

HACKATHON CONSULT IT CHALANGE

ANALIZA PROGRAMU PRACOWNICZYCH PLANÓW KAPITAŁOWYCH

Presentacja Grupy 3

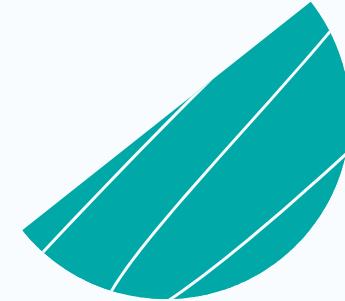


HIPOTEZA

Największe powodzenie programu PPK będzie w dużych firmach, w miejscowościach bardziej zurbanizowanych oraz w sektorze III.



4 kroki

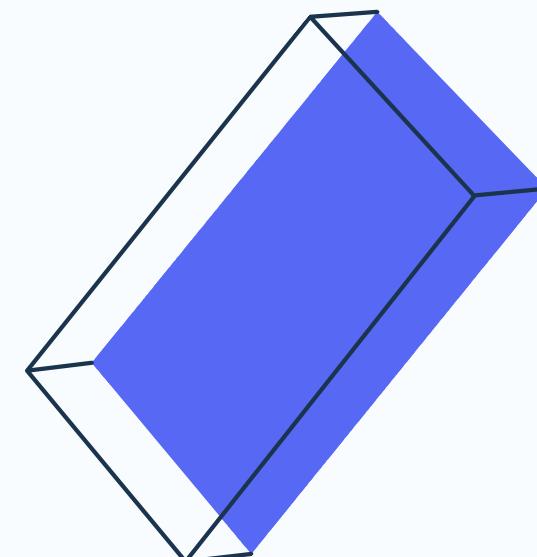


1. Oczyszczenie
danych

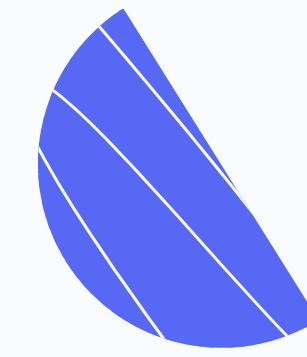
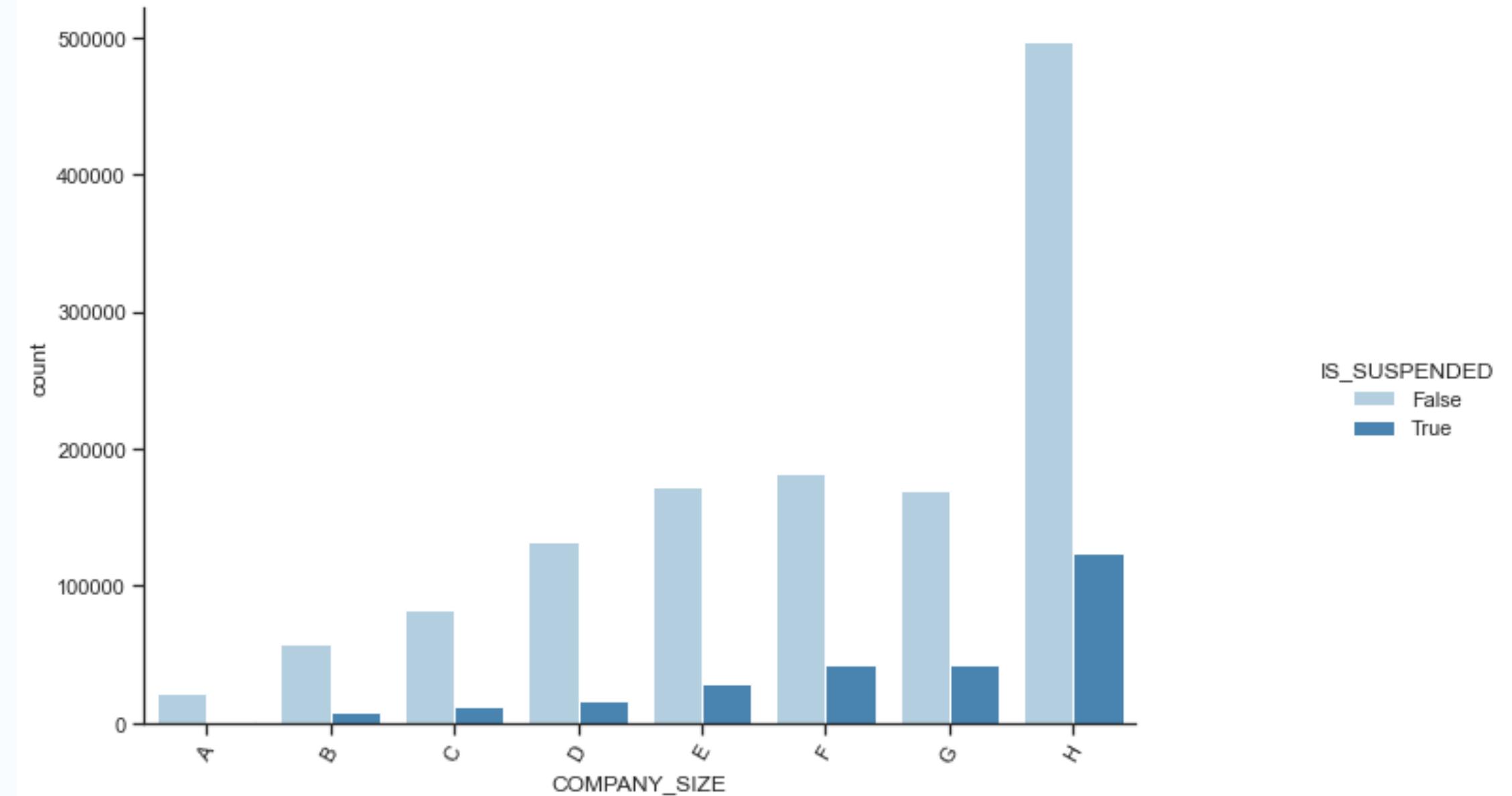
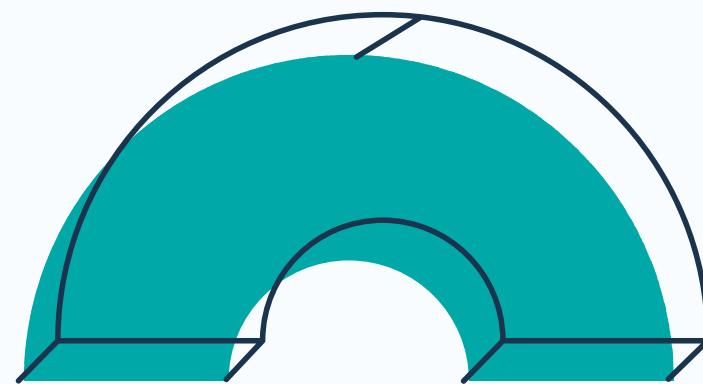
2. Symulacja modeli
przy użyciu różnych
hiperparametrów.

