R - projekt - prezentacja

2018-06-03

Kurs Junior Data Scientist Zaoczne 1 (JDSZ1)

Raczki

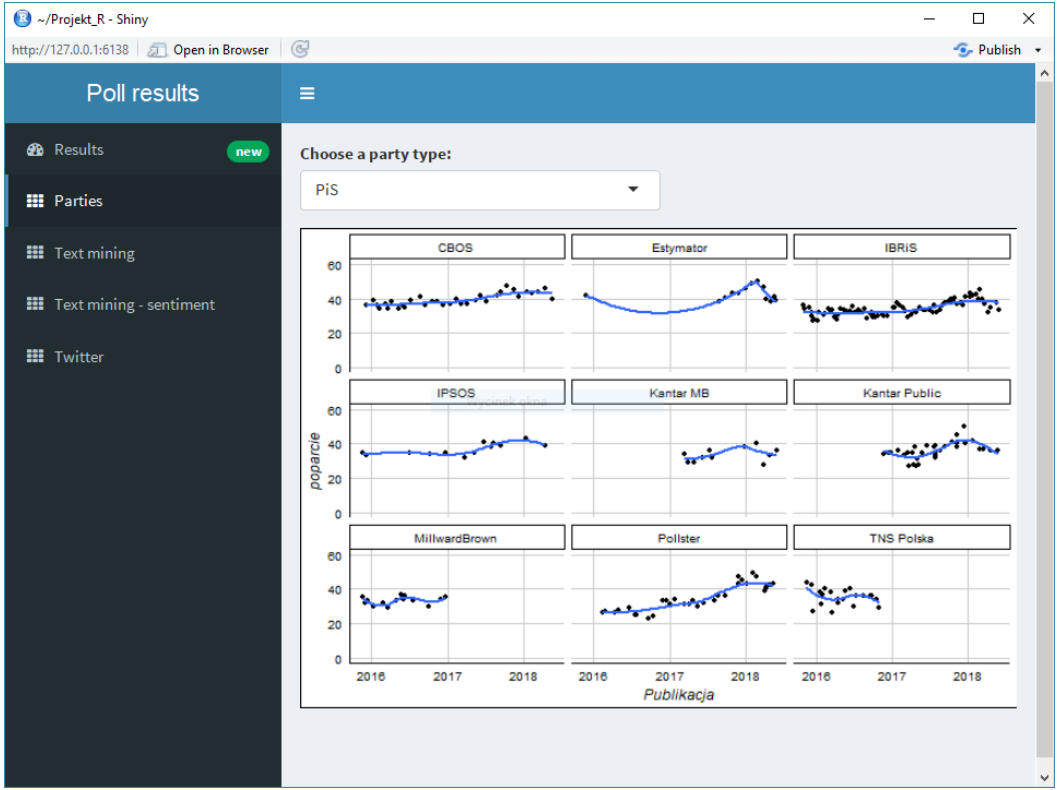
Bartosz, Filip, Monika Kucal, Piotr

1. Aplikacja R Shiny
2. Regresja liniowa
3. Regresja logistyczna

1. Aplikacja R Shiny

Aplikacja R Shiny

- Sondaże wyborcze
- Text mining
- Tweet'y o partii



2. Regresja liniowa

Regresja liniowa

- Zbiór danych z Kaggle: Weather in Szeged 2006-2016
- Dane - 96 453 obserwacji
- Jaka jest zależność temperatury od pozostałych parametrów pogodowych?

Zmienna	Przykładowa wartość	Zmienna	Przykładowa wartość
Formatted date	2006-04-01 00:00:00.000 +0200	Wind Speed [km/h]	14.1197
Summary	Partly Cloudy	Wind Bearing [º]	251.0
Precip type	rain	Visibility [km]	15.83
Temperature [ºC]	9.47	Loud Cover	0
Apparent temperature [ºC]	7.39	Pressure [millibars]	1015.13
Humidity	0.89	Daily Summary	Partly cloudy throughout the day.

