Raport z Projektu EDA

Oliwier Kolbusz

1. Wstęp

Celem projektu jest pozyskanie danych, dokonanie na nich inżynierii cech, eksploracyjnej analizy i użycia ich w nieskomplikowanym modelu w ujęciu ML.

2. Źródło Danych

Dane zostały pozyskane ze strony https://dane.gov.pl/. Udostępniane tam są zbiory z kategorii takich jak zdrowie, gospodarka, społeczeństwo, nauka. Dane można pobrać bezpośrednio ze strony lub ze stron pośredników. Dostępne są różne formaty: CSV, XML, XLSX, ale również dostęp poprzez REST_API czy SPARQL. Jakość danych można sprawdzić bardzo szybko w przypadku niektórych zbiorów - od razu wyświetlane są tabele, a czasami można nawet dokonać prostej wizualizacji. Dodatkowo znaleźć tam można opisy zbiorów danych, ich metadane oraz informacje o ich aktualizacjach.

3. Wybór Danych i Inżynieria Cech

Pozyskane dane pochodzą ze źródeł archiwalnych zawierających wyniki Narodowych Spisów powszechnych dla lat 1921-2011[1]. Zawierają one informacje na temat poziomu wykształcenia poszczególnych grup wiekowych. Składają się one na kilka arkuszy w programie Excel, podzielonych wg lat z ok. 10 rocznymi odstępami

Dane pochodzące sprzed 1960 roku informują tylko o poziomie analfabetyzmu, więc zostaną one pominięte. Arkusze począwszy od roku 1960 zawierają dużą ilość kolumn informujących o poziomie wykształcenia i wierszy mówiących o poszczególnych grupach wiekowych. Arkusze nie posiadają spójnej struktury, więc należy je odpowiednio wczytać i przetworzyć. Po wstępnym przetworzeniu danych (scalenie kolumn, zastąpienie brakujących wartości zerami, ekstrakcja badanych rekordów) otrzymano prosty DataFrame zawierający dane dotyczące wyłącznie osób w wieku 18-29 lat. Kolumna „Ogółem” informuje o ogólnej liczbie osób w tym przedziale wiekowym.

Rok	**Ogółem**	**Wyższe**	**Średnie**	**Zasadnicze Zawodowe**
1960	4999649	103839	335045	234443
1970	6211186	136538	1133427	1512418
1978	7932321	529323	2486996	2744772
1988	6330572	252397	2202566	2578006
2002	7426809	790533	3085389	1845865
2011	7117841	1568218	2854444	851388

Tab. 1. Ilość osób w wieku 18-29 lat w zależności od posiadanego wykształcenia

Dane pozyskano poprzez wybór odpowiednich kolumn z każdego arkusza, zastąpienie brakujących wartości zerami (można tak zrobić, ponieważ w przypadku tego zbioru oznaczają one brak osób z danym wykształceniem) oraz zsumowanie wartości wierszy odpowiadających badanym przedziałom wiekowym dla każdego roku.

Posiadając tak wyselekcjonowane dane dodano nowe kolumny, na których podstawie dokonano analizy danych. Poniższa tabela pokazuje nowe kolumny, które powstały poprzez podzielenie każdego z wierszy przez wartość kolumny „Ogółem" i przemnożenie przez 100 w celu pozyskania wartości procentowego udziału danego typu wykształcenia w populacji 18-29 lat. Pozwala ona na lepsze zobrazowanie liczb posiadając kontekst proporcji.

Rok	Wyższe_%	Średnie_%	Zasadnicze Zawodowe_%
1960	2.076926	6.701370	4.689189
1970	2.198260	18.248157	24.349907
1978	6.672990	31.352690	34.602382
1988	3.986954	34.792527	40.723113
2002	10.644316	41.543939	24.854079
2011	22.032215	40.102666	11.961324

Tab. 2. Procentowy udział osób w wieku 18-29 lat z danym wykształceniem

Dodatkowo korzystając z metody diff(), utworzono kolumny informujące o liczbowych zmianach w poszczególnych kolumnach względem poprzedzającego roku. Mają one na celu lepsze zobrazowanie postępujących zmian w strukturze poziomu wykształcenia.

Rok	Wyższe_delta	Średnie_delta	Zasadnicze Zawodowe_Delta
1960	0	0	0.0
1970	32699	798382	1277975
1978	392785	1353569	1232354
1988	-276926	-284430	-166766
2002	538136	882823	-732141
2011	777685	-230945	-994477

Tab. 3. Liczbowe różnice pomiędzy ilością osób z danym wykształceniem pomiędzy kolejnymi latami

4. Eksploracyjna Analiza Danych

Posiadając nowe kolumny, można stworzyć wykresy obrazujące konkretne zależności, co ułatwi wybór danych, które mają być przewidywane.

