PREDIKCIÓS MODELLEZÉS ÉS KÖVETKEZTETÉSEI EGY SALES ADATBÁZISON

OTP beadandó feladat

Készítette: Menyhért Kristóf
menyhert.kristof@gmail.com
https://krinya.github.io/

2018.05.22.

ELŐZMÉNYEK, A PROBLÉMA MEGFOGALMAZÁSA

- Az OTP Data Scientist pozícióra pályázva első körben egy predikciós feladatot kaptam. Ennek a folyamatát és eredményeit ismertetem a prezentációban.
- Egy **adatbázist** kaptam egy marketing kampányról:
  - egyéni tulajdonságok + megvette-e az ajánlott terméket
  - Valószínűleg telefonos sales tevékenységet folytattak egy a bank által forgalmazott kötvény eladása érdekében.
- **Az elemzésem célja**, egy olyan modell megalkotása az adatokon, ami az adatbázisban található egyéni tulajdonságok alapján előre jelzi, hogy kik azok az egyének és milyen tulajdonság jellemző rájuk, aki nagy arányban vásárolni fognak az ajánlott kötvényből.
- Az eredmények ismeretesé után **üzleti ajánlásokat** is teszek valamint kijelölök további lehetséges fejlesztési irányokat

AZ ADATBÁZIS BEMUTATÁSA

- Általános képalkotás az adatbázisról
- Folyamatos változók az adatbázisban
  - Hisztogram
  - Módosításaik
- Kategorikus változók az adatbázisban
  - Hisztogram
    - Y magyarázott változó
  - Módosítások

AZ ADATBÁZIS BEMUTATÁSA: ÁLTALÁNOS KÉP

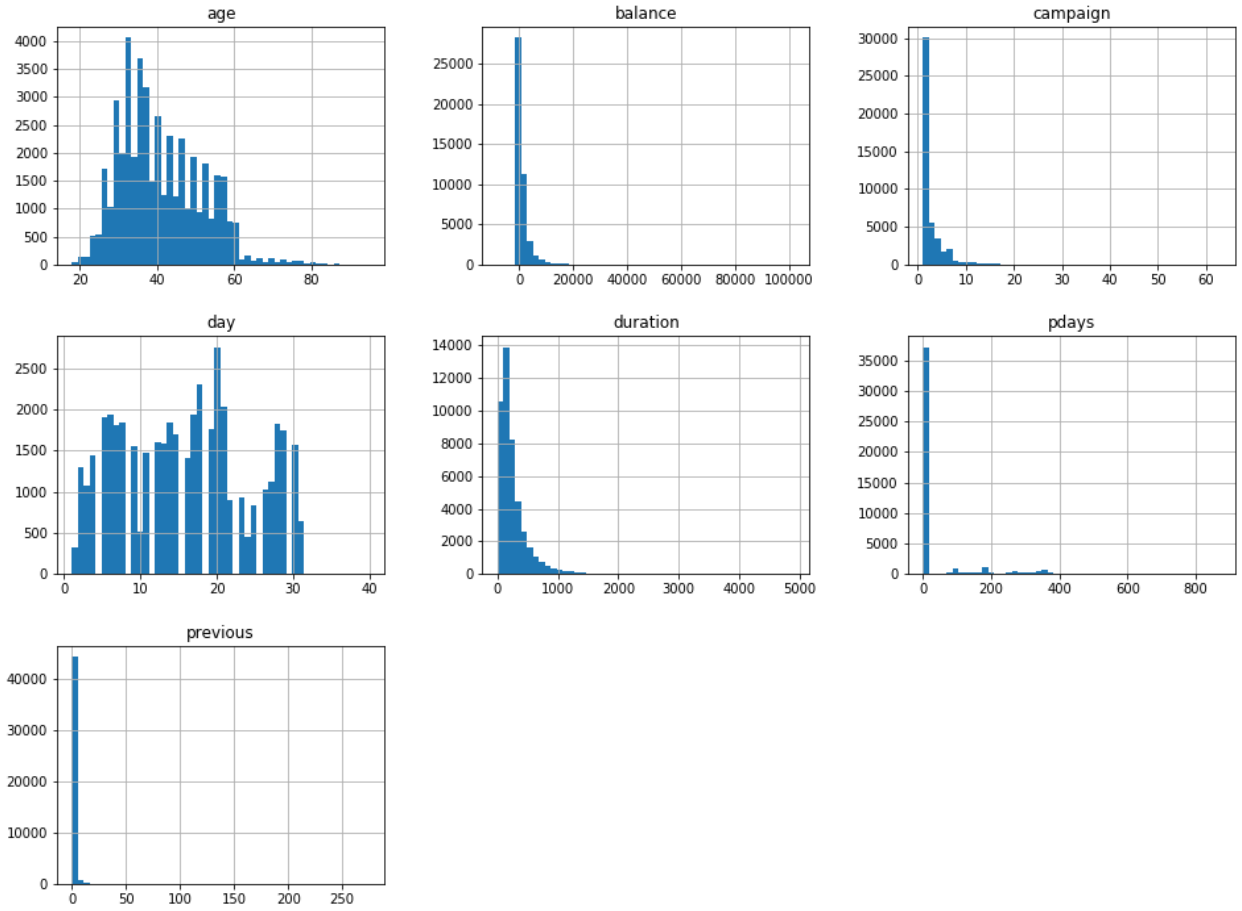
- Az adatbázis kezdetben 45 211 db megfigyelést és 17db változót tartalmazott → néhány hiányos
- **végül: 45202 db megfigyelés és 17 változó**
- Minden **sora** egy megfigyelés
- Minden **oszlopa** egy az **adott egyénre jellemző** adatot tartalmaz (lásd a képet alul):
  - **Magyarázó változók**: age, job, marital, education, default, balance, housing, loan, contact, day, month, duration, campaign, pdays, previous, poutcome
  - **Magyarázott változó**: y – megvette e a kötvényt vagy nem (yes/no)

age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
58	management	married	tertiary	no	2143	yes	no	unknown	5	may	261	1	-1	0	unknown	no
44	technician	single	secondary	no	29	yes	no	unknown	5	may	151	1	-1	0	unknown	no
33	entrepreneur	married	secondary	no	2	yes	yes	unknown	5	may	76	1	-1	0	unknown	no
47	blue-collar	married	unknown	no	1506	yes	no	unknown	5	may	92	1	-1	0	unknown	no
33	unknown	single	unknown	no	1	no	no	unknown	5	may	198	1	-1	0	unknown	no

