

Systemy Analityczne

**Projekt narzędzia wspomagającego analizę
kosztów hospitalizacji z zastosowaniem
analizy predykcyjnej dla wybranej placówki
medycznej**

Grupa: Poniedziałek TN 13:15-15:00

Autor:
Damian Kędzierski 260493

Prowadzący: dr inż. Marek Lubicz

Spis treści

A) PROBLEM BIZNESOWY	3
PROBLEMY ANALITYCZNE.....	3
B) LISTA PROBLEMÓW ANALITYCZNYCH ANALITYKI PREDYKCYJNEJ:	3
TYP:	3
ZMIENNA OBJAŚNIANIA:	3
PREDYKTORY:.....	3
C) CHARAKTERYSTYKA DANYCH.....	4
DANE ŹRÓDŁOWE	4
DANE WYNIKOWE	5
D) PROCES ANALITYCZNY (CRISP-DM)	6
KROK 1: ZROZUMIENIE PROBLEMU BIZNESOWEGO.....	6
KROK 2: ZROZUMIENIE DANYCH	6
KROK 3: PRZYGOTOWANIE DANYCH.....	6
KROK 4: MODELOWANIE	6
KROK 5: OCENA JAKOŚCI MODELI	6
KROK 6: WIZUALIZACJA I RAPORTOWANIE WYNIKÓW	6
KROK 7: PRZEDSTAWIENIE PREZENTACJI DLA DECYDENTA	6
E) EKSPLORACYJNA ANALIZA DANYCH (EDA).....	7
WSTĘPNA ANALIZA EDA:	7
SZCZEGÓŁOWA ANALIZA EDA:.....	7
F) LISTA I OMÓWIENIE STWIERDZONYCH NIEDOSKONAŁOŚCI DANYCH ŹRÓDŁOWYCH	9
DANE NIEKOMPLETNE I BŁĘDY	9
DANE RZADKIE	9
WARTOŚCI ODSTAJĄCE (OUTLIERS)	9
G) LISTA WYKONANYCH OPERACJI PREPROCESSINGU	10
H) JEDNOZNACZNIE NAZWANY KOŃCOWY PLIK/PLIKI WYNIKOWE	10
I) OMÓWIENIE ZAŁOŻEŃ I ZAPROJEKTOWANIE W ŚRODOWISKU ANALITYKI PREDYKCYJNEJ	11
DOBÓR NAJLEPSZYCH PREDYKTORÓW	11
DOBÓR ALGORYTMÓW PREDYKCYJNYCH	11
OPTIMALIZACJA PARAMETRÓW	11
DOBÓR PODEJŚCIA DO WALIDACJI	11
PROCES	12
WYNIKI MODELU	12
WYNIKI IMPLEMENTACJI	13

a) Problem Biznesowy

Klasyfikacja świadczeń medycznych na podstawie ich kosztu (WARTOSC_SKOR). Celem jest stworzenie modelu klasyfikacyjnego, który na podstawie dostępnych informacji będzie przewidywał, czy dana procedura medyczna jest tanim, średnim, drogim lub bardzo drogim świadczeniem.

Problemy analityczne

Klasyfikacja kosztów świadczeń medycznych:

- Problem klasyfikacji wieloklasowej: Zadanie polega na przewidzeniu do której kategorii (Tanie, Średnie, Drogie, Bardzo drogie) będzie należeć konkretne świadczenie medyczne na podstawie dostępnych danych. Model klasyfikacji będzie mógł pomóc w automatycznym przypisaniu odpowiedniej kategorii kosztów do nowych przypadków.

Analiza wpływu innych zmiennych na koszty:

- Analiza wpływu procedur medycznych (LISTA_PROCEDUR), głównego rozpoznania (ROZP_GLOWNE) i innych czynników na koszty świadczeń medycznych (WARTOSC_SKOR). Celem jest zidentyfikowanie czynników, które mają istotny wpływ na koszty i mogą być wykorzystane do optymalizacji procesów medycznych i zarządzania kosztami.

Wszystkie powyższe problemy analityczne mają na celu dostarczenie wiedzy i narzędzi, które mogą wesprzeć podejmowanie decyzji dotyczących zarządzania kosztami, optymalizacji procesów medycznych oraz poprawy jakości opieki zdrowotnej.

b) Lista problemów analitycznych analityki predykcyjnej:

Typ:

- Klasyfikacja

Zmienna objaśniana:

⇒ WARTOSC_SKOR (label)- wartość skorygowana.







