7 maja 2024 r.

Raport walidacyjny

Mikołaj Rowicki, Jakub Półtorak

1. Wstęp

Zespół walidacyjny nr 1, w składzie: Mikołaj Rowicki i Jakub Półtorak, nadzorował prace zespołu nr 2, w składzie: Igor Rudolf i Kacper Rodziewicz, powiązane z budową modelu Machine Learningowego do zestawu danych: Tweets Dataset. Po każdym poszczególnym kamieniu milowym walidowaliśmy napisany kod oraz udzielaliśmy niezbędnych wskazówek po zauważeniu błędów. Dodatkowo w ostatnim etapie – analizie zastosowanego modelu - uruchomiliśmy wybrany model i sprawdziliśmy jego skuteczność na specjalnie przygotowanym do tego zbiorze danych, który nie był dostępny dla zespołu budowy. Zbiór ten stanowił około 30% pierwotnego zbioru danych.

2. Cel biznesowy

Zespół budowy opisał tworzony przez siebie model następującymi słowami: „Jest on szczególnie użyteczny w miejscach, gdzie istotne jest zachowanie wysokiego poziomu pozytywnych interakcji. Możemy sobie wyobrazić jego zastosowanie na stronach skierowanych do dzieci, gdzie priorytetem jest bezpieczne i pozytywne środowisko, czy też na oficjalnych profilach politycznych, gdzie ważne jest utrzymanie pozytywnego wizerunku.” To naszym zdaniem dobry cel, który uwzględnia potencjalne komercyjne zastosowanie modelu. Z tego też powodu istotną metryką do walidacji naszego modelu będzie Precision, ale także Accuracy.

3. Eksploracyjna Analiza Danych (EDA)

W ramach walidacji eksploracyjnej analizy danych zapoznaliśmy się szczegółowo z plikiem first_step_kacper_rodziewicz_igor_rudolf.ipynb. Zwróciliśmy uwagę na wiele pozytywnych aspektów tego kodu, jak również na obszary, które wymagają poprawy.

Mocne strony:

- Profesjonalnie przeprowadzenie czyszczenie i preprocessing danych. Zespół budowy rozpoczął pracę nad przewidywaniem sentymentu tweetów od szczegółowej analizy leksykalno-gramatycznej wpisów. Usunięto wszystkie znaki specjalne i interpunkcyjne, które nie wnoszą informacji do tworzonego modelu. Usunięte zostały również wszystkie stop-wordy (typu ‘and’, ‘but’, itd.) oraz inne leksemy, które nie niosą żadnej niezbędnej informacji.

- Zespół budowy dokonał tokenizacji i lematyzacji w oparciu o aktualne standardy Natural Language Processingu z wykorzystaniem funkcji WorldNetLemmatizer z biblioteki nltk
- Na plus zasługuje także przeprowadzona analiza długości zlematyzowanych tokenów. Zespół budowy badał także poszczególne zależności między sentymentami tweetów a ich długością. Przeprowadzono także różnorodne, niezależne i zgodne z obecnymi standardami testy statystyczne: Mann-Whitney test, Mood test
- Dobrym wyborem było przeprowadzenie wizualizacji częstości występowania poszczególnych wyrazów za pomocą mapy słów. Wykresy są dość czytelne, wyraźnie widać, które słowa są najpopularniejsze, a które rzadziej używane
- Ciekawą zastosowaną techniką w eksploracyjnej analizie danych był embedding. Zespół budowy w dość interesujący sposób dokonał klasteryzacji poszczególnych pozytywnych i negatywnych słów. W części z nich następowały problemy z odczytywaniem odpowiednich przypadków, jednak w znaczącej większości nie występował taki problem.
- Zespół budowy podjął się również próby przewidywania tematu tweetu, co potencjalnie mogłoby prowadzić do skuteczniejszego przewidywania ich sentymentu. Do wizualizacji tematów, co zasługuje na podkreślenie, zastosowano również innowacyjną bibliotekę pyLDavis.

