WB XAI-2 PD1

Jakub Szypuła

22/03/2021

Cel zadania

Zadanie polega na przeanalizowaniu tego w jaki sposób zmienne wpływają na decyzje podjęte przez model ML i próbie wyjaśnienia tego. Na potrzeby zadania wykorzystałem las losowy zaimplementowany w pakiecie `ranger` na podstawie zbioru danych `german credit data`.

Pojedyncza predykcja

Zobaczmy jak model przewiduje wartości dla wybranej obserwacji, powiedzmy pierwszej.

```
predict(model, gcd[1,])$predictions
```

```
##           [,1]       [,2]
## [1,] 0.9193429 0.08065714
```

A teraz zobaczmy jak wygląda faktycznie ta wartość dla tej obserwacji.

```
gcd[1, "customer_type"]
```

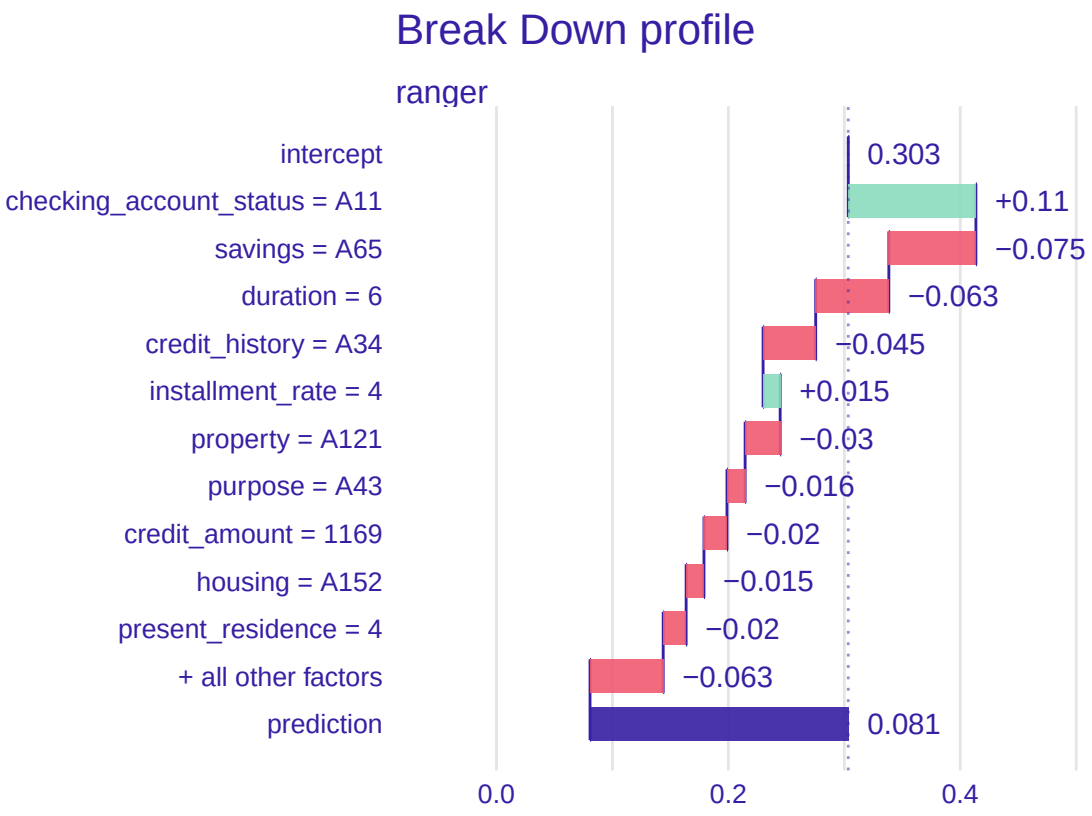
```
## [1] 1
```

Jak możemy więc zaobserwować, mówimy tutaj o prawdopodobieństwach zakwalifikowania danej obserwacji do danej klasy, gdzie 1 oznacza klasę “good”, a 2 klasę “bad”.

