

# Analiza klientów kart kredytowych

---

AUTHOR

Mateusz Waszkiewicz

## 1 Wstęp

Poniższy tekst podejmuje się analizy klientów kart kredytowych i oceny ich ryzyka. Projekt został wzbogacony o elementy shiny i dla optymalnego komfortu powinien być oglądany na stronie projektu: [Link do strony](#).

Powodem takiego rozwiązania jest fakt, że shinyapps nie akceptuje dokumentów wzbogaconych o shiny (quarto).

### 1.1 Źródło danych

---

Dane opisujące dane kart kredytowych pochodzą ze zbioru [link do kaggle](#). Składają się z 2 plików płaskich:

- application\_record.csv - tabela z danymi dotyczącymi danych aplikacyjnych klienta (18 kolumn, jeden okres, jeden wiersz per klient).
- credit\_record.csv - tabela z historycznymi zaległościami w spłatach klienta (3 kolumny, 60 różnych okresów, x wierszy).

### 1.2 Cel

---

Analiza ma dwa główne cele i stawia dwie hipotezy badawcze. Pierwszym celem jest analiza i zbudowanie profilu klienta - określenie jego podstawowych cech i zależności między cechami. Przetworzenie tych cech w zależności od poziomu ryzyka braku terminowej spłaty karty. Drugim celem jest zbudowanie modelu predykcyjnego, który by rozróżniał dobrych i złych klientów na podstawie zarówno jego cech aplikacyjnych, jak i historii spłat. W analizie stawiane są dwie hipotezy:

- Klienci o wyższych zarobkach są bardziej terminowi w spłatach.
- Klienci posiadający samochód są bardziej terminowi w spłatach.

### 1.3 Przygotowanie danych

---

#### 1.3.1 Krótkie podsumowanie

##### 1.3.1.1 Zbiór application\_record

Wszystkie kolumny w zbiorze danych to:

Lista kolumn

| Nazwa kolumny       |
|---------------------|
| ID                  |
| CODE_GENDER         |
| FLAG_OWN_CAR        |
| FLAG_OWN_REALTY     |
| CNT_CHILDREN        |
| AMT_INCOME_TOTAL    |
| NAME_INCOME_TYPE    |
| NAME_EDUCATION_TYPE |
| NAME_FAMILY_STATUS  |
| NAME_HOUSING_TYPE   |
| DAYS_BIRTH          |
| DAYS_EMPLOYED       |
| FLAG_MOBIL          |
| FLAG_WORK_PHONE     |
| FLAG_PHONE          |
| FLAG_EMAIL          |
| OCCUPATION_TYPE     |
| CNT_FAM_MEMBERS     |

Wymiary zbioru to: 438557x18. Liczba unikatowych ID w zbiorze to: 438510.

Liczba unikatowych ID nie jest równa liczbie wierszy, zatem występują duplikaty. Po usunięciu duplikatów wymiary zbioru to: 438463x18.

Podsumowanie brakujących wartości w zbiorze danych przedstawia się następująco:

| Podsumowanie brakujących wartości |                    |
|-----------------------------------|--------------------|
|                                   | Brakujące wartości |
| ID                                | 0                  |
| CODE_GENDER                       | 0                  |

| Brakujące wartości  |   |
|---------------------|---|
| FLAG_OWN_CAR        | 0 |
| FLAG_OWN_REALTY     | 0 |
| CNT_CHILDREN        | 0 |
| AMT_INCOME_TOTAL    | 0 |
| NAME_INCOME_TYPE    | 0 |
| NAME_EDUCATION_TYPE | 0 |
| NAME_FAMILY_STATUS  | 0 |
| NAME_HOUSING_TYPE   | 0 |
| DAYS_BIRTH          | 0 |
| DAYS_EMPLOYED       | 0 |
| FLAG_MOBIL          | 0 |
| FLAG_WORK_PHONE     | 0 |
| FLAG_PHONE          | 0 |
| FLAG_EMAIL          | 0 |
| OCCUPATION_TYPE     | 0 |
| CNT_FAM_MEMBERS     | 0 |