3. Predykcja na
najlepszym modelu.

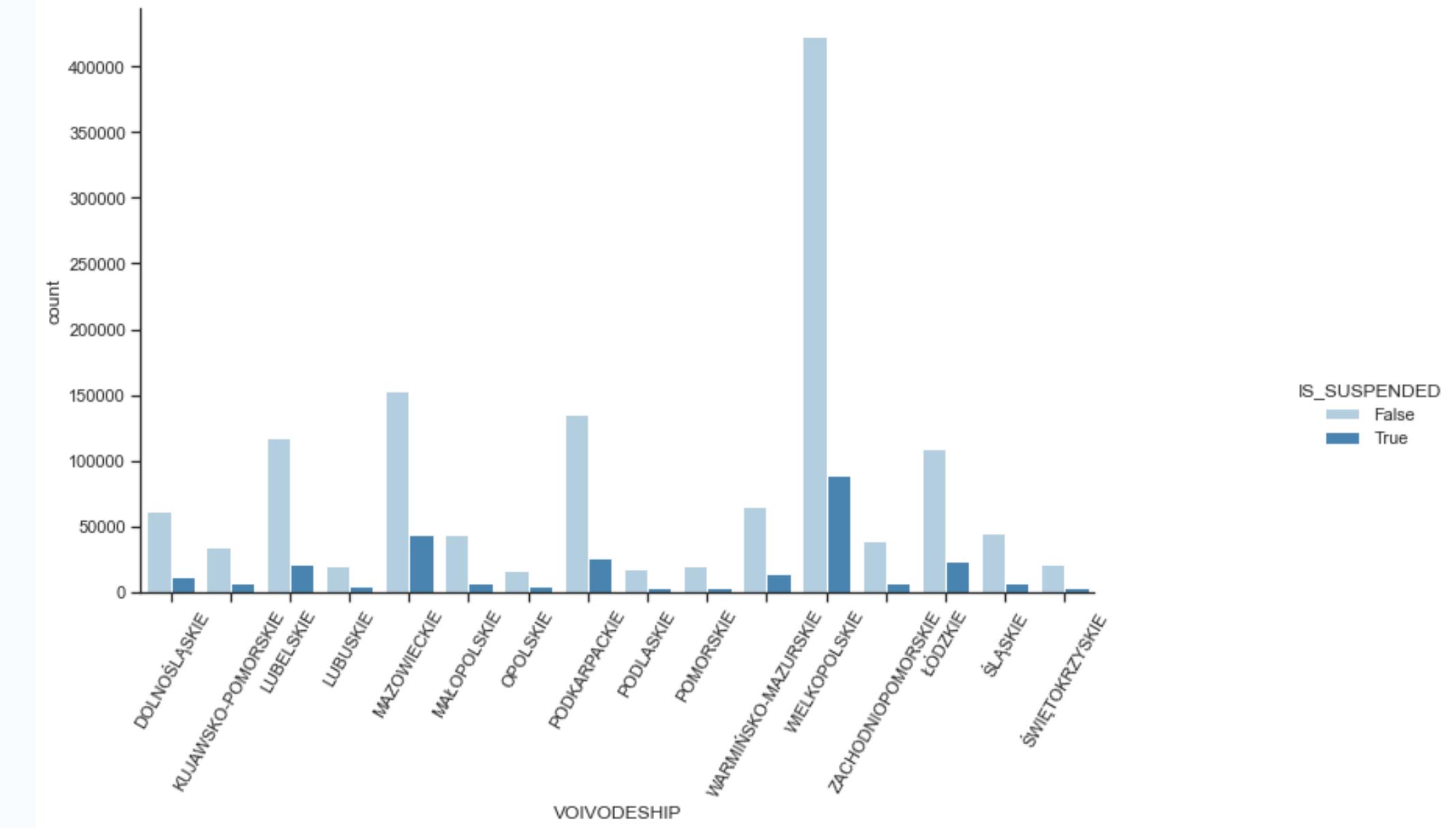
4. Sporządzenie
Najważniejszych
Wniosków.



1. Wielkość firmy jest wprost proporcjonalna do sukcesu programu



2. Największym powodzeniem cieszyło się województwo warmińsko-mazurskie



3. Powitalna wpłata:

- otrzymanie na 93% nie zrezygnuje

- brak na 71% nie zrezygnuje

Wpłata od pracodawcy:

- otrzymanie na 92% nie zrezygnuje

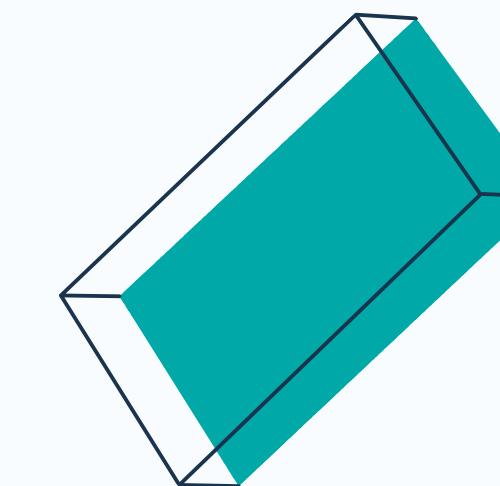
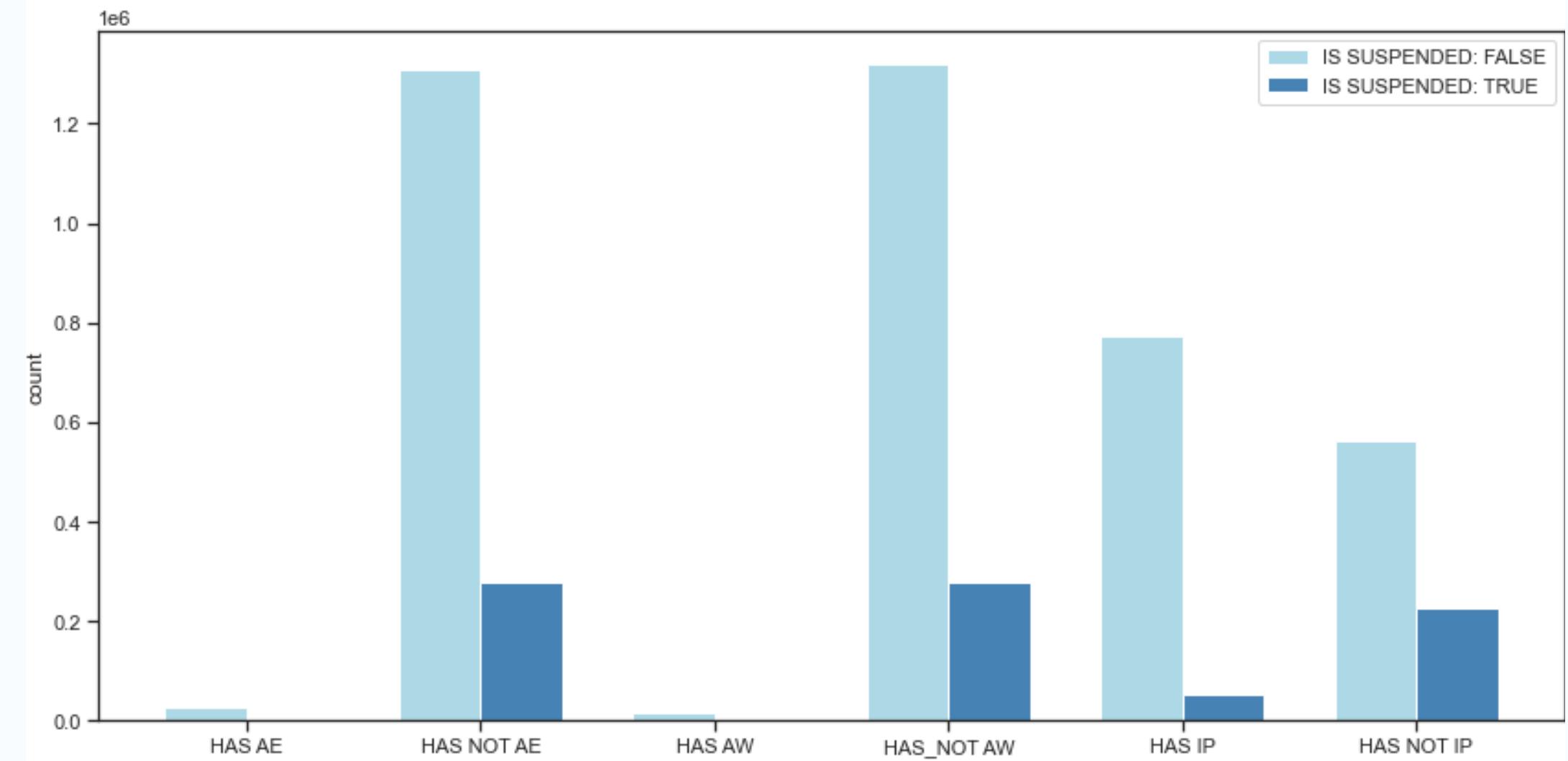
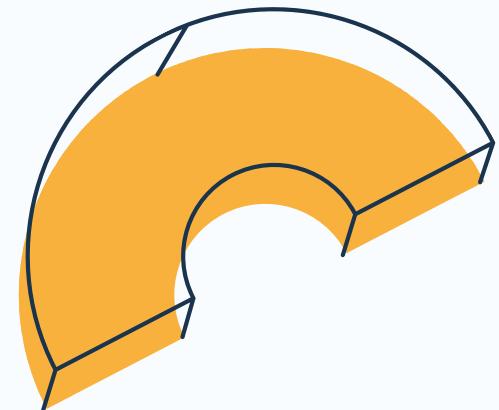
zrezygnuje

-brak na 82% nie zrezygnuje

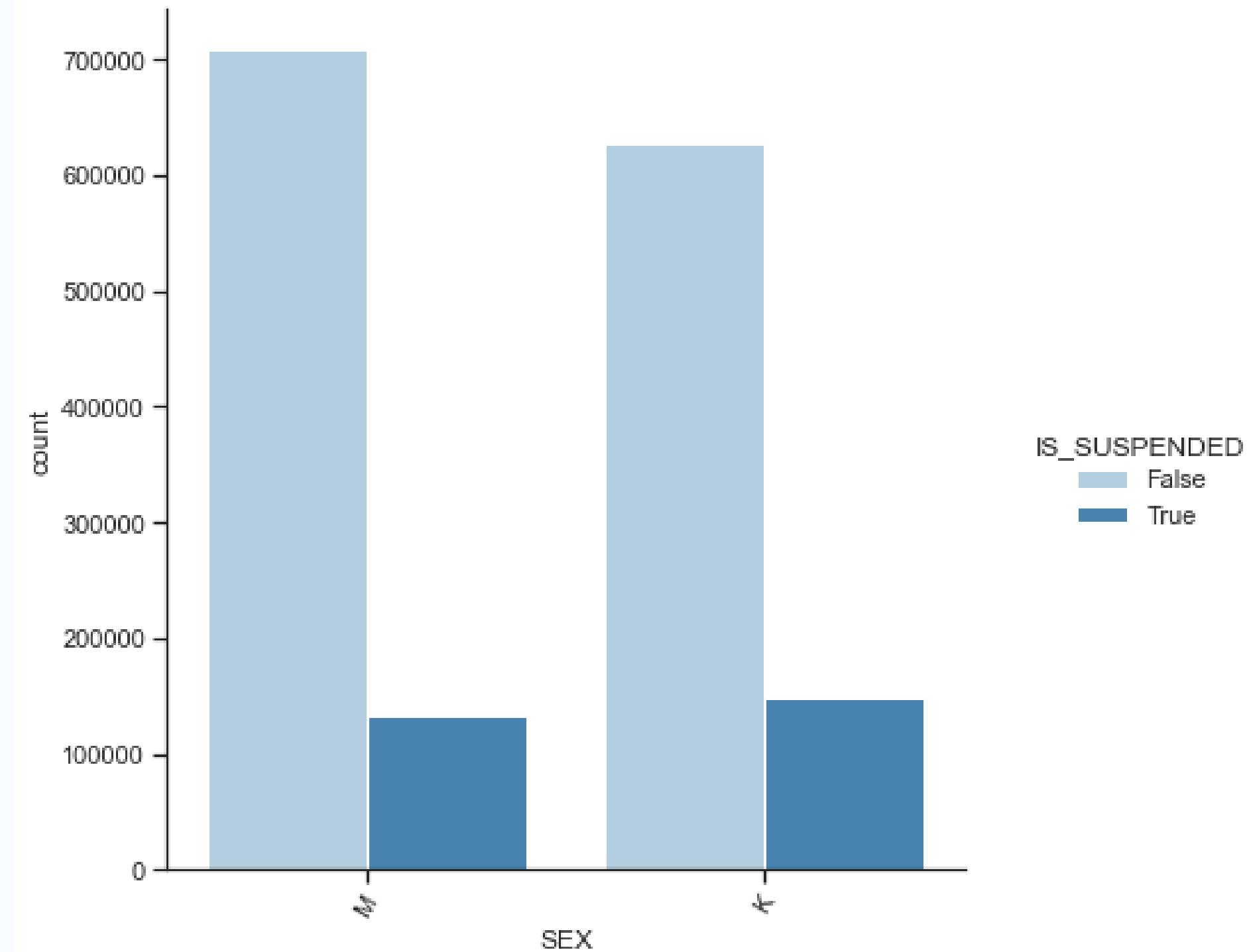
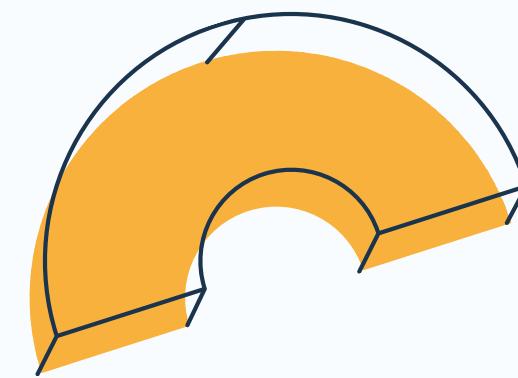
Wpłata pracownika

