**Hitelkockázati modellezés**

# Feladat

Tanácsadóként azt a feladatot kapta, hogy egy Bank számára készítsen egy hitelkockázati scorecard modellt a KKV ügyfélkör hitelbírálati folyamatának támogatására. Ehhez kapcsolódóan rendelkezésre áll egy ügyfélszintű adatbázis, amely az elmúlt évek tapasztalatai alapján került összeállításra, és tartalmazza a modellezni kívánt célváltozót („default” – kétértékű változó (I/N), az ügyfél várhatóan vissza fogja-e fizetni a hitelt?) valamint a lehetséges magyarázó változókat.

A modellezéshez felhasználandó adatbázis és a változók definíciói az alábbi fájlokban található:



Kapcsolódó feladatok:

* Készítsen több alternatív modellezési megközelítéssel olyan prediktív modelleket, amelyek a hitelt igénylő lehetséges jövőbeli ügyfelek nemfizetési valószínűségét jelzik előre.
* Mutassa be az egyes modellek jellemzőit, előnyeit / hátrányait mind elméleti, mind gyakorlati szempontból. Melyik modell bevezetését javasolná a Bank számára? Milyen szempontok alapján?
* Mutassa be részletesen a kiválasztott modell kialakításának lépéseit! A bemutatás terjedjen ki az alábbi lépésekre:
* Az adatbázison elvégzett ellenőrzések, esetleges transzformációk bemutatása
* Az alkalmazott mintavételezési technikák bemutatása
* Az egyes magyarázó változók elemzése, esetleges kategorizálásai
* Az alkalmazott modell elméleti megalapozása, kiválasztásának szempontjai
* A többváltozós modellezés lépései, eredményei
* Határozza meg az egyes ügyletekhez kapcsolódó becsült nemfizetési valószínűséget az átadott adatbázison! Ábrázolja a nemfizetési valószínűségek eloszlását egy hisztogramon, és értelmezze az eredményeket!

Megjegyzések a feladat megoldásához:

* A mintában a nemfizető (defaultos) ügyfelek száma viszonylag alacsony, ezért célszerű a modell fejlesztése során az ún. „imbalanced dataset”-re jellemző specialitásokra tekintettel lenni.
* A feladat megoldásának értékelése során előnyt jelent, ha ennek megfelelő mintavételezési technikák kerültek alkalmazásra (pl. bootstraping megoldások, over/undersampling, SMOTE eljárás, stb.), vagy olyan modell teljesítmény mérőszámok is kiszámításra kerültek, amelyek ilyen esetekben is jól kezelhetőek (pl. AUROC érték mellett a PRAUC értékek kiszámítása)

# Megoldás

A feladat megoldását Pythonban, Jupyter Notebookban készítettem el.

# Adatok előfeldolgozása

Összesen 2037 ügyfél adatai állnak rendelkezésre. Az egyetlen kategorikus változó a LOAN\_TYPE\_CATEGORY volt, melyet dummy változókra (8 db 0/1-eket tartalmazó változó, mert 8-féle hiteltípus van) cseréltem, mivel a legtöbb osztályozó algoritmus csak numerikus változók esetén alkalmazható. Bár a választott módszer (Random Forest) elméletileg működhetne kategorikus változókkal is, azonban a scikit learn könyvtár, melyet a kód írása során használtam, kizárólag numerikus változókat képes kezelni.

Az adatokat véletlenszerűen, 80%-20% arányban tanító- és teszthalmazra osztottam fel. A modelleket a 80%-nyi tanítóhalmazon tanítottam, majd a maradék 20%-nyi teszthalmazon értékeltem ki a betanított modellek teljesítményét.

A magyarázó változók többsége tartalmazott hiányos mezőket, ezeket az első két modell esetében az adott változó értékeinek átlagával helyettesítettem. A másik két modell esetén pedig az sklearn könyvtár IterativeImputer-ének segítségével pótoltam azokat, azaz egy adott változó hiányzó értékeit az összes többi változó felhasználásával, egy regressziós modellel becsültem meg.

A mintában a nemfizető (defaultos) ügyfelek száma viszonylag alacsony (2,84%), ezért két modell esetében a SMOTE eljárást (Synthetic Minority Oversampling Technique) alkalmaztam a két osztály kiegyenlítésére.

Két modell esetén az összes magyarázó változót felhasználtam, míg a másik két modellt csak az alábbi változókon tanítottam:

* DAYS\_PAST\_DUE
* NUMBER\_OF\_EMPLOYEES
* PROFIT\_AFTER\_TAX
* FORCED\_CALL\_AUTHORITY

A fenti négy magyarázó változót az egyes magyarázó változók és a célváltozó közötti kapcsolatok vizsgálata alapján választottam ki.