Podstawowe statystyki opisowe zbioru danych dla zmiennych ilościowych i jakościowych przedstawiają się następująco:

| Podstawowe statystyki zmiennych numerycznych |              |                  |            |               |
|--|--------------|------------------|------------|---------------|
|  | CNT_CHILDREN | AMT_INCOME_TOTAL | DAYS_BIRTH | DAYS_EMPLOYED |
| Min.   | 0.0          | 26100            | -25201.0   | -17531.0      |
| 1st Qu.                                      | 0.0          | 121500           | -19484.0   | -3103.0       |
| Median                                       | 0.0          | 161100           | -15630.0   | -1468.0       |
| Mean   | 0.4          | 187522           | -15997.9   | 60564.5       |
| 3rd Qu.                                      | 1.0          | 225000           | -12514.0   | -371.0        |
| Max.   | 19.0         | 6750000          | -7489.0    | 365243.0      |

| CNT_CHILDREN   | AMT_INCOME_TOTAL | DAYS_BIRTH | DAYS_EMPLOYED | CNT_FA           |
|--|------------------|------------|---------------|------------------|
| Liczby wystąpień wartości zmiennej CODE_GENDER         |                  |            |               |                  |
|  |                  |            |               | Liczba wystąpień |
| F  |                  |            |               | 294378           |
| M  |                  |            |               | 144085           |
| Liczby wystąpień wartości zmiennej FLAG_OWN_CAR        |                  |            |               |                  |
|  |                  |            |               | Liczba wystąpień |
| N  |                  |            |               | 275396           |
| Y  |                  |            |               | 163067           |
| Liczby wystąpień wartości zmiennej FLAG_OWN_REALTY     |                  |            |               |                  |
|  |                  |            |               | Liczba wystąpień |
| N  |                  |            |               | 134454           |
| Y  |                  |            |               | 304009           |
| Liczby wystąpień wartości zmiennej NAME_INCOME_TYPE    |                  |            |               |                  |
|  |                  |            |               | Liczba wystąpień |
| Commercial associate                                   |                  |            |               | 100726           |
| Pensioner  |                  |            |               | 75478            |
| State servant  |                  |            |               | 36183            |
| Student  |                  |            |               | 17               |
| Working  |                  |            |               | 226059           |
| Liczby wystąpień wartości zmiennej NAME_EDUCATION_TYPE |                  |            |               |                  |
|  |                  |            |               | Liczba wystąpień |
| Academic degree  |                  |            |               | 312              |
| Higher education                                       |                  |            |               | 117499           |

|                               | Liczba wystąpień |
|-------------------------------|------------------|
| Incomplete higher             | 14845            |
| Lower secondary               | 4051             |
| Secondary / secondary special | 301756           |

## Liczby wystąpień wartości zmiennej NAME\_FAMILY\_STATUS

|                      | Liczba wystąpień |
|----------------------|------------------|
| Civil marriage       | 36521            |
| Married              | 299768           |
| Separated            | 27249            |
| Single / not married | 55255            |
| Widow                | 19670            |

## Liczby wystąpień wartości zmiennej NAME\_HOUSING\_TYPE

|                     | Liczba wystąpień |
|---------------------|------------------|
| Co-op apartment     | 1539             |
| House / apartment   | 393748           |
| Municipal apartment | 14211            |
| Office apartment    | 3920             |
| Rented apartment    | 5973             |
| With parents        | 19072            |

## Liczby wystąpień wartości zmiennej FLAG\_MOBIL

|  | Liczba wystąpień |
|--|------------------|
|  | 438463           |

## Liczby wystąpień wartości zmiennej FLAG\_WORK\_PHONE

|   | Liczba wystąpień |
|---|------------------|
| 0 | 348083           |

|  |  | Liczba wystąpień |
|--|--|------------------|
| 1  |  | 90380            |
| Liczby wystąpień wartości zmiennej FLAG_PHONE      |  |                  |
|  |  | Liczba wystąpień |
| 0  |  | 312290           |
| 1  |  | 126173           |
| Liczby wystąpień wartości zmiennej FLAG_EMAIL      |  |                  |
|  |  | Liczba wystąpień |
| 0  |  | 391023           |
| 1  |  | 47440            |
| Liczby wystąpień wartości zmiennej OCCUPATION_TYPE |  |                  |
|  |  | Liczba wystąpień |
|  |  | 134177           |
| Accountants  |  | 15980            |
| Cleaning staff                                     |  | 5843             |
| Cooking staff                                      |  | 8074             |
| Core staff   |  | 42993            |
| Drivers  |  | 26085            |
| High skill tech staff                              |  | 17283            |
| HR staff   |  | 774              |
| IT staff   |  | 604              |
| Laborers   |  | 78225            |
| Low-skill Laborers                                 |  | 2140             |
| Managers   |  | 35477            |
| Medicine staff                                     |  | 13518            |
| Private service staff                              |  | 3455             |

|                      | Liczba wystąpień |
|----------------------|------------------|
| Realty agents        | 1041             |
| Sales staff          | 41092            |
| Secretaries          | 2044             |
| Security staff       | 7993             |
| Waiters/barmen staff | 1665             |

1.3.1.2 Zbiór credit\_record

Wszystkie kolumny w zbiorze danych to:

| Nazwy kolmn    |
|----------------|
| Nazwa kolumny  |
| ID             |
| MONTHS_BALANCE |
| STATUS         |

Wymiary zbioru to: 1048575x3. Liczba unikatowych ID w zbiorze to: 45985. Jest to liczba większa niż w pierwszym zbiorze. Będzie trzeba na to zwrócić uwagę podczas łączenia tabel.

