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

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 zasięgów 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:

$$V = \frac{s}{\overline{x}} * 100\%, \overline{x} \neq 0$$

przyjmuje dla nich wysokie wartości (kilkaset %).

W celu zwizualizowania rozrzutu zmiennych losowych, przygotowano wykresy pudełkowe (Załącznik 1).

Następnie, w celu zbadania zależności występujących pomiędzy zmiennymi, przygotowano macierz korelacji (Załącznik 2). 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). Dodatkowe środki zarobione na rynku kapitałowym to obok zwiększenia swojej wartości na rynku pracy, jeden z głównych sposobów na wzrost dochodów. Ostatnią zależność można wytłumaczyć w ten sposób, że starsze osoby zazwyczaj zajmują wyższe, lepiej opłacane stanowiska.

Nie odnotowano natomiast żadnych istotnych poziomów korelacji (przekraczających 0.50). 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.

Na początku zbadano zależność występującą między poziomem dochodów a płcią (Załącznik 3). 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 związku z tym, zmienna *sex* powinna być uwzględniona w zbiorze treningowym i testowym.

W kolejnym kroku dodano dodatkową 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). Na jego podstawie zdecydowano, że wartości przekraczające 40000 należy uznać za wartości odstające. Przedstawiono je na kolejnym wykresie (Załącznik 5). 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).

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). Największe odsetki przypadają na klasy Bachelors oraz HS-Grad (odpowiednio osoby, które ukończyły pierwszy stopień studiów oraz osoby po szkole średniej). Zbadano również, jak dużo osób z każdego poziomu wykształcenia należy do danej grupy dochodowej (Załącznik 8).

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. Pozwala interpretować zmienne kategoryczne poprzez przekształcenie ich na formę numeryczną. Nie wprowadza porządku między wartościami, co pozwala uniknąć błędnej interpretacji przez modele.

Następnie zmienne liczbowe zostały przeskalowane. Wzięto pod uwagę trzy różne sposoby skalowania danych, tj. StandardScaler, MinMaxScaler, oraz RobustScaler. Ostatecznie wybrano Robust Scaler.

Sposób rozwiązania problemu

Do wytrenowania modelu brano pod uwagę trzy różne algorytmy – Random Forest Classifier, SGD (Stochastic Gradient Descent) Classifier oraz regresję logistyczną. Wypróbowano różne hiperparametry, aż do uzyskania najlepszego wyniku. Poniżej przedstawiono krótki opis każdego z algorytmów.

Lasy losowe polegają 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.

Regresja logistyczna jest jedną z metod regresji, która jest używana, gdy zmienna zależna jest na skali dychotomicznej. Służy do szacowania prawdopodobieństwa przynależności przykładu do określonej klasy. Jeśli wyznaczone prawdopodobieństwa przekracza próg 50%, to model prognozuje, że próbka należy do klasy pozytywnej (np. pacjent jest chory). W przeciwnym razie, stwierdza się, że przykład nie stanowi części określonej klasy (pacjent jest zdrowy). Główną zaletą tej metody jest fakt, że wyniki są łatwo interpretowalne. Ponadto, metoda nie jest „czarną skrzynką".

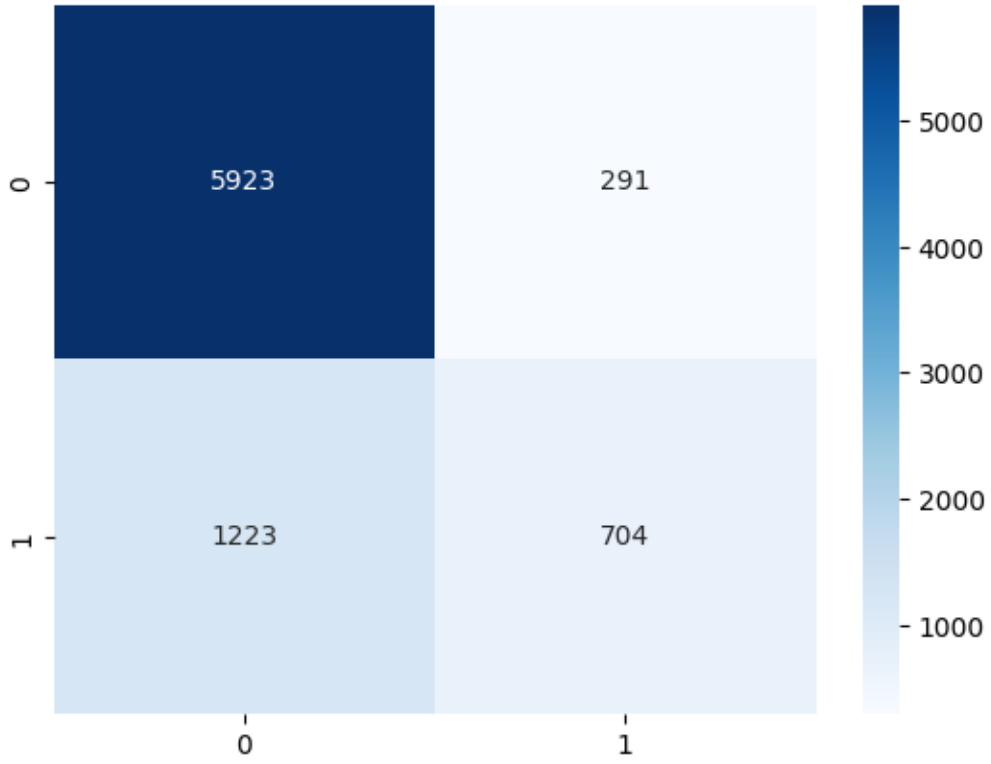
Algorytm stochastycznego spadku wzdłuż gradientu jest szybszym odpowiednikiem wsadowego algorytmu spadku wzdłuż gradientu. Oznacza to, że w każdej iteracji dobierana jest losowa próbka ucząca, za pomocą której są wyliczane gradienty. Z drugiej strony, algorytm ten jest bardziej chaotyczny, ponieważ istnieje większa szansa, że „przeskoczy" ekstremum lokalne.

Jednak zanim wybrano konkretny algorytm, wypróbowano każdy z nich i porównano wyniki. Poniżej opisano cały proces dobierania modeli oraz strojenia hiperparametrów.