W pierwszej kolejności utworzono wykres procentowego poziomu wykształcenia w zależności od czasu. Wykres podzielono kolorystycznie ze względu na typ pozyskanego dyplomu.

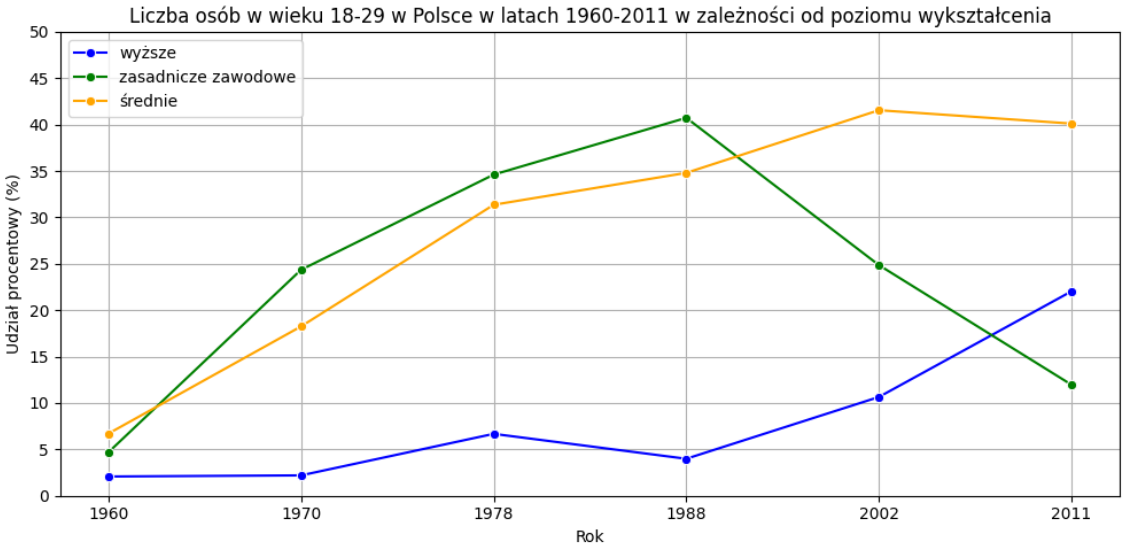

Fig. 1. Wykres zależności procentowej liczby osób w wieku 18-29 lat w zależności od poziomu wykształcenia

Na wykresie widać, że po transformacji nastąpiła zmiana trendów – znaczny spadek dla zasadniczego zawodowego i wzrost wykształcenia wyższego. Dane tego typu mogą zostać użyte do uczenia nadzorowanego przy pomocy regresji linowej/wielomianowej w celu predykcji przyszłych wartości. Można to zrobić ze względu na liczbowy charakter danych i łatwą do zobrazowania zależnością od innej cechy (czasu).

Kolejny wykres obrazuje dynamikę zmian wartości w czasie na podstawie danych z tabeli 3.

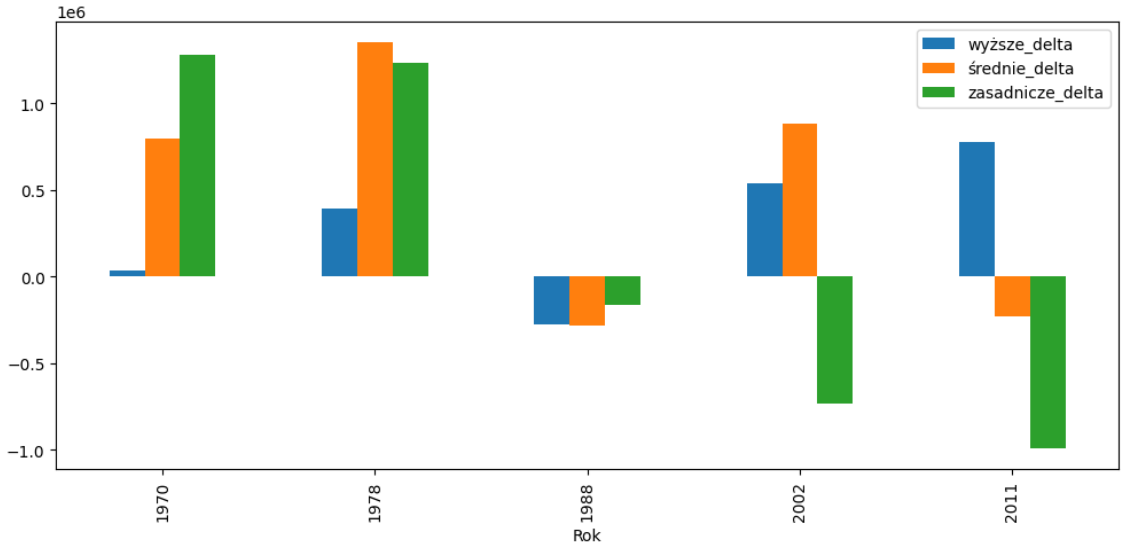

Fig. 2. Wykresy słupkowe obrazujące zmiany wartości dla poszczególnych poziomów wykształcenia względem poprzedniego roku.

