

## Feladat és a kidolgozásának ismertetése

Az alábbi feladatot választottam:

### 1. Hitelkártya fizetési késedelem/Default of credit card clients

Ebben a feladatban hitelkártya fizetési késedelmet kell előre jelezni, taiwani ügyfelek adataira építve.

R-ban dolgoztam, **döntési fa** és **logisztikus regresszió** modelleket használtam.

Döntési fa megoldásnál egy machine learning megoldást alkalmaztam, **c50** és **gmodels** package-eket használtam fel, a Haladó IT Megoldások órán vettek alapján.

A **logisztikus regresszió** modellnél **aod** package-t használtam fel.

Illetve szerettem volna látni egy másik logisztikus regressziós megoldást, ezért **Gretl**-ben készítettem egy **logisztikus regresszió** modellt és ott futattam egy predikciót. A munkám végére betettem, mint plusz kiegészítés.

## Adattábla átalakítás:

**Formai változtatás:** Kisbetűsítettem az oszlop neveket az egyszerűbb kezelhetőség miatt.

Első körös riportok SPSS-ben elvégezve:

**Missing values:**

Univariate Statistics							
	N	Mean	Std. Deviation	Missing		No. of Extremes <sup>a</sup>	
				Count	Percent	Low	High
age	22152	35.33	9.375	0	.0	0	822
pay_2	22152	-.09	1.165	0	.0	0	227
pay_3	22152	-.14	1.158	0	.0	0	128
pay_4	22152	-.21	1.122	0	.0	0	112
pay_5	22152	-.27	1.089	0	.0	0	2341
pay_6	22152	-.30	1.117	0	.0	0	2414
bill_amt1	22152	41679.30	51422.844	0	.0	0	1367
bill_amt2	22152	39477.27	49618.286	0	.0	0	1350
bill_amt3	22152	37109.33	47586.566	0	.0	0	1384
bill_amt4	22152	33843.95	44629.077	0	.0	0	1384
bill_amt5	22152	31301.94	42437.601	0	.0	0	1393
bill_amt6	22152	30134.98	41907.900	0	.0	1	1377
pay_amt1	22152	2624.65	2598.345	0	.0	0	1318
pay_amt2	22152	2519.16	2579.536	0	.0	0	1250
pay_amt3	22152	2170.14	2425.112	0	.0	0	1147
pay_amt4	22152	1969.61	2340.404	0	.0	0	1137
pay_amt5	22152	1973.63	2361.637	0	.0	0	1154

Készítette: Képesi Szilvia / I163IX

pay_amt6	22152	1908.53	2337.592	0	.0	0	1148
limit_bal	22152	143718.67	113708.970	0	.0	0	966
V12	22152	11076.50	6394.876	0	.0	0	0
sex	22152			0	.0		
education	22152			0	.0		
marriage	22152			0	.0		
default_nextmonth	22152			0	.0		

a. Number of cases outside the range (Mean - 2\*SD, Mean + 2\*SD).

Gyakorlatilag nincs hiányzó érték.

## Descriptive statistics:

### Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
limit_bal	22152	10000	720000	143718.67	113708.970
age	22152	21	75	35.33	9.375
pay_2	22152	-2	3	-.09	1.165
pay_3	22152	-2	3	-.14	1.158
pay_4	22152	-2	3	-.21	1.122
pay_5	22152	-2	3	-.27	1.089
pay_6	22152	-2	3	-.30	1.117
bill_amt1	22152	-11545	255413	41679.30	51422.844
bill_amt2	22152	-30000	255846	39477.27	49618.286
bill_amt3	22152	-46127	255846	37109.33	47586.566
bill_amt4	22152	-50616	255353	33843.95	44629.077
bill_amt5	22152	-53007	250327	31301.94	42437.601
bill_amt6	22152	-94625	253355	30134.98	41907.900
pay_amt1	22152	0	13942	2624.65	2598.345
pay_amt2	22152	0	13935	2519.16	2579.536
pay_amt3	22152	0	13959	2170.14	2425.112
pay_amt4	22152	0	13945	1969.61	2340.404
pay_amt5	22152	0	13919	1973.63	2361.637
pay_amt6	22152	0	13969	1908.53	2337.592
Valid N (listwise)	22152				

Alább a részletezés.

## Kiinduló adatok áttekintése

### Tények

default\_nextMonth: The response variable:

Készítette: Képesi Szilvia / I163IX

1 = the client defaulted 1 month after the data collection; - ügyfelek, akiknél késedelem volt

0 = the client did not default 1 month after the data collection) - ügyfelek, akiknél nem volt késedelem

A betöltött adattáblában **5378-an voltak, akiknél volt késedelem** a visszafizetésnél és 16774-en voltak, azok akiknél nem volt késedelem a visszafizetésnél a **22152 ügyfélből. 24,27%-nál volt késedelem.**

Az ügyfelek többsége, **13431 fő**, azaz a **60%-a nő**. Az ügyfelek **49%-a egyetemi végzettséggel** rendelkezik.

**Átlag életkoruk 35év**. Ebből **45,6%-uk házas**, a többiek nem házasok.

Elmondható, hogy a **számlaegyenlegük az ügyfeleknek 30- és 40000 NT dollar között** mozogott a vizsgált 6hónap során. Az **átlagos havi visszafizetések pedig 1909 és 2625 NT dollar között** mozogott a vizsgált 6hónap során. Mindkettő csökkenő tendenciát mutat. A **limit 10000 NT dollar és 720000 NT dollar között** mozog, **átlagosan 143719 NT dollar a limit.**

R-ban is megvizsgáltam, pirossal kiemelve a fontosabb részletek:

```
> summary(credit)
```

x	limit_bal	sex	education	marriage
Min. : 1	Min. : 10000	female:13431	graduate school: 7396	married:10101
1st Qu.: 5539	1st Qu.: 50000	male : 8721	high school : 3853	other : 249
Median :11076	Median :120000		other : 78	single :11802
Mean :11076	Mean :143719		university :10825	
3rd Qu.:16614	3rd Qu.:200000			
Max. :22152	Max. :720000			
age	pay_2	pay_3	pay_4	pay_5
Min. :21.00	Min. :-2.00000	Min. :-2.0000	Min. :-2.0000	Min. :-2.0000
1st Qu.:28.00	1st Qu.: -1.00000	1st Qu.: -1.0000	1st Qu.: -1.0000	1st Qu.: -1.0000
Median :34.00	Median : 0.00000	Median : 0.0000	Median : 0.0000	Median : 0.0000
Mean :35.33	Mean :-0.08997	Mean :-0.1379	Mean :-0.2132	Mean :-0.2699
3rd Qu.:41.00	3rd Qu.: 0.00000	3rd Qu.: 0.0000	3rd Qu.: 0.0000	3rd Qu.: 0.0000
Max. :75.00	Max. : 3.00000	Max. : 3.0000	Max. : 3.0000	Max. : 3.0000
pay_6	bill_amt1	bill_amt2	bill_amt3	bill_amt4
Min. :-2.0000	Min. :-11545	Min. :-30000	Min. :-46127	Min. :-50616
1st Qu.: -1.0000	1st Qu.: 2995	1st Qu.: 2301	1st Qu.: 1791	1st Qu.: 1366
Median : 0.0000	Median : 20954	Median : 19602	Median : 18591	Median : 17367
Mean :-0.2982	Mean : 41679	Mean : 39477	Mean : 37109	Mean : 33844
3rd Qu.: 0.0000	3rd Qu.: 59382	3rd Qu.: 55763	3rd Qu.: 51041	3rd Qu.: 48173
Max. : 3.0000	Max. :255413	Max. :255846	Max. :255846	Max. :255353
bill_amt5	bill_amt6	pay_amt1	pay_amt2	pay_amt3
Min. :-53007	Min. :-94625.0	Min. : 0	Min. : 0	Min. : 0
1st Qu.: 1000	1st Qu.: 727.5	1st Qu.: 626	1st Qu.: 471	1st Qu.: 296
Median : 15606	Median : 13739.5	Median : 2000	Median : 1882	Median : 1496
Mean : 31302	Mean : 30135.0	Mean : 2625	Mean : 2519	Mean : 2170
3rd Qu.: 44269	3rd Qu.: 42156.2	3rd Qu.: 3769	3rd Qu.: 3500	3rd Qu.: 3000
Max. :250327	Max. :253355.0	Max. :13942	Max. :13935	Max. :13959
pay_amt4	pay_amt5	pay_amt6	default_nextmonth	
Min. : 0	Min. : 0	Min. : 0	0:16774	
1st Qu.: 100	1st Qu.: 1	1st Qu.: 0	1: 5378	

