UNIWERSYTET WARSZAWSKI

WYDZIAŁ MATEMATYKI, INFORMATYKI I MECHANIKI

KIERUNEK: INFORMATYKA / DATA SCIENCE

PRACA DYPLOMOWA

ZASTOSOWANIE ALGORYTMÓW UCZENIA MASZYNOWEGO

W PREDYKCJI ROTACJI PRACOWNIKÓW

Z OPTYMALIZACJĄ KOSZTÓW BIZNESOWYCH

- STUDIUM PRZYPADKU

APPLICATION OF MACHINE LEARNING ALGORITHMS

IN EMPLOYEE ATTRITION PREDICTION

WITH BUSINESS COST OPTIMIZATION

- A CASE STUDY

Autor: [Twoje Imię i Nazwisko]

Nr albumu: [Numer]

Promotor: [Imię i Nazwisko Promotora]

[Tytuł naukowy, Katedra]

Warszawa, 2025

SPIS TREŚCI

STRESZCZENIE ................................................................. 5

ABSTRACT .................................................................... 6

1. WPROWADZENIE ............................................................ 7

1.1 Uzasadnienie wyboru tematu .......................................... 7

1.2 Problem badawczy i pytania badawcze ................................ 9

1.3 Cele pracy i hipotezy badawcze ..................................... 11

1.4 Struktura pracy .................................................... 13

2. PRZEGLĄD LITERATURY ..................................................... 15

2.1 Teoretyczne podstawy rotacji pracowników .......................... 15

2.1.1 Teoria dwuczynnikowa Herzberga ............................... 16

2.1.2 Model Job Demands-Resources (JD-R) ........................... 17

2.1.3 Teoria wymiany społecznej .................................... 18

2.2 HR Analytics i People Analytics ................................... 19

2.2.1 Ewolucja analityki w zarządzaniu zasobami ludzkimi .......... 19

2.2.2 Zastosowania ML w talent management .......................... 21

2.2.3 Etyczne aspekty AI w HR ...................................... 22

2.3 Algorytmy uczenia maszynowego w klasyfikacji ...................... 24

2.3.1 Modele liniowe: Logistic Regression .......................... 24

2.3.2 Modele ensemble: Random Forest, XGBoost ...................... 25

2.3.3 Modele nieparametryczne: SVM, KNN ............................ 27

2.4 Cost-sensitive learning i optymalizacja biznesowa ................. 28

2.5 Identyfikacja luki badawczej ...................................... 30

3. METODOLOGIA BADANIA ..................................................... 32

3.1 Założenia metodologiczne ........................................... 32

3.2 Opis i charakterystyka danych ...................................... 33

3.3 Proces przetwarzania danych ........................................ 35

3.3.1 Data cleaning i preprocessing ................................ 35

3.3.2 Feature engineering .......................................... 37

3.3.3 Selekcja zmiennych ........................................... 39

3.4 Wybór i konfiguracja algorytmów ML ................................. 40

3.5 Metodologia ewaluacji i optymalizacji .............................. 42

3.5.1 Metryki wydajności ........................................... 42

3.5.2 Cross-validation i hyperparameter tuning .................... 43

3.5.3 Cost-sensitive threshold optimization ........................ 44

3.6 Narzędzia i środowisko badawcze .................................... 46

4. WYNIKI BADANIA .......................................................... 48

4.1 Eksploracyjna analiza danych ....................................... 48

4.1.1 Charakterystyka rozkładów zmiennych .......................... 48

4.1.2 Analiza korelacji i zależności ............................... 52

4.1.3 Identyfikacja wzorców attrition .............................. 55

4.2 Feature engineering i preprocessing ................................ 58

4.2.1 Tworzenie nowych zmiennych ................................... 58

4.2.2 Obsługa zmiennych kategorycznych ............................. 61

4.2.3 Skalowanie i normalizacja .................................... 63

4.3 Wyniki modeli bazowych ............................................. 65

4.3.1 Porównanie 5 algorytmów ML ................................... 65

4.3.2 Ranking modeli według AUC-ROC ................................ 68

4.3.3 Analiza overfittingu ......................................... 70

4.4 Hyperparameter tuning .............................................. 72

4.4.1 Optymalizacja Logistic Regression ............................ 72

4.4.2 Porównanie baseline vs tuned models .......................... 74

4.4.3 Wybór najlepszego modelu ..................................... 76

4.5 Cost-sensitive optimization ........................................ 78

4.5.1 Definicja kosztów biznesowych ................................ 78

4.5.2 Analiza progów decyzyjnych ................................... 80

4.5.3 Optymalizacja ROI ............................................ 82

4.6 Analiza feature importance ......................................... 85

4.6.1 TOP 10 najważniejszych czynników ............................. 85

4.6.2 Kategoryzacja według teorii HR ............................... 87

4.6.3 Interpretacja biznesowa ...................................... 89