Jako pierwszy wypróbowano jeden z prostszych algorytmów – regresję logistyczną. Wykorzystano następującą listę parametrów:

{'C': 100,
 'class_weight': None,
 'dual': False,
 'fit_intercept': True,
 'intercept_scaling': 1,
 'l1_ratio': None,
 'max_iter': 100,
 'multi_class': 'auto',
 'n_jobs': None,
 'penalty': 'l2',
 'random_state': 42,
 'solver': 'lbfgs',
 'tol': 0.0001,
 'verbose': 0,
 'warm_start': False}

Dzięki nim uzyskano poniższe rezultaty:

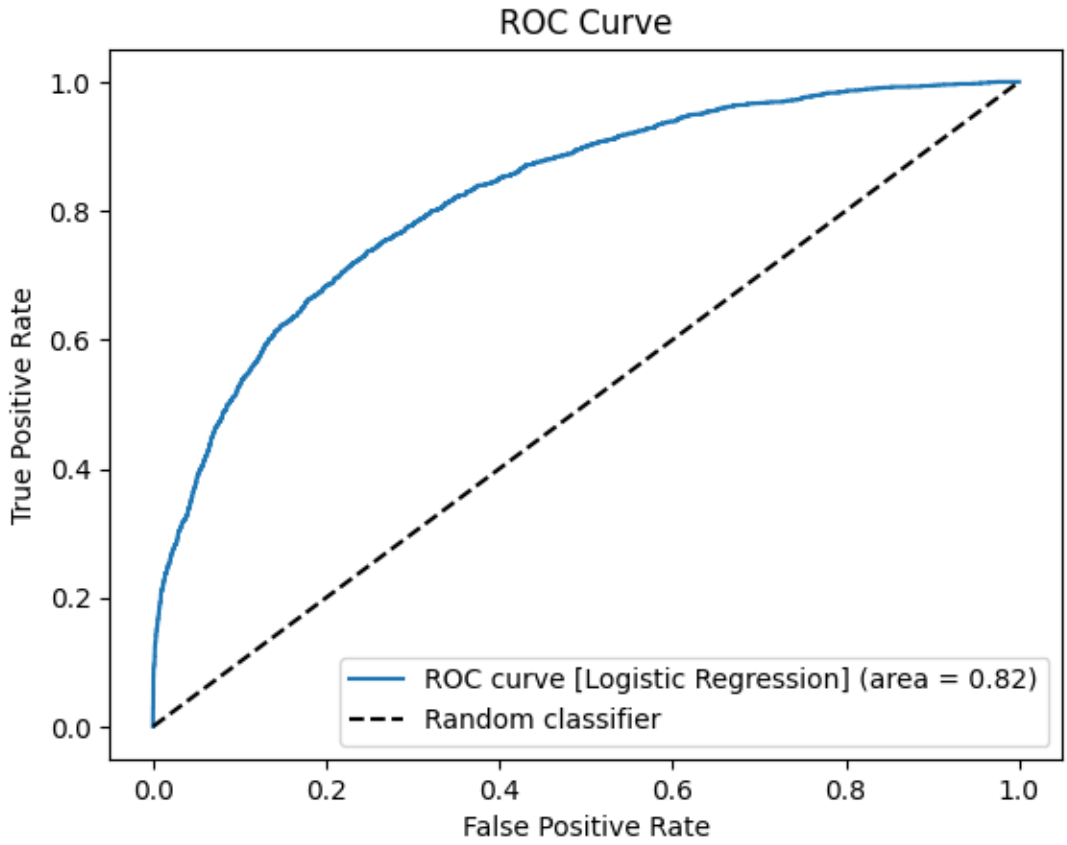


Pozostałe miary pozwalające na ocenę klasyfikatora wyglądają następująco:

- precyzja (precision): 0.829 – odsetek przykładów zaprognozowanych pozytywnie, które rzeczywiście są pozytywne,

- czułość (recall): 0.953 – prawdopodobieństwo, że klasyfikacja będzie poprawna pod warunkiem, że przypadek jest pozytywny,
- wskaźnik F1: 0.887 – średnia harmoniczna z obu powyższych wartości - precyzji i czułości.

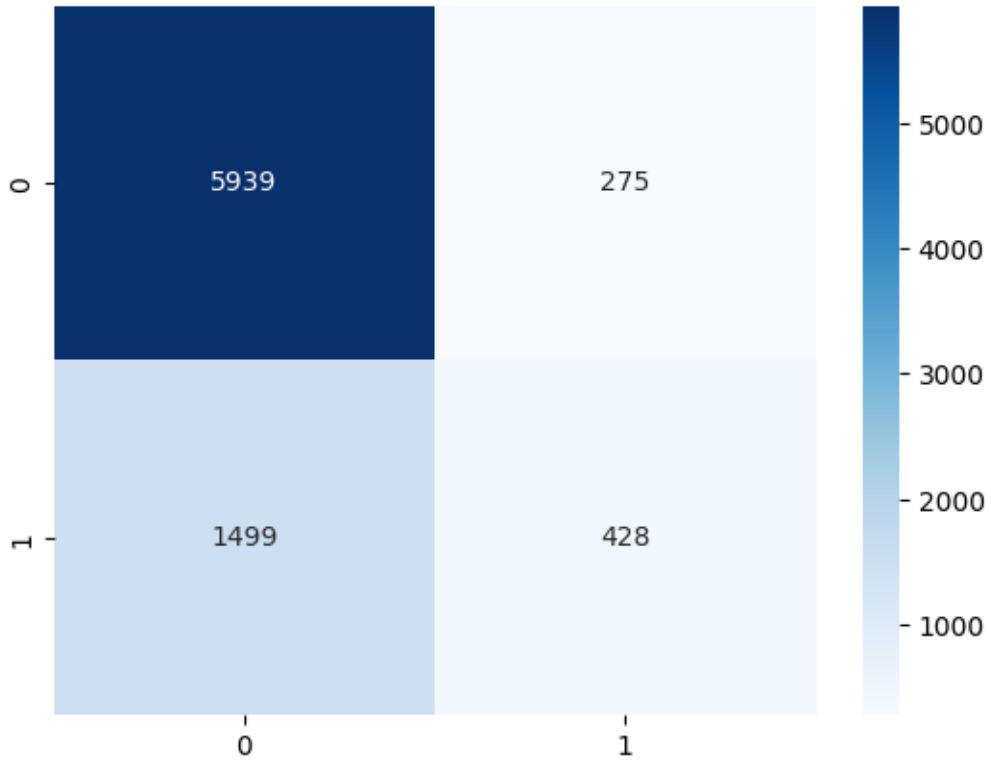
Na poniższym wykresie przedstawiono krzywą ROC, która jest graficznym narzędziem do oceny wydajności klasyfikatora binarnego. Przedstawia relację pomiędzy czułością a specyficznością. Im bliżej lewego górnego rogu, tym lepsza jest wydajność klasyfikatora. AUC-ROC (powierzchnia pod krzywą ROC) powinna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