-na 97%, jeżeli nie ma własna

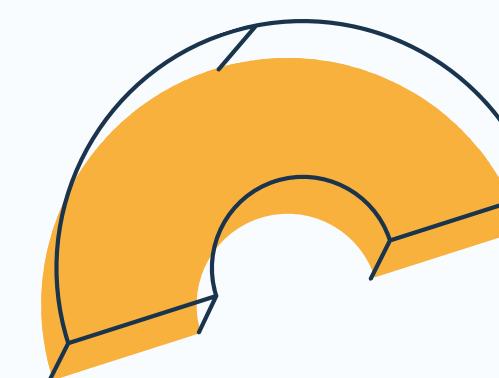
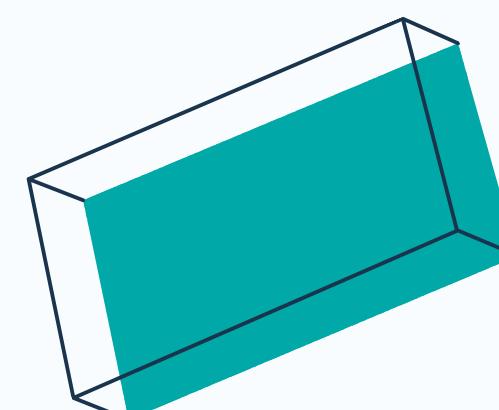
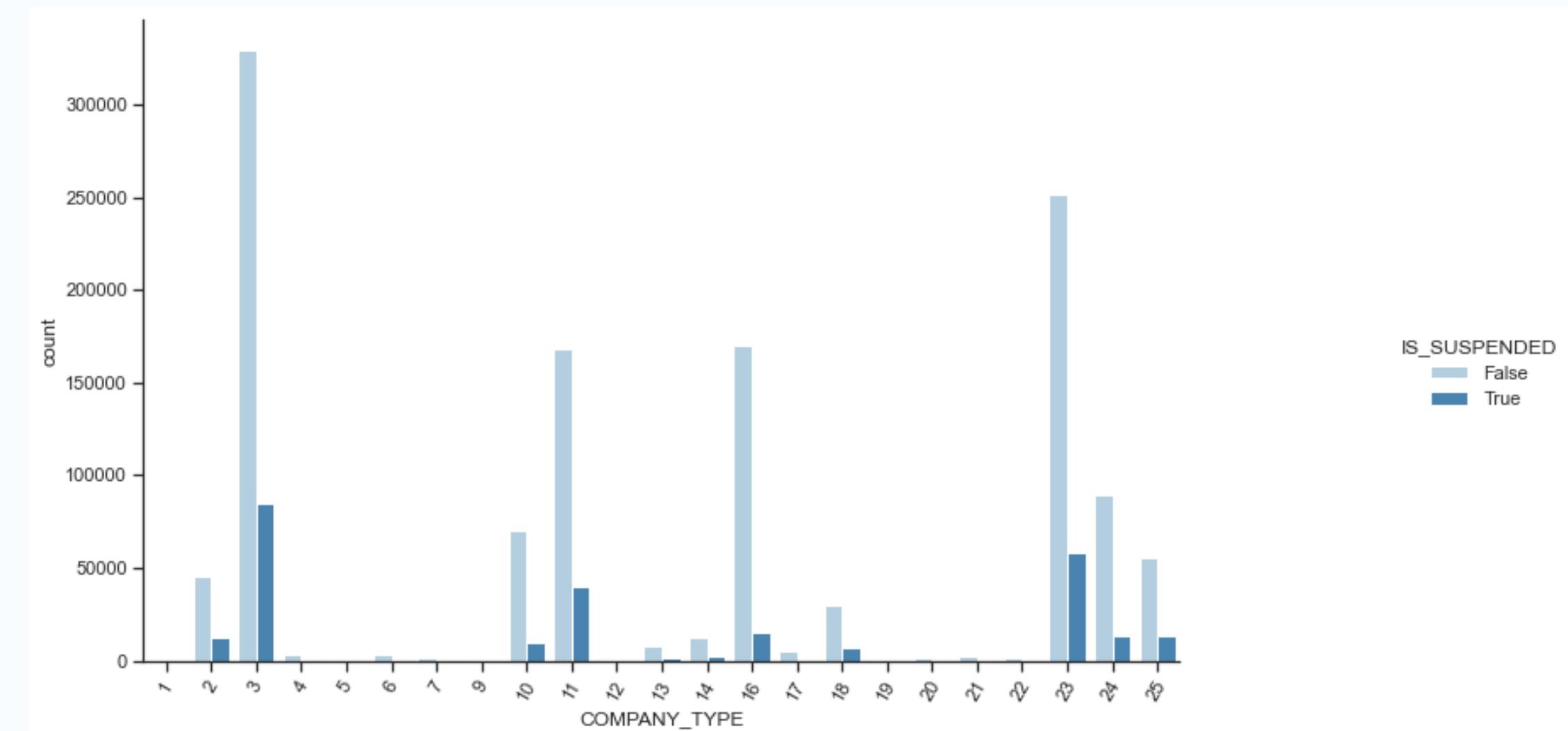
-brak na 82.5% nie zrezygnuje.