AZ ADATBÁZIS BEMUTATÁSA: VÁLTOZÓK

- A következő diákon először a **folytonos** majd a **kategorikus** változók hisztogramja látható
- Mellettük a hisztogramok alapján elvégzett **transzformációk és/vagy módosítások** láthatók szövegesen

AZ ADATBÁZIS BEMUTATÁSA: FOLYTONOS VÁLTOZÓK



age:

* log(age) → normalizálás

duration

* log(duration + 1) → normalizálás

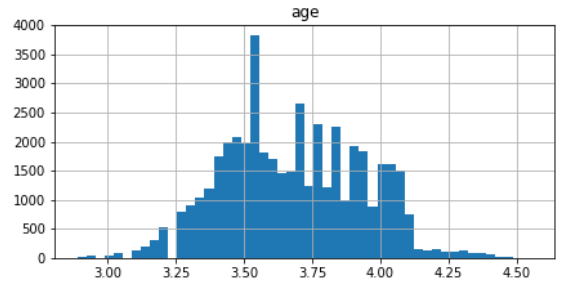
day:

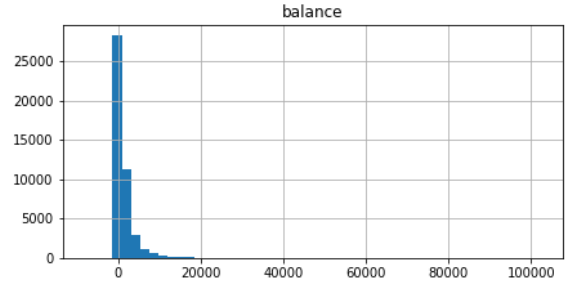
* 40. napos megfigyelés is van (3db) → nem lehet→ átrakom a 31. napra
* Kategóriák képzése: Hónap eleje; Hónap közepe; Hónap vége
*Kategorikus változóként szerepeltetem

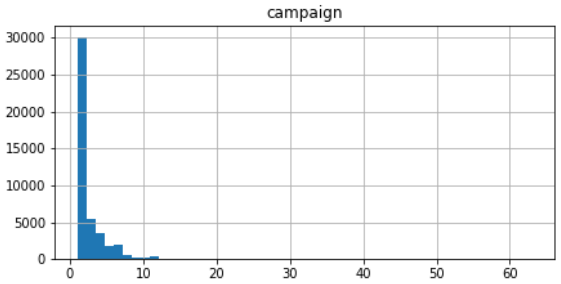
previous:

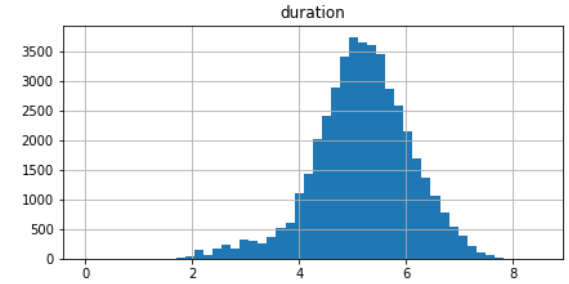
*Kategóriák kialakítása: 0; 1; 2, ;3; 4; 5+
*Kategorikus változóként szerepeltetem

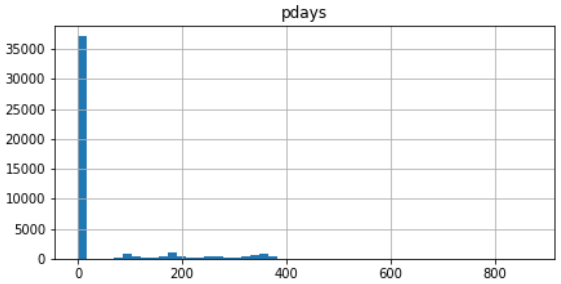
AZ ADATBÁZIS BEMUTATÁSA: FOLYTONOS VÁLTOZÓK











age:

* log(age) → normalizálás

duration

* log(duration + 1) → normalizálás

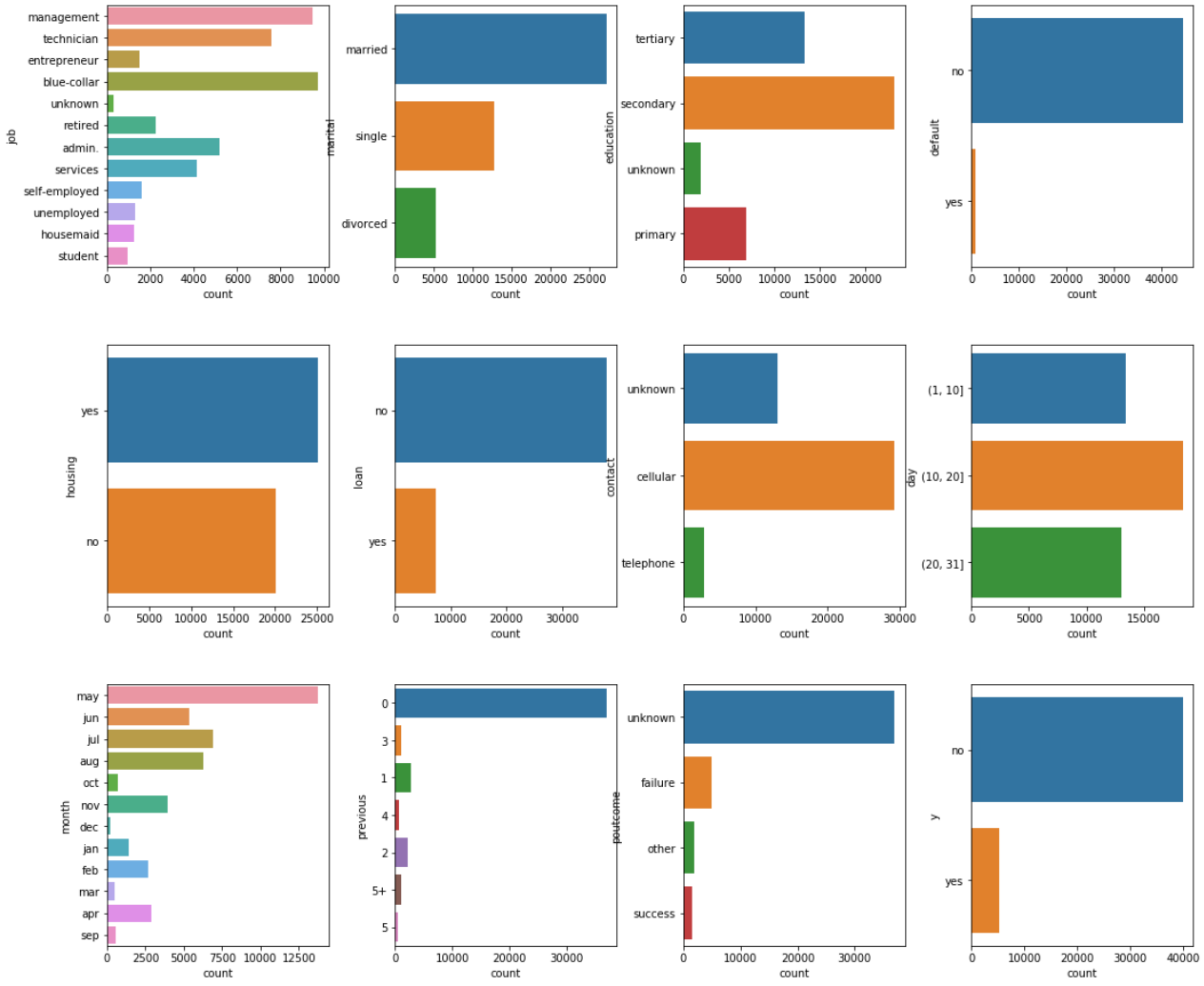
day:

* 40. napos megfigyelés is van (3db) → nem lehet→ átrakom a 31. napra
* Kategóriák képzése: Hónap eleje; Hónap közepe; Hónap vége
* *Kategorikus változóként szerepeltetem

previous:

* *Kategóriák kialakítása: 0; 1; 2, ;3; 4; 5+
* *Kategorikus változóként szerepeltetem

AZ ADATBÁZIS BEMUTATÁSA: KATEGORIKUS VÁLTOZÓK



Köztük a **prediktálni szándékozott Y** változó (jobb alsó sarok)

*Unbalanced

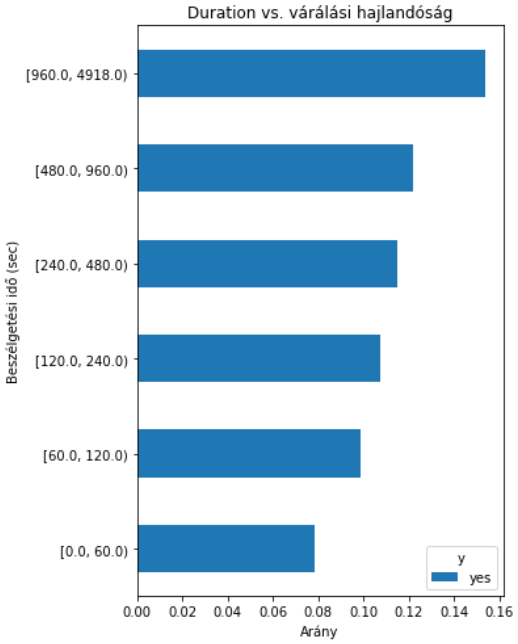
*Bináris változó (no/yes)

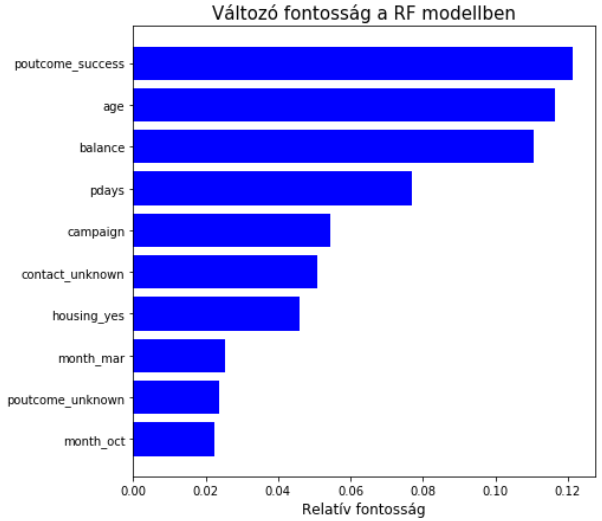
A legtöbb változó több kategóriát is tartalmaz

