Wstęp do Uczenia Maszynowego

Projekt 1 - raport

Zespół:
Jakub Jung
Agata Kaczmarek
Piotr Marciniak

16.04.2021

1 Opis problemu

Nasz zbiór danych (link) przedstawia osiągnięcia uczniów dwóch szkół w Portugalii. Są to między innymi oceny cząstkowe studentów, a także społeczne, demograficzne oraz naukowe czynniki wpływające na naukę. Naszym zadaniem jest przewidzieć ocenę końcową dla studentów, na podstawie dostarczonych danych. Problem ten rozwiążemy na dwa sposoby: raz biorąc pod uwagę oceny śródroczne, które są dość zbliżone do tych końcowych (czyli problem będzie prostszy do rozwiązania), a raz nie biorąc ich pod uwagę.

2 Zbiór danych

Nasze dane zawierają informacje dotyczące 649 studentów z dwóch różnych szkół, którzy zostali scharakteryzowani poprzez 33 różne zmienne takie jak szkoła ucznia, jego płeć, wiek a także edukacja rodziców. Nie było w nich żadnych braków danych, również brak jest powtórzeń rekordów w tym zbiorze.

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	...	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
0	GP	F	18	U	GT3	A	4	4	at_home	teacher	...	4	3	4	1	1	3	4	0	11	11
1	GP	F	17	U	GT3	T	1	1	at_home	other	...	5	3	3	1	1	3	2	9	11	11
2	GP	F	15	U	LE3	T	1	1	at_home	other	...	4	3	2	2	3	3	6	12	13	12
3	GP	F	15	U	GT3	T	4	2	health	services	...	3	2	2	1	1	5	0	14	14	14
4	GP	F	16	U	GT3	T	3	3	other	other	...	4	3	2	1	2	5	0	11	13	13

Po przeprowadzeniu eksploracyjnej analizy danych potwierdziły się nasze początkowe przypuszczenia, że rozkłady ocen śródrocznych są bardzo podobne do rozkładu ocen końcowych. Co ciekawe, w rozkładach wszystkich kolumn zawierających oceny brakowało słupka oznaczającego wartości nieznacznie mniejsze od 10, czyli prawdopodobnie 10 (z 20) punktów było progiem zaliczenia. Pozostałe nasze obserwacje na podstawie EDA:

- Około $\frac{2}{3}$ uczniów uczęszczało do tej samej szkoły - Gabriel Pereira.
- Większość uczniów było płci żeńskiej.
- Około $\frac{2}{3}$ uczniów pochodzi z miejskich obszarów.
- U większości uczniów rodzice są razem. Co więcej, rodziny w większości mają więcej niż 3 członków.
- Prawie połowa uczniów wybrała taką szkołę ze względu na oferowane przez nie kursy.
- Tylko niewielki odsetek uczniów korzysta z dodatkowej pomocy edukacyjnej, podobnie jak niewielu uczęszcza na płatne korepetycje. Pomaga im głównie rodzina.
- Znaczna większość uczniów planuje po ukończeniu szkoły kontynuować edukację na studiach. Takie osoby osiągają lepsze wyniki na teście końcowym niż osoby nieplanujące wyższej edukacji.
- Większość uczniów ma dostęp do internetu w domu.
- Rodzice dzieci mają dość często podobne wykształcenie.
- Interesujące nas oceny końcowe są mało zależne od informacji dotyczących zdrowia, jakości relacji w rodzinie, liczby wyjść ze znajomymi a także liczby absencji w szkole.
- Osoby, które otrzymują dodatkową pomoc w nauce zazwyczaj mają wyniki niezbyt oddalone od 10 pkt, natomiast wyniki pozostałych osób są bardziej rozłożone pomiędzy wartości 10 a 20.

3 Preprocessing

Jak widać powyżej, niektóre z kolumn nie mają wartości liczbowych, mają wartości typu string. Tak więc przed rozpoczęciem modelowania postanowiliśmy je zamienić na wartości numeryczne. W wyniku zastosowania takiego preprocessingu, liczba kolumn w naszych danych zwiększyła się z 33 do 46.
Tak jak już wcześniej wspominaliśmy, zdecydowaliśmy się na przygotowanie dwóch różnych modeli - dla danych zawierających oceny śródroczne oraz dla tych bez nich. Nasz preprocessing danych zależał od modeli, które staraliśmy się nauczyć.

- Dla sieci neuronowych, svm'ów zastosowaliśmy standaryzację, ponieważ modele te radzą sobie lepiej, kiedy feature'y są na jednakowej przestrzeni. Dało nam to różne efekty, dla SVR delikatnie polepszyło to wyniki.
- Dla regresji liniowej wygenerowaliśmy wielomianowe feature'y, ale uzyskaliśmy gorsze wyniki niż bez nich - najpewniej sieć przeuczała się, dlatego wybraliśmy najlepsze 26 feature'ów. Dało nam to najlepsze wyniki dla zbioru z wynikami egzaminów.

