**Uniwersytet Gdański w Sopocie**

**Wydział Nauk Zarządzania**

**Marta Jagowdzik**

nr albumu: 287048

kierunek: Informatyka i ekonometria

Raport

**Modele nieparametryczne - projekt zaliczeniowy**

Prowadzący zajęcia - dr Arkadiusz Kozłowski

Sopot 2024

Analizie poddany zostanie zbiór dotyczący płatności/transakcji kartowych. Transakcje są dwojakiego rodzaju:

- transakcje inicjalizujące,

- transakcje rekurencyjne

Baza danych zawiera transakcje rekurencyjne dokonane między 8 lipa 2020, a 7 lipca 2021, pod warunkiem, że transakcja inicjalizująca miała miejsce nie wcześniej niż 8 stycznia 2020 roku.

Celem jest stworzenie dwóch modeli predykcyjnych na bazie drzewa decyzyjnego;

-Model klasyfikacji, który pozwala ocenić, czy dana transakcja rekurencyjna jest udana,

-Model regresji, który pozwala oszacować wartość kwoty przelewu.

**ETAPY PRACY:**

Proces modelowania będzie obejmować następujące etapy:

-Konfiguracja środowiska pracy,

-Przetwarzanie wstępne danych(preprocessing),

-Tworzenie drzewa decyzyjnego do klasyfikacji, szkolenie modelu, ocena jego skuteczności oraz generowanie prognoz na zbiorze testowym,

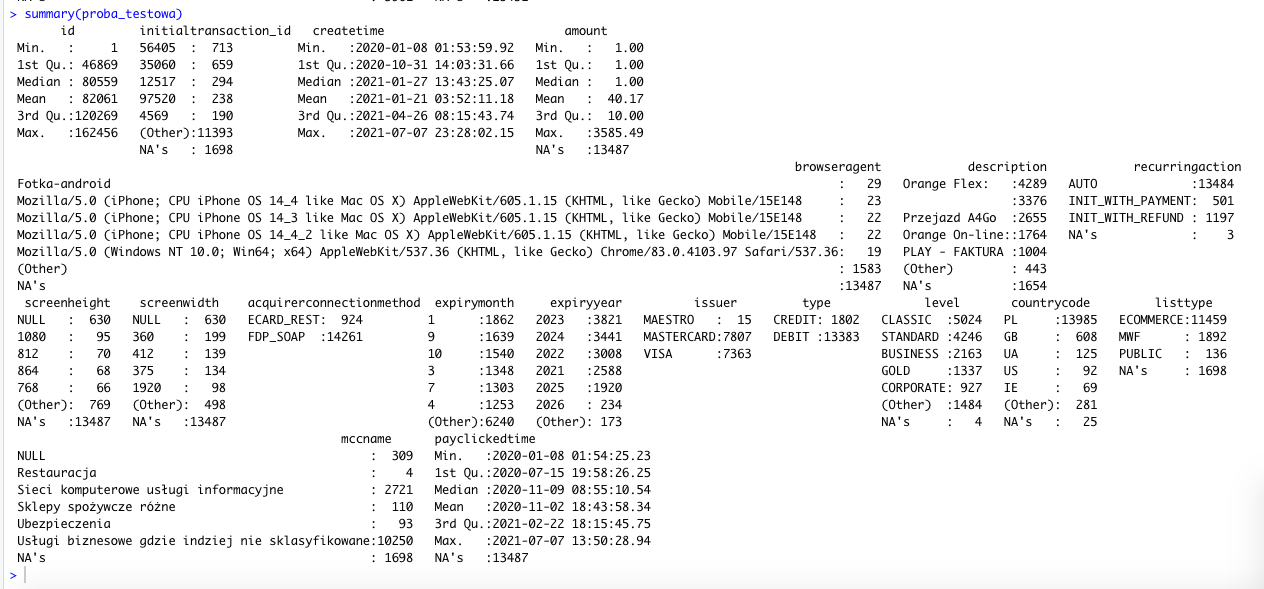
-Tworzenie drzewa decyzyjnego do regresji, szkolenie modelu, ocena jego skuteczności oraz generowanie prognoz na zbiorze testowym.

**Dokonane kroki podczas wykonywania zadania:**

1. Wstępna eksploracja danych.
2. Dokonano preprocessing na zbiorze danych „próba ucząca” oraz „próba testowa”. Użyto funkcji **factor**, która konwertuje wektor (w tym przypadku kolumnę danych) na zmienną kategoryczną.