Regresja liniowa - Pogoda w Szeged (BG)

Model liniowy 1:
- Założenie mediany dla NA

T = 34.76 - 31.00h
R^2 = 0.40 RMSE = 7.36

Model liniowy 2:
- Założenie mediany dla NA

at = 33.24 - 33.19h
R^2 = 0.36 RMSE = 8.49

Model liniowy 3:
- Założenie mediany dla NA
- Uwzględnienie miesiąca
- Dane liczbowe bez odczuwalnej temperatury
- Iteracyjne usuwanie nieistotnych zmiennych niezależnych

T = 35.34 - 0.47d(mth) - 32.95h - 0.17ws
R^2 = 0.46 RMSE = 7.03

d - Time, s - Summary, pt - Precip Type, **T** - Temperature, at - Apparent Temperature, h - humidity, ws - wind speed, wb - wind bearing, v - visibility, lc - loud cover, p - pressure, ds - daily summary

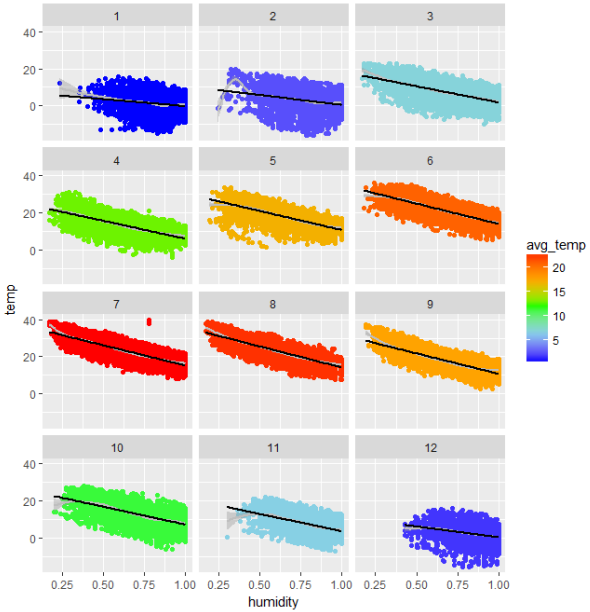
Regresja liniowa - Pogoda w Szeged (MK)

Model liniowy 1.

Wpływ wilgotności powietrza na temperaturę rzeczywistą

Temp = 34.8 - 31,1 humidity

R^2 = 40% RMSE = 7.4



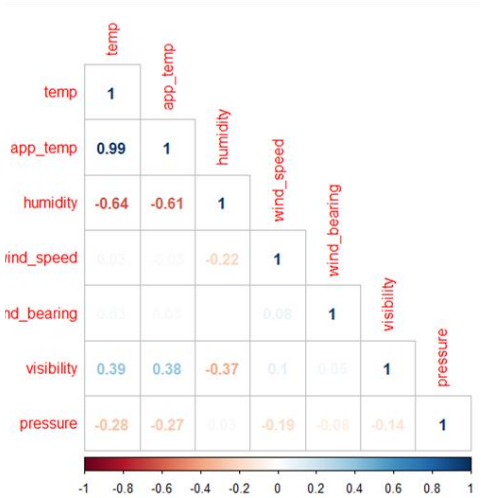
Model liniowy 2.

Wpływ wilgotności powietrza na temperaturę odczuwalną

App temp = 35.3 - 33.1 humidity

R^2 = 36% RMSE = 8.6

Macierz korelacji



Model liniowy 3.