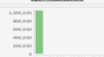




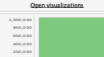

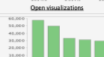
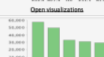



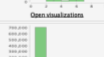
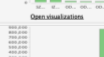
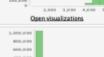
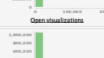
Predyktory:

- CENA - cena związana z danym świadczeniem medycznym.
- ID_PACJENT (id)- identyfikator pacjenta.
- ID_SWIADCZENIA (id)- identyfikator świadczenia medycznego.
- LISTA_PROCEDUR - lista procedur medycznych związanych ze świadczeniem.
- PKT - liczba punktów związanych ze świadczeniem.
- PROD_JEDN_KOD - kod jednostki produktowej.
- PROD_JEDN_NAZWA - nazwa jednostki produktowej.
- PROD_KONTR_KOD - kod kontrahenta produktowego.
- PROD_KONTR_NAZWA - nazwa kontrahenta produktowego.
- ROZP_GLOWNE - główne rozpoznanie.
- ROZP_GLOWNE_NAZWA - nazwa głównego rozpoznania.
- TYP_KOMORKI_OPIS - opis typu komórki.
- TYP_KOMORKI_ORG - organizacja typu komórki.

c) Charakterystyka danych

Dane źródłowe

- ⇒ nazwa(nazwy) plików - szpitale2014-2016b 2.hyper
- ⇒ liczba rekordów - 1 048 575
- ⇒ liczby zmiennych – 22
- ⇒ listy zmiennych:

szpital	Nominal	167790		Least S3 (112236)	Most S1 (288101)	Values S1 (288101), S4 (196223), S2 (143163), S5 (141062), ...[1 more] Details...	
CENA	Integer	0		Min 1	Max 21410	Average 9107.425	Deviation 5562.662
ID_PACJENT	Integer	385		Min 1856003	Max 1633569315	Average 65056597.283	Deviation 188691495.1...
ID_SWIADCZENIA	Integer	0		Min 211811875	Max 329811780	Average 232273592.8...	Deviation 11818524.178
LISTA_PROCEDUR	Nominal	0		Least Y90 (1)	Most 89.00.89.04 (21672)	Values 89.00.89.04 (21672), 89.00.89 [...] 61.99.18 (17590), 38.99.89 [...] 61.99.18 (13538), 38.99.89 [...] 521.89.61 (10879), ...[126383 more] Details...	
MC_SPROWODZAWCY	Integer	0		Min 1	Max 12	Average 6.286	Deviation 3.468
PKT	Real	0		Min 0	Max 67259.280	Average 67.791	Deviation 775.385
PROD_JEDN_KOD	Nominal	0		Least 5.54.01.0000005 (1)	Most 5.09.01.0000049 (12...)	Values 5.09.01.0000049 (125725), 5.09.01.0000041 (118205), 5.09.01.0000057 (41676), 5.09.01.0000039 (29752), ...[1324 more] Details...	
PROD_JEDN_NAZWA	Nominal	0		Least ZAPRAWNIE [...] CZEGO ... OPIKA P [...] (125725)	Most OPIKA P [...] POŁOŻNIE (125725), PORADA L [...] LEKARSKIE (118205), OCENA ST [...] TUNKOWYCH (41676), POTAS (O (29752), ...[1326 more] Details...		
PROD_KONTR_KOD	Nominal	0		Least 03.4580.930.02 (1)	Most 03.3300.008.03 (705...)	Values 03.3300.008.03 (705914), 03.4900.008.03 (110889), 03.4000.030.02 (17254), 03.4500.030.02 (17223), ...[156 more] Details...	
PROD_KONTR_NAZWA	Nominal	0		Least ORTOPEDI [...] ICZNY (1) ŚWIADCZE [...] (7059...)	Most ŚWIADCZE [...] ATUNKOWYM (705914), IZBA PRZYJĘĆ (110889), CHOROBY [...] TALIZACJA (17254), CHIRURGI [...] TALIZACJA (17223), ...[162 more] Details...		
RODZAJ_SWO_NAZWA	Nominal	0		Least Lecznict [...] (1048575)	Most Lecznict [...] (1048575)	Values Lecznictwo szpitalne (1048575) Details...	
ROK_SPROWODZAWCY	Integer	0		Min 2014	Max 2015	Average 2014.084	Deviation 0.278
ROZP_GLOWNE	Nominal	0		Least Z93.2 (1)	Most R10.4 (57424)	Values R10.4 (57424), R07.4 (49664), I10 (32898), Z51.1 (30945), ...[5536 more] Details...	
ROZP_GLOWNE_NAZWA	Nominal	0		Least Zółtak [...] niem (1)	Most Inny i n [...] a (57424)	Values Inny i n [...] brzucha (57424), Ból w kł [...] określony (49664), Samoistn [...] ciśnienie (32898), Cykle ch [...] owotworów (30945), ...[5510 more] Details...	
ROZP_WSP1	Nominal	890000		Least Z96.6 (1)	Most I10 (7177)	Values I10 (7177), C50.9 (7058), C18.9 (5278), Z51.1 (4255), ...[4137 more] Details...	
ROZP_WSP1_NAZWA	Nominal	890000		Least Zółtak [...] troby (1)	Most Samoistn [...] ie (7177)	Values Samoistn [...] ciśnienie (7177), Nowotwór [...] kreślony (7058), Nowotwór [...] kreślony (5278), Cykle ch [...] owotworów (4255), ...[4122 more] Details...	
TRYB_PRZYZJECIA_KOD	Integer	0		Min 0	Max 10	Average 3.417	Deviation 1.372
TYP_KOMORKI_OPIS	Nominal	0		Least ODDZIAŁ [...] API (56)	Most SZPITALN [...] (707533)	Values SZPITALN [...] RATUNKOWY (707533), IZBA PRZYJĘĆ SZPITALA (108389), ODDZIAŁ [...] IOTERAPI (24057), ODDZIAŁ [...] NY OGÓLNY (19145), ...[44 more] Details...	
TYP_KOMORKI_ORG	Integer	0		Min 1240	Max 4902	Average 4768.602	Deviation 304.392
WARTOSC	Real	0		Min 0	Max 194399.998	Average 613.685	Deviation 2355.455
WARTOSC_SKOR	Real	0		Min 0	Max 194399.998	Average 613.685	Deviation 2355.455

