int + grade + log(l_amnt) +
log(income) + emp_len + home + sex +
region, family = "binomial", data = train)
# Loan status predictions for all test set elements.
pred_logi_full <- predict(logi_full, newdata = test, type = "response")
# Look at the predictions range.
range(pred_logi_full)
```

```
[1] 1.422469e-09 8.544424e-01
```

Τώρα, το εύρος πρόβλεψης του pred_logi_full είναι ευρύτερο. Ένα ευρύτερο εύρος σημαίνει ότι οι προβλέψεις loan_st είναι τώρα πιο κοντά στο 1.

Αυτό είναι καλό γιατί χρειαζόμαστε το μοντέλο να μπορεί να προβλέπει τόσο μη καθυστερήσεις (0) όσο και καθυστερήσεις (1).

Ας δούμε μια σύγκριση πρόβλεψης σε σχέση με τα υπόλοιπα μοντέλα.

1.3 Πρόβλεψη και Αξιολόγηση Μοντέλου

```
pred_range <- rbind("logi_age" = range(pred_logi_age),
"logi_int" = range(pred_logi_int),
"logi_multi" = range(pred_logi_multi),
"logi_full" = range(pred_logi_full))
aic <- rbind(logi_age$aic, logi_int$aic, logi_multi$aic, logi_full$aic)
pred_range <- cbind(pred_range, aic)
colnames(pred_range) <- c("min(loan_st)", "max(loan_st)", "AIC")
```

```
> pred_range
min(loan_st) max(loan_st)       AIC
logi_age    8.417892e-02    0.1165453 13580.65
logi_int    5.019756e-02    0.3982977 13215.61
logi_multi  1.739107e-02    0.4668004 13050.30
logi_full   1.422469e-09    0.8544424 10763.67
```

1.3 Πρόβλεψη και Αξιολόγηση Μοντέλου

```
pred_logi <- data.frame(cbind(pred_logi_age, pred_logi_int,
pred_logi_multi, pred_logi_full))
pred_logi <- gather(pred_logi, key = "model", value = "pred")
```

Γραφική σύγκριση των νέων μεταβλητών πρόβλεψης pred_logi_full και pred_logi_multi.

```
ggplot(pred_logi[pred_logi$model != "pred_logi_age" &
pred_logi$model != "pred_logi_int" ,],
aes(x = pred, fill = model)) +
geom_density(alpha = 0.4) +
labs(y = "Density", x = "Default prediction") +
theme(legend.position = "bottom", legend.title = element_blank())
```

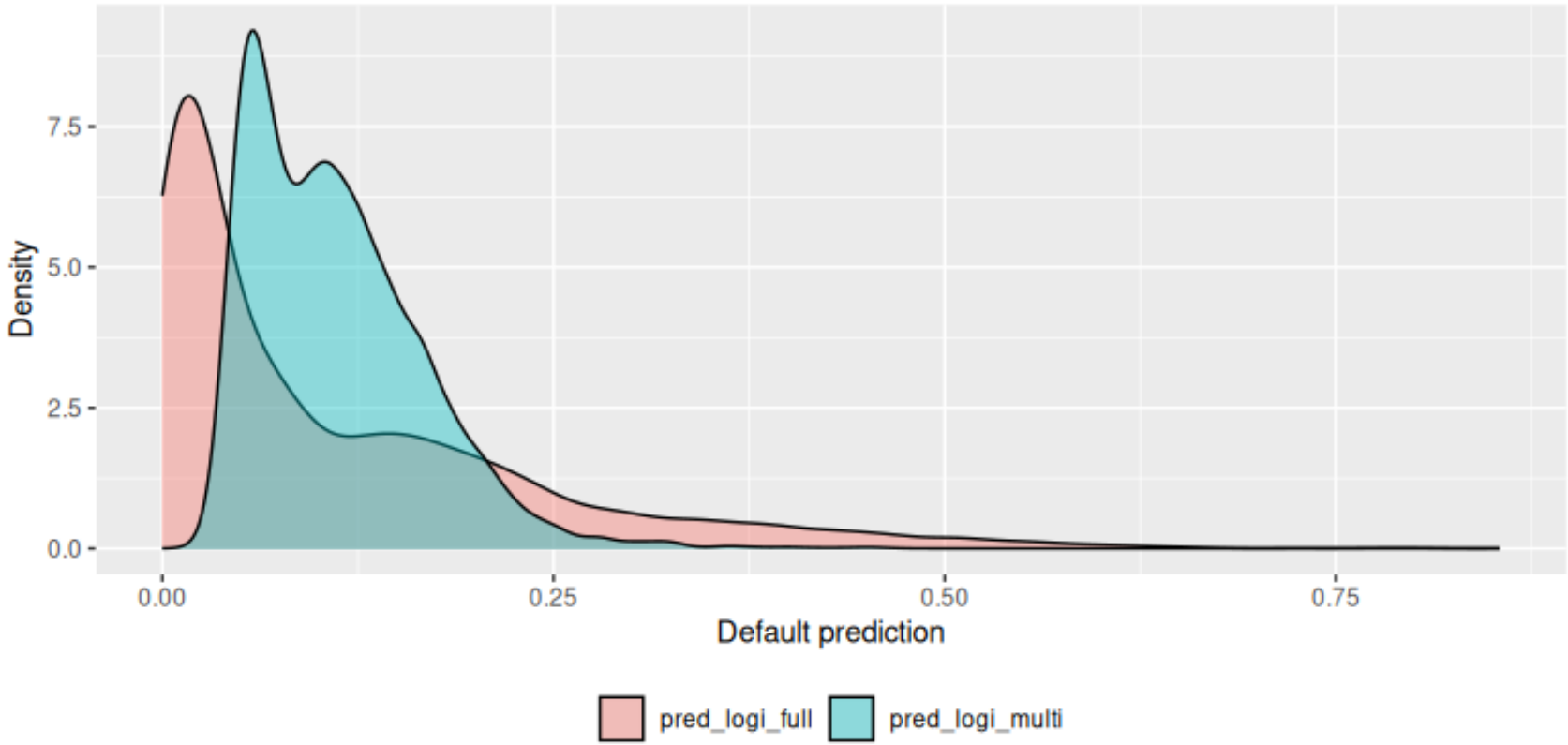
1.3 Πρόβλεψη και Αξιολόγηση Μοντέλου

```
pred_logi <- data.frame(cbind(pred_logi_age, pred_logi_int,
pred_logi_multi, pred_logi_full))
pred_logi <- gather(pred_logi, key = "model", value = "pred")
```

Γραφική σύγκριση των νέων μεταβλητών πρόβλεψης pred_logi_full και pred_logi_multi.

```
ggplot(pred_logi[pred_logi$model != "pred_logi_age" &
pred_logi$model != "pred_logi_int" ,],
aes(x = pred, fill = model)) +
geom_density(alpha = 0.4) +
labs(y = "Density", x = "Default prediction") +
theme(legend.position = "bottom", legend.title = element_blank())
```

1.3 Πρόβλεψη και Αξιολόγηση Μοντέλου



Σχήμα: Ιστόγραμμα προβλέψεων των μοντέλων Multi και Full