***One Hot Encodingra** lesz szükség

A MODELLEZÉS: PROBLÉMA ÉS FOLYAMAT

- **A modellezési probléma:**
  - **Bináris klasszifikáció:** a magyarázott változó két éréket vehet fel (jelen esetben: yes/no)
  - **Unbalanced** (kicsi a ‚yes'-ek aránya)
- Milyen modell **kiértékelési mutatószámok**at vegyünk figyelembe?
  - **AUC**-t (Nem Accuracy-t) → jobb unbalanced bináris probléma esetén → ez a fő összehasonlítási szempontom
  - Ezen kívül:
    - **Kalibrációs görbe**
    - **Confusion mátrix** a valós prédikált eredmények szemléltetésére
- **Train és Test** adatbázis szétválasztás (80% - 20%)
- **Több modell csoport futtatása** (pl. ElasticNet, RandomForest)
  - **GridSearch** + **CrossValidation** (CV=5) → a legjobb paraméterek megtalálása a train adatokon
  - **Validálás** a test adatokon





Véleményem szerint, két megközelítés lehetséges:

A) Az adatbázisban található **összes magyarázó változót felhasználhatjuk.**

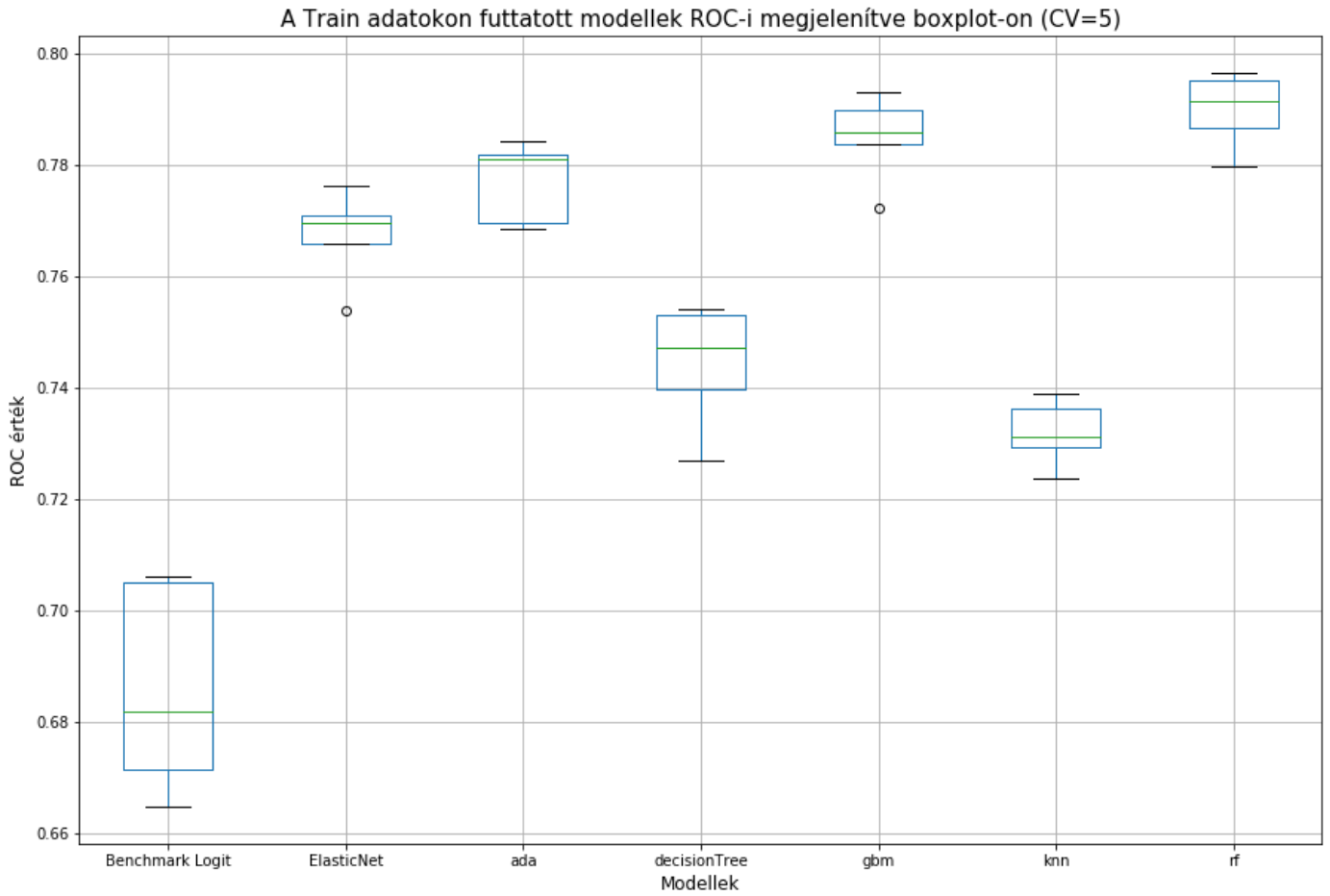
B) Azt a változót **nem használhatjuk fel,** ami a **telefonhívás előtt nem állhatott a rendelkezésre**

Én a B) opció mellett teszem le a voksom

* **Emiatt a ‚duration' (=hívás hossza) nevű változót nem használhatjuk fel:**
  * Nem tudjuk az erre vonatkozó adatot a hívás előtt
  * Nagyon nagy a magyarázó ereje ← csak ezt az egy változót használva jobb eredményt kapok mint az összes többit felhasználva (AUC ~ 0.8)

EGY KIS KITÉRŐ: VAJON MINDEN MAGYARÁZÓ VÁLTOZÓT „ÉRVÉNYES" FELHASZNÁLNUNK?