Készítette: Képesi Szilvia / I163IX

```
Median : 1100   Median : 1100   Median : 1004
Mean    : 1970   Mean    : 1974   Mean    : 1909
3rd Qu.: 3000   3rd Qu.: 3000   3rd Qu.: 3000
Max.    :13945   Max.    :13919   Max.    :13969
```

>

```
> table(credit$sex)
```

```
female   male
 13431   8721
```

```
> xtabs(~default_nextmonth+sex,data=dataset)
```

```
              sex
default_nextmonth female  male
0             10349   6425
1              3082   2296
```

(A logisztikus regressziónál kinyert kis összefoglaló tábla.)

Itt látszik, hogy:

- a hölgyek ~30%-nál fordult elő késedelmes visszafizetés
- az urak esetében, pedig a ~36 – nál fordult elő késedelem

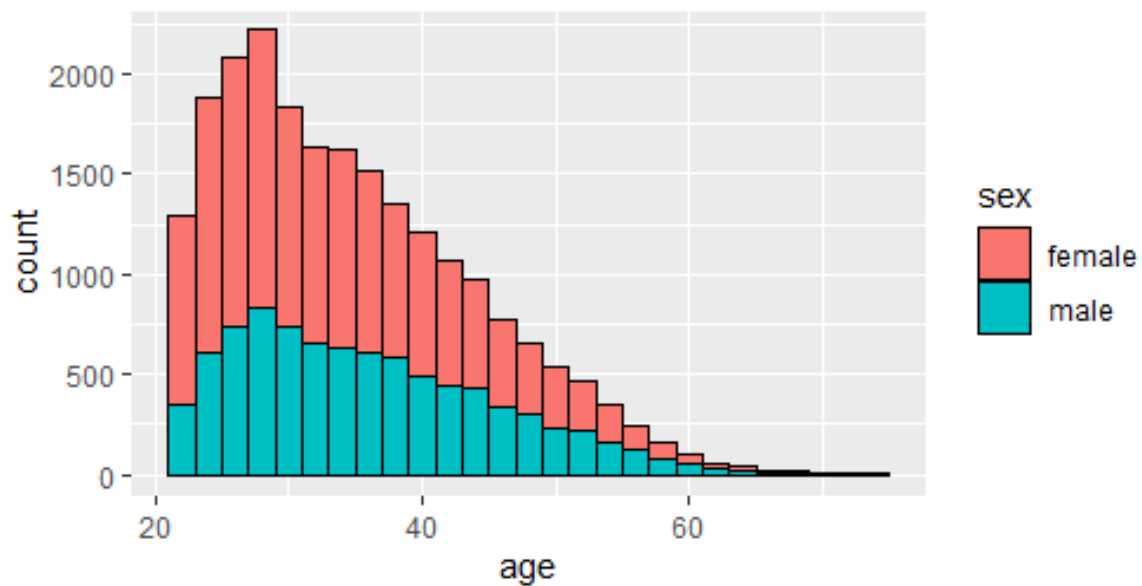
```
> table(credit$marriage)
```

```
married  other  single
 10101      249  11802
```

```
> table(credit$age)
```

```
 21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36
37  38
 60 476 759 932 947 979 1099 1036 1185 974 865 820 817 838 782 786
728 680
 39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54
55  56
 674 623 585 575 500 501 471 406 373 323 336 288 256 231 239 186
161 142
 57  58  59  60  61  62  63  64  65  66  67  68  69  70  71  72
73  75
 96  98  64  57  44  33  24  22  21  11  11  3  13  10  3  2
4 3
```

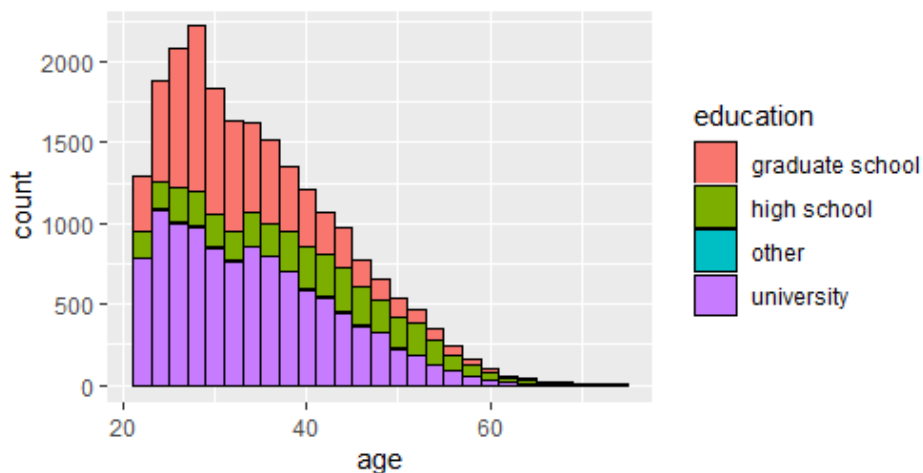
**Age és Sex arányok vizualizálva R-ban:**



Itt látszik, hogy az ügyfelek picit nagyobb arányban hölgyek, 50év felettiek körében már kiegyenlítődik.

Jól látszik, hogy a felhasználók többsége 30-as évei közepén jár.

#### Age és education arányok vizualizáva R-ban:



Itt az látszik, hogy az ügyfelek nagy arányban egyetemet végzett vagy más felsőoktatási képesítéssel rendelkezik(ha jól tudom, a graduate school felső oktatási intézményt jelent). Kis arányban rendelkeznek csak középiskolai végzettséggel(17%).