5. DYSKUSJA I INTERPRETACJA ................................................ 92

5.1 Weryfikacja hipotez badawczych ..................................... 92

5.2 Interpretacja w kontekście teorii HR ............................... 95

5.3 Porównanie z literaturą przedmiotu ................................. 98

5.4 Analiza stabilności i generalizacji ................................ 101

6. IMPLIKACJE PRAKTYCZNE ................................................... 104

6.1 Business case i analiza ROI ........................................ 104

6.2 Strategia implementacji ............................................. 107

6.3 Rekomendacje dla organizacji ....................................... 110

6.4 Zarządzanie ryzykiem i ograniczenia ................................ 113

7. OGRANICZENIA BADANIA .................................................... 116

7.1 Ograniczenia metodologiczne ........................................ 116

7.2 Ograniczenia techniczne ............................................ 118

7.3 Ograniczenia kontekstowe ........................................... 119

8. KIERUNKI PRZYSZŁYCH BADAŃ ............................................... 121

8.1 Rozszerzenia metodologiczne ........................................ 121

8.2 Nowe źródła danych ................................................. 123

8.3 Praktyczne rozszerzenia ............................................ 124

9. ZAKOŃCZENIE ............................................................. 126

9.1 Podsumowanie wyników ............................................... 126

9.2 Wkład do wiedzy .................................................... 128

9.3 Znaczenie dla przyszłości HR ...................................... 129

BIBLIOGRAFIA ............................................................... 131

ZAŁĄCZNIKI ................................................................. 137

Załącznik A: Szczegółowe wyniki modeli ................................. 137

Załącznik B: Kod źródłowy (kluczowe fragmenty) ......................... 140

Załącznik C: Dodatkowe wizualizacje .................................... 145

Załącznik D: Słownik zmiennych ......................................... 148

Załącznik E: Pipeline implementacyjny .................................. 150

SPIS TABEL ................................................................. 152

SPIS RYSUNKÓW .............................................................. 154

STRESZCZENIE

Problem rotacji pracowników stanowi jeden z kluczowych wyzwań współczesnych organizacji, generując znaczące koszty związane z rekrutacją, szkoleniem

i utratą wiedzy organizacyjnej. Niniejsza praca przedstawia kompleksowe

badanie zastosowania algorytmów uczenia maszynowego w predykcji odejść

pracowników z innowacyjnym podejściem cost-sensitive optimization.

Celem głównym badania było opracowanie modelu predykcyjnego zdolnego do

skutecznego przewidywania rotacji pracowników przy jednoczesnej optymalizacji

kosztów biznesowych związanych z błędnymi decyzjami. Wykorzystano dataset

IBM HR Analytics zawierający 1,470 obserwacji pracowników z 35 zmiennymi

opisującymi demografię, warunki pracy, satysfakcję i historię kariery.

Metodologia badania obejmowała porównanie pięciu algorytmów uczenia maszynowego:

Logistic Regression, Support Vector Machine, Random Forest, XGBoost oraz

Extra Trees. Przeprowadzono kompleksowy proces feature engineering,

tworząc 20+ nowych zmiennych biznesowych, oraz hyperparameter tuning

z wykorzystaniem 5-fold cross-validation. Kluczowym elementem było

zastosowanie cost-sensitive optimization, uwzględniającej rzeczywiste

koszty biznesowe błędów predykcji (False Negative: 80,000 PLN,

False Positive: 3,500 PLN).

Wyniki badania wykazały, że najlepszym modelem okazała się Logistic Regression

z regularyzacją Elastic Net, osiągająca AUC-ROC na poziomie 0.814.

Analiza feature importance zidentyfikowała overtime, job satisfaction

i work-life balance jako najważniejsze predyktory rotacji. Optymalizacja

progu decyzyjnego (0.030 vs domyślne 0.5) umożliwiła redukcję kosztów

o 73.8%, przekładając się na potencjalne oszczędności 3,050,000 PLN rocznie.

Business case wykazał exceptional ROI na poziomie 799.3% z okresem zwrotu

inwestycji wynoszącym jedynie 2.0 miesiące. Roczne korzyści netto oszacowano

na 1,111,000 PLN przy koszcie implementacji 181,000 PLN. Badanie potwierdza

hipotezy dotyczące skuteczności feature engineering (+8% AUC) oraz

przewagi cost-sensitive optimization nad standardowymi metrykami ML.

Praca wnosi znaczący wkład metodologiczny poprzez integrację business value

z technikami machine learning, teoretyczny poprzez empiryczną weryfikację

teorii retention oraz praktyczny przez dostarczenie gotowego do implementacji

framework'u. Wyniki mają bezpośrednie zastosowanie w organizacjach dążących

do data-driven HR management i proactive talent retention strategies.

Słowa kluczowe: machine learning, employee attrition, HR analytics,

cost-sensitive learning, business optimization, predictive modeling,

talent retention, ROI optimization.