Wpływ wielu parametrów pogodowych na temperaturę rzeczywistą w zależności od miesiąca w roku

Temp = 199 - 19.6 humidity
- 0.006 wind speed - 0.002 wind bearing
- 0.11 visibility - 0.17 pressure
+ 0.81 monthly avg temp

R^2 = 82% RMSE = 4.1

Prognoza pogody na 20.05.2018

Temp = 199 - 19.6 * 0.40
- 0.006 * 10 - 0.002 * 28
- 0.11 * 50 - 0.17 * 1010
+ 0.81 * 16.87 = **19.56**

Rzeczywista temperatura 20.05.2018

Temp = **21**

Regresja liniowa - Pogoda w Szeged (PS)

- Eksploracja danych (rozkład, gęstość, skośność itd.)
- Normalizacja cech o wysokiej skośności (abs(density)>2)
- Zastąpienie podejrzanych wartości oraz nulli (średnia, dla precip_type ratio)
- Usunięcie Loud Cover (zawierało same 0)
- Sprawdzenie macierzy korelacji (jako wynik usunięcie App. Temp.)
- Usunięcie danych godzinowych z daty
- Pokubełkowanie zmiennych kategorycznych względem temperatury (summary-5, daily summary-7)
- One-hot encoding dla PrecipType
- Podział na zbiór testowy/trenujący
- Znalezienie optymalnej wartości ilości iteracji dla XGBoosta (parametr nrounds -> 1000)
- Uruchomienie XGBoosta dla przykładowych wartości (RMSE około 3,6, R_squared około 75%)
- Optymalizacja hiperparametrów dla XGBoost (RMSE około 1,51, R_squared około 98%)

Regresja liniowa - porównanie modeli

Model	BG	MK	PS
Założenia	T ~ d(month) + h + ws + p	T ~ h + ws + wb + v + p + monthly avg temp.	T ~ d(day) + pt + h + ws + v + wb + p + s_bin + ds_bin
	- eksploracja danych - NA → mediana - outliers → no change - usunięcie odczuwalnej temperatury (silna korelacja) - usunięcie danych tekstowych	- statystyki zmiennych - wykresy rozrzutu - porównanie rozkładów zmiennych - outliers → mediana - macierz korelacji - usunięcie temperatury odczuwalnej - weryfikacja istotności parametrów modelu (p-value < 0,05)	- NA → średnia - usunięcie zmiennych silnie skorelowanych - normalizacja dla dużej skośności - grupowanie dla zmiennych kategorycznych (binning) - one-hot encoding - optymalizacja hiperparametrów (xgboost) - optymalizacja liczby iteracji (xgboost) - użycie walidacji krzyżowej - zmiana daty na przedział dzienny
	train/test: 80/20 (seed 789)	train/test: 80/20 (seed 789)	train/test: 80/20 (seed 789)
RMSE / R^2	7,0 / 46%	4,1 / 82%	1,5 / 98%

d - Time, s - Summary, pt - Precip Type, **T** - Temperature, at - Apparent Temperature, h - humidity, ws - wind speed, wb - wind bearing, v - visibility, lc - loud cover, p - pressure, ds - daily summary

3. Regresja logistyczna

Regresja logistyczna

- Zbiór danych z Kaggle: The Ultimate Halloween Candy Power Ranking
- Dane - 85 obserwacji
- Czy cukierek jest czekoladowy?

Zmienna	Przykładowa wartość	Zmienna	Przykładowa wartość
competitorname	100 Grand	hard	0
chocolate	1	bar	1
fruity	0	pluribus	0
caramel	1	sugarpercent	.73199999
peanutyalmondy	0	pricepercent	.86000001
nougat	0	winpercent	66.971725
crispedricewafer	1		

Regresja logistyczna - Cukierki czekoladowe (BG)

Model 1:
Obecność czekolady w zależności od zawartości cukru, ceny i popularności
ch ~ sug_prc + prc_prc +win_prc

	TRUE		
PRED	10	1	0
	0	6	1
	0	1	

Model 2:
Obecność czekolady w zależności od kształtu, zawartości cukru, ceny i popularności
ch ~ crw + hd + bar + plb + sg_prc + prc_prc + win_prc