4 Opis modelu i sposobu doboru parametrów

Do naszego zadania można podejść zarówno jako do problemu regresji, jak i klasyfikacji. Aby stwierdzić które podejście będzie lepiej przewidywać zmienną celu, stworzyliśmy i takie, i takie modele, jednakże w tych opierających się na regresji finalne predykcje rzutowaliśmy na liczby całkowite, jako że oceny końcowe nie mogły mieć części ułamkowej. Z powodu małej ilości danych, każdy model trenowany i sprawdzany był przy użyciu walidacji krzyżowej. Zastosowanie normalnego podziału na zbiór treningowy i testowy dałoby nam małe zbiory, zatem wynik z nich byłby nie do końca odpowiadający sytuacji rzeczywistej.
Na początek chcieliśmy nauczyć jak najwięcej modeli, aby później najlepsze z nich użyć w komitetach. Hiperparametrów dla tych modeli szukaliśmy przy użyciu GridSearchCVa. Jako miarę do porównywania jakości estymatorów użyliśmy **RMSE**, ponieważ zależało nam, aby nasze predykcje były jak najbliżej prawdziwego wyniku. Poniżej prezentujemy wyniki pojedynczych modeli.
Dla modeli z wynikami egzaminów dostajemy:

	RMSE							
	count	mean	std	min	25%	50%	75%	max
Model								
Selected regression	5.0	1.331076	0.369744	1.026570	1.130010	1.130010	1.435806	1.932986
Linear Regression	5.0	1.360112	0.349001	1.059753	1.189699	1.199359	1.408764	1.942986
SVR	5.0	1.392449	0.432248	1.003839	1.163549	1.231010	1.454436	2.109410
XGBoost Regressor	5.0	1.452704	0.296856	1.146902	1.221600	1.422349	1.588420	1.884247
XGBoost Classifier	5.0	1.461080	0.314808	1.060203	1.270978	1.531716	1.549193	1.893308
Neural Network	5.0	1.523967	0.348957	1.052470	1.361560	1.531716	1.691608	1.982481
Decision Tree	5.0	1.565333	0.267783	1.179961	1.400443	1.668717	1.771570	1.805973
Random Forest	5.0	1.618472	0.412392	1.038036	1.392286	1.718676	1.843909	2.099450
Polynomial Regression	5.0	1.697249	0.373365	1.389521	1.395046	1.521639	1.963122	2.216919

Najlepsze wyniki egzaminu osiągnęły regresja z powyżej opisanym preprocesingiem, liniowa regresja oraz SVR bez standaryzacji.
Dla modeli bez wyników egzaminów dostajemy

	RMSE							
	count	mean	std	min	25%	50%	75%	max
Model								
SVR	5.0	2.698487	0.787996	1.907475	2.424237	2.446347	2.704697	4.009678
Selected regression	5.0	2.759890	0.737478	2.123133	2.387467	2.414699	2.891632	3.982520
Linear Regression	5.0	2.762648	0.789071	2.086587	2.345208	2.452628	2.840639	4.088175
XGBoost Regressor	5.0	3.052754	0.646235	2.403523	2.710379	2.859532	3.202163	4.088175
Random Forest	5.0	3.059361	0.324533	2.728764	2.837930	2.980707	3.204564	3.544840
Polynomial Regression	5.0	3.094362	0.574515	2.585462	2.785954	2.890302	3.156190	4.053900
Neural Network	5.0	3.287775	0.607712	2.779042	2.896948	2.903579	3.712764	4.146541
Decision Tree	5.0	3.336951	0.367486	2.933887	2.956089	3.469649	3.623780	3.701351
XGBoost Classifier	5.0	3.400549	0.731764	2.548001	2.907550	3.467431	3.621517	4.458247

Jak widzimy najlepsze wyniki dla zbioru bez wyników egzaminu osiągnęły regresja z opisanym powyżej preprocessingiem, zwykła liniowa regresja i SVR ze standaryzacją. W tabelach nie ma klasyfikatorów, które po pierwszym podejściu uzyskały słabsze wyniki, uznaliśmy je za nie warte uwagi. Dla najlepszych 3 modeli zastosowaliśmy komitety (Stacking Regressor). Model w tabelach oznacza model wyjściowy. Poniżej mamy wyniki dla zbioru z wynikami egzaminów.

	RMSE							
	count	mean	std	min	25%	50%	75%	max
Model								
Linear Regression	5.0	1.359486	0.357478	1.034036	1.186462	1.196148	1.435806	1.944980
XGB Regressor	5.0	1.438033	0.313587	1.037749	1.218448	1.478044	1.640825	1.815096
SVR	5.0	1.621365	0.849156	1.026570	1.163549	1.400549	1.403293	3.112864

Poniżej mamy wyniki dla zbioru bez wyników egzaminów.

	RMSE							
	count	mean	std	min	25%	50%	75%	max
Model								
Linear Regression	5.0	2.730226	0.785716	1.978733	2.379399	2.428992	2.837930	4.026078
SVR	5.0	2.938258	1.203111	2.009592	2.390687	2.506914	2.747026	5.037072
XGB Regressor	5.0	3.290061	0.882723	2.189134	2.561550	3.536622	3.907586	4.255412

Jak widzimy komitety nie poprawiły nam wyników w obydwu przypadkach.

5 Podsumowanie

Podsumowując najlepszym modelem dla danych z wynikami egzaminu jest model nazwany **Selected Regression**. Jest to model, który generuje na początek wielomianowe feature'y stopnia 2, po czym wybiera najlepsze 26 featurerów (zastosowaliśmy **SelectKBest** z funkcją punktującą **f_regressor**). Dla taki spreparowanych danych zastosowaliśmy liniową regresję. Uzyskał on średnio **1,33** miary RMSE.
Natomiast dla danych bez wyników egzaminu najlepszym klasyfikatorem jest **ustandaryzowany SVR**, gdzie **C** ustawiliśmy na 3 (parametr regularyzacji), jako funkcję jądra ustawiliśmy **rbf** (radialna funkcja bazowa). Uzyskał

on średnio **2,70** miary RMSE.