## Fizetési magatartás szerinti besorolás

PAY\_X: the repayment status of the client, X months before the data collection (-1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight months; 9 = payment delay for nine months and above) where: X = 2, 3, ..., 6

Az adatokat megtekintve a változó értékei -2 és 3 közötti értékeket vettek fel a vizsgált időszakban. Felhívnom a figyelmet, hogy a -2 -es érték nem szerepel a változó leírásában. Feltételezem, hogy minél alacsonyabb, minél inkább negatívba hajlik az érték, annál kevésbé fordul elő fizetési késedelem és annál kevésbé volt hektikus a visszafizetés. Tehát -2 egy elfogadható/jó ügyfél magatartást képvisel.

A leírás logikáját követve, maximum 3 hónapos késedelem fordult elő.

PAY\_X értékek összefoglaló táblázata

> [pay\\_X\\_summary](#)

	pay_2	pay_3	pay_4	pay_5	pay_6
-2	2780	3118	3422	3622	3946
-1	3769	3634	3545	3523	3649
0	12042	12056	12408	12666	12143
2	3321	3215	2665	2235	2306
3	227	128	112	106	108

Alapvetően azt feltételeztem, hogy a PAY\_X értéke, ami az én értelmezésemben az ügyfelek minősítése a visszafizetésre vonatkozóan, a BILL\_AMTX és a PAY\_AMTX értékeken alapul, tehát van korreláció ezen változók között. Amennyiben erős a korreláció lenne változó páronként, akkor csak a PAY\_X változóval dolgoztam volna.

PAY\_1 nem szerepel, PAY\_2, a második hónapban történt visszafizetések alapján minősítés. Itt PAY\_2 és BILL\_AMT2 korrelációt, valamint PAY\_2 és PAY\_AMT2 korrelációt vizsgáltam SPSS-ben.

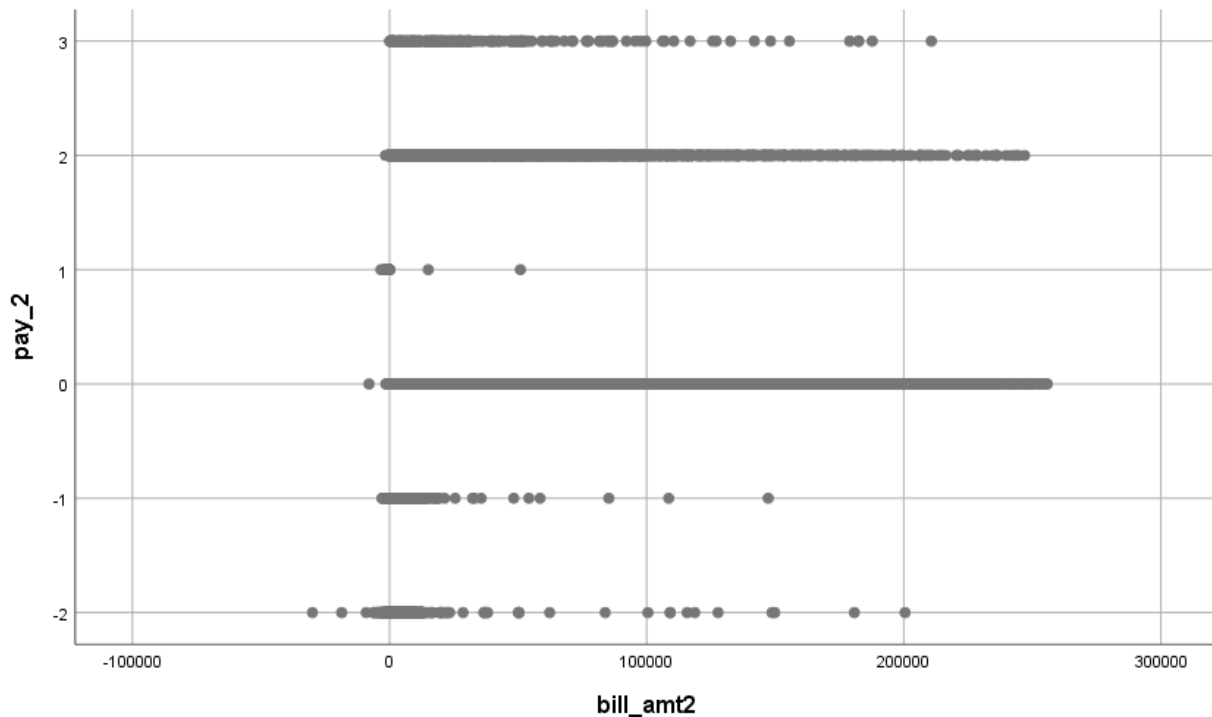
### PAY\_2 és BILL\_AMT2 vizsgálata

#### Correlations

		pay_2	bill_amt2
pay_2	Pearson Correlation	1	.287**
	Sig. (2-tailed)		.000
	Sum of Squares and Cross-products	30049.691	366905447.909
	Covariance	1.357	16563.832
	N	22152	22152
bill_amt2	Pearson Correlation	.287**	1
	Sig. (2-tailed)	.000	
	Sum of Squares and Cross-products	366905447.909	5453519354825
			9.320

Covariance	16563.832	2461974337.423
N	22152	22152

\*\* . Correlation is significant at the 0.01 level (2-tailed).