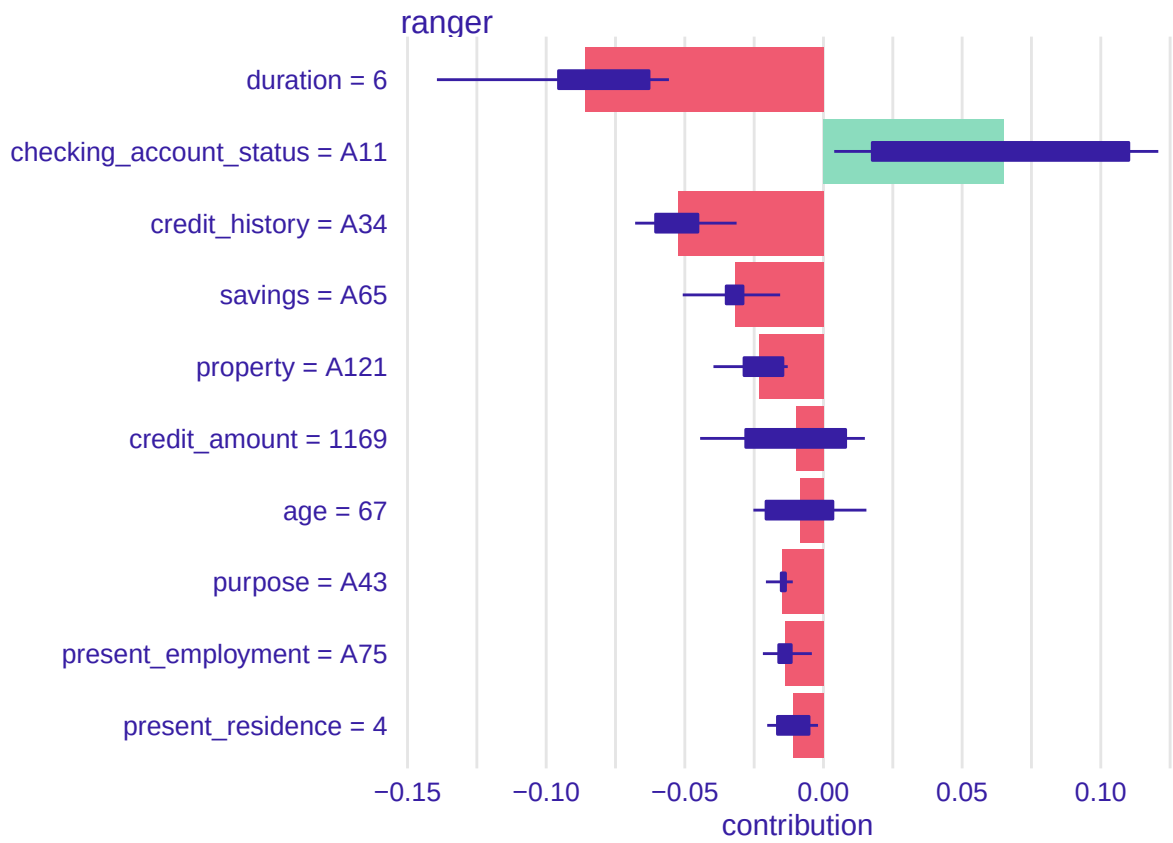
Dekompozycja

Zobaczmy, dlaczego model uznał, że taka, a nie inna wartość, pasuje w tym miejscu.

```
pp_ranger_gcd_1 <- predict_parts(explainer, new_observation = gcd[1,])
plot(pp_ranger_gcd_1)
```




```
pp_ranger_shap_gcd_1 <- predict_parts(explainer, new_observation = gcd[1,], type = "shap", B = 10)
plot(pp_ranger_shap_gcd_1)
```



"Prediction" na grafice oznacza prawdopodobieństwo zakwalifikowania obserwacji do klasy drugiej ("bad").