Podsumowanie brakujących wartości w zbiorze danych przedstawia się następująco:

| Podsumowanie brakujących wartości |                    |
|-----------------------------------|--------------------|
|                                   | Brakujące wartości |
| ID                                | 0                  |
| MONTHS_BALANCE                    | 0                  |
| STATUS                            | 0                  |

Podstawowe statystyki opisowe zbioru danych dla zmiennych ilościowych i jakościowych przedstawiają się następująco:

| Podstawowe statystyki zmiennych numerycznych |                |
|--|----------------|
|  | MONTHS_BALANCE |
| Min.   | -60.00         |
| 1st Qu.                                      | -29.00         |

|         | MONTHS_BALANCE |
|---------|----------------|
| Median  | -17.00         |
| Mean    | -19.14         |
| 3rd Qu. | -7.00          |
| Max.    | 0.00           |

Liczby wystąpień wartości zmiennej STATUS

|   | Liczba wystąpień |
|---|------------------|
| 0 | 383120           |
| 1 | 11090            |
| 2 | 868              |
| 3 | 320              |
| 4 | 223              |
| 5 | 1693             |
| C | 442031           |
| X | 209230           |

Zbiory nie zawierają braków danych, a wszystkie odpowiednie kolumny zostały zamienione na czynnik (ang. factor), kolejnym krokiem będzie połączenie zbiorów w jeden i stworzenie odpowiednich zmiennych oznaczających opóźnienia w spłatach.

## 1.4 Połączenie zbiorów

### 1.4.1 Stworzenie zmiennej celu

Przed połączeniem obu zbiorów należy dostosować zbiór **credit\_record** do formy docelowej (jeden wiersz per ID). Kolejnym krokiem będzie zatem pogrupowanie zbioru ze względu na ID, jednak aby tego dokonać najpierw należy się bliżej przyjrzeć zmiennej **STATUS**. Zmienna **STATUS** jest tutaj jedną z najważniejszych, ponieważ symbolizują opóźnienie w spłacie, jej oznaczenia są następujące:

- 0 - do 29 dni opóźnienia.
- 1 - od 30 do 59 dni opóźnienia.
- 2 - od 60 do 89 dni opóźnienia.
- 3 - od 90 do 119 dni opóźnienia.
- 4 - od 120 do 149 dni opóźnienia.
- 5 - powyżej 150 dni opóźnienia, zaległe długi albo spisanie.



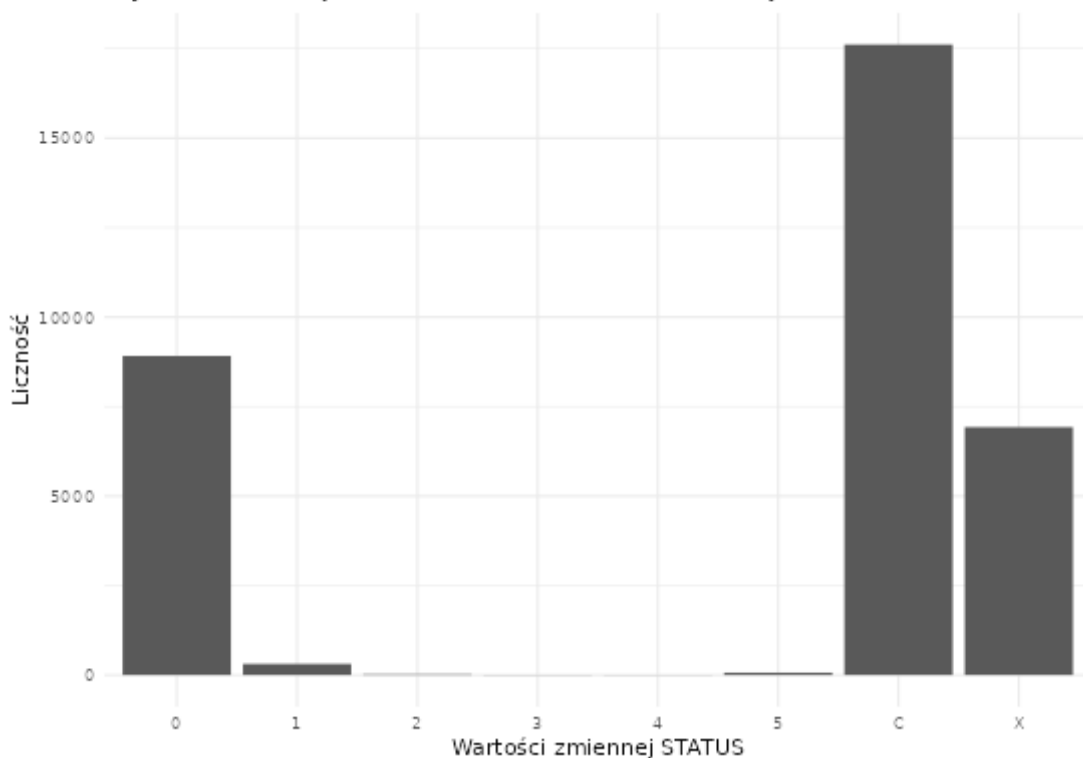
- C - w pełni spłacone.
- X - brak kredytu w danym miesiącu.

