**Projekt zaliczeniowy „Income classification” przygotowany na przedmiot Podstawy Uczenia Maszynowego (PUM)**

Autorzy:

Dominik Suszek s23396, gr. 17c

Artur Soszyński s23632, gr. 12c

Spis treści

[Opis problemu 1](#_Toc157331650)

[Dane 1](#_Toc157331651)

[Sposób rozwiązania problemu 5](#_Toc157331652)

[Dyskusja wyników i ewaluacja modelu 5](#_Toc157331653)

[Podsumowanie 7](#_Toc157331654)

[Załączniki 8](#_Toc157331655)

# Opis problemu

Poniższy raport został opracowany na podstawie zbioru danych dotyczących dochodów. Hipotetycznie, taki model mógłby zostać wykorzystany w trakcie tworzenia kampanii marketingowej, aby na podstawie danych demograficznych optymalnie przygotować grupy docelowe. Dzięki temu, poprzez zoptymalizowanie kosztów oraz zwiększenie skuteczności, koszty kampanii marketingowej uległyby zmniejszeniu.

# Dane

Zbiór danych zawierający informacje o dochodach osób pochodzących z różnych grup społecznych pochodzi ze strony <https://www.kaggle.com/datasets/lodetomasi1995/income-classification>.

Zbiór zawiera następujące zmienne kategoryczne:

* **workclass** – klasa zatrudnienia– czy osoba jest zatrudniona na etacie w sektorze prywatnym, rządowym, czy prowadzi własną działalność, etc. Jest to zmienna porządkowa,
* **education** – poziom wykształcenia osoby, to również jest zmienna porządkowa,
* **marital-status** – stan cywilny – zmienna nominalna,
* **occupation** – grupa zawodowa – zmienna porządkowa,
* **relationship** – relacje osoby w kontekście rodziny – zmienna nominalna,
* **race** – zmienna nominalna, która określa rasę danej osoby,
* **sex** – zmienna nominalna, która określa płeć danej osoby,
* **native-country** – kraj pochodzenia osoby – zmienna nominalna.

W zbiorze danych znajdują się również następujące zmienne numeryczne:

* **age** – wiek danej osoby, zmienna ciągła,
* **fnlwgt** – waga przypisana do danej grupy społecznej. Innymi słowy, jest to liczba osób, które według Urzędu Statystycznego reprezentuje daną grupę. *fnlwgt* jest zmienną ciągłą,
* **education-num** – zmienna ciągła, określająca liczbę lat edukacji,
* **capital-gain** – zmienna ciągła, określająca zyski kapitałowe,
* **capital-loss** – zmienna ciągła, określająca straty kapitałowe,
* **hours-per-week** – liczba godzin, które dana osoba przepracowała w przeciągu tygodnia. Jest to zmienna ciągła.

Przed rozpoczęciem procesu trenowania modeli, przygotowano dalsze etapy eksploracyjnej analizy danych, która pozwoliły lepiej zrozumieć poszczególne zmienne, a także możliwe zależności występujące pomiędzy nimi.

W poniższej tabeli przedstawiono statystyki opisowe zmiennych numerycznych:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **age** | **fnlwgt** | **education-num** | **capital-gain** | **capital-loss** | **hours-per-week** |
| **count** | 17923 | 17923 | 17923 | 17922 | 17922 | 17922 |
| **mean** | 38.6 | 190185.4 | 10.1 | 1050.7 | 87.8 | 40.4 |
| **std** | 13.6 | 105443.4 | 2.6 | 7281.1 | 402.5 | 12.3 |
| **V** | 35% | 55% | 26% | 693% | 458% | 30% |
| **min** | 17.0 | 12285.0 | 1.0 | 0.0 | 0.0 | 1.0 |
| **25%** | 28.0 | 118694.5 | 9.0 | 0.0 | 0.0 | 40.0 |
| **50%** | 37.0 | 178915.0 | 10.0 | 0.0 | 0.0 | 40.0 |
| **75%** | 47.5 | 237849.0 | 12.0 | 0.0 | 0.0 | 45.0 |
| **max** | 90.0 | 1484705.0 | 16.0 | 99999.0 | 4356.0 | 99.0 |

W przypadku tych zmiennych nie ma problemu z brakującymi wartościami – dla *capital-gain, capital-loss* oraz *hours-per-week­* znaleziono jedynie po jednym pustym wierszu. Najwięcej wartości relatywnie odległych od pozostałych elementów próby odnaleziono dla dwóch zmiennych – *capital-gain* oraz *capital-loss.* Współczynnik zmienności V, wyznaczony poniższym wzorem:

przyjmuje dla nich wysokie wartości (kilkaset %).