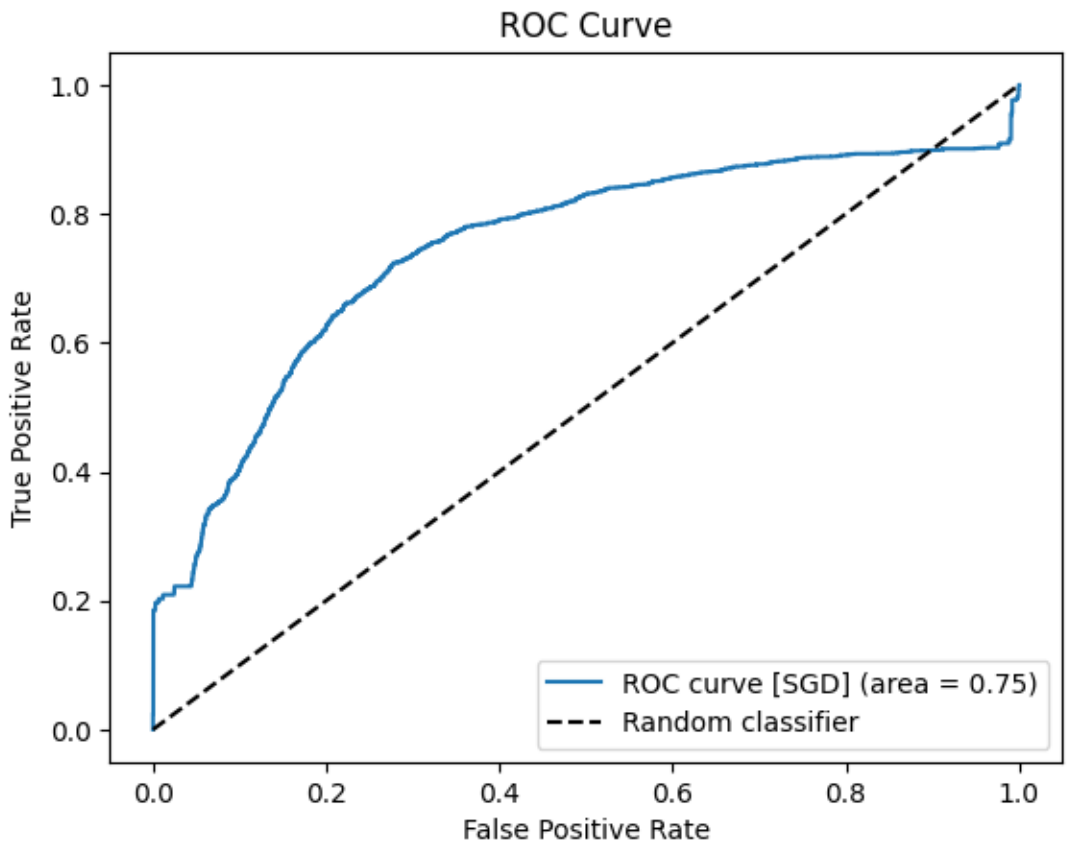


W kolejnym kroku wykorzystano algorytm stochastycznego spadku wzdłuż gradientu. Dla następujących argumentów:

{'alpha': 0.01,
 'average': False,
 'class_weight': None,
 'early_stopping': False,
 'epsilon': 0.1,
 'eta0': 0.0,
 'fit_intercept': True,
 'l1_ratio': 0.15,
 'learning_rate': 'optimal',
 'loss': 'hinge',
 'max_iter': 1000,
 'n_iter_no_change': 5,
 'n_jobs': None,
 'penalty': 'l2',
 'power_t': 0.5,
 'random_state': 42,
 'shuffle': True,
 'tol': 0.001,
 'validation_fraction': 0.1,
 'verbose': 0,
 'warm_start': False}

Uzyskano widoczne niżej wyniki:





Pozostałe metryki:

- precyzja (precision): 0.799,
- czułość (recall): 0.956,
- wskaźnik F1: 0.870.

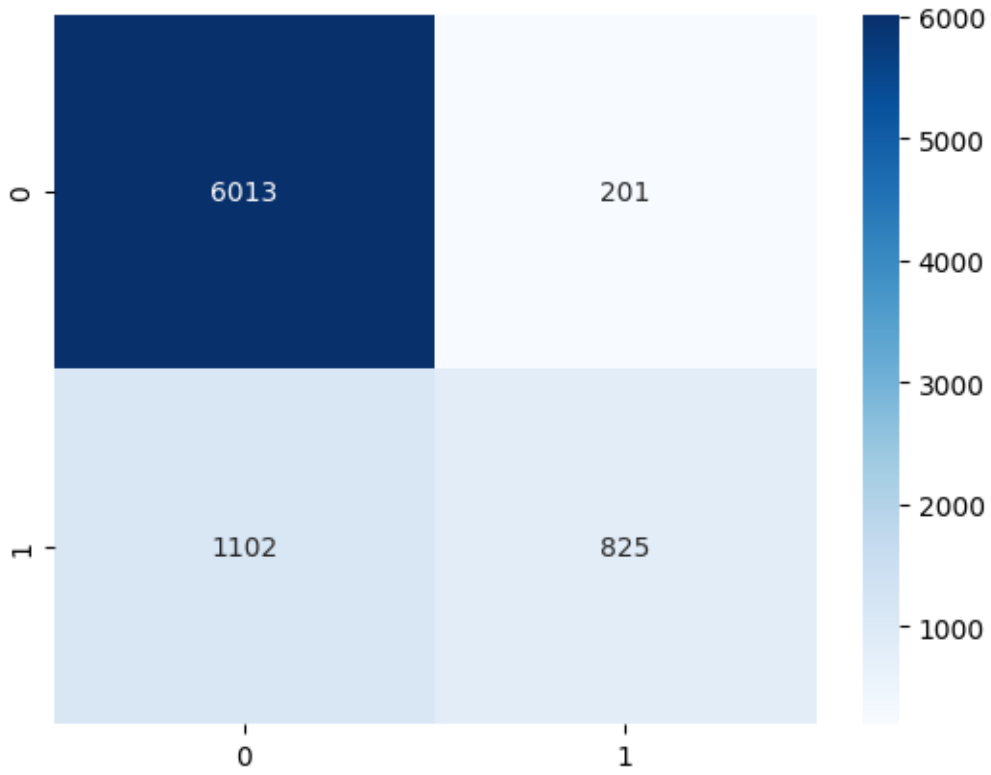
Parametry wykorzystane dla pierwszego z modeli opartych na algorytmie Random Forest Classifier:

{'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:



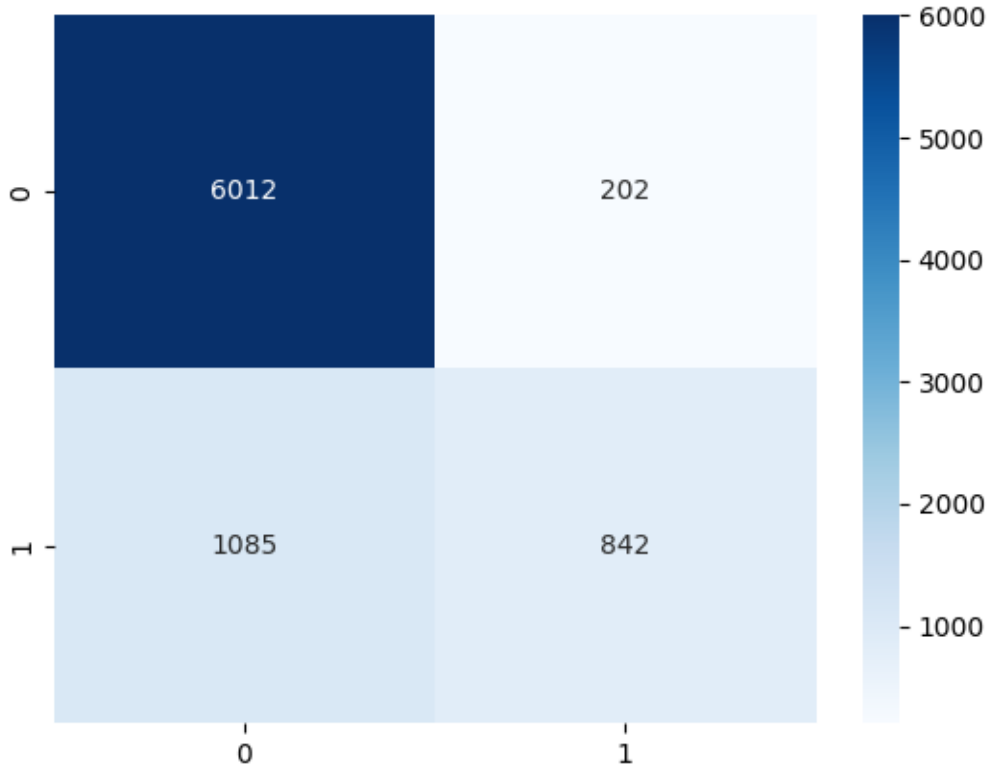
Pozostałe miary:

- precyzja (precision): 0.845,
- czułość (recall): 0.968,
- wskaźnik F1: 0.902.

Parametry wykorzystane dla drugiego z modeli opartych na algorytmie Random Forest Classifier:

{'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:

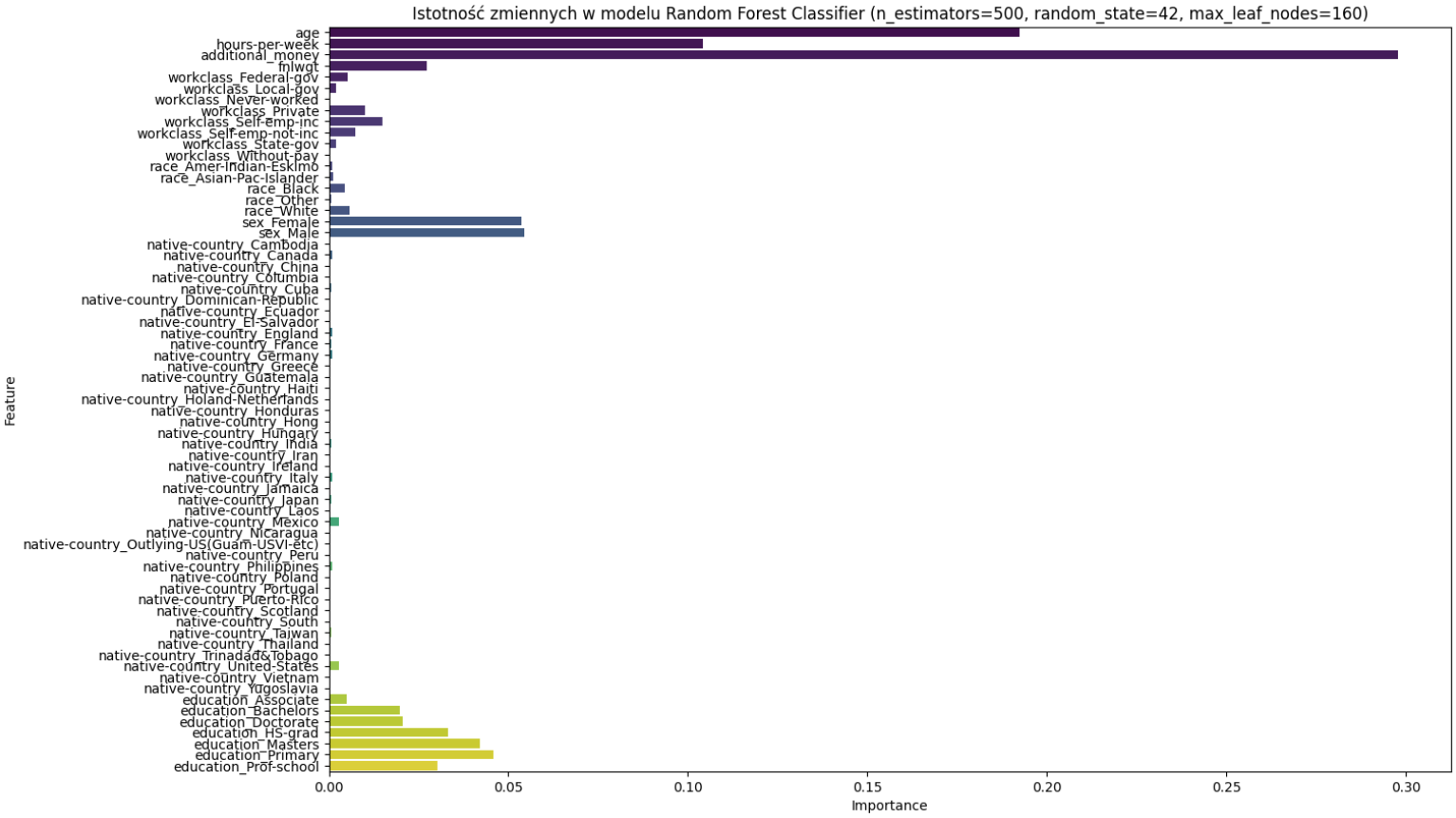


Pozostałe miary pozwalające na ocenę klasyfikatora wyglądają następująco:

- precyzja (precision): 0.847,
- czułość (recall): 0.967,
- wskaźnik F1: 0.903.

Drugi z modeli opartych na algorytmie lasów losowych jest lepszy pod względem każdej z metryk od poprzednio opisanych algorytmów (regresja logistyczna, SGD). Jest także prostszy od pierwszego z modeli opartych na lasach losowych. Został zatem wybrany jako najlepszy i ostateczny.

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



A także szczegółowo dla 10 najważniejszych atrybutów:

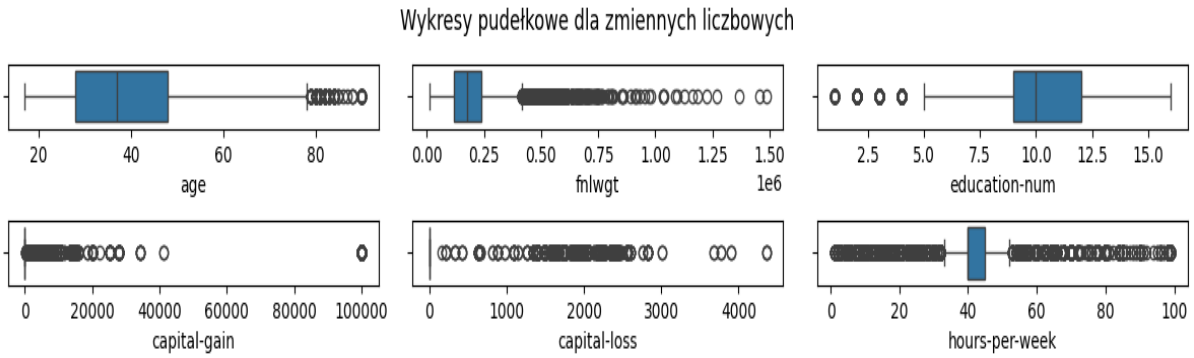
Podsumowanie

Udało się osiągnąć zdecydowanie wyższy poziom trafności (ponad 84,2%) względem zakładanego (80%). Model działa szybko na dostarczonych danych, więc nie pojawił się problem z wydajnością.

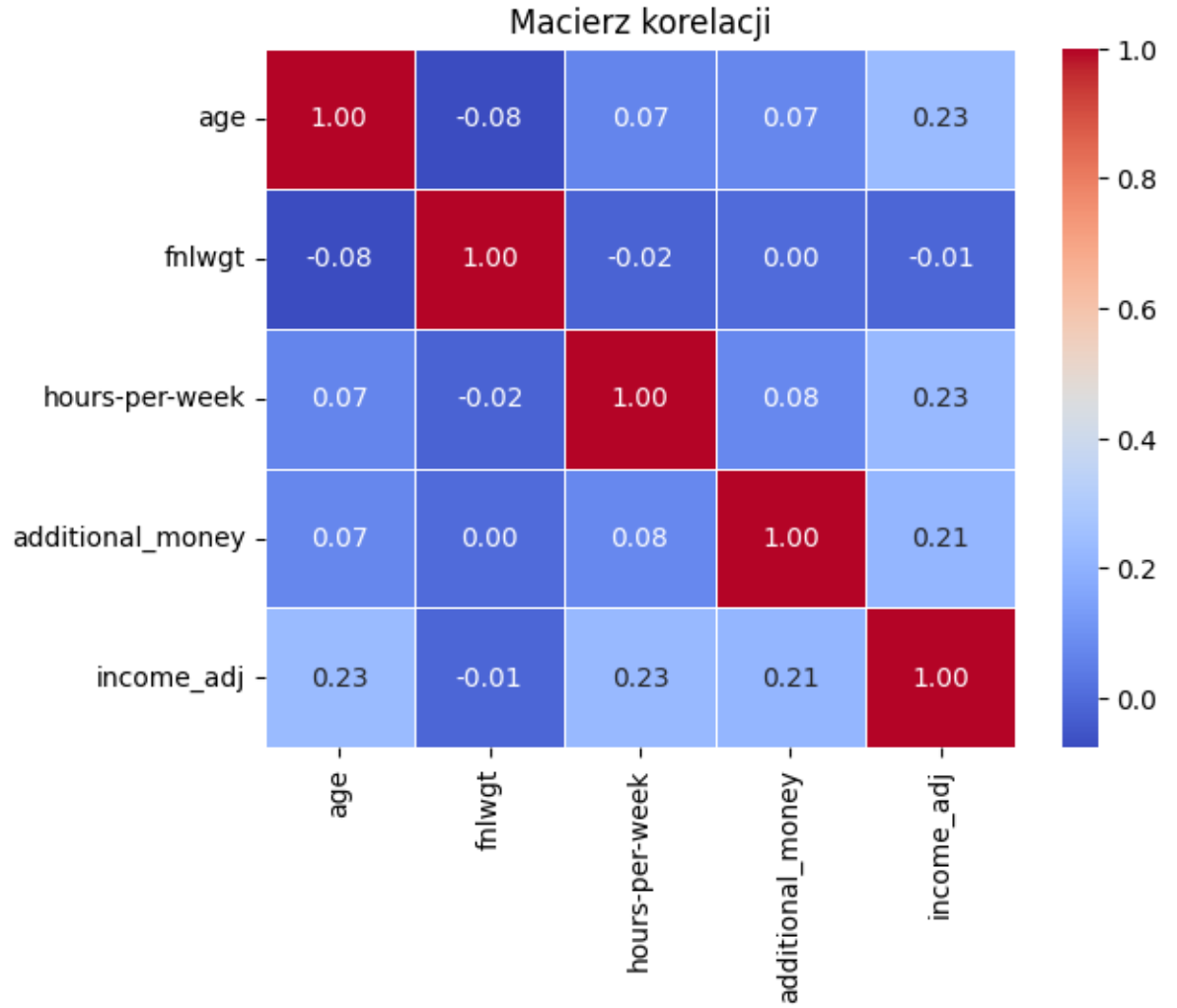
Warto odnotować, że najbardziej czasochłonnymi zadaniami było zapoznanie się z danymi oraz inżynieria cech. Koniecznym okazało się dodanie dodatkowego atrybutu, uzupełnienie brakujących wartości, uproszczenie etykiet dla zmiennej kategorycznej *education*, oraz normalizacja danych numerycznych.