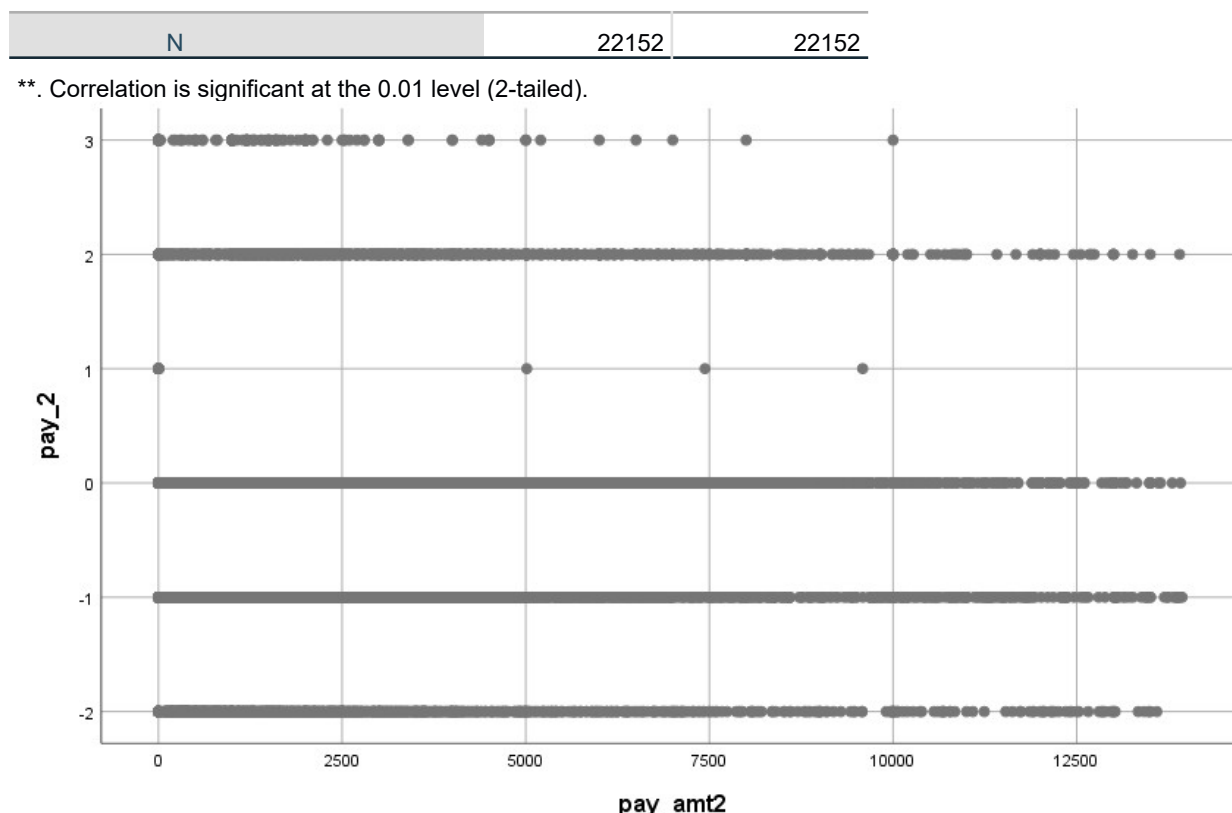


A kapcsolat szignifikáns a PAY\_2 és PAY\_AMT2 változók között, viszont gyenge a kapcsolat erőssége.

### PAY\_2 és PAY\_AMT2 vizsgálata

#### Correlations

		pay_2	pay_amt2
pay_2	Pearson Correlation	1	.084**
	Sig. (2-tailed)		.000
	Sum of Squares and Cross-products	30049.691	5567062.841
	Covariance	1.357	251.323
	N	22152	22152
pay_amt2	Pearson Correlation	.084**	1
	Sig. (2-tailed)	.000	
	Sum of Squares and Cross-products	5567062.841	147392901092.4
	Covariance	251.323	6654006.640
	N	22152	22152



A kapcsolat itt is szignifikáns a PAY\_2 és PAY\_AMT2 változók között, viszont itt nagyon gyenge a kapcsolat erőssége.

Ezt a gyakorlatot követem a további hónapokra és a következő 5hónaphoz tartozó PAY\_X BILL\_AMTX és PAY\_AMTX változók esetén. Minden hónapnál nagyjából hasonlóan alakul a korreláció erőssége. Ennél fogva azt mondanám, hogy nem lehet kizárólag a PAY\_X változóra hagyatkozni, hanem fel kell használni a BILL\_AMTX és PAY\_AMTX változókat is a modell építés során.

## Kód és eredmény kiértékelése

Maga a két R kód kommentelve található a Moodle-ben feltöltött zip fájlban.

Külön zip-ként mellékelem az egyéb próbálkozásokat, ahol próbáltam kevesebb változóval dolgozni. Ezeknek a verzióknak a kiértékelését mellőzném.

A logisztikus regresszió-ra készült munkámat csatolom azzal a céllal, hogy a visszajelzés segíthet a fennálló hibaüzenet feltárásában.

## Döntési fa

*dontesi\_fa.R*



Készítette: Képesi Szilvia / I163IX

Elemzés lépési, főbb pontok kiemelve(többi kommentelve a kódban):

**Model elkészítése a credit\_train-re(training adat):**

- 18000 esetet vettünk a training set-ben.

**- A modell a 18000-ből 3846 rekord kivételével (21,4%) helyesen sorolta be a rekordokat.**

- nyilván két class-t eredményez, ebből class „a”, akiknél nem volt késdelem, class „b”, akiknél volt késdelem.

**- 735 false positive, 1230 false negative érték van**

- A döntési fából látszik, hogy mely változókat használta( 100.00% pay\_2, 15.74% pay\_6, 9.89% limit\_bal, 9.20% pay\_3).

```
> summary(credit_model)
```

```
call:
```

```
C5.0.default(x = credit_train[-24], y = credit_train$default_nextmonth)
```

```
C5.0 [Release 2.07 GPL Edition]
```

```
Sun Apr 19 18:47:40 2020
```

```
-----  
Class specified by attribute `outcome'
```

```
Read 18000 cases (24 attributes) from undefined.data
```

```
Decision tree:
```

```
pay_2 <= 1: 0 (15166/2746)  
pay_2 > 1:  
: ... pay_6 > 0: 1 (1054/326)  
    pay_6 <= 0:  
    : ... limit_bal > 260000: 0 (124/37)  
        limit_bal <= 260000:  
        : ... pay_3 <= 1: 0 (745/328)  
            pay_3 > 1: 1 (911/409)
```

```
Evaluation on training data (18000 cases):
```

```
      Decision Tree  
-----  
Size      Errors  
  
5 3846(21.4%)  <<
```

```

      (a)   (b)   <-classified as
      ----  ----
12924   735   (a): class 0
3111   1230   (b): class 1

```

Attribute usage:

```

100.00% pay_2
 15.74% pay_6
   9.89% limit_bal
   9.20% pay_3

```

### Keresztábra elkészítése

Credit model segítségével prediktáltuk a `credit_test`-re (test adatok) kimenetelét, ezt mentettük a `credit_pred` változóba. Amelyet a keresztábra segítségével összevetjük a valós és a prediktált `default_nextmonth` értékeket.

- 4152 esetet vettünk a test set-be

- 3115db 0-as valós értékből 2934db-ot becsült helyesen 0-asnak és 181-et nem helyesen 1-esnek a model(false pos.). 1037Db 1-es valós értékből 318-at helyesen 1-esnek és 719-et nem helyesen 0-asnak(false neg.).

**Accuracy rate**= Correct/Total  $\rightarrow (2934+318)/4152=0,7832 \rightarrow \mathbf{78,32\%}$

**Error rate**=Error/Total  $\rightarrow (719+181)/4152=0,2167 \rightarrow \mathbf{21,67\%}$

**False postive rate**=181/3115=0,0581  $\rightarrow \mathbf{5,81\%}$

**False negative rate**=719/1037=0,6933  $\rightarrow \mathbf{69,33\%}$

**A modell a valós késedelemmel visszafizetők közül, ami 1037 összesen csak 318-at jelez, azaz 30,67%-ot.**

**Ebben az esetben nagyon nagy a false negatív értékek aránya!**

```

> #create crosstable
> CrossTable(credit_test$default_nextmonth, credit_pred, prop.chisq =
FALSE, prop.c = FALSE, prop.r = FALSE, dnn = c('actual default',
'predicted default'))

```

```

Cell Contents
|-----|
|                                     N |
|      N / Table Total               |
|-----|

```

Total observations in Table: 4152

	predicted default		
actual default	0	1	Row Total
0	2934	181	3115
	0.707	0.044	
1	719	318	1037
	0.173	0.077	
Column Total	3653	499	4152

### Boosted model előállítása

Trial	Decision Tree	
	Size	Errors
0	5 3846(21.4%)	
1	2 4122(22.9%)	
2	3 4196(23.3%)	
3	2 5711(31.7%)	
4	5 7158(39.8%)	
5	4 4553(25.3%)	
6	7 5167(28.7%)	
7	4 4851(26.9%)	
8	2 3989(22.2%)	
9	2 4129(22.9%)	
boost	3857(21.4%)	<<

(a)	(b)	<-classified as
12910	749	(a): class 0
3108	1233	(b): class 1

### Attribute usage:

100.00% pay\_2  
 100.00% pay\_3  
 100.00% pay\_4  
 100.00% pay\_5  
 100.00% bill\_amt1  
 100.00% pay\_amt1  
 95.24% pay\_6  
 81.34% limit\_bal  
 51.28% bill\_amt3  
 49.72% pay\_amt3

30.87% age

49.72% pay\_amt3

30.87% age

### Boosted model keresztábrája

- 3115db 0-as valós értékből 2932db-ot becsült helyesen 0-asnak és 183-et nem helyesen 1-esnek a model. 1037Db 1-es valós értékből 300-at helyesen 1-esnek és 737-et nem helyesen 0-asnak.