Dane wynikowe

- ⇒ nazwa(nazwy) plików – 260493_szpital.xlsx
- ⇒ liczba rekordów – 5 137
- ⇒ liczby zmiennych – 15
- ⇒ listy zmiennych:

<div><div></div><div>Id</div></div> <div>ID_PACJENT</div>	Nominal	0	<div><div></div><div>Open visualizations</div></div>	Least 99907928 (1)	Most 113859760 (14)	Values 113859760 (14), 2599193 (12), 6007123 (12), 16935480 (11), ...[3654 more] Details...	
<div><div></div><div>Label</div></div> <div>WARTOSC_SKOR</div>	Nominal	0	<div><div></div><div>Open visualizations</div></div>	Least Drogie (3st Qu.) (311)	Most Tanie (1st Qu.) (1797)	Values Tanie (1st Qu.) (1797), Srednie (2st Qu.) (1746), Bardzo drogie (4st Qu.) (1283), Drogie (3st Qu.) (311) Details...	
<div><div></div><div>Id_2</div></div> <div>ID_SWIADCZENIA</div>	Nominal	0	<div><div></div><div>Open visualizations</div></div>	Least 295353477 (1)	Most 276527529 (5)	Values 276527529 (5), 220682089 (3), 222387170 (3), 225971272 (3), ...[4967 more] Details...	
<div><div></div><div>CENA</div></div>	Real	0	<div><div></div><div>Open visualizations</div></div>	Min 1	Max 52.000	Average 42.251	Deviations 20.056
<div><div></div><div>PKT</div></div>	Real	0	<div><div></div><div>Open visualizations</div></div>	Min 1.006	Max 33311.260	Average 414.745	Deviations 2171.652
<div><div></div><div>TRYB_PRZYJECIA_KOD</div></div>	Real	0	<div><div></div><div>Open visualizations</div></div>	Min 0	Max 8	Average 5.549	Deviations 1.008
<div><div></div><div>TYP_KOMORKI_ORG</div></div>	Real	0	<div><div></div><div>Open visualizations</div></div>	Min 4100.000	Max 4640.000	Average 4397.320	Deviations 166.879
<div><div></div><div>LISTA_PROCEDUR</div></div>	Polynomial	0	<div><div></div><div>Open visualizations</div></div>	Least 99.04:99.25 (1)	Most 89.02:99.25 (1035)	Values 89.02:99.25 (1035), 89.00 (515), 100.51:13.49:13.71 (413), 95.415:95.436 (406), ...[188 more] Details...	
<div><div></div><div>PROD_JEDN_KOD</div></div>	Polynomial	0	<div><div></div><div>Open visualizations</div></div>	Least 5.53.01.0004044 (1)	Most 5.08.05.0000010 (551)	Values 5.08.05.0000010 (551), 5.51.01.0013020 (491), 5.52.01.0000261 (416), 5.08.10.0000047 (344), ...[136 more] Details...	
<div><div></div><div>PROD_JEDN_NAZWA</div></div>	Polynomial	0	<div><div></div><div>Open visualizations</div></div>	Least VINORELB [...] 1 MG (1)	Most HOSPITAL [...] WYM (5...	Values HOSPITAL [...] NODNIOWYM (551), N20 NOWO [...] E] OPIEKI (491), BADAНИЕ [...] O 4 R.2.) (416), ONDANSET [...] U - 1 MG (344), ...[135 more] Details...	
<div><div></div><div>PROD_KONTR_KOD</div></div>	Polynomial	0	<div><div></div><div>Open visualizations</div></div>	Least 03.4580.030.02 (1)	Most 03.4610.030.02 (633)	Values 03.4610.030.02 (633), 03.4421.160.02 (603), 03.0000.112.02 (551), 03.0001.112.02 (482), ...[50 more] Details...	
<div><div></div><div>PROD_KONTR_NAZWA</div></div>	Polynomial	0	<div><div></div><div>Open visualizations</div></div>	Least ORTOPEDI [...] ZACJA (1)	Most OTORYNOL [...] CJA (6...	Values OTORYNOL [...] TALIZACJA (633), NEONATOL [...] N24, N25 (603), CHEMIOTE [...] OJARZONYM (551), SUBSTANC [...] OJARZONYM (482), ...[50 more] Details...	
<div><div></div><div>ROZP_GLOWNE_NAZWA</div></div>	Polynomial	0	<div><div></div><div>Open visualizations</div></div>	Least Złamanie [...] rzowe (1)	Most Cykle ch [...] ów (1576)	Values Cykle ch [...] owotworów (1576), Pojedync [...] szpitalu (491), Zaczma wiktajaca (306), Obserwac [...] i stanów (277), ...[155 more] Details...	
<div><div></div><div>ROZP_GLOWNE</div></div>	Polynomial	0	<div><div></div><div>Open visualizations</div></div>	Least 572.1 (1)	Most Z51.1 (1576)	Values Z51.1 (1576), Z38.0 (491), H26.2 (306), Z03.8 (277), ...[155 more] Details...	
<div><div></div><div>TYP_KOMORKI_OPIS</div></div>	Polynomial	0	<div><div></div><div>Open visualizations</div></div>	Least ODDZIAŁ [...] ICZNY (4)	Most ODDZIAŁ [...] II (1796)	Values ODDZIAŁ [...] IOTERAPII (1796), ODDZIAŁ [...] OLOGICZNY (633), ODDZIAŁ NEONATOLOGICZNY (603), ODDZIAŁ OKULISTYCZNY (479), ...[13 more] Details...	