# Modellek

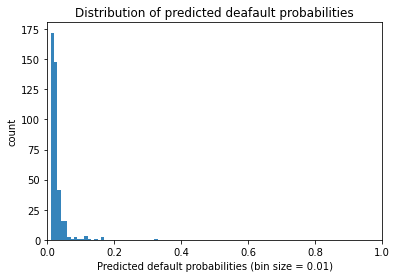
A cél az volt, hogy megbecsüljük annak a valószínűségét, hogy egy adott ügyfél vissza fogja-e fizetni a hitelt. Ennek a valószínűségnek a megbecsléséhez random forest módszert alkalmaztam, 100 db döntési fával. A túltanulás elkerülése érdekében a döntési fák mélységét 5-re, míg a levélcsúcsok számát 10-re korlátoztam. A becsült valószínűségek a nemfizetést tippelő döntési fák aránya.

Összesen négy modellt tanítottam:

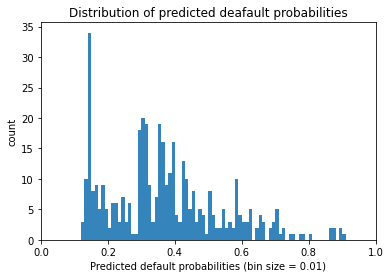
1. A hiányzó adatok az átlaggal pótolva. Nincs kiegyensúlyozva a két osztály.
2. A hiányzó adatok az átlaggal pótolva. Az osztályok kiegyensúlyozása SMOTE-tal.
3. A hiányzó adatok IterativeImputer-rel pótolva. A leghasznosabbnak tűnő magyarázó változók használata. Nincs kiegyensúlyozva a két osztály.
4. A hiányzó adatok IterativeImputer-rel pótolva. A leghasznosabbnak tűnő magyarázó változók használata. Az osztályok kiegyensúlyozása SMOTE-tal.

# Eredmények

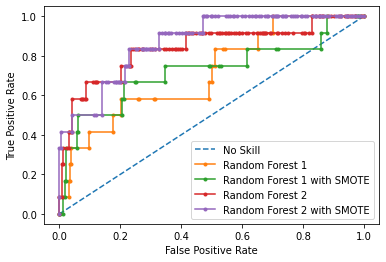
A négy modell közül az első minden ügyfélre, a harmadik pedig egy kivételével minden ügyfélre azt jósolta, hogy vissza fogja fizetni a hitelt, így ennek a két modellnek a teljesítménye messze nem kielégítő. Ennél a két modellnél, a SMOTE eljárás alkalmazásának a hiánya miatt, a tanítóhalmazban sokkal több fizető ügyfél található, mint nem fizető, így nem meglepő, hogy a nemfizetés becsült valószínűsége nemcsak, hogy a tényleges kimenetet eldöntő 0,5-ös határ alatt, hanem az esetek többségében 0,1 alatt maradt. Az alábbi hisztogram az első modell által a teszthalmaz 408 mintájára jósolt valószínűségeinek az eloszlását szemlélteti.

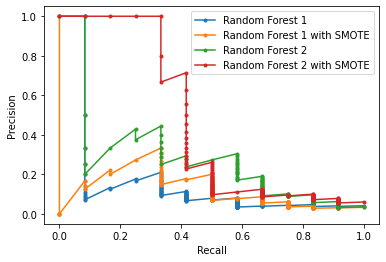


A második és negyedik modell esetében, tehát amikor a tanítóhalmazban a két osztály mintáinak a számát SMOTE eljárás segítségével kiegyensúlyoztuk, akkor a teszthalmazban található ügyfelekre becsült nemfizetési valószínűségek is jelentősen magasabbnak adódtak. A két modell a teszthalmazban található 12 nemfizető ügyfél közül 6-ot, illetve 8-at már helyesen nemfizetőnek jósolt, ugyanakkor a 396 fizető ügyfél közel 20%-át is nemfizetőnek gondolta. Az alábbi hisztogram a negyedik modell által a teszthalmaz 408 mintájára jósolt valószínűségeinek az eloszlását szemlélteti.



A négy modellt számos kiértékelési metrika szerint is összehasonlítottuk, az alábbiakban az egyes modellek ROC, illetve Precision-Recall görbéje látható.





A bank számára a négy modell közül a negyediket javasolnám, ugyanis a ROC és a Precision-Recall görbe alatti területet tekintve is ez bizonyult a legjobbnak, továbbá a nemfizető ügyfeleket is ez a modell találta meg a leghatékonyabban, a teszthalmaz 12 nemfizető ügyfele közül nyolcat is. Ha a bank a negyedik modell alapján dönti el, hogy kinek ad hitelt, akkor a teszthalmaz 408 ügyfele közül összesen 321-en kapnának hitelt, akik közül mindössze négyen (1,25%) nem fizetnék azt vissza. Ezzel szemben az első modell mind a 408 ügyfélnek megadná a kért hitelt, akik közül azonban 12-en (2,94%) nem fizetnék azt vissza.