Dla modelu najważniejszą zmienną (poza interceptem, który wynosi 0.304) jest `checking_account_status` o wartości A11. Następnie jest `savings` o wartości A65, duration równe 6 oraz `credit_history` równe A34 i `property` równe A121. Po sprawdzeniu dokumentacji zbioru[1], te enigmatyczne wartości stają się mniej enigmatyczne, w nawiasach wpływ na predykcję modelu:

- A11 w `checking_account` oznacza mniej niż 0 Marek niemieckich na rachunku bieżącym (+0.108)
- 6 w `duration` oznacza liczbę miesięcy (-0.074)
- A34 w `credit_history` oznacza, że jest to "critical account" lub ma kredyty w innych bankach (-0.039)
- A121 w `property` oznacza, że osoba posiada nieruchomość (-0.032)

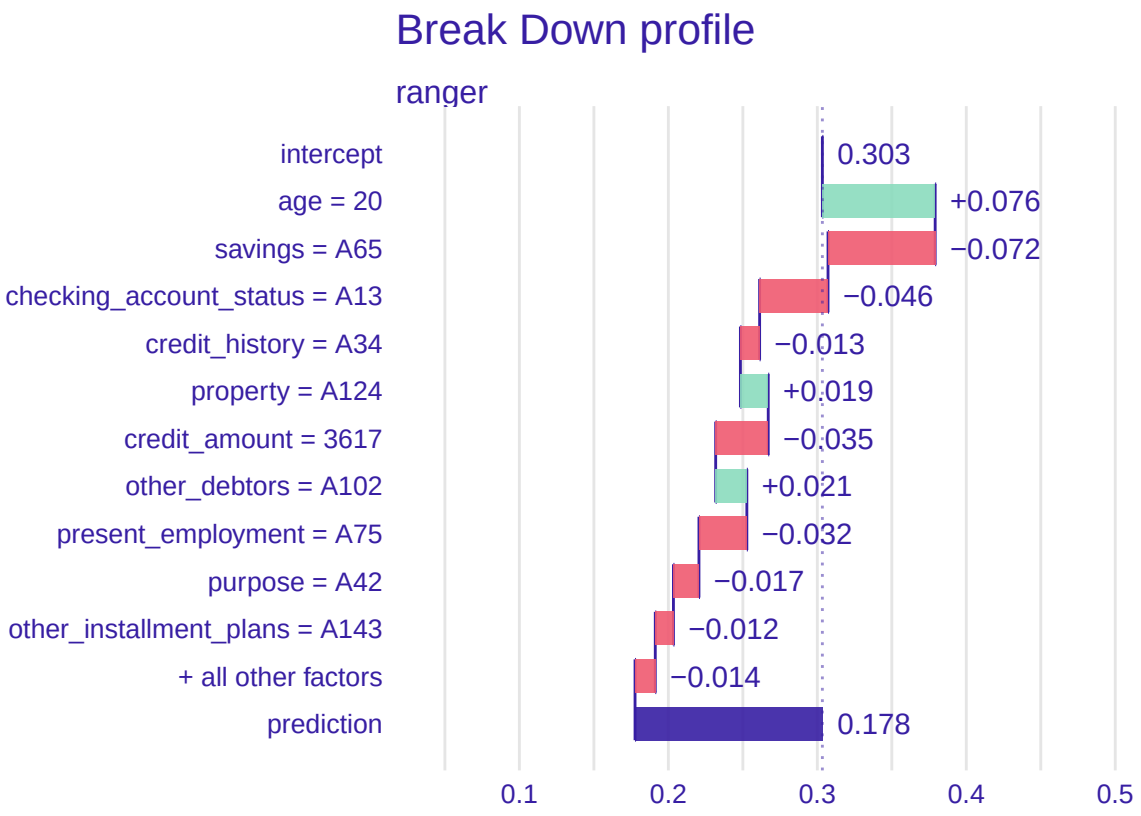
Można zauważyć, że są to sensowne wpływy, spodziewalibyśmy się pozytywnych i negatywnych wpływów po tych wartościach. Wydają się "osadzone" w świecie rzeczywistym.