MODELLEZÉS: MODELLEK ÉS EREDMÉNYEIK



Az alábbi **modelleket** futtattam:

* Egyszerű logit (**Benchmark Logit)**
* **KNN** (knn)
* **ElasticNet**
* Egyszerű **döntési fa** (decisonTree)
* **RandomForest** (rf)
* **GradientBoosting** Machine (gbm)
* **AdaBoost** (ada)

Eredmények a legjobb paraméterek felhasználásával boxploton ábrázolva (lásd. balra)

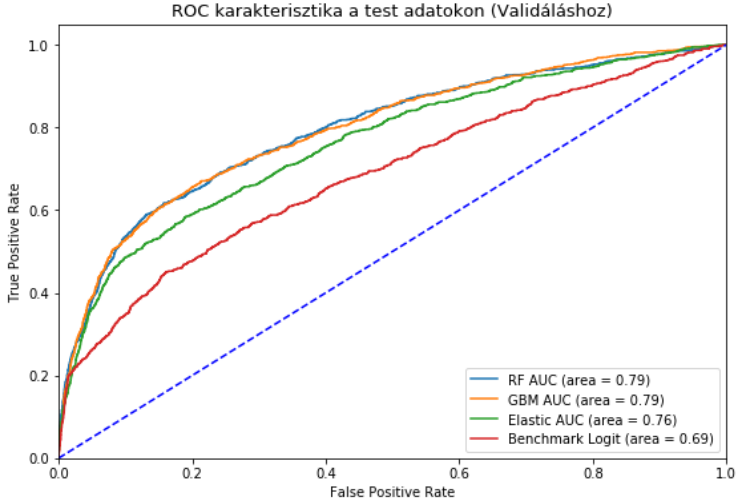
*Legjobb 3 általam választott modell további értékelésre:

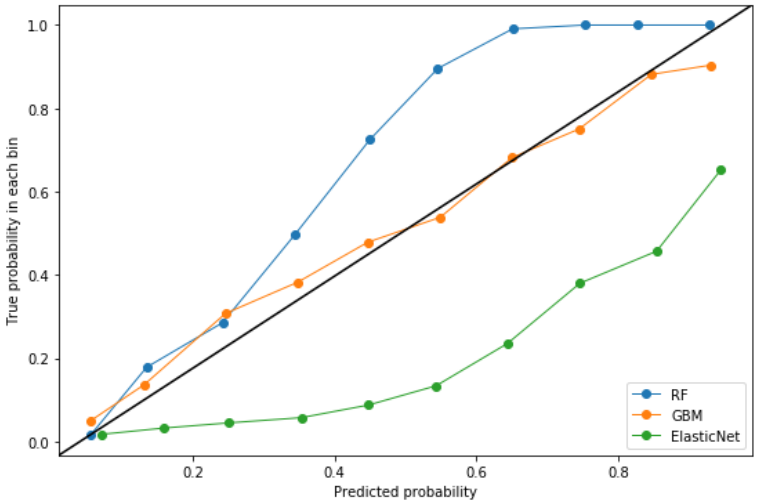
*RF – legjobb ROC

*GBM – közel legjobb ROC

*ElasticNet – legjobb nem tree alapú modell

MODELLEZÉS: AZ ÁLTALAM VÁLASZTOTT 3 MODELL TOVÁBBI JELLEMZŐI – VALIDÁLÁS, KALIBRÁCIÓ





- Modell **validálásához** és az **ROC görbék** szemléleteséhez az ROC görbéket a test adatokon jelenítettem meg:
  - Validálás: ha **nagyon hasonló az eredmény mint a boxplotnak → elfogadjuk a választott modelleket ← elfogadom → nem overfit/underfit a modell**
  - **AUC = a görbe alatti terület**
- Miután a legjobb 3 modellt kiválasztottam, **a kalibrációs görbéit is megnéztem.** A kalibrációs görbe minél közelebb húzódik az 1 meredekségű fekete egyeneshez annál jobb kalibrálásra utal.
- **<u>Ezek alapján: A GBM modellt válaszoltom a legjobb modellnek</u>**
  - ROC/AUC szerint az egyik legjobb
  - Legjobb kalibrációs görbével rendelkezik

MODELLEZÉS: GBM VALÓS EREDMÉNYEINEK SZEMLÉLTESE CONFUSION MATRIX-OT HASZNÁLVA

Train adatokon (80%):

Alap küszöbérték (0.5)

	Predicted y no	Predicted y yes
True y no	31485	478
True y yes	3089	1109

Módosított küszöbérték (0.15)

	Predicted y no	Predicted y yes
True y no	28671	3292
True y yes	1782	2416

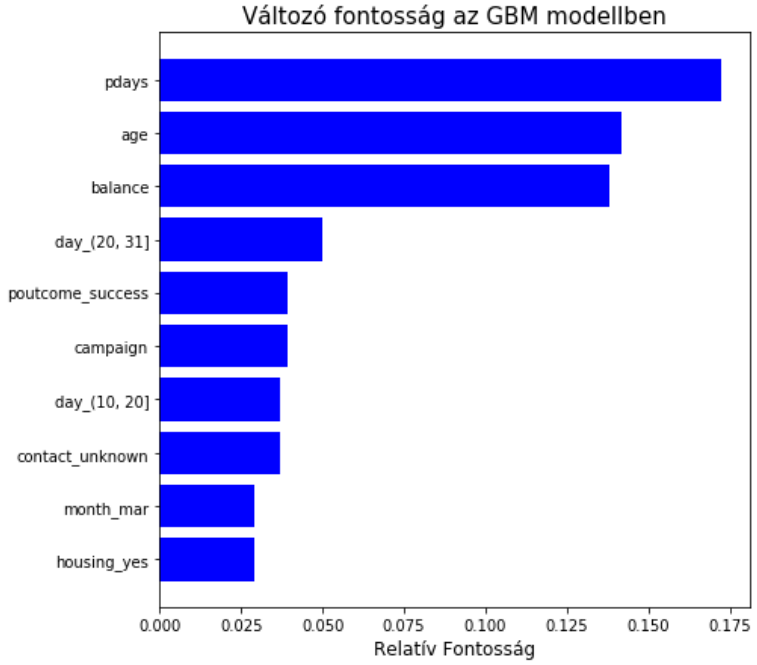
Test adatokon (20%):