Obraz zawierający paragon, tekst, algebra

Opis wygenerowany automatycznie



1. Analiza wstępna zestawu danych wraz z ilustracjami.

* Wizualizacja dla zmiennej "status"

Obraz zawierający zrzut ekranu, Prostokąt, kwadrat, diagram

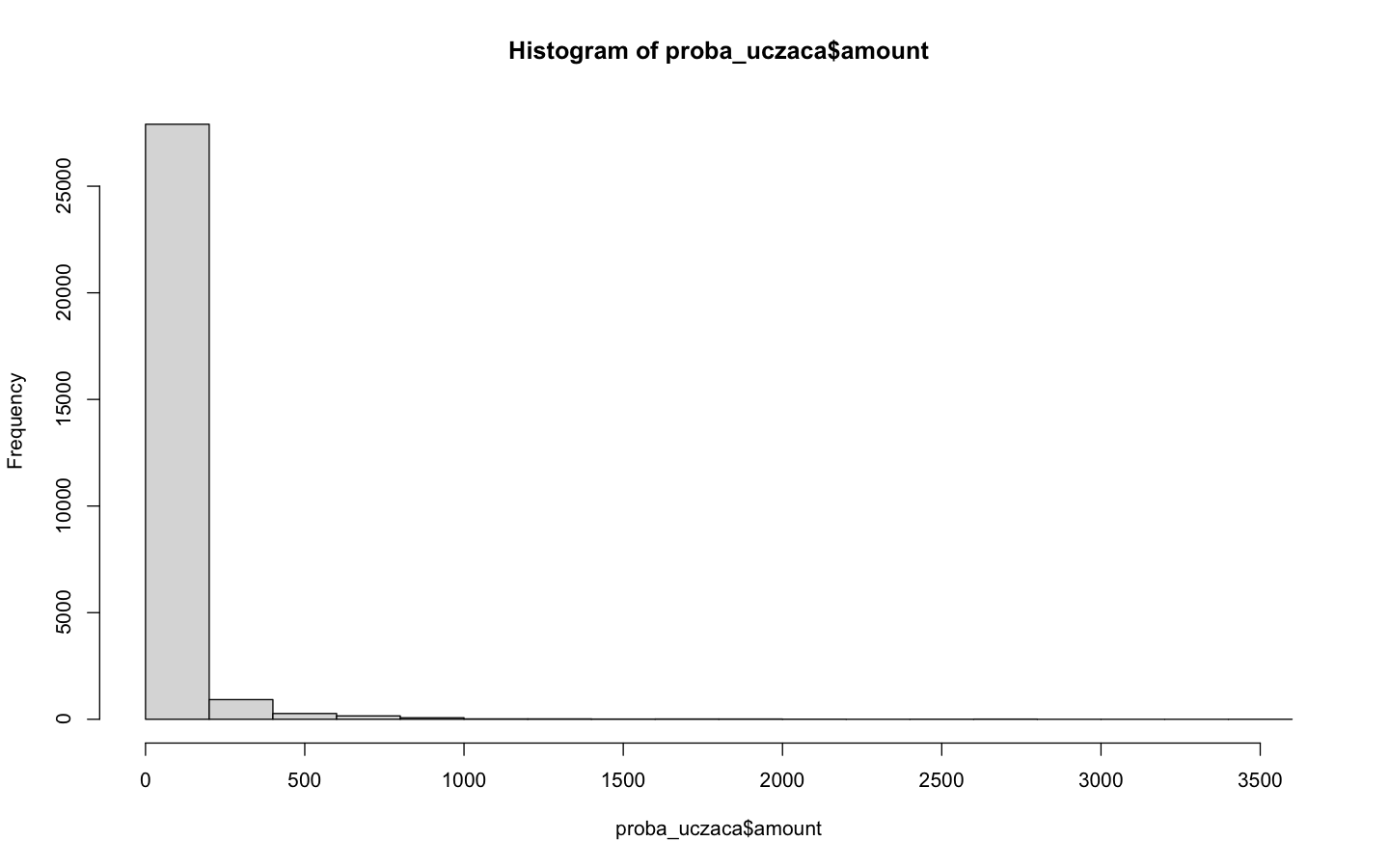
Opis wygenerowany automatycznie

* Wizualizacja dla zmiennej " recurringaction"

Obraz zawierający zrzut ekranu, Prostokąt, tekst, kwadrat

Opis wygenerowany automatycznie

* Wizualizacja dla zmiennej " amount"