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

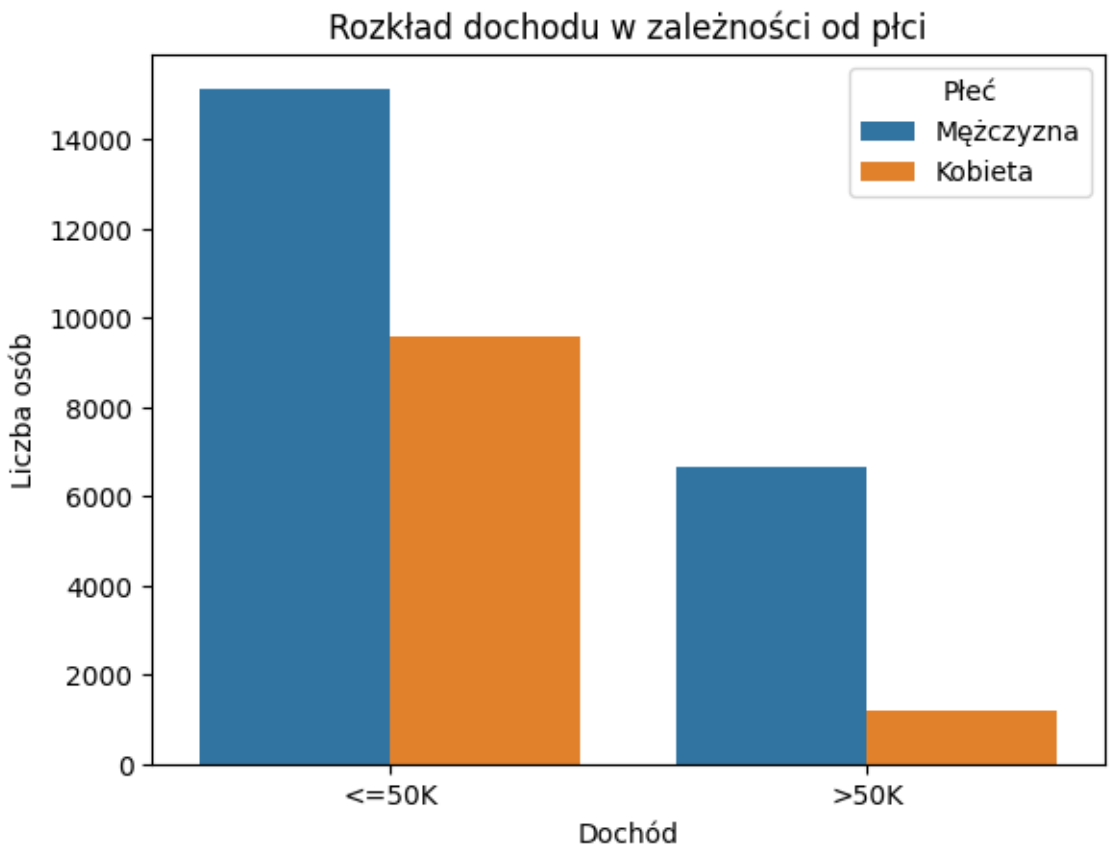
Załączniki



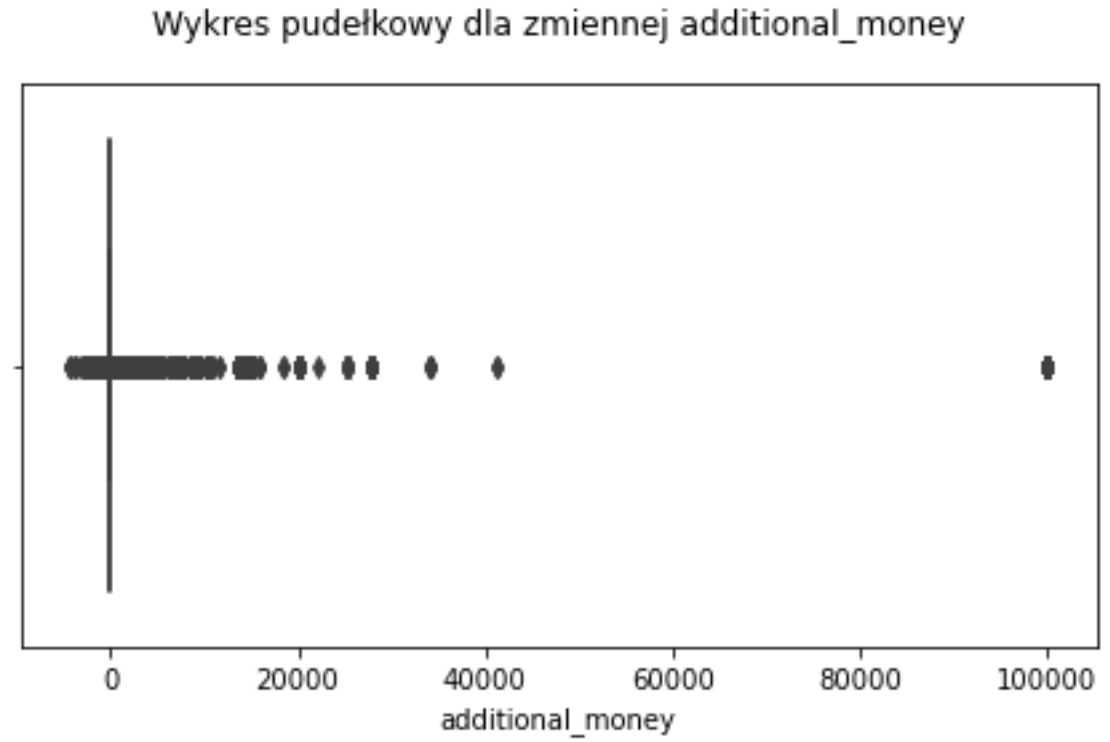
Załącznik 1. Wykresy pudełkowe dla zmiennych liczbowych