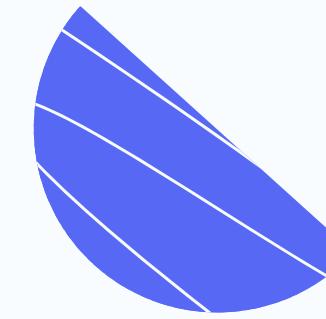
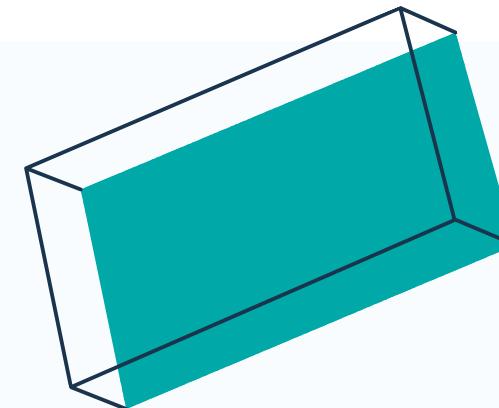
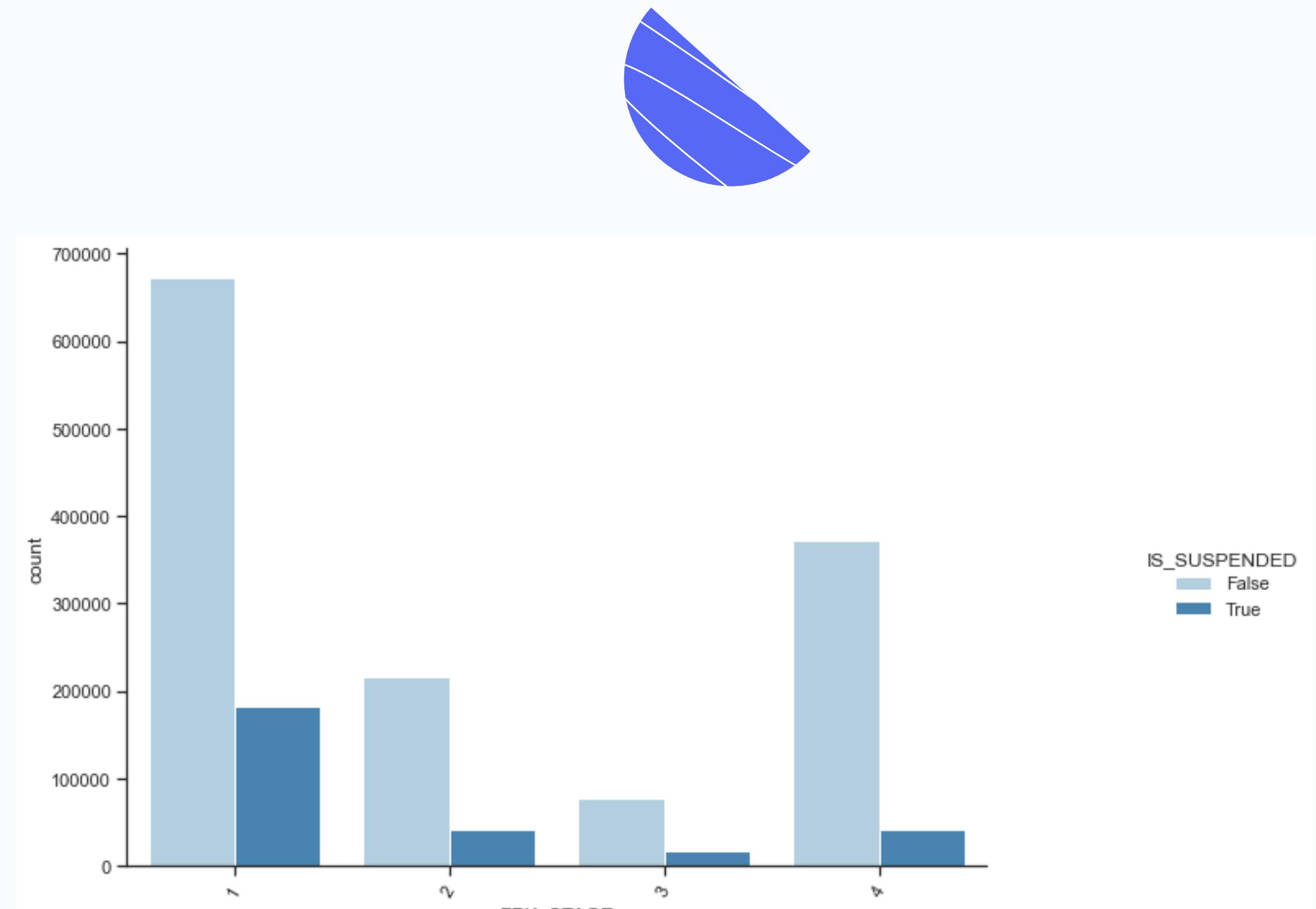
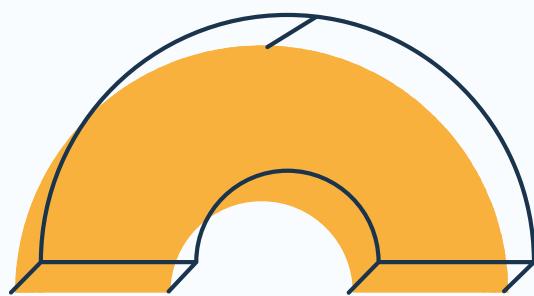
**4. Powodzenie
programu jest
słabo zależne od
płci - 84%
mężczyzn zostaje
w programie w
stosunku do 80%
kobiet**