d) Proces Analityczny (CRISP-DM)

Krok 1: Zrozumienie problemu biznesowego

- Celem biznesowym jest stworzenie modelu klasyfikacyjnego, który przewiduje koszt świadczeń medycznych na podstawie dostępnych informacji.

Krok 2: Zrozumienie danych

- Dane obejmują informacje takie jak cena, identyfikatory pacjenta i świadczenia medycznego, lista procedur medycznych, liczba punktów, kod jednostki produktowej, kod kontrahenta produktowego, główne rozpoznanie, kod trybu przyjęcia, opis typu komórki i organizacja typu komórki.

Krok 3: Przygotowanie danych

- Przygotowanie danych będzie obejmować oczyszczenie danych, usuwanie wartości odstających i brakujących, skalowanie zmiennych oraz inżynierię cech.

Krok 4: Modelowanie

- Zastosowanie modelu klasyfikacji - drzewa decyzyjnego do przewidywania kosztów świadczeń medycznych.
- Przygotowanie danych treningowych i testowych.
- Dopasowanie modelu decyzyjnego do danych treningowych.
- Ocena wydajności modelu na danych testowych.

Krok 5: Ocena jakości modeli

- Wykorzystanie miar jakości, takich jak precyzja, czułość, specyficzność, AUC, do oceny jakości modelu klasyfikacyjnego.
- Porównanie wyników miar jakości modelu dla różnych parametrów lub algorytmów.

Krok 6: Wizualizacja i raportowanie wyników

- Stworzenie wykresów i tabel, które przedstawiają informacje o kosztach świadczeń medycznych i innych zmiennych.
- Przygotowanie raportu zawierającego wyniki analizy predykcyjnej i wnioski.

Krok 7: Przedstawienie prezentacji dla decydenta

- Przygotowanie prezentacji, która zawiera opis problemu biznesowego, podejście do analizy danych, wyniki analizy predykcyjnej i wnioski.
- Omówienie różnych opcji i zaleceń dotyczących zmniejszenia kosztów świadczeń medycznych i optymalizacji procesów medycznych.

e) Eksploracyjna analiza danych (EDA)

Wstępna analiza EDA:

CENA	PKT	TRYB_PRZYJECIA_KOD	TYP_KOMORKI_ORG	LISTA_PROCEDUR	PROD_JEDN_KOD	PROD_JEDN_NAZWA
Min. : 1.00	Min. : 1.01	Min. : 0.000	Min. : 4100	Length:5137	Length:5137	Length:5137
1st Qu.:52.00	1st Qu.: 9.00	1st Qu.:6.000	1st Qu.:4242	Class :character	Class :character	Class :character
Median :52.00	Median : 20.00	Median :6.000	Median :4421	Mode :character	Mode :character	Mode :character
Mean :42.25	Mean : 414.74	Mean :5.549	Mean :4397			
3rd Qu.:52.00	3rd Qu.: 49.00	3rd Qu.:6.000	3rd Qu.:4600			
Max. :52.00	Max. :33311.26	Max. :8.000	Max. :4640			

PROD_KONTR_KOD	PROD_KONTR_NAZWA	ROZP_GLOWNE_NAZWA	ID_PACJENT	ID_SWIADCZENIA	WARTOSC_SKOR
Length:5137	Length:5137	Length:5137	113859760: 14	276527529: 5	Min. : 1.01
Class :character	Class :character	Class :character	2599193 : 12	220682089: 3	1st Qu.: 468.00
Mode :character	Mode :character	Mode :character	6007123 : 12	222387170: 3	Median : 1040.00
			16935480 : 11	225971272: 3	Mean : 1764.93
			2433732 : 11	226309173: 3	3rd Qu.: 2184.00
			2649154 : 11	226892785: 3	Max. :33311.26
			(Other) :5066	(Other) :5117	

ROZP_GLOWNE	TYP_KOMORKI_OPIS
Length:5137	Length:5137
Class :character	Class :character
Mode :character	Mode :character

