Wstęp do Uczenia Maszynowego

Projekt 1

Mikołaj Spytek
Artur Żółkowski

19 kwietnia 2021

Spis treści

1 Opis problemu

Problem, którym zajmowaliśmy się w ramach tego projektu dotyczył predykcji ocen końcoworocznych uczniów z dwóch portugalskich szkół. Dane pochodzą ze strony `https://www.apispreadsheets.com/datasets/110`. Zadanie potraktowaliśmy jako problem regresji (z jednym małym wyjątkiem, któremu poświęcimy krótki rozdział tego raportu). Zmienna celu to liczba naturalna z zakresu 0-20. Zmienne wyjaśniające natomiast zostały zebrane za pomocą ankiety przeprowadzonej wśród uczniów w badanych szkołach. Dotyczą one głównie czynników społecznych i ekonomicznych.

Dwie ze zmiennych opisują oceny uzyskane przez uczniów w poprzednich semestrach, jednak jak zaznaczył sam autor zbioru danych model korzystający z tych kolumn jest dużo mniej przydatny. Dlatego też postanowiliśmy zająć się tylko modelami, które z nich nie korzystają.

2 Eksploracyjna analiza danych

EDA rozpoczęliśmy od zapoznania się ze zmiennymi obecnymi w zbiorze danych. Dotyczą one bardzo różnych aspektów życia takich jak wykształcenie i zawód rodziców, miejsce zamieszkania, czas dojazdu do szkoły, czas poświęcony na naukę, zdrowie, czy ilość nieobecności. Kilka przykładowych wierszy z ramki danych przedstawia rysunek 1.

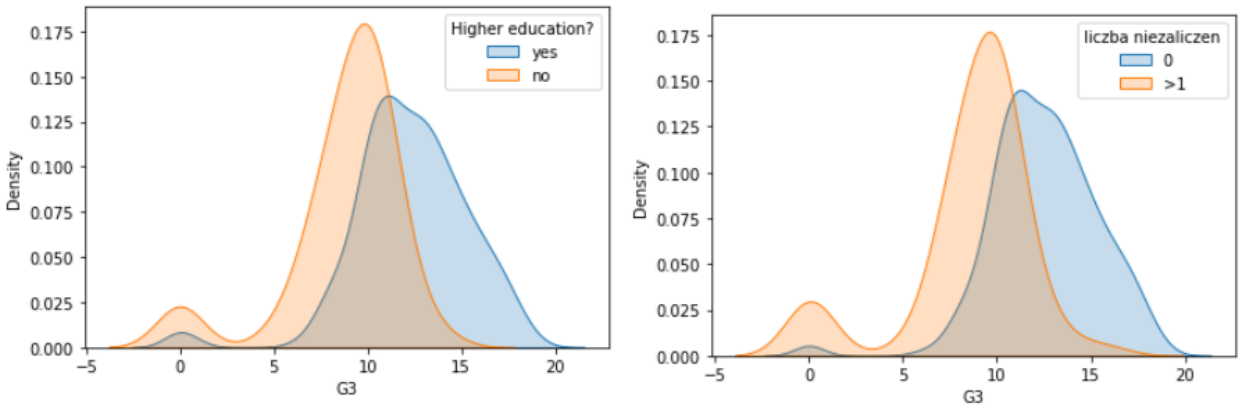
	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason	guardian	traveltime	studytime	failures	schoolsup	famsup	paid	activities	nursery	higher	internet	romantic	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
0	GP	F	18	U	GT3	A	4	4	at_home	teacher	course	mother	2	2	0	yes	no	no	no	yes	yes	no	no	4	3	4	1	1	3	4	0	11	11
1	GP	F	17	U	GT3	T	1	1	at_home	other	course	father	1	2	0	no	yes	no	no	no	yes	yes	no	5	3	3	1	1	3	2	9	11	11
2	GP	F	15	U	LE3	T	1	1	at_home	other	other	mother	1	2	0	yes	no	no	no	yes	yes	yes	no	4	3	2	2	3	3	6	12	13	12
3	GP	F	15	U	GT3	T	4	2	health	services	home	mother	1	3	0	no	yes	no	yes	yes	yes	yes	yes	3	2	2	1	1	5	0	14	14	14
4	GP	F	16	U	GT3	T	3	3	other	other	home	father	1	2	0	no	yes	no	no	yes	yes	no	no	4	3	2	1	2	5	0	11	13	13

Rysunek 1: Przykładowe rekordy

Po zapoznaniu się ze zmiennymi sprawdziliśmy, czy zbiór nie zawiera wartości pustych - okazało się że nie, więc nie trzeba przeprowadzać imputacji.