Ważność zmiennych

Czy jednak bycie pod kreską i długość trwania kredytu są zawsze najważniejszymi zasadami, którymi model się posługuje?

```
pp_ranger_gcd_1_oth <- predict_parts(explainer, new_observation = gcd[94,])
plot(pp_ranger_gcd_1_oth)
```

[1] https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)



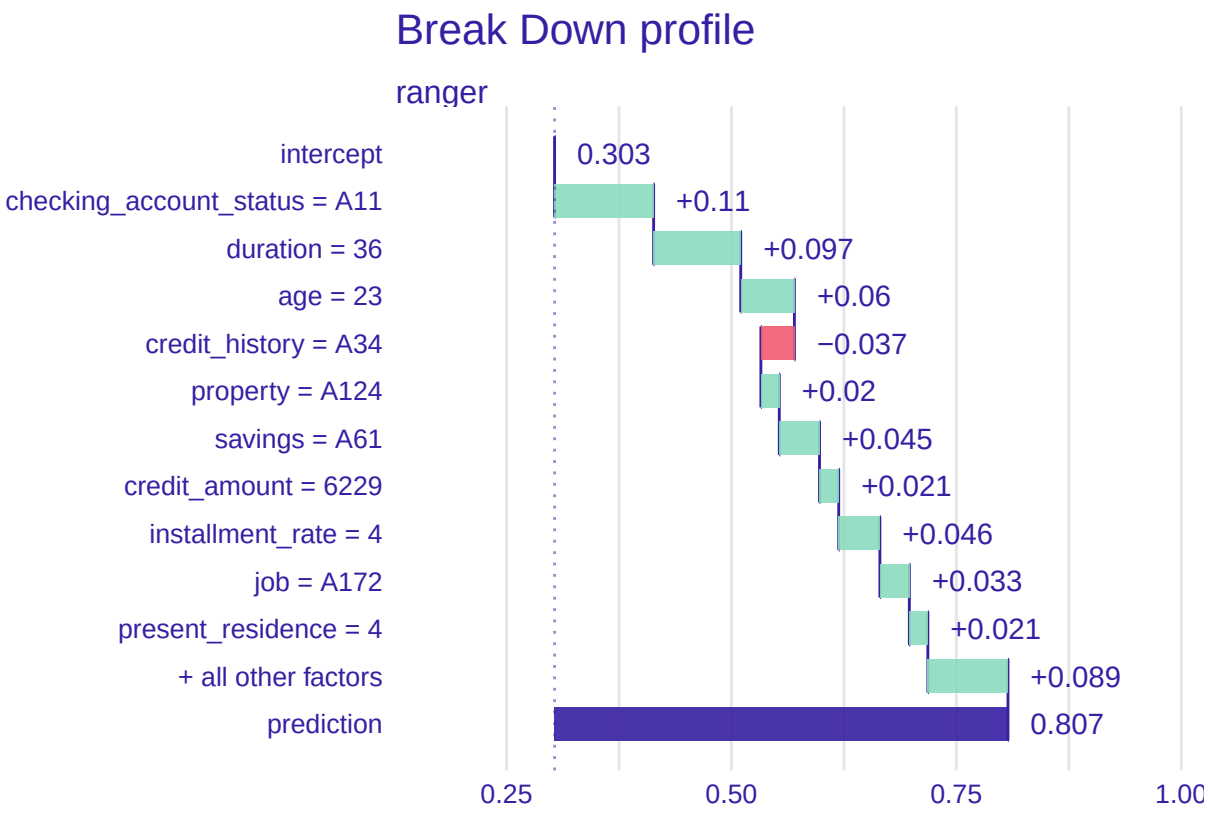
O ile wcześniej najważniejszymi zmiennymi były `checking_account` i `duration`, to teraz jest to `age` równe 20 (+0.075) oraz `savings` równe A65 (-0.069), czyli brak konta oszczędnościowego (bądź nieznane konto oszczędnościowe). Ponownie, wydaje się to sensowne biorąc pod uwagę fakt, że młodzi dorośli, świeży na rynku pracy, mogą być gorszymi kredytobiorcami. Informacja o rachunku oszczędnościowym mogła zostac uznana za ważniejszą np. dlatego, że osoba bez takiego rachunku będzie mieć do dyspozyjci większą część swojej pensji.