Wykres dokładnie pokazuje, że od lat 1960 do 1978 panował znaczny wzrost osób z wykształceniem średnim i zasadniczym zawodowym. Trend zmienił się po transformacji, gdzie zyskało wykształcenie wyższe, zaś mocno straciło zasadnicze zawodowe. Pozwala to określić, jakich wyników można się spodziewać po modelowaniu.

5. MODEL REGRESJI

Po wizualizacji danych zdecydowano się na predykcję przyszłych wartości poziomu wykształcenia za pomocą regresji. Zastosowane zostały bezwzględne wartości danych – na wizualizacji wyglądają podobnie jak procentowe, a dają nieco lepsze wyniki. Jako zmienną TARGET użyto ilości osób z poszczególnym typem wykształcenia. Stworzono 3 osobne modele dla wszystkich uwzględnionych typów wykształcenia w celu predykcji każdego z nich. Za zmienną FEATURES posłużyła zmienna „rok", ponieważ badane są zmiany w zależności od czasu. Zastosowano 2 modele z różnymi stopniami wielomianu – jeden będący regresją liniową i drugi będący regresją wielomianową stopnia drugiego.

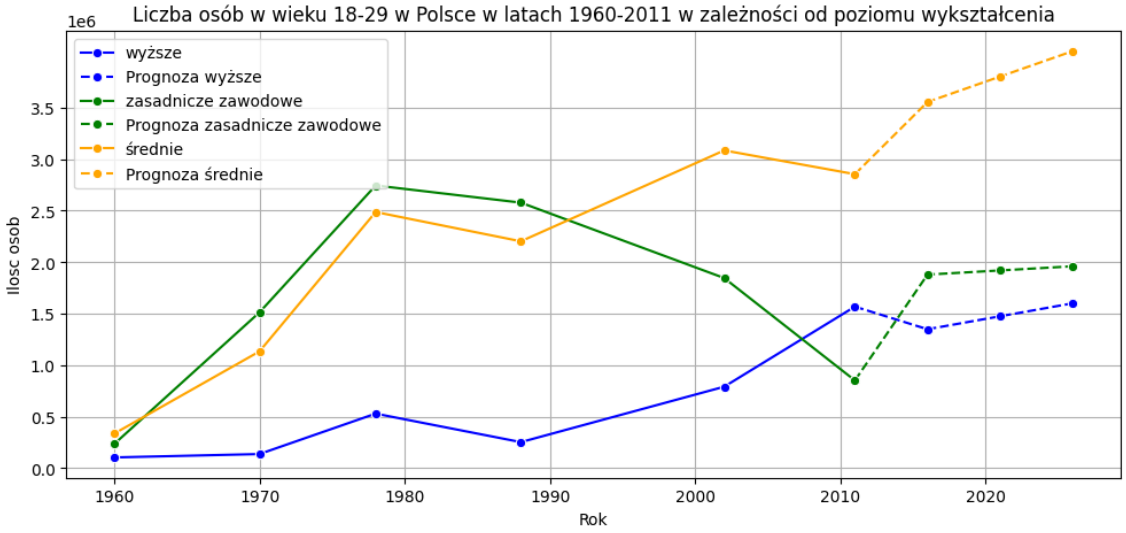

Fig. 3. Wykres zależności liczby osób w wieku 18-29 lat w zależności od poziomu wykształcenia z predykcją przy pomocy regresji liniowej