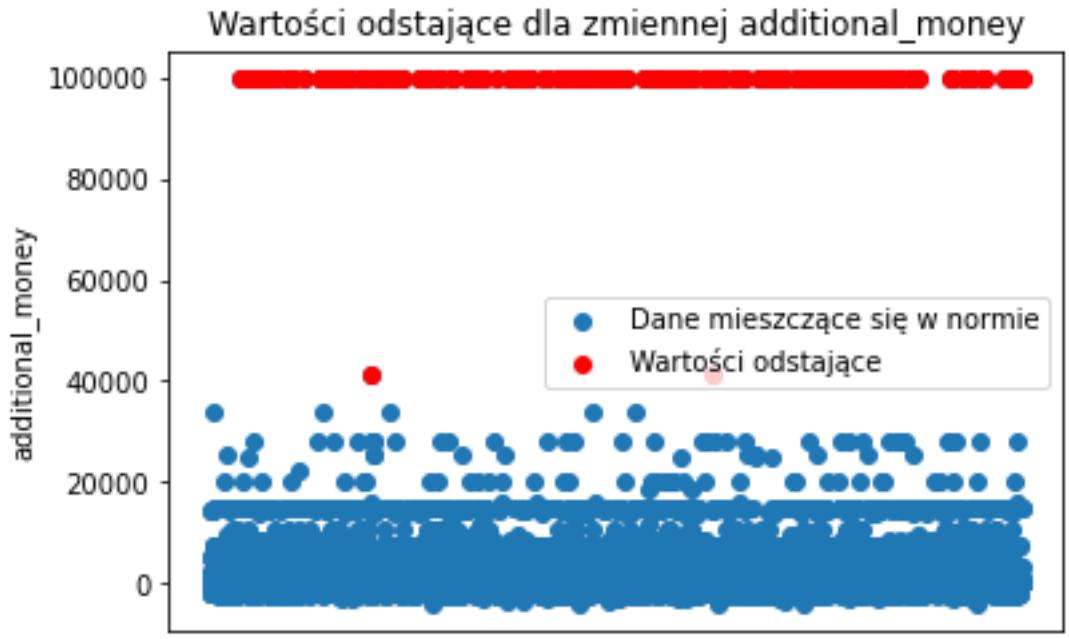
Załącznik 2. Macierz korelacji



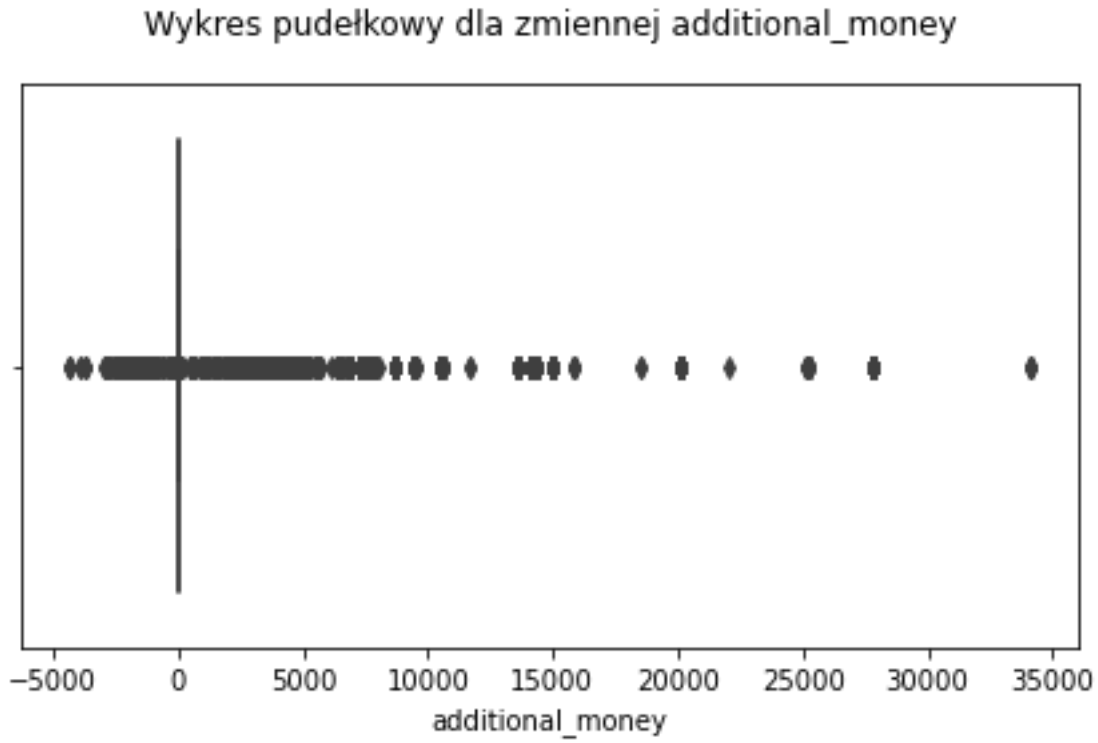
Załącznik 3. Wykres słupkowy przedstawiający zależność pomiędzy płcią, a poziomem dochodu.