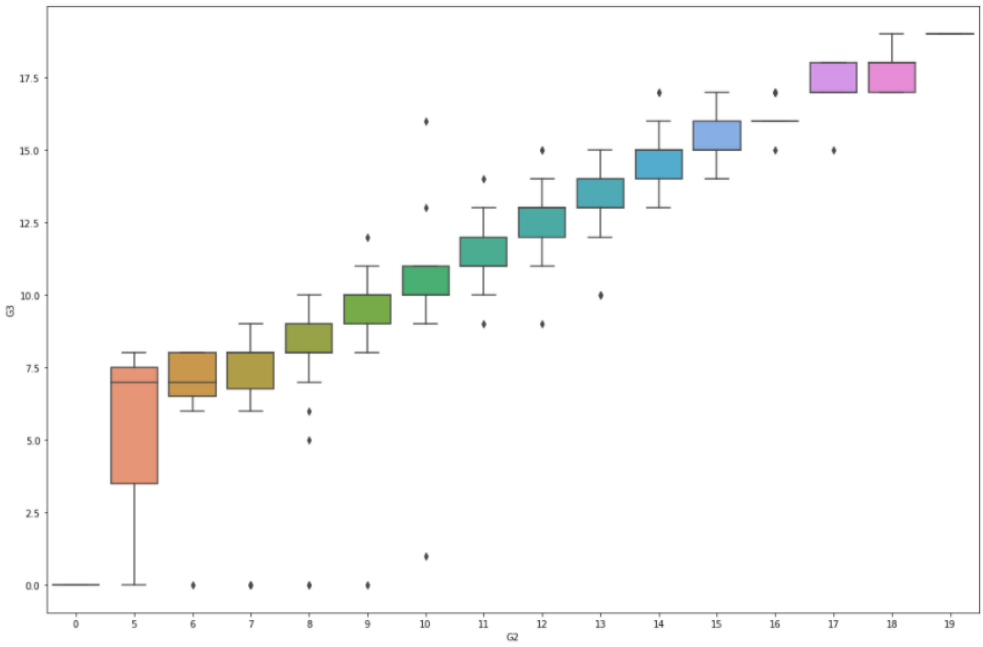
Następnie stworzyliśmy wykresy zmiennych liczbowych, oraz wyznaczyliśmy macierz korelacji zmiennych między sobą. Dzięki temu mogliśmy zobaczyć, które zmienne są od siebie zależne oraz które najbardziej wpływają na zmienną celu. Stworzyliśmy również wykresy rozkładów niektórych zmiennych, przyjmując jako parametr zmienną celu. Dzięki nim już na pierwszy rzut oka mogliśmy zobaczyć, które z nich dobrze rozdzielają ocenę końcową. Dwie zmienne które robią to naszym zdaniem najlepiej zaprezentowane są na rysunku 2. Oprócz tego zaobserwowaliśmy kilka innych ciekawych faktów:

- Wykształcenie matki i ojca jest silnie ze sobą skorelowane - do modelu wystarczy wziąć jedną z tych zmiennych.
- Dużą korelację z oceną końcową mają zmienne oznaczające: chęć podjęcia nauki wyższej, szkoła, liczba niezaliczeń, czas poświęcony na naukę, spożycie alkoholu.
- Niektóre zmienne wykazywały większy wpływ w połączeniu ze sobą niż osobno (np. płeć + informacja o byciu w związku)
- Zmienne które logicznie powinny być ze sobą powiązane - były. Na przykład adres zamieszkania poza miastem z dłuższym czasem dojazdu - dzięki temu mogliśmy przypuszczać że korelacja jest dobrze wyliczona.



