Raport walidacyjny

Etap EDA, raport numer 1

Korekty wymagają następujące elementy:

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.

Etap FEATURE ENGINEERING, raport numer 2

1. 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.
2. Brak standaryzacji zmiennych lub chociaż komentarz że robienie tego jest w tym przypadku niepotrzebne. Tylko niektóre zmienne zostały znormalizowane
3. modeling\_validation.ipynb - W komórce 4 brakuje importu StratifiedKFold, W komórce 8 jest omyłkowo wpisany x przy nawiasie. W komórce 10. i 11. jest używana biblioteka xgboost, choć nie ma o niej informacj w pliku requirements.txt.

Etap MODELING, raport numer 3

1. feature\_engineering\_pipeline – wszystkie operacje sekwencyjne, można by część z nich robić równolegle. Część z transformacji polega na hard-codowanych wartościach, nie wiadomo skąd zostały wzięte. Do deployu produkcyjnego brakuje zapisania pipeline z możliwością wczytania go z innego miejsca, aktualna struktura (funkcyjna) transformatorów nie pozwoli na łatwy zapis.
2. Dla przejzystosci, odnosi się to tak naprawdę do wszystkich notebookow, logikę pythona odpowiadającą za rysowanie wykresow itp. Wypadało by przezucic do plikow .py i jedynie importować. Notebook bylby prostszy w odszyfrowywaniu jakby był bardziej biznesowy
3. Brakuje jakiś wnioskow – np. dla Multinomial Logistic Regression widzimy ze tuning nic nie daje oraz ze model gorzej radzi sobie z klasą 1. To samo można powiedzieć o pozostałych modelach. Dodatkowo przestrzeń parametrów dla większości z nich jest przeszukiwania pobieżnie, jedynie dla paru hard-codowanych wartości. Fajnie to zoptymalizować i szukać w przestrzeni parametrow ciągłych według jakieś funkcji kosztu pozwalającej określić minimum (tj najlepszą wartość parametrów), może być to droższe obliczeniowo chodź niekoniecznie – nieobiecujące przypadki można wywalać zawczasu.
4. Podkreślenie punktu 3 – cały ModelTuning.ipynb jest strasznie powtarzalny i mało czytelny – brakuje zestawienia modeli ze sobą, wniosków z ich wyników, propozycji który model ostatecznie wybrać czy tłumaczenia dlaczego optymalizujemy akurat ten – optymalizacja parametrów jest decyzją z punktu widzenia biznesu często drogą, po co „forcem” przeszukiwac parametry wszystkich dostępnych sklearnowych modeli jak w przypadku niektórych z góry widać że nie będą one „optymalne”?
5. 3 i 4 tak samo dla Model Selection.ipynb + czemu jest podział na 2 pliki o podobnych nazwach? Domyślam się ze robily to inne osoby ale prowadzi to do niespójności
6. Jest parę fajnych analiz np. SHARP, ale cały step 3 projektu nie prowadzi do jego biznesowego zamkniecia – odnosi się wrazenie ze z calosci idziemy w wiele stron, żadnej strony nie komentujemy i nie faworyzujemy – poszukiwania bez celu
7. Nazewnicto plików nie spójne. Jako data-scientist walidujący pracę zespołu trzeba się domyslac co było po czym + brak finalnego produktu i prezentacji. Projekt robi wrazenie work-in-progress czy tez playgroundu