- Cena (CENA) świadczeń medycznych w analizowanych danych oscyluje głównie wokół wartości 52, jednak średnia cena wynosi 42,25. Oznacza to, że istnieją również niższe ceny. Wartość minimalna to 1, a maksymalna to 52.
- PKT (Punkty) są zróżnicowane, od minimalnej wartości 1,01 do maksymalnej 33 311,26. Średnia liczba punktów wynosi 414,74. Mediana wynosi 20 punktów, co sugeruje skośność rozkładu.
- TRYB_PRZYJECIA_KOD (Kod trybu przyjęcia) ma wartości od 0 do 8, z medianą wynoszącą 6. Najczęściej występująca wartość to 6, co może wskazywać na dominujący tryb przyjęcia.
- TYP_KOMORKI_ORG (Typ komórki organizmu) ma przeważającą wartość 4397, ale występują również inne wartości, takie jak 4100, 4242 i 4421.
- LISTA_PROCEDUR (Lista procedur) w analizowanych danych zawiera informacje o różnych procedurach medycznych, które są związane ze świadczeniami. Długość listy procedur wynosi 5137.
- PROD_JEDN_KOD (Kod jednostki produktu), PROD_JEDN_NAZWA (Nazwa jednostki produktu), PROD_KONTR_KOD (Kod kontraktu produktu) i PROD_KONTR_NAZWA (Nazwa kontraktu produktu) zawierają informacje o różnych kodach i nazwach jednostek i kontraktów produktów.
- ROZP_GLOWNE_NAZWA (Nazwa głównego rozpoznania) ma zróżnicowane wartości, co sugeruje różnorodność rozpoznań w analizowanych danych.
- ID_PACJENT (ID pacjenta) i ID_SWIADCZENIA (ID świadczenia) są identyfikatorami pacjentów i świadczeń medycznych, które mogą być wykorzystane do analizy, segmentacji pacjentów i śledzenia historii świadczeń.

Szczegółowa analiza EDA:

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
CENA	1	5137	42.25	20.06	52	46.18	0.00	1.00	52.00	51.00	-1.57	0.47	0.28
PKT	2	5137	414.74	2171.65	20	30.71	19.27	1.01	33311.26	33310.25	9.44	119.18	30.30
TRYB_PRZYJECIA_KOD	3	5137	5.55	1.01	6	5.84	0.00	0.00	8.00	8.00	-2.53	6.46	0.01
TYP_KOMORKI_ORG	4	5137	4397.32	166.88	4421	4400.03	265.39	4100.00	4640.00	540.00	-0.05	-1.39	2.33
ID_PACJENT*	5	5137	1734.33	1031.78	1696	1713.42	1286.90	1.00	3658.00	3657.00	0.14	-1.12	14.40
ID_SWIADCZENIA*	6	5137	2469.28	1431.16	2458	2465.61	1832.49	1.00	4971.00	4970.00	0.02	-1.20	19.97
WARTOSC_SKOR	7	5137	1764.93	2858.96	1040	1199.56	1002.24	1.01	33311.26	33310.25	5.66	47.72	39.89

CENA:

- Średnia cena wynosi 42.25, a mediana 52. To sugeruje, że większość świadczeń medycznych ma cenę zbliżoną do 52.
- Wartość minimalna to 1, a maksymalna 52. Zakres cen wynosi 51.

- Współczynnik skośności wynosi -1.57, co wskazuje na asymetrię rozkładu w lewo. Kurtoza wynosi 0.47, co oznacza, że rozkład jest nieco spłaszczony w porównaniu do rozkładu normalnego.

PKT:

- Średnia liczba punktów wynosi 414.74, a mediana 20. Wartość średnia jest wyższa od mediany ze względu na kilka obserwacji o bardzo wysokich wartościach punktów.
- Wartość minimalna to 1.01, a maksymalna 33311.26. Zakres liczby punktów wynosi 33310.25.
- Współczynnik skośności wynosi 9.44, co wskazuje na wyraźną asymetrię rozkładu w prawo. Kurtoza wynosi 119.18, co oznacza, że rozkład ma długie ogony i jest bardziej skupiony wokół średniej niż rozkład normalny.

TRYB_PRZYJECIA_KOD (Kod trybu przyjęcia):

- Średnia wartość kodu trybu przyjęcia wynosi około 5,55, z odchyleniem standardowym wynoszącym 1,01. Mediana wynosi 6, co wskazuje na przewagę wartości bliskich 6.
- Skośność jest ujemna (-2,53), co sugeruje, że rozkład jest skośny w lewo. Wyższy kurtoza (6,46) wskazuje na większe ogony rozkładu.

TYP_KOMORKI_ORG (Organizacja typu komórki):

- Średnia wartość organizacji typu komórki wynosi około 4397,32, z odchyleniem standardowym wynoszącym 166,88. Mediana wynosi 4421.
- Rozkład ma niewielką skośność (-0,05) i ujemną kurtozę (-1,39), co wskazuje na lekkie odchylenie od rozkładu normalnego.

ID_PACJENT:

- Średnia wartość identyfikatora pacjenta wynosi 1734.33, a mediana 1696.
- Wartość minimalna to 1, a maksymalna 3658. Zakres identyfikatora pacjenta wynosi 3657.
- Współczynnik skośności wynosi 0.14, co wskazuje na niewielką asymetrię rozkładu w prawo. Kurtoza wynosi -1.12, co oznacza, że rozkład jest nieco spłaszczony w porównaniu do rozkładu normalnego.

