Credit Card Fraud Detection - Zaawansowane uczenie maszynowe

Rębacz Gabriel, Belniak Michał

12 grudnia 2020

Spis treści

1 Interpretacja tematu projektu **2**

2 Analiza danych **2**

3 Przetwarzanie danych **3**

4 Przygotowanie, uczenie i strojenie wybranych modeli **4**
4.1 Drzewo klasyfikacji 4
4.2 Regresja logistyczna 4
4.3 Sieć neuronowa 4

5 Ocena jakości modeli **4**
5.1 Drzewo klasyfikacji 4
5.2 Regresja logistyczna 4
5.3 Sieć neuronowa 4

1. Interpretacja tematu projektu

Wybranym zadaniem jest analiza problemu klasyfikacji na zbiorze danych *Credit Card Fraud Detection*[1] dostępnym w serwisie Kaggle.com.

Zakres projektu obejmuje 4 podstawowe sfery:

- wnikliwą analizę zbioru danych,
- przetworzenie zbioru danych do postaci odpowiednich dla poszczególnych modeli,
- przygotowanie, uczenie i strojenie modeli wybranych spośród dostępnych w języku R,
- ocenę i porównanie modeli wraz z adekwatnymi wnioskami.

2. Analiza danych

Zbiór danych należy przeanalizować pod kątem czynników które mają lub mogą mieć wpływ na sposób ich przygotowania oraz dobór parametrów modeli.

Zbiór zawiera informacje o transakcjach kartami kredytowymi, w tym informacje o kwocie transakcji, czasie transakcji oraz 28 parametrach o nieznanym znaczeniu, które zostały poddane transformacji poprzez analizę głównych składowych.

```
## % latex table generated in R 4.0.3 by xtable 1.8-4 package
## % Sat Dec 12 18:07:09 2020
## \begin{table}[ht]
## \centering
## \begin{tabular}{rrrrrrrrrrrrrrrrrrrrrrrrrrrrrr}
##   \hline
##  & Time & V1 & V2 & V3 & V4 & V5 & V6 & V7 & V8 & V9 & V10 & V11 & V12 & V13 & V14 & V15 & V16 & V17
##   \hline
## 1 & 0.00 & -1.36 & -0.07 & 2.54 & 1.38 & -0.34 & 0.46 & 0.24 & 0.10 & 0.36 & 0.09 & -0.55 & -0.62 & -
##   2 & 0.00 & 1.19 & 0.27 & 0.17 & 0.45 & 0.06 & -0.08 & -0.08 & 0.09 & -0.26 & -0.17 & 1.61 & 1.07 &
##   3 & 1.00 & -1.36 & -1.34 & 1.77 & 0.38 & -0.50 & 1.80 & 0.79 & 0.25 & -1.51 & 0.21 & 0.62 & 0.07 &
##   4 & 1.00 & -0.97 & -0.19 & 1.79 & -0.86 & -0.01 & 1.25 & 0.24 & 0.38 & -1.39 & -0.05 & -0.23 & 0.1
##   5 & 2.00 & -1.16 & 0.88 & 1.55 & 0.40 & -0.41 & 0.10 & 0.59 & -0.27 & 0.82 & 0.75 & -0.82 & 0.54 &
##   6 & 2.00 & -0.43 & 0.96 & 1.14 & -0.17 & 0.42 & -0.03 & 0.48 & 0.26 & -0.57 & -0.37 & 1.34 & 0.36 &
##    \hline
## \end{tabular}
## \end{table}
```

Rysunek 1: Pierwsze 5 rekordów danych

Przeprowadzona podstawową analizę statystycznę zbioru i jego wyniki znajdują się na rysunku **??**.

```
## NOTE: Output requires \usepackage{booktabs} in your preamble.
```

	N	Mean	SD	Min	Q1	Median	Q3	Max
Time	284807	94813.86	47488.15	0.00	54201.50	84692.00	139320.50	172792.00
V1	284807	0.00	1.96	−56.41	−0.92	0.02	1.32	2.45
V2	284807	0.00	1.65	−72.72	−0.60	0.07	0.80	22.06
V3	284807	−0.00	1.52	−48.33	−0.89	0.18	1.03	9.38
V4	284807	0.00	1.42	−5.68	−0.85	−0.02	0.74	16.88
V5	284807	0.00	1.38	−113.74	−0.69	−0.05	0.61	34.80
V6	284807	0.00	1.33	−26.16	−0.77	−0.27	0.40	73.30
V7	284807	−0.00	1.24	−43.56	−0.55	0.04	0.57	120.59
V8	284807	0.00	1.19	−73.22	−0.21	0.02	0.33	20.01
V9	284807	−0.00	1.10	−13.43	−0.64	−0.05	0.60	15.59
V10	284807	0.00	1.09	−24.59	−0.54	−0.09	0.45	23.75
V11	284807	0.00	1.02	−4.80	−0.76	−0.03	0.74	12.02
V12	284807	−0.00	1.00	−18.68	−0.41	0.14	0.62	7.85
V13	284807	0.00	1.00	−5.79	−0.65	−0.01	0.66	7.13
V14	284807	0.00	0.96	−19.21	−0.43	0.05	0.49	10.53
V15	284807	0.00	0.92	−4.50	−0.58	0.05	0.65	8.88
V16	284807	0.00	0.88	−14.13	−0.47	0.07	0.52	17.32
V17	284807	−0.00	0.85	−25.16	−0.48	−0.07	0.40	9.25
V18	284807	0.00	0.84	−9.50	−0.50	−0.00	0.50	5.04
V19	284807	0.00	0.81	−7.21	−0.46	0.00	0.46	5.59
V20	284807	0.00	0.77	−54.50	−0.21	−0.06	0.13	39.42
V21	284807	0.00	0.73	−34.83	−0.23	−0.03	0.19	27.20
V22	284807	−0.00	0.73	−10.93	−0.54	0.01	0.53	10.50
V23	284807	0.00	0.62	−44.81	−0.16	−0.01	0.15	22.53
V24	284807	0.00	0.61	−2.84	−0.35	0.04	0.44	4.58
V25	284807	0.00	0.52	−10.30	−0.32	0.02	0.35	7.52
V26	284807	0.00	0.48	−2.60	−0.33	−0.05	0.24	3.52
V27	284807	−0.00	0.40	−22.57	−0.07	0.00	0.09	31.61
V28	284807	−0.00	0.33	−15.43	−0.05	0.01	0.08	33.85
Amount	284807	88.35	250.12	0.00	5.60	22.00	77.16	25691.16
Class	284807	0.00	0.04	0.00	0.00	0.00	0.00	1.00

3. Przetwarzanie danych

W rozdziale tym opisane zostaną sposoby przygotowania danych dla poszczególnych modeli.

4. Przygotowanie, uczenie i strojenie wybranych modeli

4.1. Drzewo klasyfikacji

4.2. Regresja logistyczna

4.3. Sieć neuronowa

5. Ocena jakości modeli

5.1. Drzewo klasyfikacji

5.2. Regresja logistyczna

5.3. Sieć neuronowa

Bibliografia

[1] *Credit Card Fraud Detection*. Dostęp zdalny (15.11.2020): https://www.kaggle.com/mlg-ulb/creditcardfraud. 2016.

[2] Wikipedia contributors. *Decision tree — Wikipedia, The Free Encyclopedia*. [Online; accessed 16-November-2020]. 2020. URL: https://en.wikipedia.org/w/index.php?title=Decision_tree&oldid=983253586.