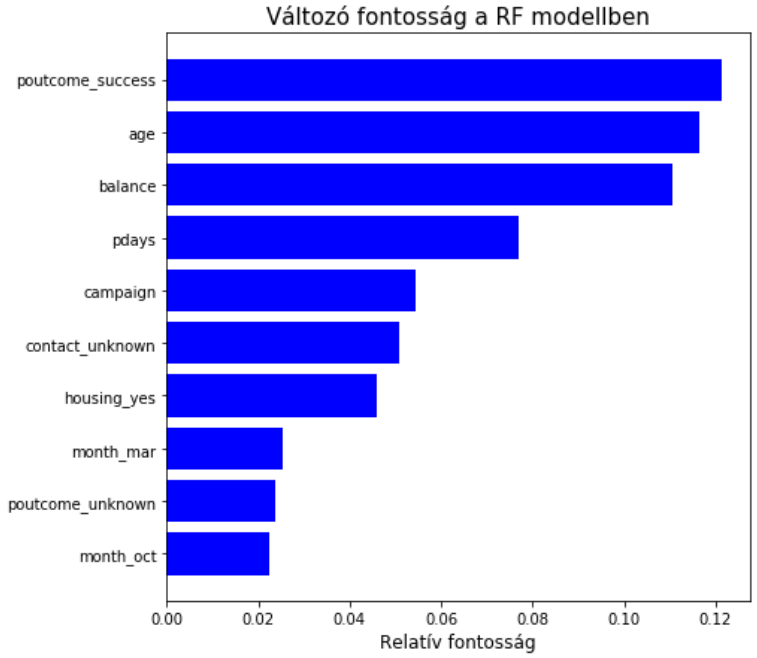
Alap küszöbérték (0.5)

	Predicted y no	Predicted y yes
True y no	7812	139
True y yes	851	239

Módosított küszöbérték (0.15)

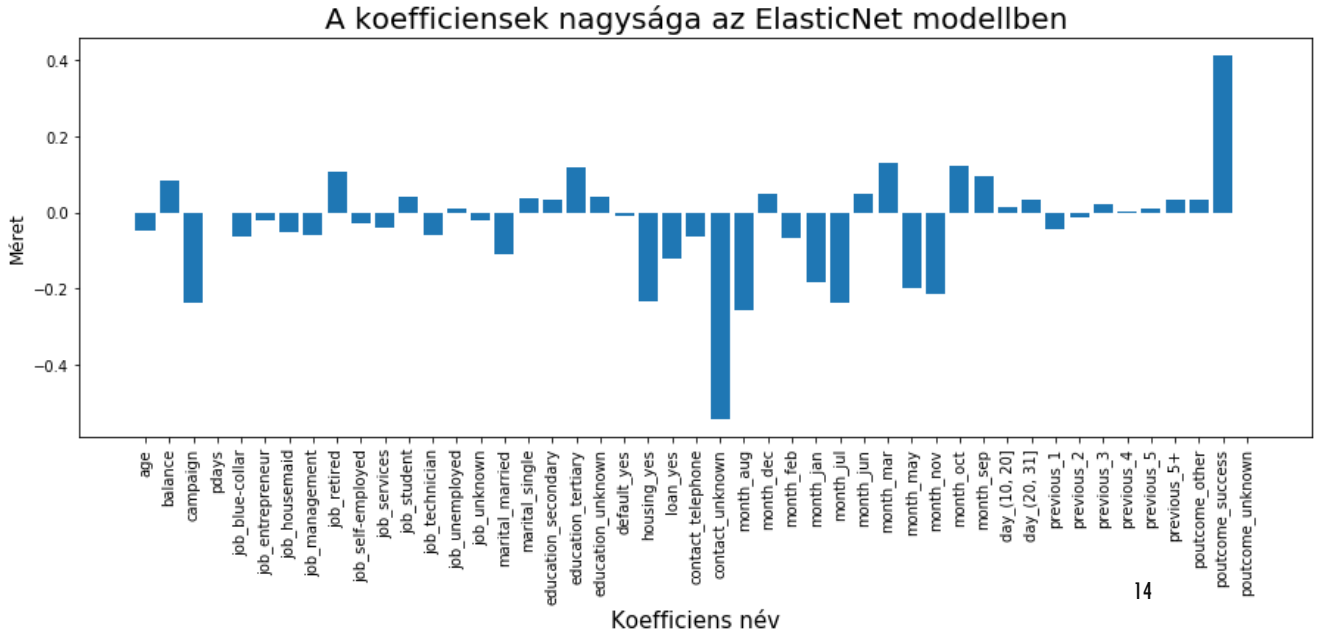
	Predicted y no	Predicted y yes
True y no	7099	852
True y yes	505	585



