Politechnika Wrocławska
Wydział Informatyki i Telekomunikacji


# Regresja danych dotyczących ilości miejsc pracy

Sprawozdanie z laboratorium

**Wojciech Gruba**
nr albumu: **259170** kierunek: **Informatyka Stosowana**

*07 czerwca 2022*

**Streszczenie**

Praca przedstawia program do obliczania regresji liczby zajętych miejsc pracy w wybranych rodzajach przemysłu oraz regionu w jakim dana firma sie znajduje. Program działa dzięki danym pobranym ze strony data.ny.gov. Dane te zostały oczyszczone ze zbędnych wierszy nieposiadających kompletnych informacji, do danych została dodana kolumna "PrevJobs" zawierająca informację o stanie miejsc pracy sprzed roku. Następnie ze zbioru danych zostały wylosowane wiersze mające służyć predykcji miejsc pracy na podstawie daty, rodzaju przemysłu oraz regionu. W programie zostały wykorzystane modele regresji liniowej, model customowy wtkorzystujący funkcję curve fit z modułu scipy oraz SVR z modułu sklearn. Następnie dzięki funkcjom mean squared error oraz mean absolute percentage error z modułu sklearn zostały obliczone błędy kwadratowe oraz procentowe modeli regresji. Na zakończenie pracy program wyświetla wykres z oryginalnymi wartościami oraz wartościami wyliczonymi dzięki modelom regresji.

# 1 Wstęp – sformułowanie problemu

Autor chce przewidzieć ilość miejsc pracy w poszczególnych rejonach oraz typach przemysłu. Pozwoli mu to na ocenę rozwoju poszczególnych typów przemysłu w przyszłych latach.

# 2 Opis danych

Wielkość datasetu 2079 wierszy. Kolumna "Year" - zmienna całoliczbowa, określa rok z którego dane zostały pobrane.
Zbiór wartości: 2012 - 2020

Kolumna "Region" - zmienna kategoryczna, określająca nazwę lokalizacji dla jakiej dane zostały przygotowane.
Zbiór wartości to Capital Region, Finger Lakes, Mid-Hudson, New York City, North Country, Southern Tier, Mohawk Valley, New York, Central New York

Kolumna "NAICS code" - zmienna całoliczbowa, Północno-amerykański system klasyfikacji przemysłu.
Zbiór wartości: 11-99

Kolumna "Industry" - zmienna kategoryczna, określająca nazwę przemysłu.
Zbiór wartości: Retail Trade, Wholesale Trade, Finance and Insurance, Arts, Accomodation and Food Services, Information, Agriculture, Mining, Other, Technical Services, Educational Services, Goverment, Transportation, Real Estate, Administrative and Support and Waste Managment, Utilites, Health Care, Construction, Manufacturing, Managment of Companies and Enterprises, Unclassifed.

Kolumna "Jobs" - zmienna całoliczbowa, określająca ilość wszystkich zatrudnionych ludzi w określonym rodzaju przemysłu.
Zbiór wartości: 27 -703,838

# 3 Opis rozwiązania

Dane dotyczące ilości pracowników zostały pobrane ze strony
`https://data.ny.gov/Economic-Development/Jobs-By-Industry-Beginning-2012/pxa9-czw8`.

Baza została zapisana w postaci ramki danych biblioteki `Pandas`. Zawiera ona informacje o 3 cechcach określających możliwości zatrudnienia w danych gałęziach przemysłu. Po dodaniu wiersza PrevJobs program wybiera wiersze do dalszej pracy. Kożystając z modeli linear model, SVR oraz własnego modelu regresji, na wybranych danych program wylicza przewidywaną liczbę zatrudnionych pracowników. Następnie dzięki bibliotece matplotlib, prawdziwe i wyliczone dane nanoszone są na wykres i wyświetlane uzytkowikowi.

# 4 Rezultaty obliczeń

## 4.1 Plan badań

Zbiór danych zostanie podzielony na dwie części: treningową i testową w stosunku 80:20.

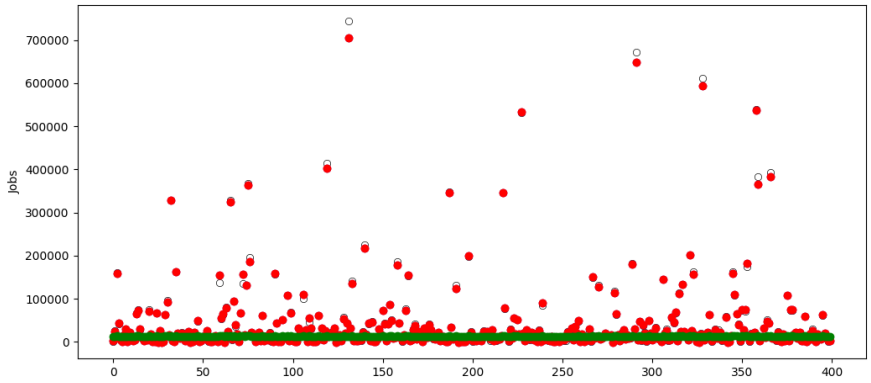
## 4.2 Wyniki obliczeń

Model wyliczania liczby miejsc pracy można przedstawić następującym wzorem:

$$\det Jobs = \alpha * Year + \beta * get\_dummies(Region) + \gamma * get\_dummies(Industry) + \delta * PrevJobs \quad (1)$$

gdzie $get\_dummies()$ to funkcja mapująca dane kategoryczne na reprezentację one-hot.

Na rys. 1 pokazany jest przykładowy wykres.




Rysunek 1: Przewidywane wartości dla modeli

Funkcje mean squared error oraz mean absolute percentage error z modułu sklearn wyliczają błędy kwadratowe oraz procentowe dla poszczególnych sesji, aby sprawidzć ich skuteczność

# 5 Wnioski

Przedstawiony program pozwala na dobranie optymalnego modelu regresji do przewidzenia ilości zatrudninych ludzi w danej gałęzi przemysłu. Po próbach z różnymi wielkościami danych skuteczność modelu SVR niezależnie od ilości danych wypada najgorzej, natomiast własny model oraz

model liniowy przewidują nieznacznie różniące się wartości. Zależnie od liczby wierszy wziętych do stworzenia modelu, model linowy oraz custom model zazwyczaj osiągają błąd procentowy rzędu 0,5 natomiast SVR rzędu kilku procent. Wszystkie modele dla danych bez dodatkowej kolumny PrevJobs osiągały błąd procentowy rzędu 20 procent niezależnie od typu modeu. Model własny potrzebuje danych posiadających kilkaset wierszy aby optymalnie działał, ponieważ liczba kolumn musi być mniejsza niż liczba wierszy, z tego powodu gdy posiadamy dane mające 40 wierszy, model customowy sie nawet nie wykona, a dla wartości rzędu 70 wierszy błąd procentowy będzie wyższy niż dla dużej ilości danych.

# A Dodatek

Kody źródłowe(utrzymane w konwencji języka Python) umieszczone zostały w repozytorium github: `https://github.com/wgruba/MSID`.