W celu stworzenia optymalnej zmiennej celu oznaczającą złego klienta, została dokonana dodatkowa analiza zmiennej **TARGET** w zależności od wybranego okna obserwacji. Rozważane okna to - aktualny miesiąc, ostatnie 3 miesiące, ostatnie pół roku, ostatni rok, cała historia. Zbudowana próbka do analizy jest oparta na tym czy kiedykolwiek w takim oknie wystąpiło opóźnienie spłaty dla danego ID.

Wybrany okres

ostatni miesiąc ▼

Wykres zmiennej STATUS w okresie: ostatni miesiąc



W oparciu o powyższy wykres dość jasno można stwierdzić, że opóźnienia powyżej 30 dni zdarzają się bardzo sporadycznie i zbudowanie zmiennej celu na ich podstawie nie ma większego sensu. W efekcie została stworzona nowa zmienna pomocnicza, która definiuje dobrych i złych klientów w sposób następujący:

$$isBad = \begin{cases} 1 & \text{jeśli STATUS} \in \{1, 2, 3, 4, 5\} \\ 0 & \text{w innym przypadku} \end{cases}$$

Jest to niezgodne z ogólnoświatowym standardem bankowym, w którym opóźnienie w spłacie powyżej 90 dni oznacza wejście w stan default, jednak jak widać tylko nieznaczna część klientów ma opóźnienia > 90, co może utrudnić późniejsze analizy. Zostały stworzone następujące zmienne pomocnicze:

- *isBad* - czy klient jest aktualnie w stanie default.
- *isBad\_3* - czy klient był w stanie default w ciągu ostatnich 3 miesięcy.
- *isBad\_6* - czy klient był w stanie default w ciągu ostatnich 6 miesięcy.
- *isBad\_12* - czy klient był w stanie default w ciągu ostatnich 12 miesięcy.

- *isBad\_ever* - czy klient kiedykolwiek był w stanie default.