Elementy do poprawy:

- Mapy słów pokazujące częstość występowania poszczególnych wyrazów z podziałem na sentyment tweetów zawierają wiele niepotrzebnych wyrazów. Na obu wykresach widzimy, że bardzo często w tweetach, niezależnie od ich sentymentu, pojawiają się słowa takie jak ‘i`m’, ‘go’, ‘get’, ‘time’. To oznacza, że nie wnoszą one wiele do predykcyjności modelu, a zatem być może warto rozważyć ich usunięcie.
- Porównanie najczęściej występujących słów w pozytywnych i negatywnych tweetach na wykresie nic nie wnosi. Obserwujemy znaczącą dominację poszczególnych słów w negatywnych tweetach, jednak ich ogólny odsetek jest znacznie większy niż w pozytywnych tweetach, przez co jakakolwiek konstruktywna analiza nie ma znaczenia. Należałoby przeskalować te dane i ewentualnie wtedy wyciągnąć jakieś ciekawe zależności.
- Jedna z wizualizacji grafowych ilustrujących poszczególne wyrazy była nieczytelna i nie wnosiła nic do dalszych operacji na zbiorze danych

Odpowiedź zespołu budowy

Propozycje zmian zostały pozytywnie zaakceptowane przez zespół budowy. Błędna wizualizacja została usunięta z projektu, a sugestie dotyczące bliższego przyjrzenia się niektórym słowom (w kontekście tego, czy mają one wartość predykcyjną) wywarły duży wpływ na kolejne części projektu

4. Feature Engineering

W ramach Feature Engineeringu zespół walidacyjny również nie napotkał rażących błędów oraz znaczących niedociągnięć w działaniu zespołu budowy. Sprawdzano jakość kodu w pliku

second_third_step_kacper_rodziewicz_igor_rudolf, który jest konkatenacją efektów prac po drugim i trzecim kamieniu milowym. Oto pozytywne i negatywne aspekty kodu napisanego przez zespół budowy:

Mocne strony:

- Wprowadzenie prostych statystyk opisujących poszczególne zależności pod kątem liczby słów takie jak: średnia liczba słów, wariancja liczby słów, odchylenie standardowe liczby słów we wszystkich tweetach. Wielkości nie odegrały znaczącej roli w dalszych analizach, jednak na plus zasługuje fakt, że pamiętano i wprowadzono takie zależności.
- Użycie nietrywialnego podejścia bazującego na bigramach. Wyznaczono wszystkie unikalne pary poszczególnych tagów. Dodatkowo skorzystano także z podejścia one-hot encoding, co wydaje się dość zasadne.
- Ciekawie przeprowadzona analiza tematów tweetów. Wybrano 35 najbardziej trafnych wielkości. Zastosowano nietrywialny model LDA, który używany był już wcześniej (podczas eksploracyjnej analizy danych). Badano również spójność poszczególnych tematów w przypadkach pozytywnych i negatywnych. W tym celu zastosowano Coherence Model.

Elementy do poprawy:

- Zespół budowy zdecydował się na usuwanie kolumn z niską wariancją i uznał za wartość graniczną wariancję równą 0,2. Naszym zdaniem nie można stosować stałego, niezmiennego progu dla wszystkich cech, ponieważ w niektórych przypadkach kolumna o nawet tak małej wariancji może mieć duży wpływ na predykcyjność modelu. Należy usuwać kolumny o wariancji równej dokładnie 0 lub mniejszej od odpowiednio dostosowanego do danej cechy progu
- W kodzie brakowało wniosków, w szczególności nie wiadomo było, w jaki sposób wykonywane przez zespół walidacyjny przekształcenia wpływają na predykcyjność modelu. Zasugerowaliśmy, by sprawdzić to za pomocą wielu niezależnych testów statystycznych (na przykład chi-kwadrat dla zmiennych binarnych, czy też testu Anova-F) lub ewentualnie bezpośrednio na tworzonych modelach, licząc jednoczynnikowe Gini.
- Podczas walidacji drugiego kamienia w kodzie panował lekki chaos, który jednak udało się opanować

Odpowiedź zespołu budowy

Wszystkie z wysuniętych przez nas sugestii zostały rozważone przez zespół budowy. W kontekście usuwania kolumn z niską wariancją, zespół tworzący model podjął decyzję o obniżeniu progu do wariancji równej 0,001, co w kontekście tego zbioru danych powinno być wystarczająco nisko wartością. Jednakże, nasza sugestia jest niezmienna, by usuwać wyłącznie kolumny z wariancją równą 0 lub naprawdę bliską tej wartości. Stworzone cechy oraz przeprowadzone przez zespół budowy ich przekształcenia zostały przetestowane przez kilka innowacyjnych metod selekcji. Zespół budowy wykorzystał metody SelectBySingleFeaturePerformance, SelectKBest oraz Tree Based Feature Selection, które pozwoliły wybrać podzbiór najbardziej istotnych cech. Sam kod został natomiast uporządkowany i odpowiednio udokumentowany, co ułatwiło pracę zespołowi walidacji.