**Accuracy rate**= Correct/Total  $\rightarrow (2932+300)/4152=0,7784 \rightarrow 77,84\%$

**Error rate**=Error/Total  $\rightarrow (737+183)/4152=0,2215 \rightarrow 22,15\%$

**False postive rate**=183/3115=0,0587  $\rightarrow 5,87\%$

**False negative rate**=737/1037=0,7107  $\rightarrow 71,07\%$

**A modell a valós késedelemmel visszafizetők közül, ami 1037 összesen csak 300-at jelez, azaz 28,92%-ot.**

**A boostolás hatására növekedett a false negatív értékek aránya!**

**> #boost result**

```
> credit_boost_pred10 <- predict(credit_boost10, credit_test)
> CrossTable(credit_test$default_nextmonth, credit_boost_pred10,
+ prop.chisq = FALSE, prop.c = FALSE, prop.r = FALSE,
+ dnn = c('actual default', 'predicted default'))
```

Cell Contents

	N
N / Table Total	

Total Observations in Table: 4152

	predicted default		
actual default	0	1	Row Total
0	2932 0.706	183 0.044	3115
1	737 0.178	300 0.072	1037
Column Total	3669	483	4152

### Cost matrix bevezetése és utána a kereszttábla alakulása

- 3115db 0-as valós értékből 1710db-ot becsült helyesen 0-asnak(54%) és 1405-et nem helyesen 1-esnek a model. 1037Db 1-es valós értékből 760-at helyesen(73%) 1-esnek és 277-et nem helyesen 0-asnak.

**Accuracy rate**= Correct/Total  $\rightarrow (1710+760)/4152=0,5948 \rightarrow$  **59,48%**

**Error rate**=Error/Total  $\rightarrow (277+1405)/4152=0,4051 \rightarrow$  **40,51%**

**False postive rate**=1405/3115=0,4510  $\rightarrow$  **45,10%**

**False negative rate**=277/1037=0,2671  $\rightarrow$  **26,71%**

**A modell a valós késedelemmel visszafizetők közül, ami 1037 összesen csak 760-at jelez, azaz 73,28%-ot.**

**A cost matrix alkalmazásának hatására 26,71%-ra csökkent a false negative becslések aránya!**

```
> credit_cost_pred <- predict(credit_cost, credit_test)
> CrossTable(credit_test$default_nextmonth, credit_cost_pred,
+ prop.chisq = FALSE, prop.c = FALSE, prop.r = FALSE,
+ dnn = c('actual default', 'predicted default'))
```

Cell Contents

		N
N / Table Total		

Total Observations in Table: 4152

actual default	predicted default		Row Total
	0	1	
0	1710 0.412	1405 0.338	3115
1	277 0.067	760 0.183	1037
Column Total	1987	2165	4152

>

### Döntési fa módszerrel elért eredmények konklúziója

A **model 1** nagyon szép arányban eltalálja a valós adatokat, viszont sajnos nagyon nagy, 69,33% a false negative becslések aránya. Azaz **315 késedelmes fizetést jelzett előre az 1037-ből.**

A boosted verzió rontott az eredményen.

A **cost matrix**os verziónál a helyesen eltalált eredmények aránya nagyban csökkent viszont 760-at jelzett előre az 1037 késedelmes fizetésből, ami **73,28%.**

**Mivel a bedőlt hiteleknek van a legnagyobb költsége, így az a cél, hogy minél nagyobb pontossággal előrejelezzük a késedelmet, ezt pedig a cost matrix-os verzióval tudjuk elérni.**

A 3 variációból ezt választanám, viszont nem vagyok teljesen megelégedve a 73,28%-kal, úgyhogy valós munka esetén valós piaci helyzetben biztosan több modellt készítenék. A jelenlegi beadandó munka keretén belül a másik modelnek a logisztikus regressziót választottam, az eddig tanultak alapján.

### Logisztikus regresszió

*logisztikus\_reg.R*

*Elemzés lépési, főbb pontok kiemelve(többi kommentelve a kódban):*

#### Model elkészítése:

Az output mutatja a koefficienseket, sztenderd hibájukat, a z-statisztikákat, coefficients, standard errors, the z-statistic (Wald z-statistic), és a hozzájuk tartozó p értékeket.

A logisztikai regressziós koefficiensek jelzik a változást a log esélyekben (függő változó) a prediktor változók egységnyi növekedése esetén.

Első körben a p érték alapján szignifikáns koefficienseket szűrtem, majd felül vizsgáltam a z -score alapján.

(Ha a p érték nagyobb, mint 0,05, az nem azt jelenti, hogy a változó nem játszik szerepet, hanem azt, hogy kicsi a minta, vagy nagy a minta szórása. A Z-score az adott koefficiens értéke osztva annak szórásával, ha abszolút értéke nagy, akkor a koefficiens jelentősen különbözik a nullától. Ha egy koefficiens nem szignifikáns, de a z-score abszolút értéke nagy, akkor figyelembe kell venni. Itt az elfogadott szabály az, hogy ha  $z > 2$ , vagy  $z < -2$ , akkor figyelembe kell venni. A Z előjel és a p érték nincs kapcsolatban, egy koefficiens lehet pozitív is, és negatív is, a p pedig azt fejezi ki, hogy ez az érték mennyire szignifikáns. Ha  $p < 0,05$ , akkor a koefficiens elfogadjuk, mert  $> 0,95$  valószínűséggel nem teljesül az a hipotézis, hogy az értéke nulla.)

Alább a sorok közé írva elemzem az egyes változók koefficienseit és azok hatását:

Coefficients:

Estimate Std. Error z value Pr(>|z|)

Készítette: Képesi Szilvia / I163IX

(Intercept)	-4.995e-01	9.083e-02	-5.499	3.82e-08	***
sexmale	1.032e-01	3.408e-02	3.028	0.00246	**

A sexmale koefficiensének egységnyi változásával a késedelmes fizetés esélyének logaritmus 1.032e-01-vel nő.

educationhigh school	-1.299e-01	5.116e-02	-2.540	0.01109	*
----------------------	------------	-----------	--------	---------	---

A educationhigh school koefficiensének egységnyi változásával a késedelmes fizetés esélyének logaritmus 1.299e-01 csökken.

educationother	-9.116e-01	4.310e-01	-2.115	0.03441	*
----------------	------------	-----------	--------	---------	---

A educationother koefficiensének egységnyi változásával a késedelmes fizetés esélyének logaritmus 9.116e-01 csökken.

educationuniversity	-9.949e-02	3.885e-02	-2.561	0.01044	*
---------------------	------------	-----------	--------	---------	---

A educationuniversity koefficiensének egységnyi változásával a késedelmes fizetés esélyének logaritmus 9.949e-02 csökken.