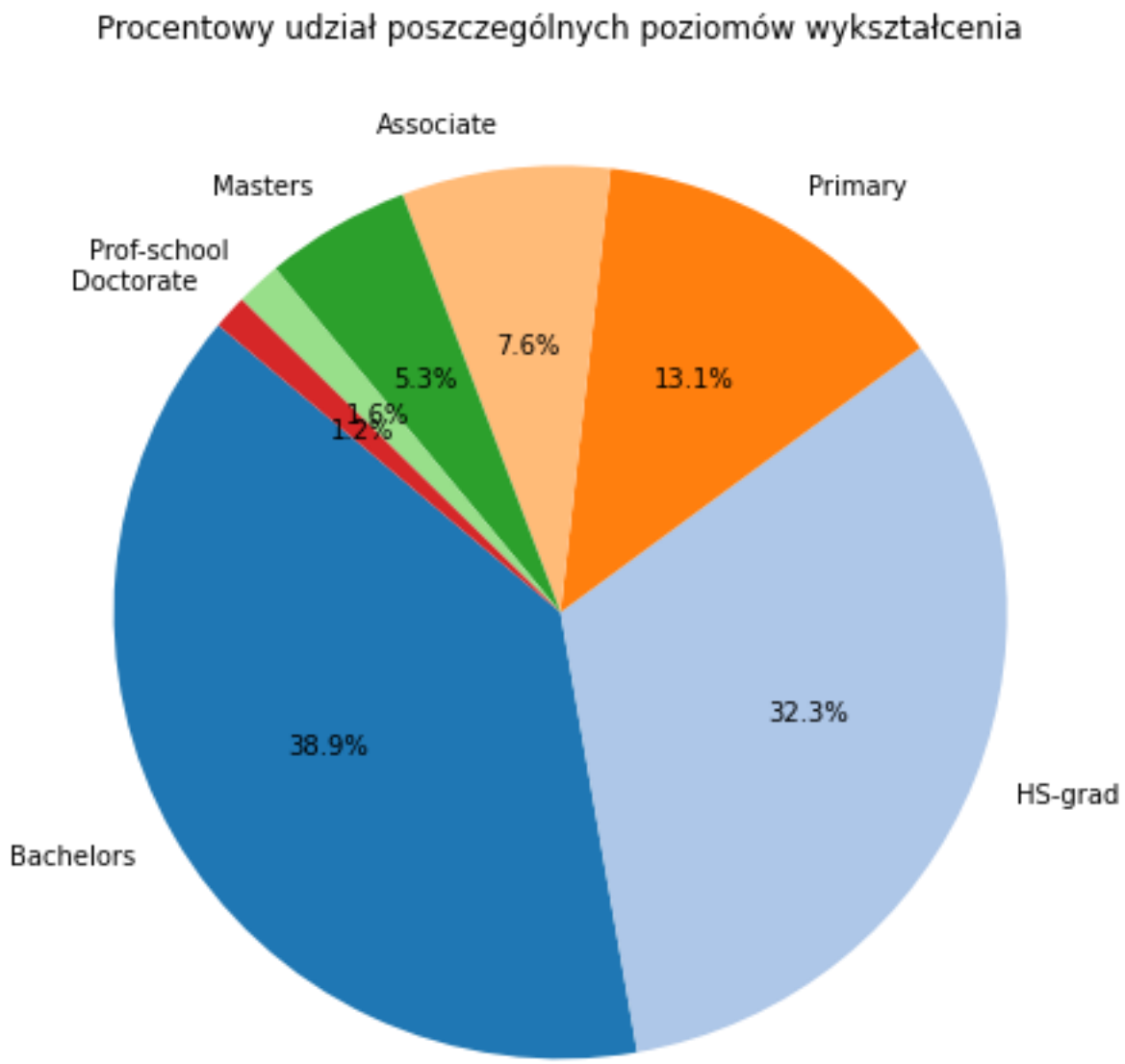
Załącznik 4. Wykres pudełkowy przygotowany dla zmiennej *additional_money.*



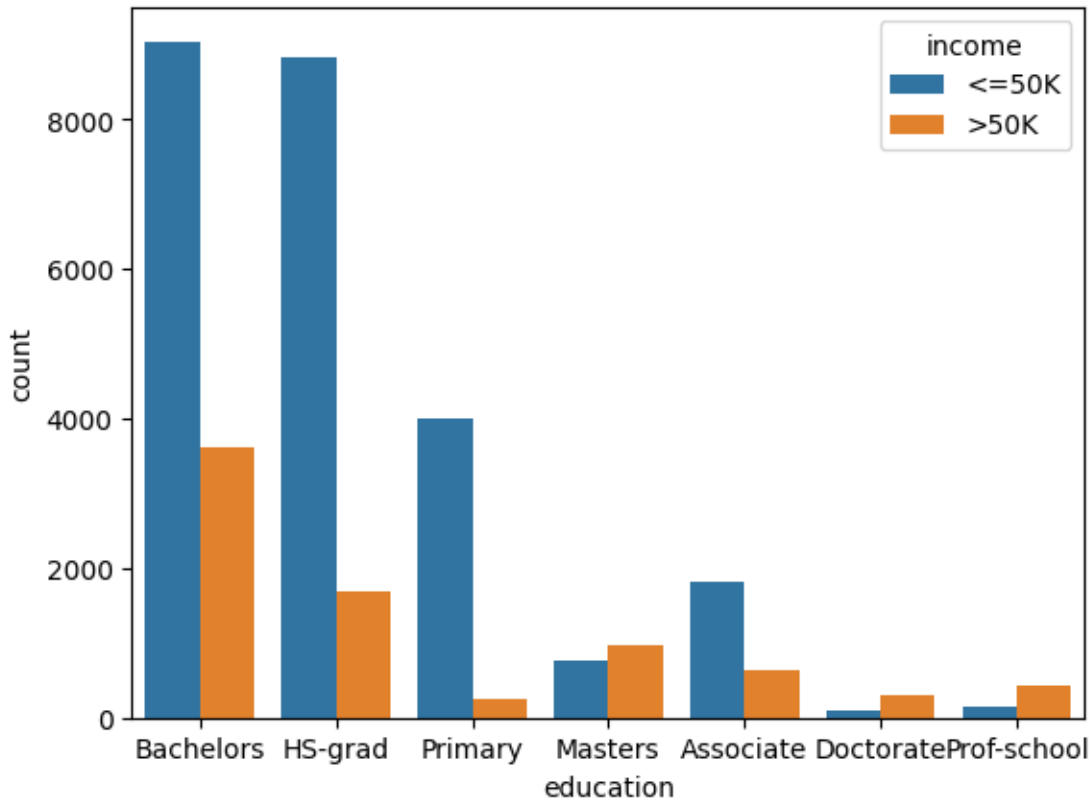
Załącznik 5. Wykres punktowy przedstawiający wartości odstające dla zmiennej *additional_money.*



Załącznik 6. Wykres pudełkowy przygotowany dla zmiennej *additional_money* po modyfikacji.



Załącznik 7. Wykres kołowy dla zmiennej *education.*



Załącznik 8. Wykres słupkowy pokazujący podział osób z poszczególnymi poziomami wykształcenia na grupy dochodowe.