5. Analiza zastosowanego modelu

Zespół budowy przygotował kilka modeli dla prezentowanego celu biznesowego. Zespół walidacyjny przeanalizował poszczególne wyniki analiz różnych metryk oraz sam uruchomił wybrane modele na specjalnie do tego przeznaczonym zbiorze danych, który nie był dostępny dla zespołu budowy (zbiorze testowym). Wspólnie ustalono, że najbardziej kluczowym aspektem będzie zmaksymalizowanie wartości metryki precision dla wszystkich tweetów. Zespół budowy podkreślił jednak, że dla niego bardzo ważną metryką było również Accuracy. Zespół walidacyjny z kolei zwrócił uwagę na niebagatelne znaczenie metryki Gini przy ocenie jakości modelu uczenia maszynowego. Poniżej przedstawiano wartości różnych metryk dla kolejnych testowanych modeli.

Pierwsze trzy modele były wywoływane na zbiorach niezbalansowanych, stąd też wartość niektórych metryk, np. Accuracy, może być mało wiarygodna. Po lewej stronie mamy przedstawione wyniki uzyskane na zbiorze walidacyjnym (z którego korzystał zespół budowy), zaś po prawej wyniki na zbiorze testowym (z którego korzystał zespół walidacyjny)

Na zbiorze walidacyjnym	Na zbiorze testowym
XGBClassifier Metrics: Accuracy: 0.82 Precision: 0.65 Recall: 0.52 F1 Score: 0.58 ROC AUC: 0.85	XGBClassifier Metrics: Accuracy: 0.81 Precision: 0.62 Recall: 0.49 F1 Score: 0.55 ROC AUC: 0.83
LGBMClassifier Metrics: Accuracy: 0.84 Precision: 0.72 Recall: 0.50 F1 Score: 0.59 ROC AUC: 0.88	LGBMClassifier Metrics: Accuracy: 0.84 Precision: 0.72 Recall: 0.50 F1 Score: 0.59 ROC AUC: 0.88
CatBoostClassifier Metrics: Accuracy: 0.83 Precision: 0.65 Recall: 0.63 F1 Score: 0.64 ROC AUC: 0.88	CatBoostClassifier Metrics: Accuracy: 0.83 Precision: 0.65 Recall: 0.63 F1 Score: 0.64 ROC AUC: 0.88

Następnie te same modele zostały wywołane na zbiorach zbalansowanych. Można zauważyć spadek w metryce Accuracy, ale za to wzrost w bardzo istotnym Precision, metryce Recall oraz F1. Balansowanie zbioru nie wpłynęło natomiast na wartość metryki ROC AUC czy też Gini. Poniżej wywołanie tych modeli na zbiorze zbalansowanym:

Na zbiorze walidacyjnym	Na zbiorze testowym
XGBClassifier Metrics: Accuracy: 0.77 Precision: 0.74 Recall: 0.82 F1 Score: 0.78 ROC AUC: 0.85	XGBClassifier Metrics: Accuracy: 0.75 Precision: 0.72 Recall: 0.80 F1 Score: 0.76 ROC AUC: 0.83
LGBMClassifier Metrics: Accuracy: 0.79 Precision: 0.79 Recall: 0.80 F1 Score: 0.80 ROC AUC: 0.88	LGBMClassifier Metrics: Accuracy: 0.80 Precision: 0.79 Recall: 0.80 F1 Score: 0.80 ROC AUC: 0.88
CatBoostClassifier Metrics: Accuracy: 0.79 Precision: 0.76 Recall: 0.86 F1 Score: 0.80 ROC AUC: 0.88	CatBoostClassifier Metrics: Accuracy: 0.79 Precision: 0.76 Recall: 0.85 F1 Score: 0.80 ROC AUC: 0.88

Gołym okiem widać, że wyniki na zbiorze testowym i zbiorze walidacyjnym są bardzo do siebie zbliżone. Nie obserwujemy żadnych niepokojących zjawisk takich jak underfitting lub overfitting. To bardzo dobry prognostyk na przyszłość, gdyż śmiało możemy założyć, że przy ewentualnym następnym testowaniu na zupełnie innym zbiorze, otrzymamy równie podobne wyniki. Oczywiście wartości metryk accuracy i precision prezentują się dość dobrze zarówno dla niezbalansowanego jak i zbalansowanego zbioru danych, jednak jest tu znaczące pole do poprawy, dlatego nie zastosowano powyższych modeli jako ostatecznego rozwiązania.