marriagesingle	-1.748e-01	3.842e-02	-4.549	5.39e-06	***
----------------	------------	-----------	--------	----------	-----

A marriagesingle koefficiensének egységnyi változásával a késedelmes fizetés esélyének logaritmus 1.748e-01 csökken.

pay_2	3.192e-01	2.251e-02	14.180	< 2e-16	***
-------	-----------	-----------	--------	---------	-----

A pay\_2 koefficiensének egységnyi változásával a késedelmes fizetés esélyének logaritmus 3.192e-01 nő.

pay_4	1.369e-01	2.951e-02	4.640	3.49e-06	***
-------	-----------	-----------	-------	----------	-----

A pay\_4 koefficiensének egységnyi változásával a késedelmes fizetés esélyének logaritmus 1.369e-01 nő.

pay_6	3.664e-02	2.583e-02	1.419	0.15598	
-------	-----------	-----------	-------	---------	--

A pay\_6 koefficiensének egységnyi változásával a késedelmes fizetés esélyének logaritmus 3.664e-02 nő.

bill_amt1	-1.223e-05	1.933e-06	-6.328	2.49e-10	***
-----------	------------	-----------	--------	----------	-----

A bill\_amt1 koefficiensének egységnyi változásával a késedelmes fizetés esélyének logaritmus 1.223e-05 csökken.

bill_amt2	5.143e-06	2.621e-06	1.963	0.04967	*
-----------	-----------	-----------	-------	---------	---

A bill\_amt2 koefficiensének egységnyi változásával a késedelmes fizetés esélyének logaritmus 5.143e-06 nő.

bill_amt3	5.734e-06	2.381e-06	2.409	0.01602	*
-----------	-----------	-----------	-------	---------	---

Készítette: Képesi Szilvia / I163IX

A bill\_amt3 koefficiensének egységnyi változásával a késedelmes fizetés esélyének logaritmus 5.734e-06 nő.

bill\_amt5 -8.173e-06 3.231e-06 -2.529 0.01142 \*  
A bill\_amt5 koefficiensének egységnyi változásával a késedelmes fizetés esélyének logaritmus 8.173e-06 csökken.

bill\_amt6 1.215e-05 2.738e-06 4.439 9.02e-06 \*\*\*  
A bill\_amt6 koefficiensének egységnyi változásával a késedelmes fizetés esélyének logaritmus 1.215e-05 nő.

pay\_amt1 -9.262e-05 1.117e-05 -8.292 < 2e-16 \*\*\*  
A pay\_amt1 koefficiensének egységnyi változásával a késedelmes fizetés esélyének logaritmus 9.262e-05 csökken.

pay\_amt2 -8.740e-05 1.122e-05 -7.794 6.51e-15 \*\*\*  
A pay\_amt2 koefficiensének egységnyi változásával a késedelmes fizetés esélyének logaritmus 8.740e-05 csökken.

pay\_amt3 -6.713e-05 1.159e-05 -5.793 6.91e-09 \*\*\*  
A pay\_amt3 koefficiensének egységnyi változásával a késedelmes fizetés esélyének logaritmus 6.713e-05 csökken.

pay\_amt4 -3.674e-05 1.161e-05 -3.165 0.00155 \*\*  
A pay\_amt4 koefficiensének egységnyi változásával a késedelmes fizetés esélyének logaritmus 3.674e-05 csökken.

pay\_amt5 -3.337e-05 1.083e-05 -3.082 0.00206 \*\*  
A pay\_amt5 koefficiensének egységnyi változásával a késedelmes fizetés esélyének logaritmus 3.337e-05 csökken.

pay\_amt6 -4.348e-05 1.048e-05 -4.149 3.34e-05 \*\*\*  
A pay\_amt6 koefficiensének egységnyi változásával a késedelmes fizetés esélyének logaritmus 4.348e-05 csökken.

Az én értelmezésemben a pirossal megjelölt változók tűnnek a legbefolyásosabbaknak: pay\_2, bill\_amt1, pay\_amt1, pay\_amt2, pay\_amt3

SEXMALE koefficiense növelő hatású, az összes EDUCATION változó koefficiensei csökkentő, valamint a MARRIGAE SINGLE szintén csökkentő hatású.

A PAY\_X változó koefficiensei mindig növelő hatással bírnak. Ebből a PAY\_2 szignifikanciája tűnik a legmagasabbnak.

A BILL\_AMTX, kivéve az első hónaphoz tartozó változó, mindig növelő hatással bírnak.

A PAY\_AMTX változó koefficiensei mindig csökkentő hatással bírnak.



```
> model1<-
glm(default_nextmonth~sex+education+marriage+age+pay_2+pay_3+pay_4+pay_5+
pay_6+bill_amt1+bill_amt2+bill_amt3+bill_amt4+bill_amt5+bill_amt6+pay_amt
1+pay_amt2+pay_amt3+pay_amt4+pay_amt5+pay_amt6,
data=dataset, family="binomial")
> summary(model1)
```