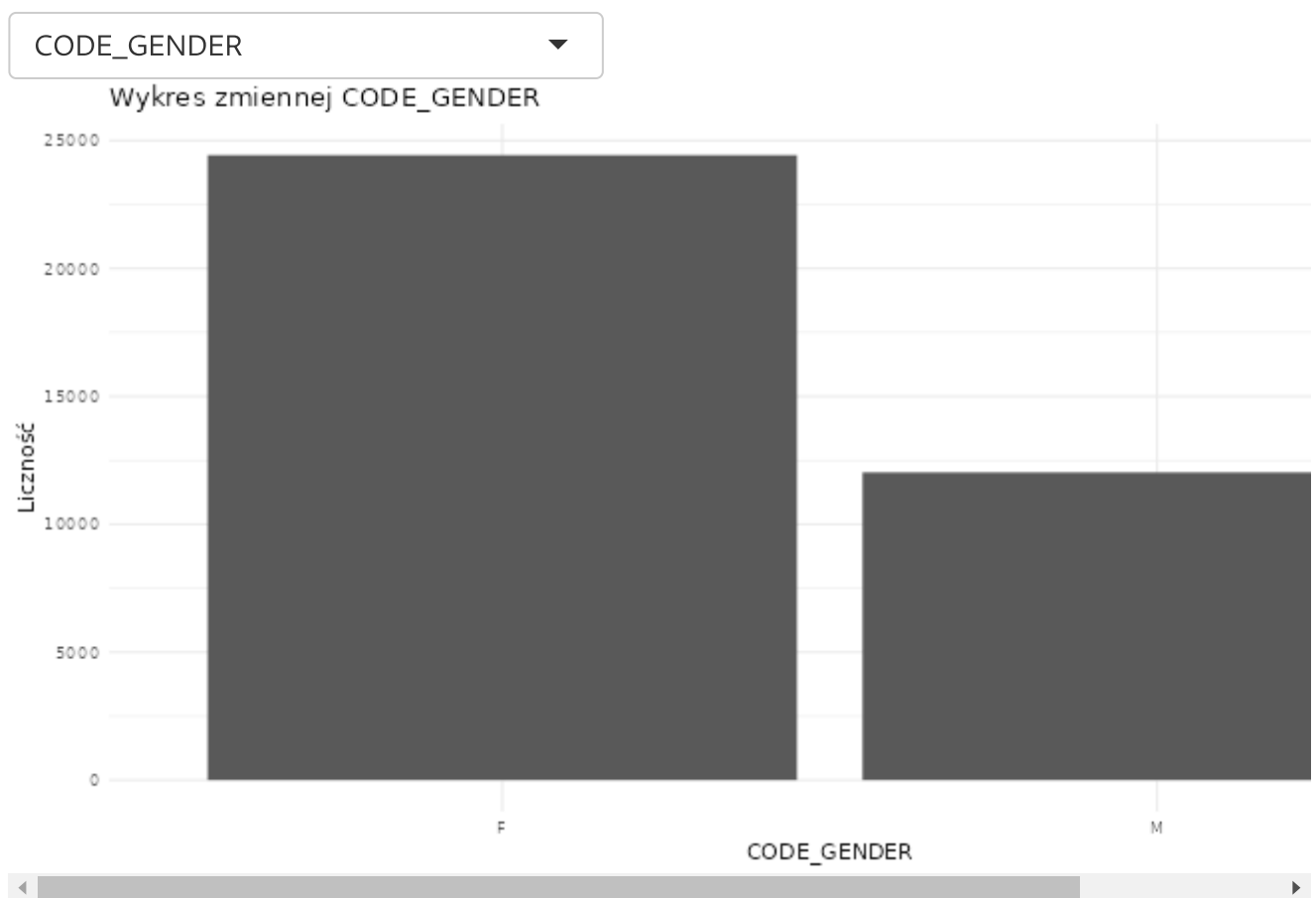
W idealnych warunkach takie zmienne wykorzystane w modelowaniu powinny patrzeć w przyszłość, a nie w przeszłość, natomiast tylko na taką możliwość pozwalają wybrane dane.

## 1.4.2 Sprawdzenie ID

Oba zbiory zostały połączone za pomocą kolumny **ID**. Tylko klienci znajdujący się w obu początkowych zbiorach znajdują się w połączonej tabeli. Nowy zbiór ma 36457 wierszy. Znaczne ograniczenie liczby wierszy wynika z małego pokrycia **ID** z tabeli `application_record` przez tabelę `credit_record`.

## 2 Profil klienta

Zmienna



Postawowe wykresy zmiennych opisujących cechy klientów nie są bardzo interesujące. Parę ciekawszych wniosków ogólnych dotyczących danych: - Wszyscy klienci mają telefon komórkowy (ta zmienna została wykluczona z późniejszego modelowania). - W zbiorze jest 2 razy więcej kobiet niż mężczyzn. - Zdecydowana większość nie ma dzieci. - Zdecydowana większość ma średnią edukację. - Telefony domowe (zmienna **FLAG\_PHONE**) są mało popularne w badanym zbiorze. - Zmienna dotycząca posiadania samochodu (istotna z punktu widzenia stawianych hipotez) ma relatywnie równy rozkład.

## 2.1 Profil klienta ze względu na poziom ryzyka

Poniższy wykres prezentuje profil klienta ze względu na nowo stworzone zmienne ryzyka. Wykres ma możliwość pokazania wartości absolutnych, jak i wartości procentowych udziałów zmiennej ryzyka. Dzięki temu można łatwiej werfikować proporcję złych i dobrych klientów w ramach poszczególnych zmiennych.