ID_SWIADCZENIA:

- Średnia wartość identyfikatora świadczenia wynosi 2469.28, a mediana 2458.
- Wartość minimalna to 1, a maksymalna 4971. Zakres identyfikatora świadczenia wynosi 4970.
- Współczynnik skośności wynosi 0.02, co wskazuje na niewielką asymetrię rozkładu w prawo. Kurtoza wynosi -1.20, co oznacza, że rozkład jest nieco bardziej spłaszczony w porównaniu do rozkładu normalnego.

WARTOSC_SKOR:

- Średnia wartość skorygowana wynosi 1764.93, a mediana 1040. Oznacza to, że średnia wartość skorygowana jest wyższa od mediany, co sugeruje występowanie wartości odstających lub niestabilność w rozkładzie.
- Wartość minimalna to 1.01, a maksymalna 33311.26. Zakres wartości skorygowanych jest bardzo szeroki, co wskazuje na duże zróżnicowanie kosztów świadczeń medycznych.
- Współczynnik skośności wynosi 5.66, co wskazuje na wyraźną asymetrię rozkładu w prawo. Skośność dodatnia oznacza, że rozkład ma długie ogony w prawo i większą koncentrację wartości w lewej części rozkładu.
- Kurtoza wynosi 47.72, co oznacza, że rozkład wartości skorygowanych ma długie ogony i jest bardziej skupiony wokół średniej niż rozkład normalny. Wysoka wartość kurtozy wskazuje na występowanie wartości odstających lub nietypowych w danych.

g) lista wykonanych operacji preprocessingu

1. Załadowanie danych
2. Przefiltrowanie danych
 - a. Usunięcie wartości niekompletnych w ID_PACJENT i szpital
 - b. wybranie tylko jednego szpitala – S1,
 - c. WARTOSC_SKOR i CENA > 0 (w celu skupienia się tylko na hospitalizacjach kosztownych)
3. Wzięcie próbki danych (20% z powodu braku możliwości uruchomienia algorytmu wykrywania wartości odstających na całym zbiorze)
4. Zamiana typu wartości numerycznych id i zmiennej objaśnianej na zmienne nominalne
5. Ustanowienie ról
 - a. Id – ID_PACJENT, ID_SWIADCZENIA
 - b. Label – WARTOSC_SKOR
6. Wykluczenie atrybutów RODZAJ_SWD, ROZP_WSP1, ROZP_WSP1_NAZWA, WARTOSC, szpital
7. Zidentyfikowanie wartości rzadkich, które miały mniej niż 5 wystąpień i zmiana na wartość Other
8. Zdeklarowanie wartości Other jako wartość missing
9. Usunięcie wartości missing
10. Sprawdzenie wartości wag predyktorów za pomocą 2 algorytmów (Weight by information gain i weight by information gain ratio)
11. Wykluczenie atrybutów miesiąca i roku sprawozdawczego z powodu zbyt małych wag wpływu na zmienną objaśnianą
12. Normalizacja danych
13. Segmentacja algorytmem X-means
14. Zidentyfikowanie wartości odstających za pomocą operatora Detect Outlier (LOF)
15. Denormalizacja i zastosowanie modelu
16. Usunięcie wartości odstających (z wartością powyżej 5)
17. Zapisanie wartości odstających do pliku txt
18. Zapisanie oczyszczonego zbioru danych do pliku.xlsx

h) jednoznacznie nazwany końcowy plik/pliki wynikowe

1. plik końcowy – 260493_szpital.xlsx
2. preprocessing – 260493_preprocessing.rmp
3. analiza EDA – 260493_EDA.R
4. analiza atrybutów - 260493_Weights.res
5. wartości odstające – 260493_outliers.res

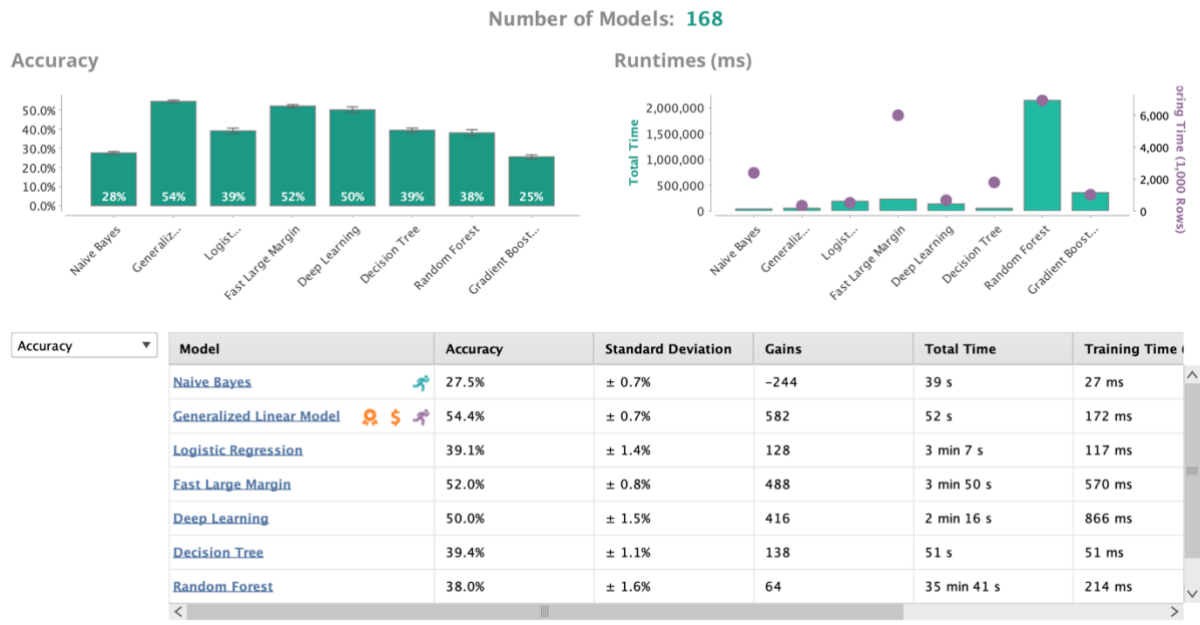
i) omówienie założeń i zaprojektowanie w środowisku analityki predykcyjnej