call:

```
glm(formula = default_nextmonth ~ sex + education + marriage +
age + pay_2 + pay_3 + pay_4 + pay_5 + pay_6 + bill_amt1 +
bill_amt2 + bill_amt3 + bill_amt4 + bill_amt5 + bill_amt6 +
pay_amt1 + pay_amt2 + pay_amt3 + pay_amt4 + pay_amt5 + pay_amt6,
family = "binomial", data = dataset)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.6599	-0.7572	-0.5877	-0.2298	3.1491

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )	
(Intercept)	-4.995e-01	9.083e-02	-5.499	3.82e-08	***
sexmale	1.032e-01	3.408e-02	3.028	0.00246	**
educationhigh school	-1.299e-01	5.116e-02	-2.540	0.01109	*
educationother	-9.116e-01	4.310e-01	-2.115	0.03441	*
educationuniversity	-9.949e-02	3.885e-02	-2.561	0.01044	*
marriageother	-2.545e-03	1.489e-01	-0.017	0.98636	
marriagesingle	-1.748e-01	3.842e-02	-4.549	5.39e-06	***
age	3.984e-03	2.032e-03	1.961	0.04991	*
pay_2	3.192e-01	2.251e-02	14.180	< 2e-16	***
pay_3	4.721e-02	2.737e-02	1.725	0.08457	.
pay_4	1.369e-01	2.951e-02	4.640	3.49e-06	***
pay_5	5.098e-02	3.172e-02	1.607	0.10803	
pay_6	3.664e-02	2.583e-02	1.419	0.15598	
bill_amt1	-1.223e-05	1.933e-06	-6.328	2.49e-10	***
bill_amt2	5.143e-06	2.621e-06	1.963	0.04967	*
bill_amt3	5.734e-06	2.381e-06	2.409	0.01602	*
bill_amt4	2.822e-06	2.335e-06	1.208	0.22690	
bill_amt5	-8.173e-06	3.231e-06	-2.529	0.01142	*
bill_amt6	1.215e-05	2.738e-06	4.439	9.02e-06	***
pay_amt1	-9.262e-05	1.117e-05	-8.292	< 2e-16	***
pay_amt2	-8.740e-05	1.122e-05	-7.794	6.51e-15	***
pay_amt3	-6.713e-05	1.159e-05	-5.793	6.91e-09	***
pay_amt4	-3.674e-05	1.161e-05	-3.165	0.00155	**
pay_amt5	-3.337e-05	1.083e-05	-3.082	0.00206	**
pay_amt6	-4.348e-05	1.048e-05	-4.149	3.34e-05	***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

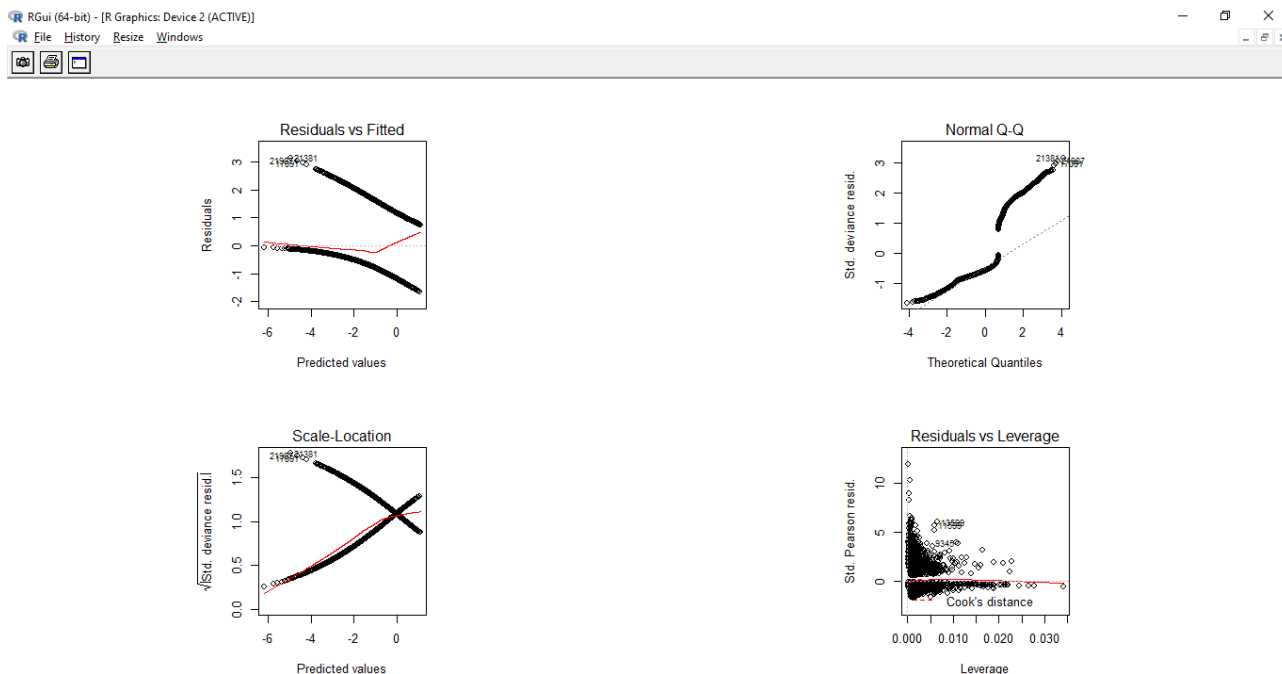
Null deviance: 24556 on 22151 degrees of freedom  
Residual deviance: 22265 on 22127 degrees of freedom

Készítette: Képesi Szilvia / I163IX

AIC: 22315

Number of Fisher Scoring iterations: 5

## Model1 vizualizáció:



Afenti ábrák közül a feladathoz szorosabban kapcsolódó a Q-Q plot ábra. A többi ábrára nem térnék ki.

A q-q plot az elméleti kvantilisok és a mintából számolt kvantilisok pont ábrája. Akkor lesz a modell teljesen pontos, ha az ábrán egy 45 fokos egyenest látunk. Az eltérés mértéke azt jelzi, hogy a modell mennyire pontatlan. A modellhez tartozó alsó ponthalmaz jól illeszkedik, viszont a felső szakasz pont halmazai jelentősen távol helyezkednek a 45 fokos szöveget ábrázoló egyenestől, így feltételezhetően nem teljesen pontos becslést nyújt a modell.

## Odd ratios

Az odds ratio megmutatja, hogy az adott változó egységnyi változása hányszorosára növeli a késedelmes fizetés esélyét. Ez egy kicsit beszédesebb, mint a koefficiensek értelmezése.

A kapott eredményeket rendeztem az odds ratios csökkenő sorrendjében:

```
> ORC_SUM[order(ORC_SUM$OR, decreasing=TRUE),]
      OR      X2.5..      X97.5..
pay_2  1.3760705  1.3166698  1.4381581
pay_4  1.1467413  1.0821913  1.2149339
sexmale 1.1087054  1.0370129  1.1852293
pay_5  1.0522975  0.9887900  1.1197171
pay_3  1.0483442  0.9934541  1.1059955
```

Készítette: Képesi Szilvia / I163IX

pay_6	1.0373203	0.9860563	1.0911241
age	1.0039915	0.9999958	1.0079919
bill_amt6	1.0000122	1.0000069	1.0000176
bill_amt3	1.0000057	1.0000011	1.0000104
bill_amt2	1.0000051	1.0000000	1.0000103
bill_amt4	1.0000028	0.9999982	1.0000074
bill_amt5	0.9999918	0.9999854	0.9999981
bill_amt1	0.9999878	0.9999839	0.9999915
pay_amt5	0.9999666	0.9999452	0.9999877
pay_amt4	0.9999633	0.9999403	0.9999858
pay_amt6	0.9999565	0.9999358	0.9999769
pay_amt3	0.9999329	0.9999099	0.9999554
pay_amt2	0.9999126	0.9998904	0.9999344
pay_amt1	0.9999074	0.9998853	0.9999291
marriageother	0.9974583	0.7411403	1.3296132
educationuniversity	0.9053014	0.8389632	0.9769817
educationhigh school	0.8781575	0.7942127	0.9705884
marriagesingle	0.8396564	0.7787331	0.9053090
(Intercept)	0.6068518	0.5079270	0.7251643
educationother	0.4018747	0.1544254	0.8617524

A koefficiens, p value és z score alapján ugye én ezt gondoltam: **pay\_2, bill\_amt1, pay\_amt1, pay\_amt2, pay\_amt3**

Ehhez képest az odds ratios szerint ezek a legbefolyásosabbak: **pay\_2, pay\_4, sexmale, pay\_5, pay\_3**

## Predikció

A predict függvény és a model 1 segítségével prediktáltam a default\_nextmonth értékeit. A leadott anyagban CSV formájában becsatoltam(log\_reg\_probpred.csv)

## Model tesztelése null modellhez hasonlítva

A Chi-square: 2290,587,df: 24, p<0,05, LR: 11132.67(df:25)

A null modellhez képest mindenképp egy jó választás! Nagyon magas a likelihood rate.