Zmienna

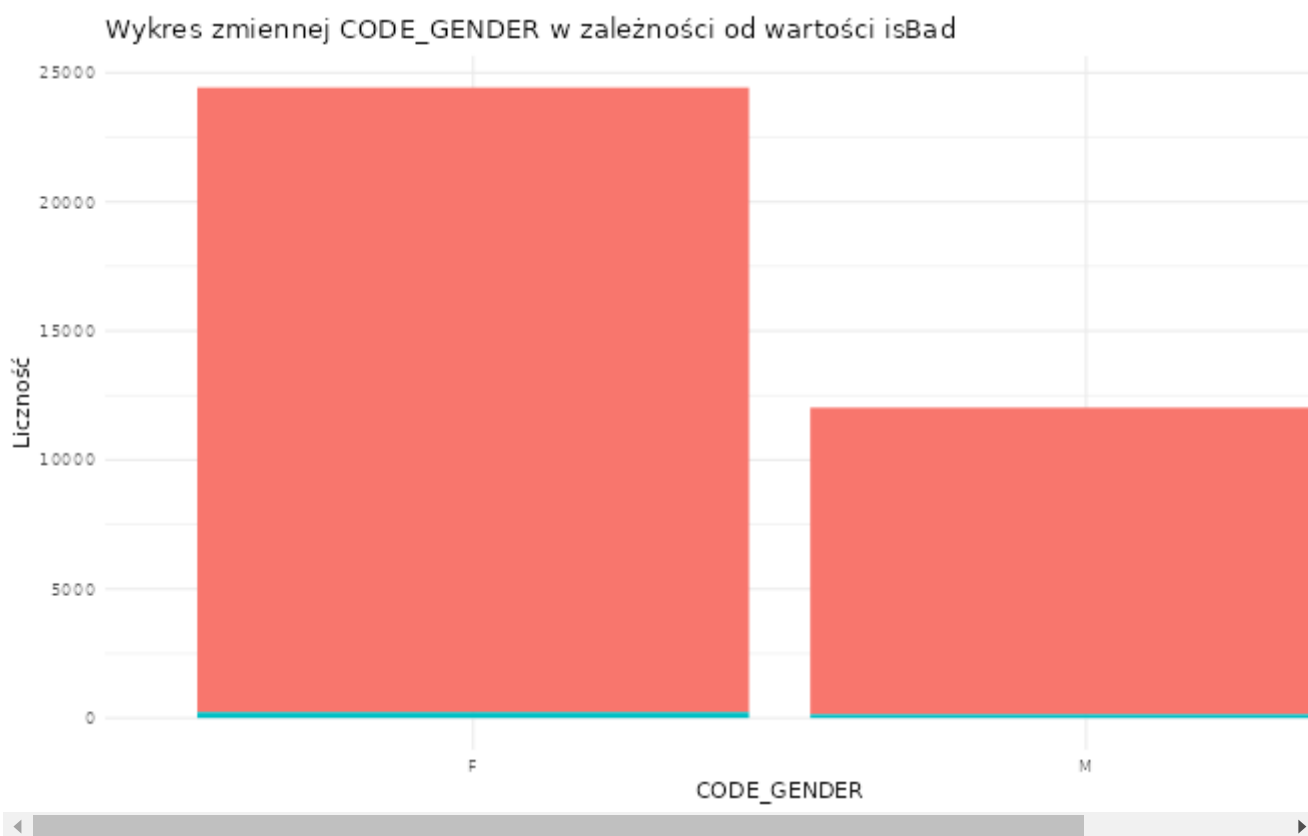
CODE\_GENDER ▼

Zmienna ryzyka

isBad ▼

Typ wykresu kolumnowego

Wartości absolutne ▼



Większość zmiennych zdaje się mieć podobną proporcję ryzykownych klientów w ramach grup. Posiadający samochód jednak zdają się być mniej ryzykowni od klientów, którzy samochodu nie posiadają. Parę innych obserwacji dotyczących poszczególnych w ramach swoich dziedzin: - Studenci są najbardziej ryzykowną grupą. - Ludzie z wyższą edukacją są najmniej ryzykowną grupą. - Single są zdecydowanie najbardziej ryzykowną grupą.

### 3 Modelowanie

Zmienną wybraną jako zmienna celu w modelowaniu została zmienna **isBad\_12**, 12 miesięczne okna obserwacji są dość klasycznym podejściem w ryzyku kredytowym. Taki okres gwarantuje wystarczająco dużą liczbę złych klientów do analizy. Dane zostały podzielone na zbiór treningowy i testowy w stosunku 7 do 3.

### 3.1 Regresja logistyczna

Modelem, który będzie służył do weryfikacji hipotez będzie prosty model regresji logistycznej. Zostały stworzone 2 modele - jeden prosty ze wszystkimi zmiennymi, drugi - model stepwise, w którym nastąpi automatyczna selekcja zmiennych na podstawie kryterium AIC.