Przeciwne wpływy

Czasami zdarza się, że przy różnych wartościach innych zmiennych, ta sama wartość ma negatywny, bądź pozytywny wpływ na predykcję. Zobaczmy:

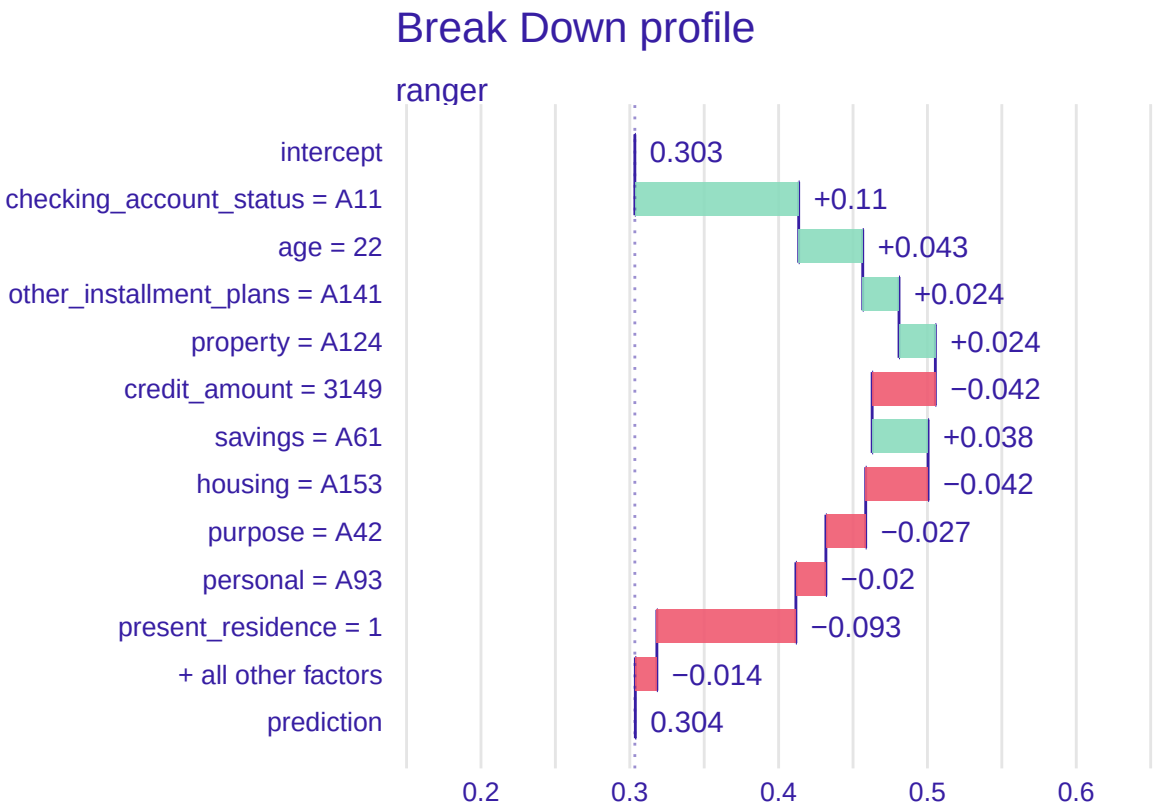
Obserwacja nr 60

```
pp_ranger_gcd_inv_1 <- predict_parts(explainer, new_observation = gcd[60,])
plot(pp_ranger_gcd_inv_1)
```



Obserwacja nr 230

```
pp_ranger_gcd_inv_2 <- predict_parts(explainer, new_observation = gcd[230,])
plot(pp_ranger_gcd_inv_2)
```