1. Wzbogacenie danych
2. Tworzenie dodatkowej zmiennej, która ukazuje dzień tygodnia

Obraz zawierający tekst, Czcionka, zrzut ekranu, algebra

Opis wygenerowany automatycznie

1. Utworzenie dodatkowej zmiennej związanej z rodzajem karty kredytowej w zbiorze uczącym i zbiorze testowym

Zbiór **testowy**:

2 3 4 5 6

15 1075 6732 727 6636

Zbiór **uczący**:

2 3 4 5 6

26 2569 11031 1599 14187

1. Zmiana zmiennej o nazwie status

**porażka** **sukces**

5064 24348

1. Eliminacja zmiennych, które nie są istotne lub nieprzydatne w kontekście prognozowania
2. Wykluczanie obserwacji dotyczących transakcji nierekurencyjnych z zestawów treningowych i testowych jest konieczne, aby skupić się na przewidywaniu transakcji rekurencyjnych
3. Tworzenie i prezentacja graficzna drzewa decyzyjnego do klasyfikacji
4. Drzewo klasyfikacyjne i jego wizualizacja

Obraz zawierający zrzut ekranu, tekst, diagram, design

Opis wygenerowany automatycznie

1. Weryfikacja poprawności przetworzenia zmiennych; porównanie oczekiwanej liczby obserwacji z liczbą obserwacji w zestawie testowym. Są takie same, co potwierdza zgodność.
2. Tworzenie prognoz dla zestawu testowego.
3. Analiza skuteczności prognoz generowanych przez drzewo decyzyjne poprzez ocenę miar jakości predykcji
4. Optymalizacja drzewa decyzyjnego poprzez przycinanie, przy użyciu optymalnej wartości opartej na parametrze "cp”

Obraz zawierający tekst, zrzut ekranu, diagram, design

Opis wygenerowany automatycznie

1. Analiza istotności zmiennych w modelu oraz graficzne przedstawienie

Obraz zawierający tekst, numer, zrzut ekranu, linia

Opis wygenerowany automatycznie

1. Tworzenie prognoz dla drzewa po przycięciu, wraz z miarami predykcji, obejmującą macierz błędów.
2. Macierz błędnych klasyfikacji (confiusion matrix)
3. Obliczanie miary predykcji "F1 score"

**[1] 0.239113**

1. Obliczanie miary precyzji predykcji "accuracy" i wyrażanie wyniku w formie procentowej do interpretacji

**[1] 81.39562**

1. Regresja
2. Eliminacja zmiennej "predykcja\_status" - niezbędny krok przed konstrukcją drzewa regresyjnego
3. Konstrukcja i prezentacja graficzna drzewa regresji

Drzewo regresji:

Obraz zawierający diagram, Plan, Rysunek techniczny, design

Opis wygenerowany automatycznie

1. Rozpoznanie istotności zmiennych
2. Analizowanie wartości "parametr złożoności"

Obraz zawierający tekst, paragon, zrzut ekranu

Opis wygenerowany automatycznie

1. Graficzne przedstawienie estymacji "parametru złożoności"

Obraz zawierający linia, diagram, Wykres, Równolegle

Opis wygenerowany automatycznie

1. Optymalizacja drzewa poprzez przycinanie, uwzględniając najniższą wartość parametru 'cp'

Obraz zawierający diagram, Plan, design

Opis wygenerowany automatycznie

1. Eliminowanie zmiennej, która jest celem ewaluacji rezultatów, z danych używanych do prognozowania
2. Wyliczenie miary błędu średniokwadratowego (MSE) dla drzewa regresji nr 1
3. Wyliczenie miary MSE

**[1] 8885.823**

1. Wyliczenie miary MSE dla drzewa regresyjnego nr 2
2. Wyliczenie miary MSE

**[1] 8885.823**

1. Tworzenie ostatecznych prognoz dla zestawu testowego
2. Wyniki prognoz dla estymacji zmiennej 'amount' na zbiorze testowym, bazujące na wskaźnikach statystyki opisowej

**Min. 1st Qu. Median Mean 3rd Qu. Max.**

24.41 38.22 38.22 128.31 50.43 2037.18

1. Przeniesienie wyników predykcji do tabeli o nazwie "predykcje\_testowa"

**WNIOSKI:**

Model **drzewa klasyfikacji** osiągnął:

* wartość miary f1 na poziomie 0.239113
* precyzję predykcji wynoszącą 81,4%.

Z kolei **model drzewa regresji**:

* uzyskał 8885.82 jako wartość miary MSE (Mean Square Error).