Dobór najlepszych predyktorów

Wykluczenie atrybutów ceny, miesiąca i roku sprawozdawczego z powodu zbyt małych wag wpływu na zmienną objaśnianą.

Dobór algorytmów predykcyjnych

Overview



- Najlepszym modelem w kontekście dokładności (accuracy) jest Generalized Linear Model, osiągając wynik 0,5. Oznacza to, że model poprawnie sklasyfikował 50% obserwacji. Wartości dokładności dla innych modeli wahają się między 0,3 a 0,5.

Optymalizacja parametrów

Optymalizacja parametrów obejmowała zmianę trzech kluczowych parametrów:

- Maximum_number_of_threads: Zwiększono liczbę wątków do 11, co może przyspieszyć obliczenia i przetwarzanie modelu.
- Family: Wybrano automatyczne dobranie rodziny funkcji do modelu, co pozwala na elastyczne dopasowanie do danych.
- Solver: Wybrano solver L-BFGS, który jest efektywny dla dużej ilości danych.

Dzięki optymalizacji parametrów modelu Generalized Linear Model osiągnięto minimalne poprawki wyników, co może przyczynić się do lepszej predykcji i generalizacji modelu na nowych danych.

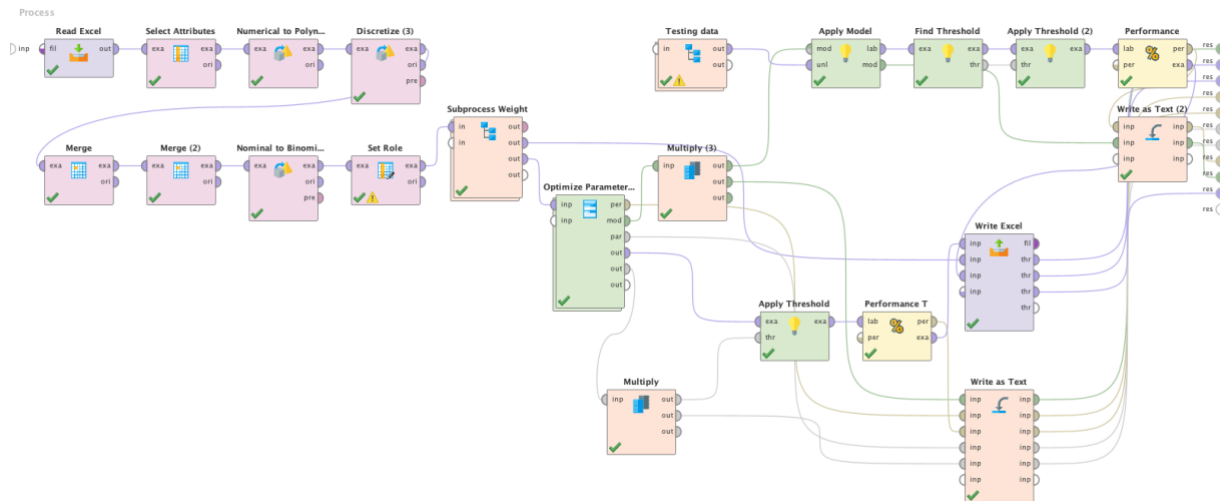
Dobór podejścia do walidacji

- Wybór walidacji krzyżowej jako podejścia do walidacji. Pozwala na dokładne ocenienie wydajności modelu oraz zapewnia solidne oszacowanie jego zdolności do generalizacji na nowe dane.