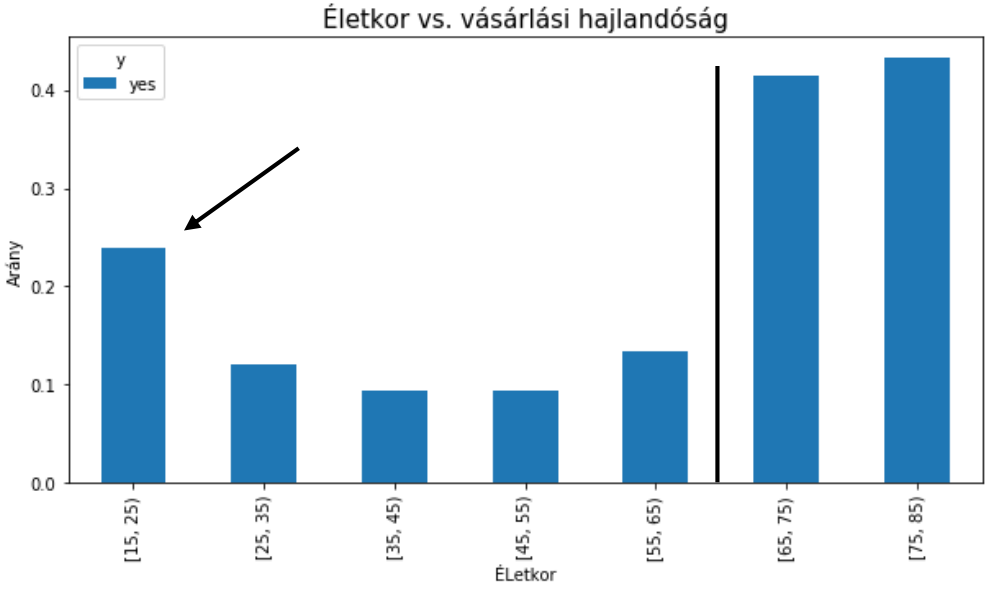


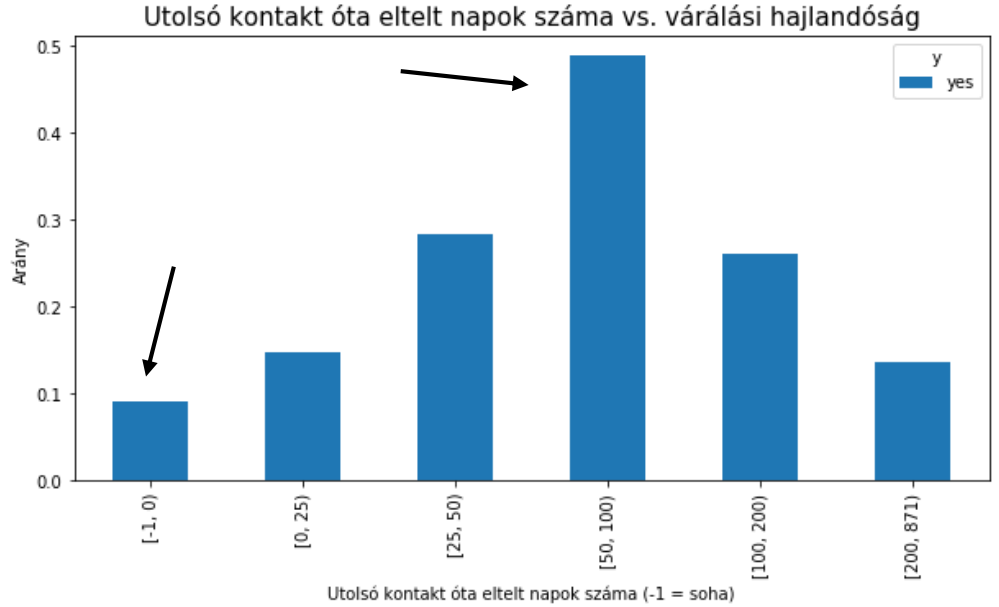
MODELLEZÉS: AZ EGYES MODELLEK FONTOSNAK ÍTÉLT VÁLTOZÓI

- AZ alábbi dián az egyes, általam legjobbnak ítélt modellek fontosnak ítélt változói láthatók:
- Néhány fontosnak ítélt változó pl:
  - pdays
  - poutcome (succes)
  - age
  - balance
- Ezek azok változók azok, amik valószínűleg nagyban befolyásolják a predikció végeredményét, vagyis hogy valaki vásárol e a termékből vagy sem