```
#testing model1 against null model
> with(model1, null.deviance - deviance)
[1] 2290.587
> with(model1, df.null - df.residual)
[1] 24
> with(model1, pchisq(null.deviance - deviance, df.null - df.residual,
lower.tail = FALSE))
[1] 0
> logLik(model1)
'log Lik.' -11132.67 (df=25)
```

Készítette: Képesi Szilvia / I163IX

Sajnos a döntési fa modellel nem sikerült összehasonlíttanom a fenti módszerrel, mert a fenti ad függvények nem kezelik a c50-es objektumokat.

## + Kiegészítés

### Logisztikus regresszió – Gretl verzió

Szerettem volna egy másik megoldást is látni logisztikus regresszióra ezért Gretl-ben is létrehoztam egy logisztikus regressziós modellt.

Gretl: <http://gretl.sourceforge.net/>

#### Model elkészítése

A létrehozott logisztikus regresszió model 77,9%-ban képes korrektül megbecsülni a a default\_nextmonth értékét.

**Accuracy rate**= Correct/Total  $\rightarrow (16201+1058)/22152=0,7791 \rightarrow 77,91\%$

**Error rate**=Error/Total  $\rightarrow (4320+573)/22152=0,2208 \rightarrow 22,08\%$

Model 1: Logit, using observations 1-22152

Dependent variable: default\_nextmonth

Standard errors based on Hessian

	coefficient	std. error	z	p-value	
-----					
const	-0.428786	0.126198	-3.398	0.0007	***
limit_bal	-2.10273e-07	1.80099e-07	-1.168	0.2430	
sex	0.101048	0.0340370	2.969	0.0030	***
education	-0.0504510	0.0194768	-2.590	0.0096	***
marriage	-0.145973	0.0355737	-4.103	4.07e-05	***
age	0.00443782	0.00191421	2.318	0.0204	**
pay_2	0.316271	0.0225982	14.00	1.66e-044	***
pay_3	0.0454140	0.0274045	1.657	0.0975	*
pay_4	0.136065	0.0295404	4.606	4.10e-06	***
pay_5	0.0479947	0.0317551	1.511	0.1307	
pay_6	0.0350932	0.0259159	1.354	0.1757	
bill_amt1	-1.21402e-05	1.93564e-06	-6.272	3.57e-010	***
bill_amt2	5.04799e-06	2.62251e-06	1.925	0.0542	*

Készítette: Képesi Szilvia / I163IX

bill_amt3	5.64637e-06	2.38602e-06	2.366	0.0180	**
bill_amt4	2.84929e-06	2.33915e-06	1.218	0.2232	
bill_amt5	-8.05005e-06	3.23307e-06	-2.490	0.0128	**
bill_amt6	1.21341e-05	2.73774e-06	4.432	9.33e-06	***
pay_amt1	-9.23768e-05	1.11881e-05	-8.257	1.50e-016	***
pay_amt2	-8.69745e-05	1.12297e-05	-7.745	9.56e-015	***
pay_amt3	-6.62729e-05	1.15986e-05	-5.714	1.10e-08	***
pay_amt4	-3.59876e-05	1.16114e-05	-3.099	0.0019	***
pay_amt5	-3.23303e-05	1.08342e-05	-2.984	0.0028	***
pay_amt6	-4.22394e-05	1.05031e-05	-4.022	5.78e-05	***

Mean dependent var 0.242777 S.D. dependent var 0.428771  
McFadden R-squared 0.092881 Adjusted R-squared 0.091007  
Log-likelihood -11137.58 Akaike criterion 22321.17  
Schwarz criterion 22505.30 Hannan-Quinn 22381.11

Number of cases 'correctly predicted' = 17259 (77.9%)

f(beta'x) at mean of independent vars = 0.169

Likelihood ratio test: Chi-square(22) = 2280.77 [0.0000]

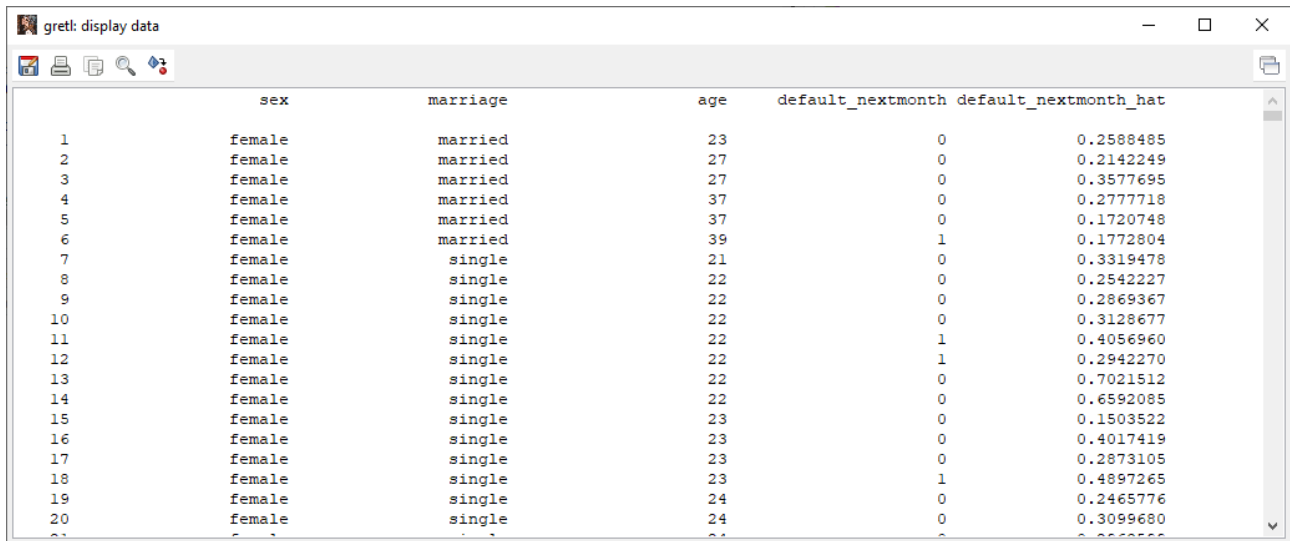
		Predicted	
		0	1
Actual	0	16201	573
	1	4320	1058

Excluding the constant, p-value was highest for variable 1 (limit\_bal)

## Prediction

A prediktált default\_nextmonth\_hat elmenthető változóként, majd az eredeti változókkal(vagy azok egy részével) megjeleníthető. Így látszik egymás mellett az eredeti default\_nextmonth és a prediktált default\_nextmonth\_hat.

Készítette: Képesi Szilvia / I163IX



	sex	marriage	age	default_nextmonth	default_nextmonth_hat
1	female	married	23	0	0.2588485
2	female	married	27	0	0.2142249
3	female	married	27	0	0.3577695
4	female	married	37	0	0.2777718
5	female	married	37	0	0.1720748
6	female	married	39	1	0.1772804
7	female	single	21	0	0.3319478
8	female	single	22	0	0.2542227
9	female	single	22	0	0.2869367
10	female	single	22	0	0.3128677
11	female	single	22	1	0.4056960
12	female	single	22	1	0.2942270
13	female	single	22	0	0.7021512
14	female	single	22	0	0.6592085
15	female	single	23	0	0.1503522
16	female	single	23	0	0.4017419
17	female	single	23	0	0.2873105
18	female	single	23	1	0.4897265
19	female	single	24	0	0.2465776
20	female	single	24	0	0.3099680

Kapcsolódó pdf fájlokat, illetve a gretl source fájlt csatoltam a leadandó anyagok között. Abban a részletek megtekinthetők.