W celu zwizualizowania rozrzutu zmiennych losowych, przygotowano wykresy pudełkowe ([Załącznik 1](#_Załącznik_1._Wykresy)).

Następnie, w celu zbadania zależności występujących pomiędzy zmiennymi, przygotowano macierz korelacji ([Załącznik 2](#_Załącznik_2._Macierz)). Wprowadzono nową zmienną *income\_adj*, która powstała po zmodyfikowaniu *income.* Wartości „<=50K” zmieniono na wartości liczbowe 0, a „>50K” na wartości 1. Wyniki są zgodne z oczekiwaniami – wystąpiła słaba dodatnia korelacja pomiędzy poziomem dochodu, a:

* liczbą godzin przepracowanych w tygodniu przez daną osobę (0.23),
* dodatkowymi środkami zarobionymi na rynku kapitałowym (0.21),
* wiekiem (0.23).

Powyższe zależności łatwo wytłumaczyć – im więcej godzin pracujemy, tym wyższe będzie nasze wynagrodzenie (przy pozostałych warunkach niezmienionych). Starsze osoby zazwyczaj zajmują wyższe, lepiej opłacane stanowiska.

Nie odnotowano natomiast żadnych istotnych poziomów korelacji. Oznacza to, że potrzebna będzie inżynieria cech w celu odnalezienia ukrytych zależności – na przykład poprzez połączenie dwóch cech w jedną, albo zamianę zmiennych kategorycznych na numeryczne.

Następnie zbadano zależność występującą między poziomem dochodów a płcią ([Załącznik 3](#_Załącznik_2._Wykres)). Dysproporcje pomiędzy udziałami kobiet i mężczyzn w poszczególnych grupach dochodowych są ewidentne. W grupie zarabiającej poniżej 50 tysięcy udział kobiet wynosi 38,8%, natomiast mężczyzn 61,2%. Natomiast w drugiej grupie, zarabiającej powyżej 50 tysięcy, udział mężczyzn jest zdecydowanie wyższy i wynosi 85%. Można z tego wysnuć wniosek, że płeć danej osoby ma znaczący wpływ na to, czy jej zarobki przekroczą wcześniej zdefiniowany próg.

W kolejnym kroku dodano dodatkowe zmienną *additional\_money*. Jest to różnica między zmiennymi *capital gain*, oraz *capital loss*. Informuje o tym, czy dana osoba odniosła zyski, czy straty kapitałowe. Przygotowano wykres pudełkowy dla tej zmiennej ([Załącznik 4](#_Załącznik_4._Wykres_1)). Na jego podstawie zdecydowano, że wartości przekraczające 40000 należy uznać za wartości odstające. Przedstawiono je na kolejnym wykresie ([Załącznik 5](#_Załącznik_4._Wykres)). Jest ich jedynie 161, co stanowi niecałe 0,5% wszystkich dostępnych rekordów. W związku z tym, podjęto decyzję o przycięciu tych wartości – w przypadku przekroczenia wartości 40000 dla tej zmiennej, przypisywano jej wartość 40000. Dla porównania załączono wykres pudełkowy dla zmiennej po wprowadzeniu modyfikacji ([Załącznik 6](#_Załącznik_6._Wykres_1)).

Jako ostatni krok procesu inżynierii cech, zmodyfikowano zmienną *education.* Zmniejszono liczbę etykiet poprzez zastosowanie następującego mapowania:

* wartości “10th", "11th", "12th", "1st-4th", "5th-6th", "7th-8th", "9th", oraz "Preschool" zastąpiono jedną etykietą “Primary”,
* wartości “Bachellors” oraz “Some-college” zastąpiono “Bachelors”,
* “Assoc-acdm”, oraz “Assoc-voc” zostały zmienione na „Associate”.

