Data 06.05.2024

Małgorzata Mokwa
Sebastian Pergała
Zespół nr 4

Raport walidacyjny

Projekt 1 - klasyfikacja
Walidacja zespołu 1

Spis treści

1 Wstęp

Przeprowadziliśmy walidację dla zespołu numer 1, który pracował na zestawie danych numer 1. Zbiór ten zawierał informacje dotyczące decyzji o zatwierdzeniu pożyczki (oznaczonej wartością logiczną w kolumnie FINALIZED_LOAN, gdzie 0 oznaczało brak zatwierdzenia). Zespół przystąpił do próby przewidzenia zdolności do uzyskania pożyczki za pomocą metod klasyfikacji. Proces walidacji został podzielony na trzy etapy, związane z oddzielnymi kamieniami milowymi, aby uzyskać ocenę skuteczności modelu.

2 Eksploracyjna analiza danych

Na tym etapie dokładnie przyjrzeliśmy się w analizie zbioru danych, z którym pracował zespół numer 1. Ich analiza była bardzo gruntowna, rozpoczynając od ogólnych wniosków dotyczących liczby i rodzaju kolumn, ich typów danych, aż po generowanie wykresów i bardziej zaawansowane techniki analizy. Poznaliśmy ogólną strukturę danych, identyfikując dostępne kolumny oraz dokładnie przyjrzeliśmy się rozkładowi poszczególnych zmiennych, szczególnie w kontekście zmiennej celu. Dzięki ich starannym analizom ujawniły się istotne zależności między różnymi zmiennymi oraz ich wpływ na siebie nawzajem. W trakcie pracy na zbiorze testowym zauważyliśmy, że wszystkie ich wnioski i obserwacje pokrywają się z naszymi własnymi spostrzeżeniami, co potwierdza dokładność przeprowadzonych przez nich analiz.

Po dokładnym zapoznaniu się z zawartością plików odnoszących się do etapu inżynierii cech zastały znalezione elementy wymagające poprawy, które są wymienione poniżej.

Walidacja pliku **eda_processing.ipynb**.

1. Brak wskazanych kolumn o znaczącej korelacji.
2. W przypadku analizy cech pod kątem korelacji zabrakło wniosków, jak silnie jedna zmienna zależy od drugiej. (np. “Age Distribution by Marital Status”, “Income Distribution by Education Level”).
3. W podrozdziale ‘Learning data correlation and column engineering’ brak wniosków co do zależności kolumn w macierzy korelacji (np. pomiędzy zmiennymi finalized_loan i lenght_relationship _with_client czy debit_card i salary_account).
4. Ramka danych “x_train.csv” ma nazwy kolumn zawierające spacje.
5. Brak informacji o niezrównoważeniu zmiennej do przewidywania i brak informacji o korekcie tego w ramce danych do trenowania.
6. (Detale) „Count of Categorical Variables”: podwykresy są zbyt małe, przez co są nieczytelne. „Count of Categorical Variables”: “Most have finishes a university.” – błąd koniugacji. “Income Distribution by Education Level”: “We may be able to see that Uniwersity graduates[...]”, “# and map it as half of possible numbers or 2000 ti indicate highest one” literówki.

3 Inżynieria cech

W drugiej sekcji, zespół przeprowadził analizę zbioru danych pod kątem korelacji między zmiennymi oraz dokonał oceny ważności poszczególnych kolumn, w celu wyboru najbardziej informatywnych cech (takich jak LENGTH_RELATIONSHIP_WITH_CLIENT, AGE), które są kluczowe dla modelu. Jednakże zgłoszono zastrzeżenia dotyczące tego, dlaczego tylko te dwie cechy zostały wybrane, pomimo obecności innych równie istotnych zmiennych. Dodatkowo, pojawiła się uwaga odnośnie faktu, że tylko jedna zmienna została znormalizowana, bez wyjaśnienia, czy istnieje potrzeba normalizacji lub standaryzacji pozostałych zmiennych oraz jakie są tego powody.

Po dokładnym zapoznaniu się z zawartością plików odnoszących się do etapu inżynierii cech zastały znalezione elementy wymagające poprawy, które są wymienione poniżej.

Walidacja pliku **feature_engineering.ipynb**.