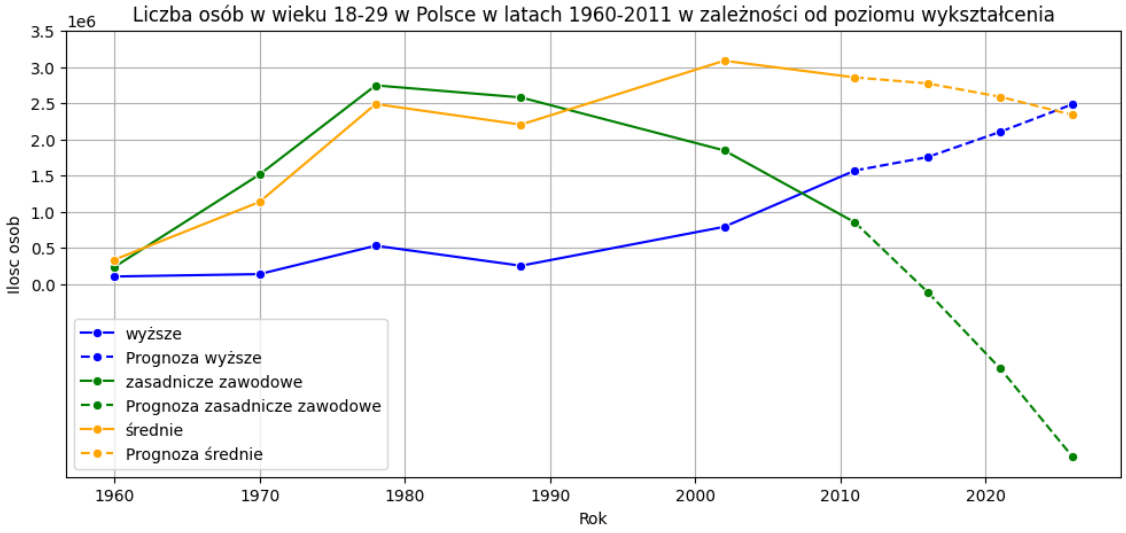

Fig. 4. Wykres zależności liczby osób w wieku 18-29 lat w zależności od poziomu wykształcenia z predykcją przy pomocy regresji wielomianowej drugiego stopnia

W obu przypadkach widać wadliwość tej prostej metody przewidywań. W przypadku regresji liniowej wszystkie przewidywane wartości zaczęły rosnąć, co nie zgrywa się z obserwowanym trendem spadku popularności wykształcenia zasadniczego zawodowego. Z kolei w przypadku wielomianu drugiego stopnia otrzymano predykcję, która nie może być spełnialna w rzeczywistości (-2391161 osób z wykształceniem zasadniczym zawodowym w 2026). Dopiero przy wielomianie 46 stopnia predykcja nie daje wartości ujemnych.

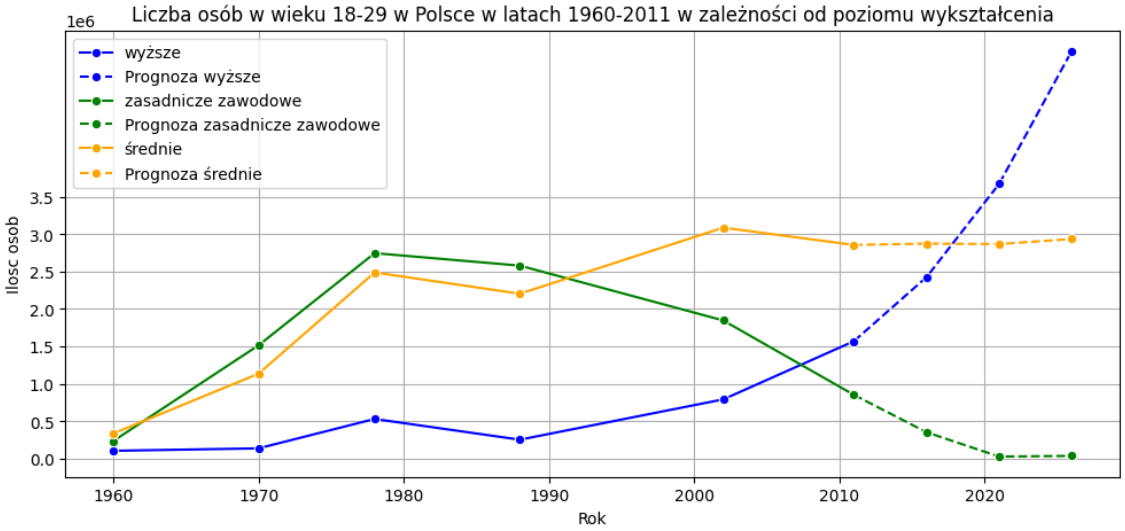

Fig. 5. Wykres zależności liczby osób w wieku 18-29 lat w zależności od poziomu wykształcenia z predykcją przy pomocy regresji wielomianowej 46-tego stopnia (26683 osób z wykształceniem zasadniczym zawodowym w 2021)

6. Wnioski

Projekt pokazał wadliwość prostej metody regresji w przypadku tego typu danych. Predykcje są błędne i niespełnialne w rzeczywistości. Możliwą przyczyną ze strony danych może być zbyt mała ilość lat potrzebnych do cech FEATURES, które pozwalają na przewidywanie. Alternatywą może być próba przyjęcia bardziej skomplikowanego modelu lub ograniczenie wyników do konkretnych liczb.

Źródła

[1] Portal dane.gov.pl, *Ludność według poziomu wykształcenia 1921–2011*, https://dane.gov.pl/pl/dataset/1572,ludnosc-wedug-poziomu-wyksztacenia-1921-2011, dostęp: 19.04.2025