Rysunek 2: Przykłady rozkładów zmiennych rozdzielających zmienną celu

Sprawdziliśmy również jak rozkładają się zmienne, których postanowiliśmy nie używać ze względu na przydatność modelu tj. oceny z poprzednich semestrów. Na rysunku 3 widać, że zależność jest prawie liniowa. Potwierdza to fakt, że taka zmienna bardzo ułatwia wytrenowanie modelu, jednak sprawia, że jest mało przydatny.



Rysunek 3: Rozkład zmiennej oznaczającej ocenę z poprzedniego okresu

3 Podejście klasyfikacyjne

Przez pewien czas rozważyliśmy podejście klasyfikacyjne do tego problemu. Jednak wydawało nam się, że klasyfikacja 20-klasowa, w szczególności z użyciem miar oceniających modele klasyfikacyjne nie ma sensu. (W naszych klasach jest porządek, więc ma sens ocenianie odległości pomiędzy nimi). Dlatego też zastosowaliśmy model do przewidzenia, czy dany uczeń zdał (ocena $\geqslant 10$). Korzystając z przeprowadzonej analizy

dobraliśmy kilkanaście według nas najbardziej istotnych zmiennych i stworzyliśmy przykładowy model. Zauważyliśmy jednak, że podział klas (zdał - nie zdał) jest nierówny. Uczniowie, którzy zdali stanowili około 84% wszystkich uczniów. Miara accuracy naszego modelu była na poziomie 0.89, więc model nie był znacznie lepszy od zwykłego przypisania wszystkich obserwacji do klasy 1 (uczeń zdał egzamin). Biorąc to pod uwagę, stwierdziliśmy, że model tak dopasowujący obserwacje nie satysfakcjonuje nas, chcielibyśmy, by nasz model niósł ze sobą więcej informacji, a więc zdecydowaliśmy się na rozważanie problemu regresyjnego i przewidywanie ostatecznego wyniku ucznia.

4 Preprocessing i feature engineering

Po ostatecznej decyzji, że traktujemy zadanie jako problem regresyjny, przygotowaliśmy dane tak, aby modele pracowały na nich optymalnie.

W tym celu wszystkie kolumny tekstowe zakodowaliśmy do wartości liczbowych stosując One-Hot-Encoding. Zmienne liczbowe natomiast przeskalowaliśmy przez maksimum tak, aby wszystkie liczby były z przedziału $[0; 1]$, co pomaga lepiej nauczyć niektóre z wykorzystanych przez nas modeli.

Postanowiliśmy skorzystać z automatycznych sposobów wybierania najbardziej znacznych feature'rów. Zastosowaliśmy kilka metod i wybraliśmy najlepsze z nich. Do oceny modelu zastosowaliśmy miarę RMSE, jako że jest naszym zdaniem odpowiednia do problemu oraz intuicyjna - mówi średnio jak daleko była ocena predykowana od rzeczywistej.

Zaczęliśmy od wygenerowania kombinacji wielomianowych odpowiednio przekształconych już zmiennych. Spośród nich wybieraliśmy 12 najlepszych za pomocą narzędzi automatycznego wyboru. Ponieważ zbiór danych nie jest dużych rozmiarów (zawiera 649 rekordów, z których 20% zostało odłożonych jako zbiór testowy), postanowiliśmy trenować model na dość małej ilości zmiennych. Z pewnością pomoże to zmniejszyć ryzyko przeuczenia (które np. w lasach losowych jest dość wysokie) jednocześnie pozwalając osiągnąć zadowalające wyniki.