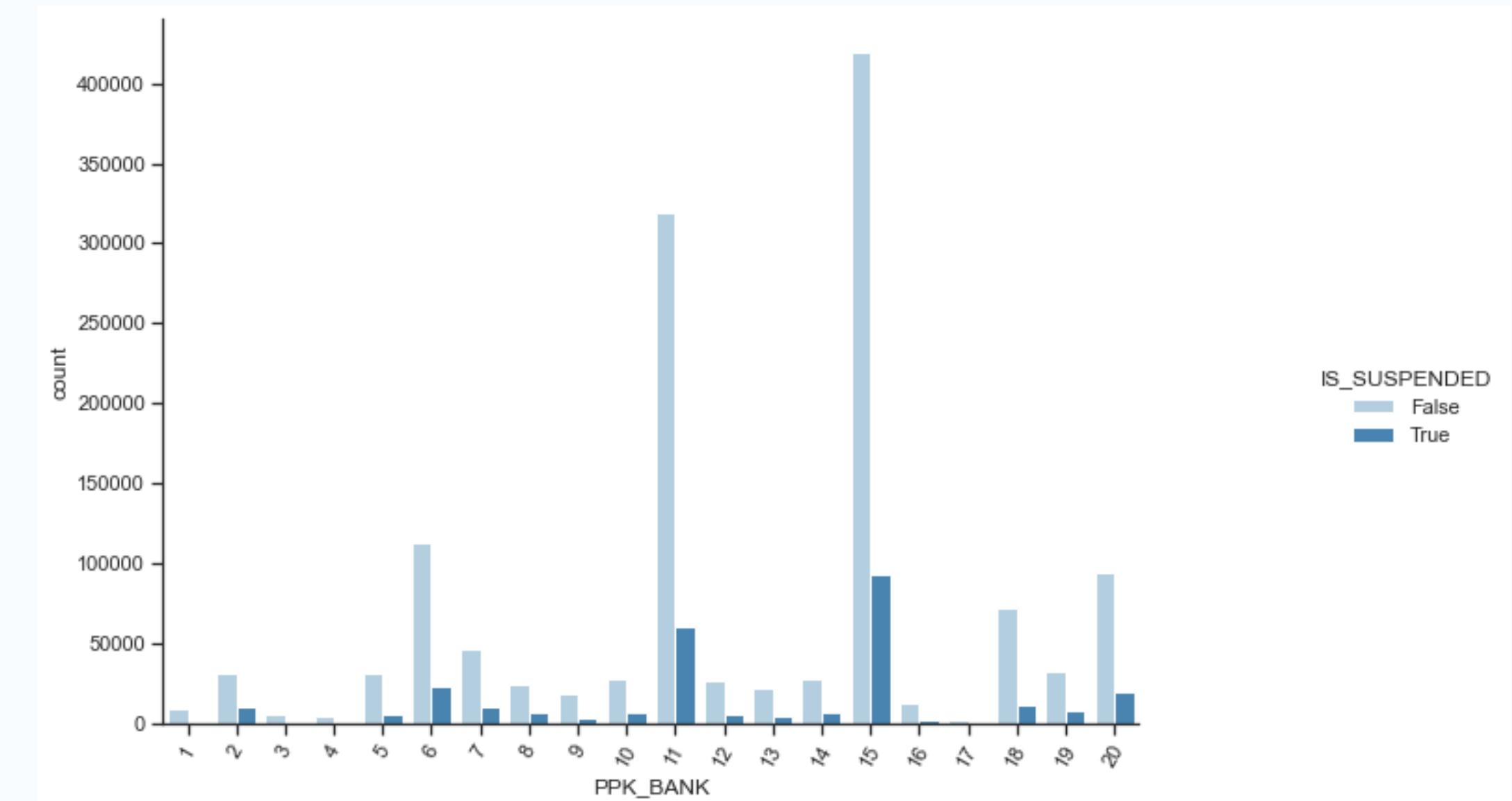
**5. Istnieje
duża
zależność
pomiędzy
typem spółki a
powodzeniem
programu**



6. Najwięcej osób skorzysta z programu w pierwszej jego fazie. Najniższy chern - ostatnia faza



7. Bank 15 oraz bank 11 są najbardziej popularnymi bankami.



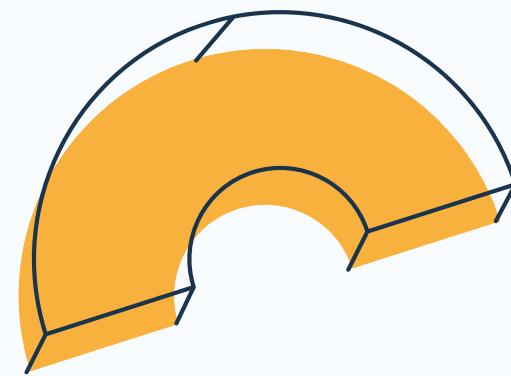
Preprocessing danych

- 1) normalizacja AGE poprzez odjęcie średniej i podzielenie przez standard deviation
- 2) tworzymy osobne klasy dla najczęst najczęstsze 30 NATIONALITY, pozostałe wartości do klasy "OTHER"
- 3) usuwamy kolumny z datami
- 4) usuwamy kolumnę NUMERICAL_VALUE ze względu na dużo NaN
- 5) w tabeli pracodawca przekształcamy region na Powiat i Gmina
- 6) zamieniamy PKD na PKD GROUP (pierwsze 2 cyfry PKD)
- 7) dla każdego pracodawcy tworzymy dodatkową kolumnę suspended ratio informującą, czy u danego pracodawcy większość zatrudnionych rezygnuje z programu, czy nie