Na początku tego nie widać, ale zagłębmy się do wszystkich wartości:

```
gcd_inv_1 <- data.frame(
  pp_ranger_gcd_inv_1$variable_name,
  pp_ranger_gcd_inv_1$contribution,
  pp_ranger_gcd_inv_1$variable_value)

gcd_inv_2 <- data.frame(
  pp_ranger_gcd_inv_2$variable_name,
  pp_ranger_gcd_inv_2$contribution,
  pp_ranger_gcd_inv_2$variable_value)

gcd_inv_1 <- gcd_inv_1[1:22,]
gcd_inv_2 <- gcd_inv_2[1:22,]

colnames(gcd_inv_1) <- c("Variable_name", "Contribution1", "Value1")
colnames(gcd_inv_2) <- c("Variable_name", "Contribution2", "Value2")

comp <- merge(gcd_inv_1, gcd_inv_2, by = "Variable_name")
comp
```

```
##             Variable_name Contribution1 Value1 Contribution2 Value2
## 1                     age  0.0596210533     23  0.0430442815     22
## 2  checking_account_status  0.1101013026    A11  0.1101013026    A11
## 3           credit_amount  0.0213310881   6229 -0.0424657155   3149
## 4          credit_history -0.0372598066    A34 -0.0050261063    A32
## 5           customer_type  0.0000000000      2  0.0000000000      1
```

```
## 6               dependents  0.0023436651      1 -0.0038740746      1
## 7                 duration  0.0968444589     36  0.0090565663     24
## 8         existing_credits  0.0122134016      2 -0.0170014127      1
## 9           foreign_worker  0.0023186960   A201  0.0014514897   A201
## 10                 housing  0.0161248627   A151 -0.0419372003   A153
## 11        installment_rate  0.0457783690      4  0.0181231624      4
## 12               intercept  0.3033881592      1  0.3033881592      1
## 13                     job  0.0327269984   A172 -0.0099358238   A173
## 14           other_debtors  0.0043824640   A102  0.0010833730   A101
## 15 other_installment_plans  0.0006592325   A143  0.0243627571   A141
## 16                personal  0.0170666381    A92 -0.0201569952    A93
## 17      present_employment  0.0162878463    A72 -0.0001446966    A72
## 18       present_residence  0.0205984484      4 -0.0934100881      1
## 19                property  0.0204770865   A124  0.0244236972   A124
## 20                 purpose  0.0015399865    A42 -0.0267517603    A42
## 21                 savings  0.0447340024    A61  0.0375987496    A61
## 22               telephone  0.0160990310   A192 -0.0082153794   A191
```

Część zmiennych ma różny wpływ przy różnych wartościach (np. `job`), natomiast wyróżnia się jedna - `purpose`, która ma w obu obserwacjach wartość A42 (meble/wyposażenie) która dla obserwacji 60-tej zwiększa ryzyko bycia w złej kategorii o 0.07%, a dla obsweracji 230-tej zmniejsza je o 2.2%!

Nie jest to takie nieoczywiste. Weźmy za przykład przewidywanie cen mieszkań. Jeżeli zachowamy taki sam metraż, ale zwiększymy liczbę przedpokoi o jeden, to cena naturalnie spadnie, co samo w sobie wydaje się nieintuicyjne (jak dodanie pokoju miałoby obniżyć cenę?). Uwzględnienie tej zmiany w kontekście danych jest ważne dla zrozumienia wpływu na predykcję.

W tym konkretnym przypadku mówimy o osobach, które mają wiele wspólnego, więc skupię się na różnicach. 60-tka ma już 2 kredyty w tym banku, wynajmuje mieszkanie w którym mieszka od 4 lat i jest niewyszkolonym pracownikiem 230-tka tylko jeden kredyt, mieszka za darmo, mieszka tam od roku i jest wyszkolonym pracownikiem. Z perspektywy banku ma to sens, że osoba, która nie jest wyszkolonym pracownikiem, ma już dwa kredyty i mieszka w wynajmowanym mieszkaniu od 4 lat jest bardziej ryzykowna jeśli chce zakupić nowy mebel bądź wyposażenie. Natomiast osoba, która mieszka w obecnym miejscu zamieszkania od niedawna, jest wyszkolonym pracownikiem i ma tylko jeden kredyt może potrzebować tych mebli, więc szansa na spłacenie kredytu będzie nieznacznie wyższa, niż gdyby celem było coś innego. Oddaje to też ogólne prawdobieństwo zakwalifikowania do "złej" klasy - 60-tka ma aż 80%, zaś 230-tka tylko 28%

Podsumowanie

Jak widać, wpływy zmiennych i ich wartości na predykcję nie są takie oczywiste jak mogłyby się wydawać. Jak pokazuje ostatni przykład, wpływy te nie mogą być rozważane osobno, także jak pokazuje pierwsze porównanie, nie można z góry zakładać jakie zmienne są "najważniejsze". Przydatne w tym wypadku okazało się osadzenie danych w łatwym do zrozumienia i dosyć intuicyjnym kontekście, co pozwoliło na wysnucie wniosków i hipotez, które mogą pomóc wytłumaczyć przyczyny takiego, a nie innego zachowania modelu.