Do wyboru zmiennych zastosowaliśmy:

- algorytm `selectKBest` z selectorem `chi2`
- algorytm `selectKBest` z selectorem `mutual information`
- recursive feature selection
- l1 based feature selection

Dodatkowo zdecydowaliśmy się także ręcznie przygotować i wybrać zmienne korzystając z przeprowadzonej wcześniej EDA, rezultatów automatycznych narzędzi, a także z własnej wiedzy i intuicji (w końcu nie tak dawno sami byliśmy uczniami liceum).

selectKBest z selectorem chi2		selectKBest z selectorem mutual information		recursive feature selection		l1 based feature selection		ręcznie wybrane zmienne	
failures	school_MS Fjob_health	failures	school_MS goout	failures	absences	age	failures	fail	reason
address_U failures	Pstatus_T failures	Pstatus_T failures	Fjob_health Walc	school_MS Pstatus_T	school_MS guardian_father	guardian_mother failures	higher_yes Medu	higher	school
Mjob_at_home Fjob_healt	Mjob_services failures	higher_yes studytime	higher_yes Dalc	schoolsup_yes nursery_yes	nursery_yes failures	failures^2	failures Medu	age	goout
Fjob_health reason_home	Fjob_services failures	age failures	failures^2	higher_yes Fedu	higher_yes famrel	failures traveltime	failures studytime	Pedu	Gender Relations
reason_other failures	guardian_mother failures	failures Fedu	failures studytime	internet_yes failures	Medu famrel	failures famrel	failures Dalc	address	internet
internet_yes failures	romantic_yes failures	failures Dalc	failures Walc	failures Fedu	Fjob_health reason_home	failures Walc	traveltime Dalc	Mjob	studytime

Rysunek 4: Wybrane zmienne dla zastosowanych metod

5 Końcowe modele

Zdecydowaliśmy się na sprawdzenie i porównanie między sobą czterech modeli. Dwa z nich, to modele proste: regresja logistyczna oraz regresor opierający się na SVM. Natomiast kolejne dwa opierają się na komitetach. Las losowy oparty na baggingu, oraz model oparty na Gradient Boostingu.

Hiperparametry modeli stroiliśmy ręcznie bez użycia narzędzi automatycznych. W wyniku naszych eksperymentów ustaliliśmy, że wartościami pozwalającymi osiągnąć najlepsze wyniki są:

Dla modelu opartego na Gradient Boostingu:

- `learning_rate=0.045`,
- `n_estimators=100`,
- `criterion='mse'`

Dla lasu losowego:

- `n_estimators=20`,
- `max_features=0.5`,
- `min_samples_split=3`

Każdy z tych modeli wytrenowaliśmy na wcześniej przygotowanym zbiorze danych. Wyniki (Miarę RMSE) przedstawia tabela 1.