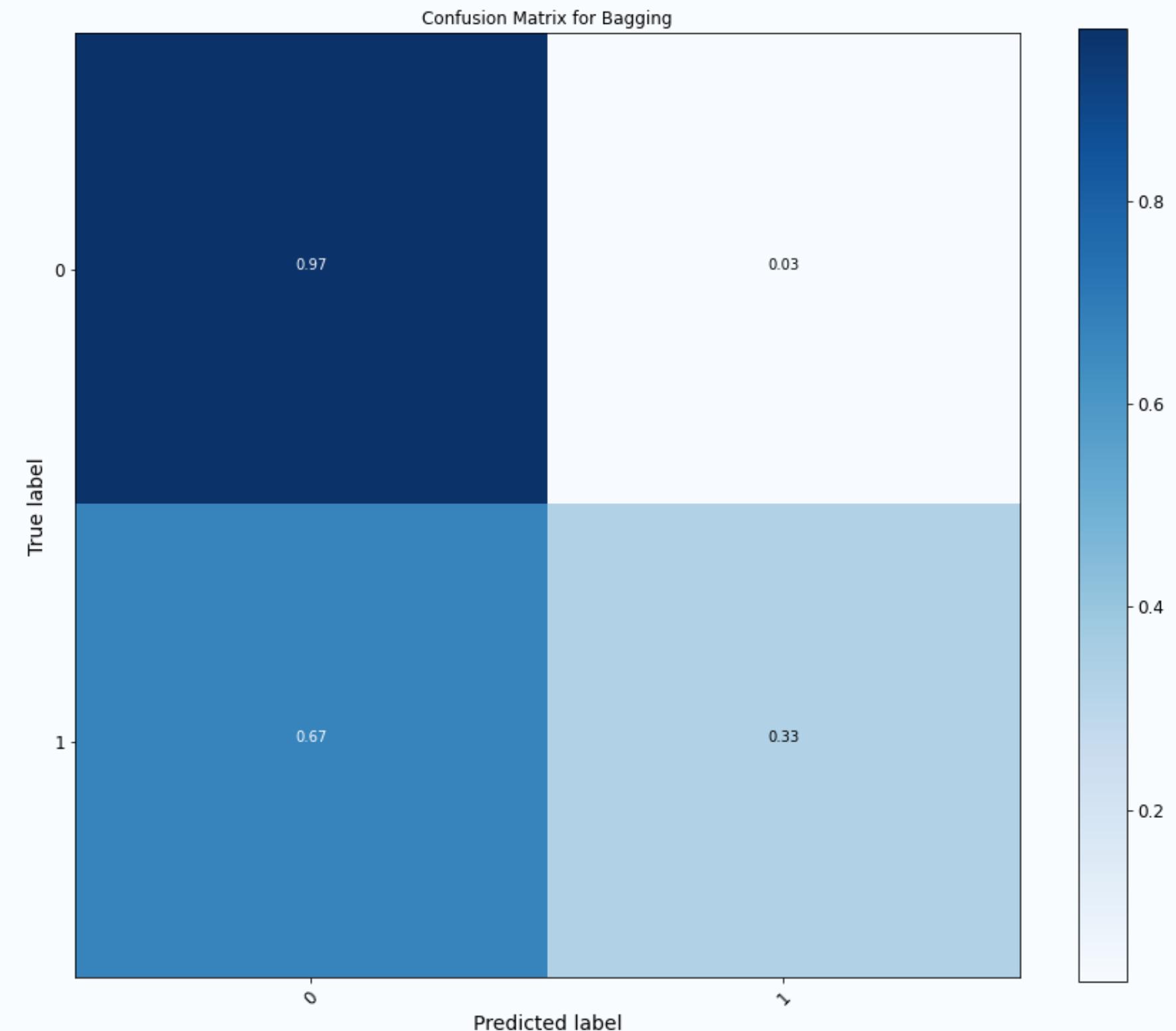
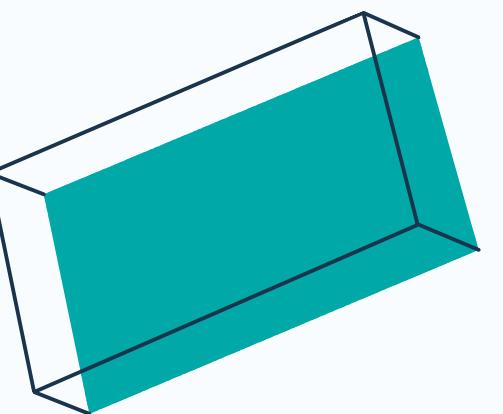
Przetestowane modele

	LINEAR REGRESSION	SVN: LINEAR KERNEL	SVN: SIGMOID KERNEL	SIECI NEURONOWE	BAGGING ON DECISION TREE	XGBOOST
RECALL	0.3295773748723187	0.09331797235023041	0	0.3440266758652409	0.4470046082949309	0.5632207942609898
PRECISION	0.6972817828802972	0.9050279329608939	0	0.704995287464656	0.6615515771526002	0.756448476992871
ACCURACY ON TESTING SET	85.79 %	84.09 %	82.64 %	86.09 %	86.4300 %	89.2849 %