Procentowe udziały wartości dla zmiennej *education* przedstawiono na wykresie ([Załącznik 7](#_Załącznik_7._Wykres)). Największe odsetki przypadają na klasy Bachelors oraz HS-Grad (odpowiednio osoby, które ukończyły pierwszy stopień studiów oraz osoby po szkole średniej).

Ostatecznie, do procesu trenowania modelu, wybrano następujące zmienne:

*age, workclass, education, race, sex, hours-per-week, native-country, additional\_money, income, fnlwgt.*

Zostały one podzielone na dwa podzbiory: zmiennych kategorycznych i liczbowych. Do pierwszego z nich zaliczamy zmienne: *workclass, race, sex, native-country, education.* Do drugiego wszystkie pozostałe, tj.: *age, hours-per-week, additional\_money, fnlwgt.* Dla zmiennych kategorycznych zastosowano OneHotEncoder.

Następnie zmienne liczbowe zostały przeskalowane. Wzięto pod uwagę trzy różne sposoby skalowania danych, tj. StandardScaler, MinMaxScaler, oraz RobustScaler. Wszystkie z nich są dostępne w pakiecie scikit-learn.

# Sposób rozwiązania problemu

Do wytrenowania modelu brano pod uwagę trzy różne algorytmy – Random Forest Classifier, SGD (Stochastic Gradient Descent) Classifier oraz regresję logistyczną. Na każdym z nich wypróbowano różne hiperparametry, aż do uzyskania najlepszego wyniku. Ostatecznie wybrano algorytm losowych lasów, który polega na uczeniu wielu drzew decyzyjnych skonstruowanych za pomocą różnych podzbiorów cech. Następnie otrzymane prognozy są uśredniane. Dzięki temu można uzyskać jeszcze lepszą wydajność. Innymi zaletami lasu losowego są:

* odporność na przeuczenie,
* możliwość dokładniejszego odtworzenia zależności pomiędzy zmiennymi, niż byłyby to w stanie zrobić drzewa decyzyjne,
* odporność na różnorodne problemy związane z danymi.

Pierwszym etapem realizacji projektu było zapoznanie się z danymi i próba zrozumienia zależności występujących pomiędzy zmiennymi. Do tego celu wykorzystano wykres przedstawiający macierz korelacji, statystyki opisowe zmiennych, a także pozostałe wykresy przedstawiające związki pomiędzy nimi. Ostatecznie wybrano następujące zmienne: *age, workclass, education, race, sex, hours-per-week, native-country, additional\_money, income, fnlwgt.*

Poniżej opisano cały proces dobierania modeli oraz strojenia hiperparametrów.

Parametry wykorzystane dla modelu rand\_for\_1:

{'bootstrap': False,

'ccp\_alpha': 0.0,

'class\_weight': None,

'criterion': 'gini',

'max\_depth': None,

'max\_features': 'auto',

'max\_leaf\_nodes': 100,

'max\_samples': None,

'min\_impurity\_decrease': 0.0,

'min\_samples\_leaf': 5,

'min\_samples\_split': 3,

'min\_weight\_fraction\_leaf': 0.0,

'n\_estimators': 100,

'n\_jobs': None,

'oob\_score': False,

'random\_state': 42,

'verbose': 0,

'warm\_start': False}

Dla powyższych wartości uzyskano przedstawione poniżej wyniki.

Macierz pomyłek:

|  |  |
| --- | --- |
| **TP**  6013 | **FP**  201 |
| **FN**  1095 | **TN**  832 |

Pozostałe miary pozwalające na ocenę klasyfikatora wyglądają następująco:

* precyzja (precision): 0.846 – odsetek przykładów zaprognozowanych pozytywnie, które rzeczywiście są pozytywne,
* czułość (recall): 0.968 – prawdopodobieństwo, że klasyfikacja będzie poprawna pod warunkiem, że przypadek jest pozytywny,
* wskaźnik F1: 0.903 – średnia harmoniczna z obu powyższych wartości - precyzji i czułości.