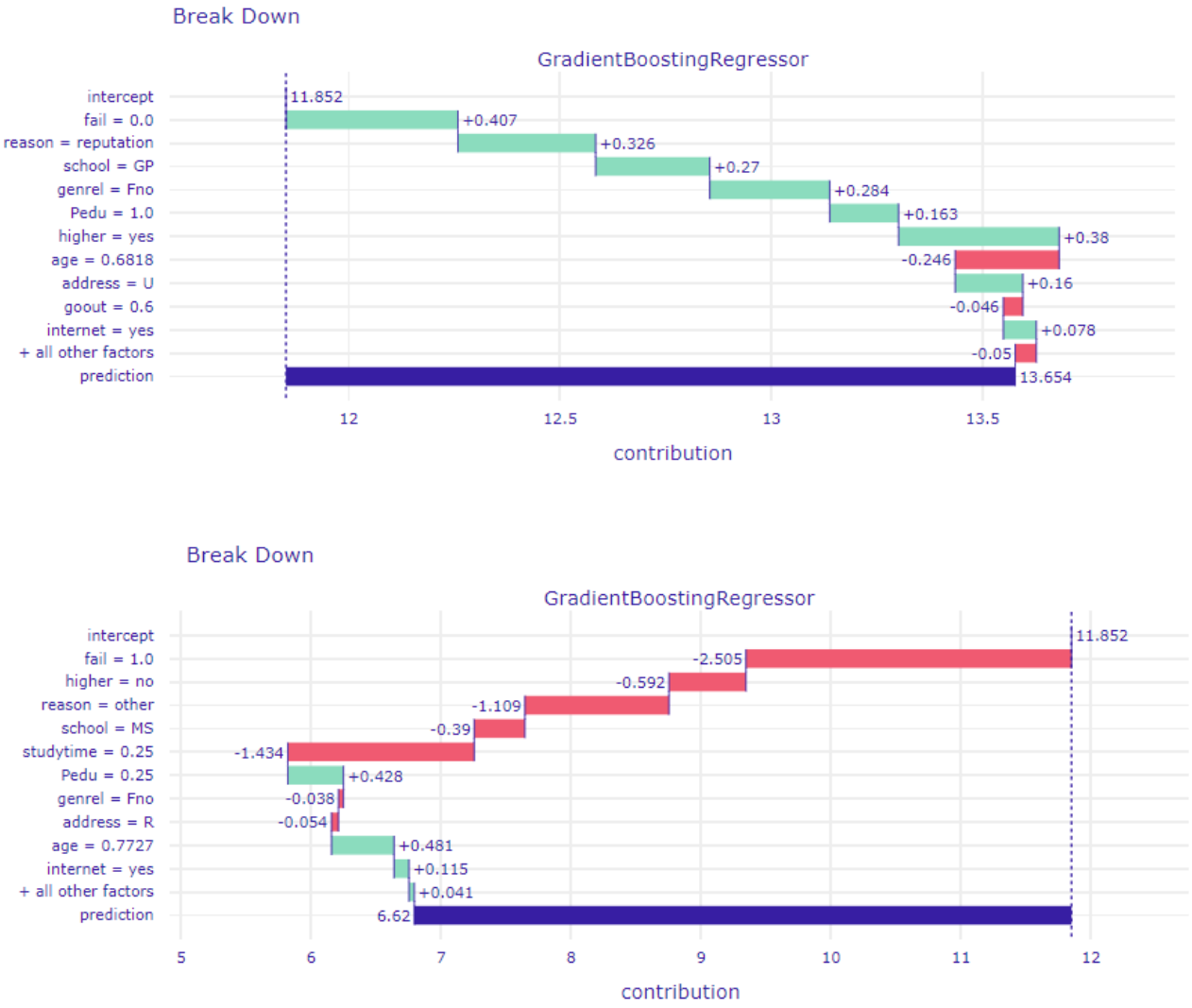
	Logistic Regression	**SVR**	**Random Forest**	**Gradient Boosting**
SelectKBest (chi2)	3.359258	2.956194	2.979676	3.063093
Mutual information	3.336281	2.981460	3.036491	2.957216
RFE	3.181678	2.940780	2.932803	2.940329
L1 based Model Selection	3.080959	2.979083	3.169495	2.999660
Hand-prepared features	3.124346	2.782860	2.819433	2.656182

Tabela 1: Miary RMSE osiągnięte przez modele na poszczególnych podzbiorach zmiennych

6 Wyniki

Spośród wszystkich przetestowanych przez nas modeli najlepszy wynik osiągnął model oparty na Gradient boostingu uzyskując miarę **RMSE** równą **2.65**.

Uznaliśmy, że ciekawe będzie sprawdzenie, czy nasze przeczucia z etapu eksploracji danych były właściwe - czy wytypowane przez nas zmienne rzeczywiście mają największy wpływ na predykcję modelu. Dla najlepszego modelu zastosowaliśmy explainer z biblioteki `DALEX`. Na rysunku 5 znajduje się wyjaśnienie przewidzianych ocen dla dwóch obserwacji - wyniku przeciętnego i słabego.



Rysunek 5: Wyjaśnienie predykcji

Z wykresów tych wynika, że największy wpływ ma liczba uprzednich niezaliczeń, powód wybrania danej szkoły, konkretna szkoła, w której uczeń się uczy oraz chęć podjęcia dalszej edukacji. Większość z powyższych zmiennych pokrywa się z naszymi przewidywaniami z pierwszej fazy projektu. Warto również zauważyć, że np. w przypadku zmiennej `genrel` łączącą w sobie informacje o płci i statusu związku danego ucznia wpłynęła odmiennie dla prezentowanych rekordów, mimo takiej samej wartości. Może to świadczyć o wyłapaniu przez model bardziej złożonej zależności tej zmiennej, którą podejrzewaliśmy, po analizie EDA.