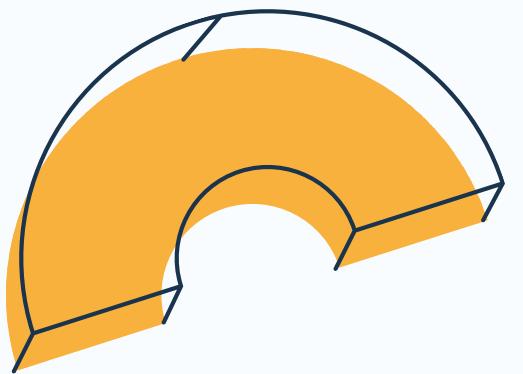
Linear regression

POZIOM DOKŁADNOŚCI - 85.79%

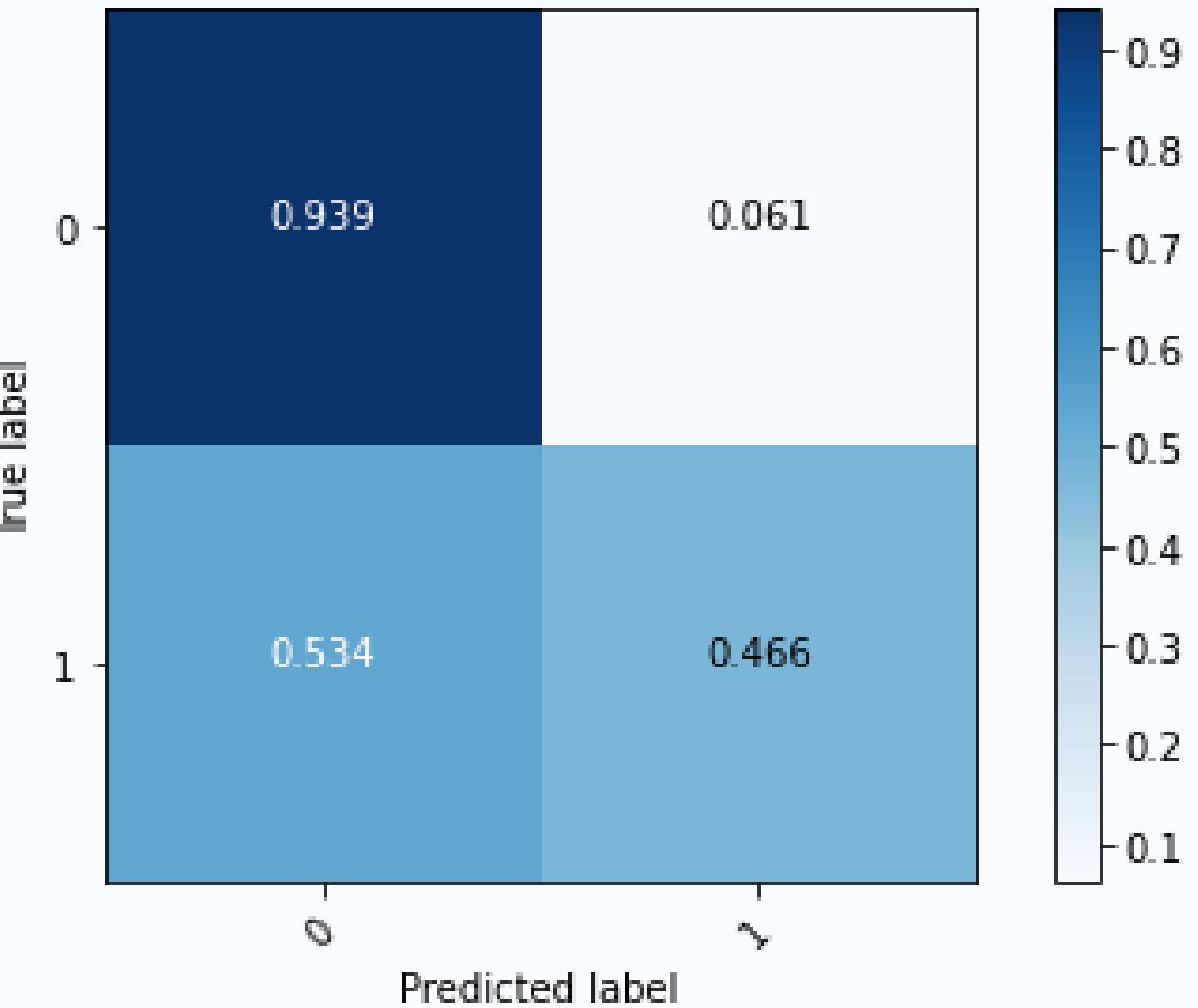


XGBOOST

NAJWYŻSZY POZIOM DOKŁADNOŚCI
- 97%.

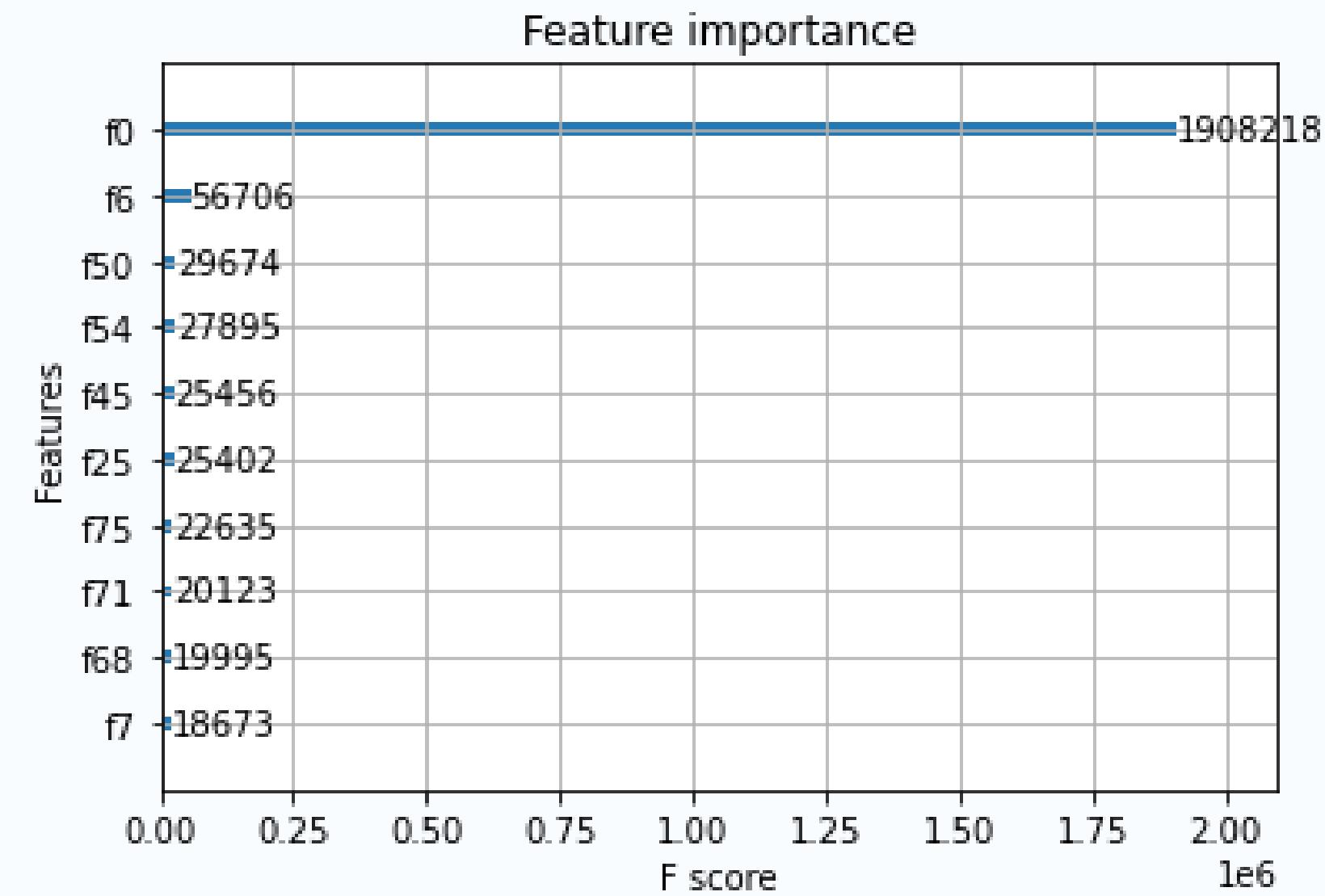


Confusion Matrix for XGBClassifier



XGBOOST

ISTOTA CECH
CHARAKTERYSTYCZNYCH



	score	name_features
f0	1908218	RATIO_IS_SUSPENDED_0.0
f6	56706	HAS_IP
f50	29674	RATIO_IS_SUSPENDED_1.0
f54	27895	AGE
f45	25456	PPK_BANK_6

Dziękujemy za uwagę