Parametry wykorzystane dla modelu rand\_for\_2:

{'bootstrap': True,

'ccp\_alpha': 0.0,

'class\_weight': None,

'criterion': 'gini',

'max\_depth': None,

'max\_features': 'sqrt',

'max\_leaf\_nodes': 160,

'max\_samples': None,

'min\_impurity\_decrease': 0.0,

'min\_samples\_leaf': 1,

'min\_samples\_split': 2,

'min\_weight\_fraction\_leaf': 0.0,

'n\_estimators': 1000,

'n\_jobs': None,

'oob\_score': False,

'random\_state': 42,

'verbose': 0,

'warm\_start': False}

Macierz pomyłek dla modelu rand\_for\_2:

|  |  |
| --- | --- |
| **TP**  6009 | **FP**  205 |
| **FN**  1089 | **TN**  838 |

Pozostałe miary pozwalające na ocenę klasyfikatora wyglądają następująco:

* precyzja (precision): 0.847 – odsetek przykładów zaprognozowanych pozytywnie, które rzeczywiście są pozytywne,
* czułość (recall): 0.967 – prawdopodobieństwo, że klasyfikacja będzie poprawna pod warunkiem, że przypadek jest pozytywny,
* wskaźnik F1: 0.903 – średnia harmoniczna z obu powyższych wartości - precyzji i czułości.

# Dyskusja wyników i ewaluacja modelu

Jako główny punkt odniesienia do oceny jakości modelu wykorzystano poziom trafności (accuracy) – założono, że docelowo powinien on przekroczyć próg 80%.

Na poniższym wykresie przedstawiono istotność w generowaniu prognoz dla poszczególnych atrybutów:

A graph with different colored bars

Description automatically generated with medium confidence

A graph of a bar graph

Description automatically generatedA także szczegółowo dla 10 najważniejszych atrybutów:

Na poniższym wykresie przedstawiono krzywą ROC, która jest graficznym narzędziem do oceny wydajności klasyfikatora binarnego. Przedstawia relację pomiędzy czułością a specyficznością. Im bliżej lewego górnego rogu, tym lepsza jest wydajność klasyfikatora. AUC-ROC (powierzchnia pod krzywą ROC) powinna być jak największa, ponieważ 1 oznacza klasyfikator idealny, a wartość 0.5 oznacza klasyfikator losowy.

A graph of a curve

Description automatically generated

# Podsumowanie

Udało się osiągnąć zdecydowanie wyższy poziom trafności (ponad 84,2%) względem zakładanego (80%).

Najbardziej czasochłonnymi zadaniami było zapoznanie się z danymi oraz inżynieria cech. Koniecznym okazało się dodanie dodatkowego atrybutu, uzupełnienie brakujących wartości, uproszczenie etykiet dla zmiennej kategorycznej *education*, oraz normalizacja danych numerycznych.

Inspiracją do prac nad modelem oraz niniejszym raportem był podręcznik „Uczenie maszynowe z użyciem Scikit-Learn, Keras i TensorFlow. Wydanie II” autorstwa Aurélien Géron.

# Załączniki

A graph with numbers and a bar

Description automatically generated

## Załącznik 1. Wykresy pudełkowe dla zmiennych liczbowych

## A blue and red squares with white text Description automatically generatedZałącznik 2. Macierz korelacji

A graph with blue and orange bars

Description automatically generated

## Załącznik 3. Wykres słupkowy przedstawiający zależność pomiędzy płcią, a poziomem dochodu.

A screen shot of a screen

Description automatically generated

## Załącznik 4. Wykres pudełkowy przygotowany dla zmiennej *additional\_money.*

A screen shot of a computer screen

Description automatically generated

## Załącznik 5. Wykres punktowy przedstawiający wartości odstające dla zmiennej *additional\_money.*

## A black and white screen with a black border Description automatically generated

## Załącznik 6. Wykres pudełkowy przygotowany dla zmiennej *additional\_money* po modyfikacji.

A pie chart with numbers and a black background

Description automatically generated

## Załącznik 7. Wykres kołowy dla zmiennej *education.*