Problem oceny sentymentu tweetu zespół budowy próbował rozwiązać również za pomocą sieci neuronowej. Wytrenowano i wywołano ją na niezbalansowanym zbiorze danych. Poniżej przedstawiamy wartości metryk dla tego modelu wywołanego odpowiednio na zbiorze walidacyjnym i testowym:

Na zbiorze walidacyjnym **Na zbiorze testowym**

	precision	recall	f1-score	support
0	0.88	0.90	0.89	82343
1	0.64	0.59	0.61	25554
accuracy			0.82	107897
macro avg	0.76	0.74	0.75	107897
weighted avg	0.82	0.82	0.82	107897

```
[[73800  8543]
 [10512 15042]]
```

	precision	recall	f1-score	support
0	0.86	0.92	0.89	82343
1	0.67	0.53	0.59	25554
accuracy			0.83	107897
macro avg	0.77	0.72	0.74	107897
weighted avg	0.82	0.83	0.82	107897

```
[[75747  6596]
 [12120 13434]]
```

Widać, że przez niezbalansowanie znacznie lepsze wyniki metryk osiągane są dla tweetów o negatywnym sentymencie. Cieszyć mogą jednak ponownie bardzo podobne wyniki dla różnych zbiorów dla metryki precision. Można zauważyć także nieduży wzrost dla metryki precision w porównaniu z poprzednimi modelami, jednak biorąc pod uwagę to, że sieć neuronowa znacznie traci na interpretowalności względem innych modeli, to ten wzrost nie jest na tyle atrakcyjny, aby wybrać tę sieć neuronową jako ostateczne rozwiązanie.

Zespół budowy zdecydował się również na użycie transformerów do przewidywania sentymentu tweetów. Wartości poszczególnych metryk uzyskanych na obu zbiorach przedstawiają ilustracje poniżej:

Na zbiorze walidacyjnym **Na zbiorze testowym**

	precision	recall	f1-score	support
Class 0	0.91	0.93	0.92	82343
Class 1	0.76	0.71	0.73	25554
accuracy			0.88	107897
macro avg	0.83	0.82	0.83	107897
weighted avg	0.87	0.88	0.88	107897

	precision	recall	f1-score	support
Class 0	0.91	0.93	0.92	117710
Class 1	0.75	0.71	0.73	36430
accuracy			0.88	154140
macro avg	0.83	0.82	0.82	154140
weighted avg	0.87	0.88	0.87	154140

Wynik metryki precision, a dokładniej średniej ważonej tej metryki dla obu klas jest w tym modelu zdecydowanie najlepszy i wynosi 0,87. Wynik pozostałych metryk, takich jak Recall czy F1 jest jeszcze lepszy i wynosi (przy średniej ważonej) 0,88. Ze względu na te wyniki zespół budowy uznał ten model jako najlepszy ze wszystkich. My, jako zespół walidacyjny, zgadzamy się z tym. Zwróciliśmy jednak uwagę na brak metryki Gini czy też ROC-AUC przy ocenie poprawności tego modelu. Zespół budowy przyznał nam rację i dodał tę metrykę do ostatecznej wersji projektu. Podkreśliliśmy też, że model jest wymagający obliczeniowo i jego wywołanie w sensownym czasie wymaga GPU Nvidii. Próba wytrenowania tego modelu na CPU zwykłego procesora mogłaby zająć ponad 20 godzin.

6. Wnioski

Wybrany przez zespół budowy model jest zdecydowanie najlepszy ze wszystkich pod względem metryki Precision. Zgadzamy się zatem, że ten właśnie model powinien zostać

ostatecznie wybrany. Dobre wyniki w poszczególnych metrykach uzyskane zostały jednak kosztem interpretowalności modelu. Zespół budowy uznał, że w przypadku transformerów nie jest możliwe zinterpretowanie wpływu poszczególnych zmiennych na działanie modelu. Zespół walidacyjny zaakceptował to, stąd też brak w projekcie wartości Shapleya czy jakiejś innej formy interpretacji. Zespół walidacyjny zauważył również, że wśród metryk stosowanych do oceny ostatniego modelu brakuje ROC-AUC czy też Gini, co zostało bardzo szybko skorygowane przez zespół budowy.

Podsumowując, model spełnia swoje biznesowe założenia i może z powodzeniem być wykorzystywany do przewidywania sentymentu tweetów, w szczególności w miejscach, gdzie priorytetem jest bezpieczne i pozytywne środowisko albo utrzymanie pozytywnego wizerunku. Jako zespół walidacyjny potwierdzamy jego skuteczność, prawidłowość i zgodność z przepisami prawa europejskiego, w tym jego najnowszymi regulacjami zawartymi w AI Act.