|   | <i>Dependent variable:</i> |             |
|---|----------------------------|-------------|
|   | isBad_12                   |             |
|   | (1)                        | (2)         |
| CODE_GENDERM                                      | 0.075                      |             |
| FLAG_OWN_CARY                                     | -0.134**                   | -0.129**    |
| FLAG_OWN_REALTY                                   | -0.109*                    | -0.125**    |
| CNT_CHILDREN                                      | -2.112***                  | -2.155***   |
| AMT_INCOME_TOTAL                                  | 0.00000***                 | 0.00000***  |
| NAME_INCOME_TYPEPensioner                         | 2.885***                   | 2.879***    |
| NAME_INCOME_TYPEState servant                     | -0.083                     | -0.034      |
| NAME_INCOME_TYPEStudent                           | -11.025                    | -10.990     |
| NAME_INCOME_TYPEWorking                           | -0.119*                    | -0.109*     |
| NAME_EDUCATION_TYPEHigher education               | 10.765                     | 10.877      |
| NAME_EDUCATION_TYPEIncomplete higher              | 10.928                     | 11.033      |
| NAME_EDUCATION_TYPERLower secondary               | 10.392                     | 10.525      |
| NAME_EDUCATION_TYPERSecondary / secondary special | 10.906                     | 11.023      |
| NAME_FAMILY_STATUSSMarried                        | -0.049                     | -0.058      |
| NAME_FAMILY_STATUSSSeparated                      | 2.061***                   | 2.085***    |
| NAME_FAMILY_STATUSSSingle / not married           | 2.114***                   | 2.147***    |
| NAME_FAMILY_STATUSSWidow                          | 1.870***                   | 1.897***    |
| NAME_HOUSING_TYPEHouse / apartment                | 0.971*                     |             |
| NAME_HOUSING_TYPERMunicipal apartment             | 0.900                      |             |
| NAME_HOUSING_TYPEROffice apartment                | 1.398**                    |             |
| NAME_HOUSING_TYPERRented apartment                | 1.086*                     |             |
| NAME_HOUSING_TYPERWith parents                    | 1.144*                     |             |
| DAYS_BIRTH  | 0.00004***                 | 0.00004***  |
| DAYS_EMPLOYED                                     | -0.00001***                | -0.00001*** |
| FLAG_WORK_PHONE1                                  | -0.123*                    | -0.129**    |
| FLAG_PHONE1                                       | -0.004                     |             |
| FLAG_EMAIL1                                       | 0.207**                    | 0.205**     |
| OCCUPATION_TYPEAccountants                        | 0.002                      |             |
| OCCUPATION_TYPECleaning staff                     | 0.337                      |             |
| OCCUPATION_TYPECooking staff                      | 0.467***                   |             |
| OCCUPATION_TYPECore staff                         | 0.187*                     |             |
| OCCUPATION_TYPERDrivers                           | -0.020                     |             |
| OCCUPATION_TYPEHigh skill tech staff              | 0.197                      |             |

|                                      |                                     |            |
|--------------------------------------|-------------------------------------|------------|
| OCCUPATION_TYPEHR staff              | 1.167***                            |            |
| OCCUPATION_TYPEIT staff              | 0.546                               |            |
| OCCUPATION_TYPELaborers              | 0.135                               |            |
| OCCUPATION_TYPELow-skill Laborers    | 0.469                               |            |
| OCCUPATION_TYPEManagers              | 0.184                               |            |
| OCCUPATION_TYPEMedicine staff        | 0.266*                              |            |
| OCCUPATION_TYPEPrivate service staff | -0.586                              |            |
| OCCUPATION_TYPEDealors agents        | 0.307                               |            |
| OCCUPATION_TYPESales staff           | 0.005                               |            |
| OCCUPATION_TYPESecretaries           | 0.376                               |            |
| OCCUPATION_TYPESecurity staff        | 0.305                               |            |
| OCCUPATION_TYEWaiters/barmen staff   | 0.353                               |            |
| CNT_FAM_MEMBERS                      | 2.104***                            | 2.143***   |
| Constant                             | -18.237                             | -17.265    |
| Observations                         | 25,591                              | 25,591     |
| Log Likelihood                       | -6,082.103                          | -6,102.805 |
| Akaike Inf. Crit.                    | 12,258.210                          | 12,249.610 |
| Note:                                | $p < 0.1$ ; $p < 0.05$ ; $p < 0.01$ |            |

Wnioski z regresji: - Klienci posiadający samochód mają istotnie mniejsze ryzyko. - Rodzaj zarobków nie ma większego znaczenia o ile nie jest to emerytura, natomiast wielkość zarobków już ma. Podobnie prezentuje się podsumowanie zawodów - konkretne nie mają znaczenia, ale już czas na rynku pracy owszem. - Edukacja okazała się zmienną nieistotną. - Zmienne związane z rodziną były istotnymi czynnikami - zarówno liczba dzieci, liczba członków rodziny, jak i status matrymonialny były istotnymi zmiennymi.

## 3.2 Drzewa decyzyjne

Kolejnym rozważanym modelem były drzewa decyzyjne. Zostały skonstruowane 3 drzewa decyzyjne o różnej wielkości (ze względu na **Complexity Parameter**).

Wielkość drzewa

małe





Drzewa decyzyjne okazały się słabym wyborem dla tego zbioru danych. Małe drzewa w ogóle nie mają podziałów.