Proces

Plik – 260493_proces

Wyniki (performance) – 260493_wyniki



Wyniki modelu

W opisanym wyniku modelu klasyfikacji hospitalizacji przedstawiono różne miary oceny wydajności modelu. Oto niektóre z głównych wniosków:

Metryki modelu:

- MSE (Mean Squared Error): 0.042181417 - wskazuje na średnią kwadratową różnicę między wartościami przewidywanymi a rzeczywistymi.
- RMSE (Root Mean Squared Error): 0.20538116 - jest pierwiastkiem kwadratowym z MSE i mierzy przeciętną odległość między przewidywanymi wartościami a rzeczywistymi w tych samych jednostkach co zmienna docelowa.
- R^2 (R-squared): 0.7860879 - jest miarą dopasowania modelu i wskazuje, jak dużo zmienności w danych jest wyjaśniane przez model. Wartość zbliżona do 1 oznacza lepsze dopasowanie.
- AUC (Area Under the Curve): 0.9960712 - to miara skuteczności klasyfikatora, która ocenia zdolność modelu do rozróżniania między pozytywnymi i negatywnymi przykładami. Wartość bliska 1 oznacza wysoką skuteczność klasyfikacji.
- pr_auc (Precision-Recall AUC): 0.99835336 - to miara precyzji i czułości modelu, która jest szczególnie przydatna w przypadku nie zrównoważonych zbiorów danych.

Macierz pomyłek (Confusion Matrix):

- Przedstawia wyniki klasyfikacji modelu dla poszczególnych klas.
- W tym przypadku, dla klasy "Drogie_Bardzo drogie", model poprawnie sklasyfikował 1343 przypadki, a 45 przypadków zostało błędnie sklasyfikowanych jako "Tanie_Srednie".
- Dla klasy "Tanie_Srednie", model poprawnie sklasyfikował 3692 przypadki, a 57 przypadków zostało błędnie sklasyfikowanych jako "Drogie_Bardzo drogie".

Threshold:

- Określa wartość progu, powyżej którego klasa "Tanie_Srednie" jest przewidywana przez model. W tym przypadku, jeśli pewność przewidywanej klasy "Tanie_Srednie" jest większa niż 0.6361296676649265, to zostanie przypisana ta klasa, w przeciwnym razie przypisana zostanie klasa "Drogie_Bardzo drogie".

Wyniki walidacji:

- Wartości dokładności, błędu klasyfikacji, współczynnika kappa i AUC są przedstawione dla różnych metryk.
- Średnia dokładność wynosi 94.28%, a błąd klasyfikacji wynosi 5.72%, co wskazuje na ogólnie dobrą wydajność modelu.
- Wartość kappa wynosi 0.844, co wskazuje na znaczne zgodności w klasyfikacji.
- AUC wynosi 0.995, co oznacza wysoką skuteczność w rozróżnianiu między klasami.

Podsumowując, model klasyfikacji hospitalizacji wydaje się być skuteczny i dobrze radzi sobie w rozróżnianiu między klasami. Wartości metryk potwierdzają wysoką dokładność i skuteczność modelu. Macierz pomyłek pokazuje niewielką liczbę błędów klasyfikacji. Jednak dokładniejszą interpretację i ostateczne wnioski można wyciągnąć, uwzględniając kontekst i cele analizy.

Wyniki implementacji

Oto wnioski dla nowych danych na podstawie implementacji modelu:

Metryki modelu:

- Dokładność (accuracy): 97.20% - odsetek poprawnych klasyfikacji modelu.
- Błąd klasyfikacji (classification_error): 2.80% - odsetek błędnych klasyfikacji modelu.
- Kappa: 0.930 - miara zgodności klasyfikacji modelu.
- AUC: 0.992 - miara skuteczności klasyfikatora w rozróżnianiu między klasami.
- MSE (Mean Squared Error): 0.042181417 - średnia kwadratowa różnica między wartościami przewidywanymi a rzeczywistymi.
- RMSE (Root Mean Squared Error): 0.20538116 - pierwiastek kwadratowy z MSE, mierzący przeciętną odległość między przewidywanymi wartościami a rzeczywistymi w tych samych jednostkach.

Macierz pomyłek (Confusion Matrix):

- Dla klasy "Drogie_Bardzo drogie", model poprawnie sklasyfikował 939 przypadków, a 77 przypadków zostało błędnie sklasyfikowanych jako "Tanie_Srednie".
- Dla klasy "Tanie_Srednie", model poprawnie sklasyfikował 2571 przypadków, a 24 przypadków zostało błędnie sklasyfikowanych jako "Drogie_Bardzo drogie".

Threshold:

- Określa wartość progu, powyżej którego klasa "Tanie_Srednie" jest przewidywana przez model. W tym przypadku, jeśli pewność przewidywanej klasy "Tanie_Srednie" jest większa niż 0.631654679775238, to zostanie przypisana ta klasa, w przeciwnym razie przypisana zostanie klasa "Drogie_Bardzo drogie".

Wnioski:

- Model nadal wykazuje wysoką dokładność (97.20%) i skuteczność w rozróżnianiu między klasami (AUC: 0.992).
- Błąd klasyfikacji wynosi 2.80%, co wskazuje na niewielką liczbę błędów w klasyfikacji.
- Wartość kappa wynosi 0.930, co oznacza znaczną zgodność w klasyfikacji.
- Macierz pomyłek pokazuje niewielką liczbę błędów klasyfikacji dla obu klas.

- W przypadku nowych danych model utrzymuje wysoką skuteczność i wydajność w klasyfikacji hospitalizacji.