EREDMÉNYEK: A FONTOSNAK ÍTÉLT VÁLTOZÓK HATÁSA A VÁSÁRLÁSI HAJLANDÓSÁGRA

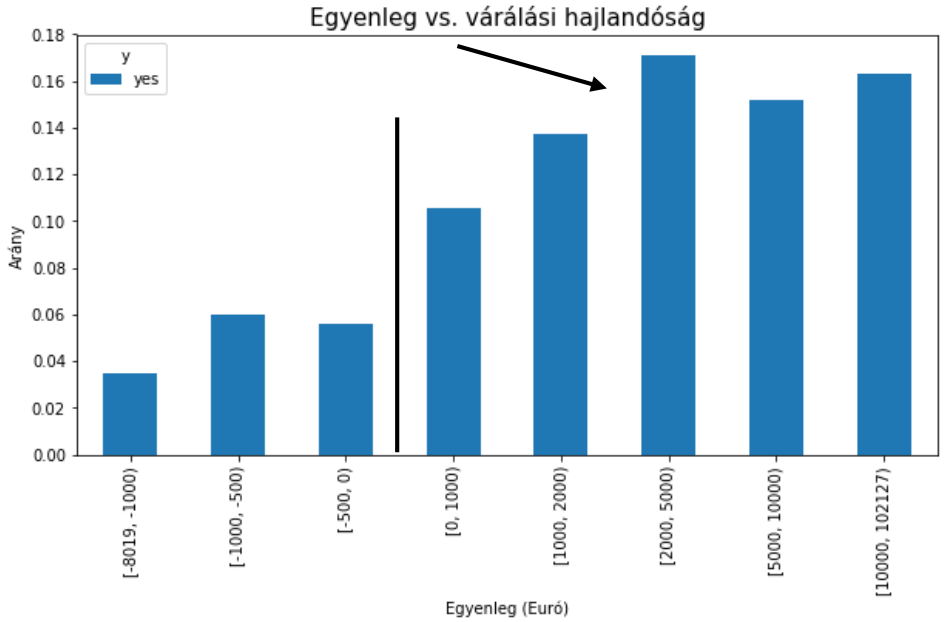


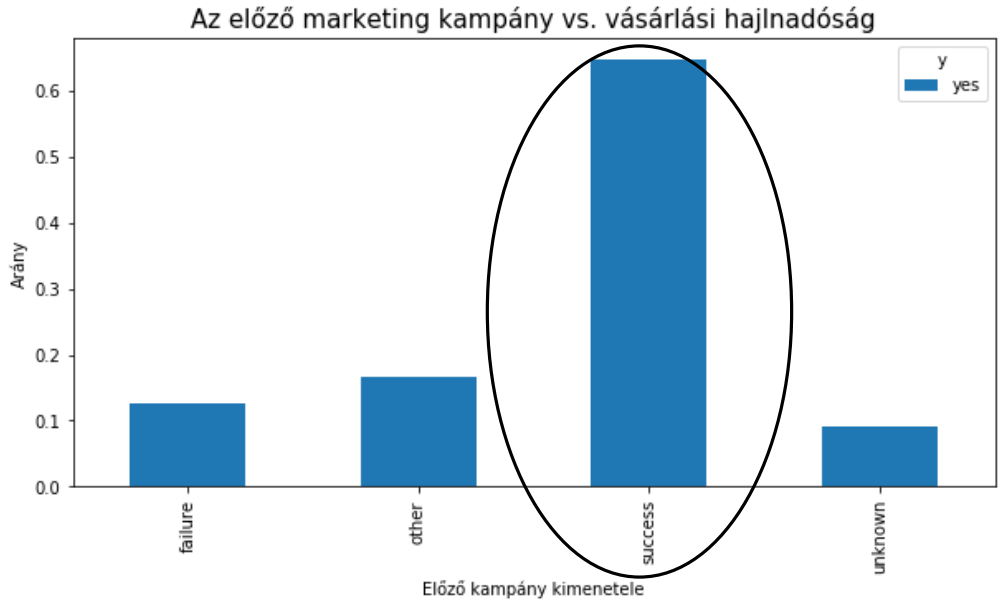


- A vásárlási hajlandóság nagy a 65+ -os korosztályban, de a fiatalok 15-25 évesek is nagy arányban vásárolnak
- 15-től 45 éves korig csökken a vásárlási hajlandóság
- 45-55 éves korig kb. konstans
- 65 + kortól egy jóval nagyobb ugrás található a vásárlási hajlandóságban, majdnem 50% ezen korcsoportnak vásárol a termékből. (Azonban a 65+-os megfigyelések száma alacsony)

- Akiket még nem hívtak (-1) ott a legalacsonyabb a vásárlási arány
- Mindenhol magasabb a vásárlási arány, akit már legalább egyszer hívtak. (Lehet hogy azért mert már ismer és jobban bízik? De lehet azért mert már eladtak nekünk előzőleg valamilyen terméket és a bank ügyfelei?)
- 100 nap eltelte közelében a legnagyobb a vásárlási arány
- 0-tól 100 napig a vásárlási arány folyamatosan nő
- 100 nap elteltével a vásárlási arány azonban csökkeni kezd

EREDMÉNYEK: A FONTOSNAK ÍTÉLT VÁLTOZÓK HATÁSA A VÁSÁRLÁSI HAJLANDÓSÁGRA





- Nagyobb egyenleg nagyobb vásárlási hajlandóság, egészen 5000 euróig, ahol konstanssá válik az arány
- azok akiknek negatív az egyenlegük sokkal kisebb arányban vesznek a termékből
- akiknek nagyon alacsony azok (szinte) semmit
- egy bizonyos összeg fölött, már nem növekszik a vásárlási hajlandóság.

- Akiknél az előző marketing kampány sikeres volt sokkal nagyobb arányban vásárolnak a termékből
- A többi csoportban a vásárlási arány kb. megegyezik

ÜZLETI AJÁNLÁS AZ EREDMÉNYEK ISMERETÉBEN

- Az alkalmazásra vonatkozó tanácsaim **attól függők, hogy milyen vállalat döntéshozóinak készíteném elő az elemzésem.** Gondolok itt a következőre:
  a) Egy **csak sales tevékenységet végző vállalat** számára a modellem által nagy valószínűséggel társító egyéneket javasolnám felhívásra/ megkeresésre → Jutalék maximalizálás.
     - életkor szerint a 15-35-as és a 65+-os korosztályt lenne érdemes felkeresni
     - olyanokat lenne érdemes felhívni, akiket már felhívtak és sikeresen eladtak nekik más terméket és kb. 100 napja volt az utolsó kontaktálás
     - akiknek pozitív a számla egyenlegük
  b) Egy **bank esetében** azonban szükséges lenne magát a marketing kampány hatását lemérni, ezt most nem teszi lehetővé az adatbázis.
     - Miért? Mivel nem találhatóak benne olyan egyénekre vonatkozó megfigyelések, akik nem vettek részt a marketing kampányban, és így nem tudjuk a marketing kampány hatását vizsgálni. Nincs mihez viszonyítanunk, így nem tudjuk a hozzáadott értékét sem. Gondoljunk bele abba, hogy lehet, hogy az adott egyén a marketing kampány nélkül is megvásárolta volta a bankfiókban a terméket, így a telefonos megkeresésnek semmi hatása nem volt a döntésére nézve.
     - Tkp. egy A/B tesztelésre lenne szükség, amit a mostani adatbázis nem, vagy csak kevésbé tesz lehetővé

TOVÁBBI FEJLESZTÉSI LEHETŐSÉGEK

- Jó irány lenne mérni a sales tevékenység hatását, így megtudhatnánk:
  - Mely egyéneket szükséges megcéloznunk ← marketing kampány hatása * populációs részarány
  - Szükséges ehhez: olyan személyek is az adatbázisba, akik nem vettek részt a marketing kampányban
    - Új változó: Treated/ Not Treated
- Milyen szoláltatása van a banknál, ha van az adott személynek:
  - Cross sell és up sell
- Ezen kívül további változók, amiket szerepeltetnék:
  - Mikor történt a hívás pontosan és mikor történt a vásárlás pontosan
    - ÉV → hétvége/ hétköznap
    - Hívás pontos ideje: reggel/este/stb.

KÖSZÖNÖM A FIGYELMET!

- Kérdésekre szívesen válaszolok
  - Személyesen most
  - Vagy: menyhert.kristof@gmail.com