### 3.3 Porównanie wyników

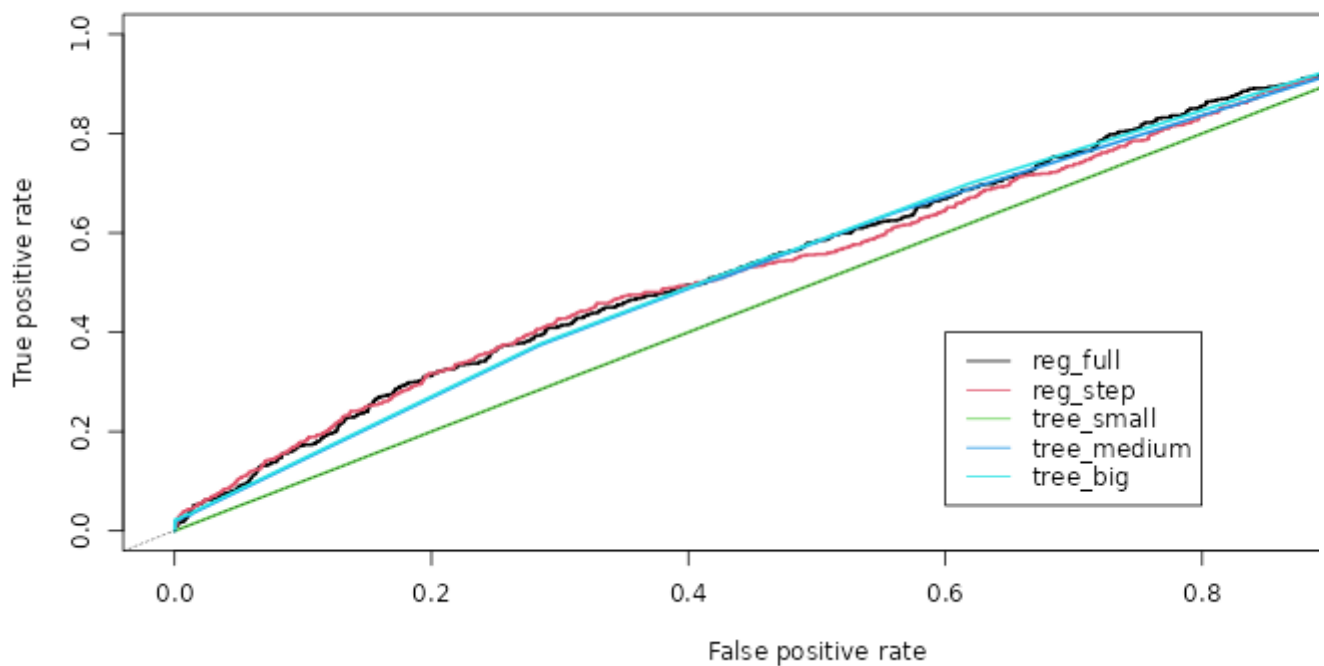
Jako ostatni etap porównano różne skonstruowane modele ze względu na własności statystyczne. Wszystkie statystyki są dostępne do wglądu zarówno na zbiorze treningowym, jak i testowym.

Zbiór do porównania wyników

testowy

|             | Pełna regresja | Regresja stepwise | Małe drzewo | Średnie drzewo | Duże drzewo |
|-------------|----------------|-------------------|-------------|----------------|-------------|
| accuracy    | 0.935          | 0.935             | 0.935       | 0.936          | 0.936       |
| MER         | 0.065          | 0.065             | 0.065       | 0.064          | 0.064       |
| precision   | 0.4            | 0.333             | NaN         | 0.875          | 0.8         |
| sensitivity | 0.003          | 0.003             | 0           | 0.02           | 0.023       |
| specificity | 1              | 1                 | 1           | 1              | 1           |
| F1          | 0.006          | 0.006             | NaN         | 0.039          | 0.044       |

#### 3.3.1 Krzywa ROC



| Model       | AUC   |
|-------------|-------|
| reg_full    | 0.570 |
| reg_step    | 0.563 |
| tree_small  | 0.500 |
| tree_medium | 0.556 |
| tree_big    | 0.562 |

Wszystkie modele prezentują bardzo słabą moc predykcyjną, jednak regresja logistyczna wydaje się radzić lepiej. Wszystkie modele mają minimalnie gorsze właściwości na zbiorze testowym, jednak tam dalej ten trend się utrzymuje - regresja logistyczna jest lepszym wyborem.

## 4 Podsumowanie

Zbiór okazał się być mocno problematyczny do modelowania - brak wymiaru czasu w **application\_record**, patrzenie wstecz zamiast do przodu w **credit\_record** i mocno ograniczony iloczyn tych zbiorów po połączeniu za pomocą ID znacząco utrudniały zadanie. Nie oznacza to, że nie da się wyciągnąć żadnych wniosków - obie hipotezy zostały potwierdzone za pomocą analizy graficznej i regresji logistycznej.