	TRUE		
PRED	11	0	0
	1	5	1
	0	1	

Regresja logistyczna - Cukierki czekoladowe (PS)

- Eksploracja danych
- Sprawdzenie ważności cech i usunięcie zmiennych mało istotnych (cumulative p-value > 80%)
- Przeskalowanie winpercent na liczbę
- Podział na zbiór testowy/trenujący
- Znalezienie optymalnej wartości ilości iteracji dla XGBoosta (parametr nrounds -> 10)
- Uruchomienie XGBoosta dla przykładowych wartości (Accuracy około 82%)
- Optymalizacja hiperparametrów dla XGBoost (Accuracy około 94%)

Metryki dla ostatecznego modelu:

- Accuracy - 94%
- AUC - 93%
- Precision - 92%
- Recall - 100%

Regresja logistyczna - Cukierki czekoladowe (MK)

Jak rozpoznać cukierki czekoladowe?

Model logistyczny: choco ~ fruit + price

Cukierki czekoladowe vs. **cukierki owocowe**

- Większość cukierków czekoladowych nie jest cukierkami owocowymi.
- Istnieją cukierku nieowocowe, które nie są czekoladowe.

Liczba cukierków (próba ucząca)		Owocowe	
		0	1
Czekoladowe	0	8	29
	1	30	1

Cukierki czekoladowe vs. **cena**

- Cukierki czekoladowe są droższe

Cukierki czekoladowe (próba ucząca)	Średnia cena
0	0.32
1	0.60

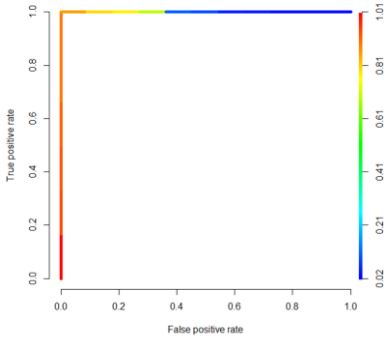
Weryfikacja modelu logistycznego - próba ucząca

Confusion matrix

Liczba cukierków (próba testowa) dla prob > 0.8		Czekoladowe Rzeczywistość	
		0	1
Czekoladowe Predykcja	0	11	0
	1	0	6

Precision = 6/6 = 100%
Recall = 6/6 = 100%
F1 Score = 100%
Accuracy = 17/17 = 100%

Krzywa ROC



Regresja logistyczna - porównanie modeli

Model	BG	PS	MK
Założenia	ch ~ sg_prc + prc_prc + win_prc	ch ~ bar + win_prc + ft + hd	ch ~ fr + prc_prc
	- założenie cukierka czekoladowego > 0.67	- usunięcie zmiennych nieistotnych (skumulowane p-value > 80%) - walidacja modelu na próbie testowej przy założenia prawdopodobieństwa cukierka czekoladowego > 0.5 - optymalizacja hiperparametrów (xgboost) - optymalizacja liczby iteracji (xgboost)	- iteracyjne usuwanie zmiennych nieistotnych (p-value > 0.05) - wybór najlepszego modelu na podstawie kryterium AIC - walidacja modelu na próbie testowej przy założenia prawdopodobieństwa cukierka czekoladowego > 0.8
	train/test: 80/20 (seed 789)	train/test: 80/20 (seed 789)	train/test: 80/20 (seed 789)
Accuracy / AUC / Precision / Recall	94% / 97% / 87% / 100%	94% / 93% / 92% / 100%	100% / 100% / 100% / 100%

cn - competitorname, ch - chocolate, ft - fruity, cr - caramel, pna - peanutyalmondy, ngt - nougat, crw - crispedricewafer, hd - hard, bar - bar, plb - pluribus, sg_prc - sugarpercent, prc_prc - pricepercent, win_prc - winpercent

Dziękujemy!

Pytania?
Slack / email