1. X_train.csv (wczytywanej w komórce 1.) nie ma kolumny CURRENT_ACCOUNT, która jest później wywoływana (w komórce 4 i 5), co powoduję wyrzucenie błędu.

2. Przed komórką 7. były wykorzystywane kolumny z ramek orig_###.csv, a po komórce 7. były wykorzystywane kolumny z ramki X_train.csv. Kod wymaga wczytania różnych ramek danych, jednak wczytywana jest tylko jedna ramka danych.

3. Jako cechy o dużym znaczeniu dla modelu zostały wybrane LENGTH_RELATIONSHIP_WITH _CLIENT i AGE, jednak były też obecne inne cechy o większym znaczeniu niż AGE. Nie zostało wyjaśnione dlaczego tylko te dwie cechy są wyróżnione.

4. W komórce 15 jest wczytywana ramka x_train.csv, a w komórce 16 jest zapisywana ramka X_train.csv. Takiego typu nazewnictwo bywa mylące. Ponaddto niektóre komendy są nieczułe na wielkość liter, np. pd.read/write_csv traktują te ramki jak tę samą.

5. Program traktuje x_train.csv i X_train.csv jako tę samą ramkę danych, więc błędem jest wczytywanie tej ramki danych w komórce 15, jej obróbka i zapis w komórce 16. Operacje wykonywane jednorazowo powinny być zakomentowane.

Walidacja pliku **modeling_validation.ipynb** .

1. W modeling_validation.ipynb nie zostały wczytane ramki danych (zmienne df i df_val są puste).

2. W komórce 3 brakuje importu RandomForestClassifier, w 4. StratifiedKFold, w 5. accuracy_score.

3. W komórce 8 jest omyłkowo wpisany x przy nawiasie.

4. W komórce 10. i 11. jest używana biblioteka xgboost, choć nie ma o niej informacj w pliku requirements.txt.

4 Modelowanie i walidacja modeli

Zapoznaliśmy się z modelami stworzonymi przez zespół pierwszy. Pracując na niezależnym zbiorze danych możemy potwierdzić większość wniosków wysnutych przez nich. To co najbardziej było podobne w wynikach to nazwane przez nich 'traditional models'. Metodą kroswalidacji wyznaczyliśmy na zbiorze testowym wartości takich miar jak f1 score, recall oraz roc auc oraz porównaliśmy je z wyznaczonymi przez grupę budującą wynikami. Różnice są bardzo małe, różnią się dopiero kilka miejsc po przecinku. Jedynym wyjątkiem jest model MLP Classifier, który osiągnął na zbiorze testowym gorsze wyniki:

Przy wybieraniu najlepszych parametrów dla modeli GaussianNB, XGB Classifier wyniki pokrywają się. Jednak na danych testowych model SVC osiąga f1 score tylko na poziomie 0.259270 (zespół budujący osiągnął wartość 0.843585). Model ten nie może zostać użyty do decydowania o zatwierdzaniu pożyczek.

Dla modeli w podrozdziale Deep Learning wyniki (recall, f1) otrzymane na zbiorze testowym są porównywalne lub nawet w niektórych przypadkach lepsze o około 1-2 procent.

Ostateczny wybrany model (Voting Classifier) osiągą na zbiorze testowym nieznacznie gorsze wyniki.

Podsumowując, wszystkie wyniki (oprócz jednego modelu) zostały potwiedzone.

Model	recall	f1	roc auc
MLPClassifier dane zespołu walidującego	0.50988408136529	0.5397029412081237	0.892760999300732
MLPClassifier dane zespołu budującego	0.5437999321566013	0.5611709666052485	0.8952207776711391

Tabela 1: Wyniki Dla Modelu MLP Classifier

	precision	recall	f1-score
0	0.93	0.93	0.93
1	0.64	0.64	0.64
accuracy			0.89
macro avg	0.78	0.79	0.79
weighted avg	0.89	0.89	0.89

Obraz 1: Classification Report Zespołu Budującego

	precision	recall	f1-score
0	0.93	0.92	0.92
1	0.60	0.65	0.62
accuracy			0.87
macro avg	0.77	0.78	0.77
weighted avg	0.88	0.87	0.87

Obraz 2: Classification Report Zespołu Walidującego