**1. WPROWADZENIE**

**1.1 Uzasadnienie wyboru tematu (2-3 strony)**

**1.1.1 Znaczenie problemu rotacji pracowników w współczesnych organizacjach**

W erze gospodarki opartej na wiedzy, gdzie kapitał ludzki stanowi kluczowy czynnik przewagi konkurencyjnej, problem rotacji pracowników (employee attrition) nabiera szczególnego znaczenia. Rotacja pracowników definiowana jako dobrowolne opuszczenie organizacji przez pracownika, stanowi jeden z najkosztowniejszych wyzwań współczesnego zarządzania zasobami ludzkimi.

Dane wykorzystane w niniejszym badaniu, pochodzące z datasetu IBM HR Analytics, wskazują na **16% stopę rotacji** w analizowanej organizacji, co oznacza, że co szósty pracownik opuszcza firmę w ciągu roku. Wskaźnik ten, choć może wydawać się umiarkowany, w kontekście dużej organizacji zatrudniającej tysiące pracowników przekłada się na setki odejść rocznie, generując ogromne koszty i zakłócenia operacyjne.

Podle badań Society for Human Resource Management (SHRM, 2022), średnia stopa rotacji w branży IT i usług technologicznych wynosi 13.2%, podczas gdy w niektórych sektorach może sięgać nawet 25-30%. Oznacza to, że problem rotacji ma charakter systemowy i dotyka większość organizacji, szczególnie w branżach o wysokiej konkurencji o talenty.

**Kluczowe konsekwencje wysokiej rotacji pracowników:**

1. **Bezpośrednie koszty finansowe** - rekrutacja, selekcja, onboarding nowych pracowników
2. **Utrata wiedzy organizacyjnej** - odchodzący pracownicy zabierają ze sobą know-how i doświadczenie
3. **Spadek produktywności** - okresy adaptacji nowych pracowników, obciążenie pozostałego zespołu
4. **Wpływ na morale** - rotacja może mieć efekt domina, wpływając na zadowolenie pozostałych pracowników
5. **Zakłócenia w relacjach z klientami** - szczególnie istotne w organizacjach usługowych

**1.1.2 Koszty rotacji pracowników - analiza ekonomiczna**

Oszacowanie rzeczywistych kosztów rotacji pracowników stanowi complex challenge, wymagający uwzględnienia zarówno kosztów bezpośrednich, jak i pośrednich. Na podstawie analizy literatury przedmiotu oraz benchmarków branżowych, w niniejszym badaniu przyjęto **koszt rotacji jednego pracownika na poziomie 80,000 PLN**.

**Dekompozycja kosztów rotacji:**

**A) Koszty bezpośrednie (40,000 PLN):**

* Proces rekrutacji i selekcji: 15,000 PLN
  + Ogłoszenia rekrutacyjne i portale pracy
  + Czas HR-owców na screening kandydatów
  + Koszty zewnętrznych firm rekrutacyjnych
* Onboarding i szkolenia: 20,000 PLN
  + Programy wprowadzające
  + Szkolenia specjalistyczne
  + Mentoring i supervision
* Koszty administracyjne: 5,000 PLN
  + Rozliczenia z odchodzącym pracownikiem
  + Dokumentacja i procedury

**B) Koszty pośrednie (40,000 PLN):**

* Utrata produktywności: 25,000 PLN
  + Okresy niepełnej wydajności nowego pracownika
  + Dodatkowa praca pozostałego zespołu
  + Możliwe błędy i opóźnienia projektów
* Utrata wiedzy organizacyjnej: 10,000 PLN
  + Know-how specjalistyczne
  + Relacje z klientami i partnerami
  + Nieudokumentowane procesy
* Wpływ na team morale: 5,000 PLN
  + Spadek zaangażowania pozostałych pracowników
  + Potencjalne kolejne odejścia (efekt domina)

Według badań Gallup (2020), koszt zastąpienia jednego pracownika wynosi od 50% do 200% jego rocznego wynagrodzenia, w zależności od poziomu stanowiska. Dla organizacji zatrudniającej 1,000 pracowników z 16% stopą rotacji, roczne koszty rotacji mogą osiągnąć **12.8 miliona PLN** (160 odejść × 80,000 PLN).

**1.1.3 Potencjał technologii i machine learning w HR**

Rozwój technologii analitycznych, w szczególności machine learning, otwiera nowe możliwości w zarządzaniu zasobami ludzkimi. **People Analytics** i **HR Analytics** stanowią dynamicznie rozwijający się obszar, który umożliwia:

1. **Predykcyjne zarządzanie talentami** - przewidywanie problemów zanim wystąpią
2. **Data-driven decision making** - podejmowanie decyzji opartych na danych, a nie intuicji
3. **Personalizacja doświadczeń pracowniczych** - dostosowanie strategii retention do indywidualnych potrzeb
4. **Optymalizacja procesów HR** - automatyzacja i usprawnienie rutynowych działań

**Machine Learning w kontekście attrition prediction oferuje:**

* **Wczesne ostrzeganie** - identyfikacja pracowników zagrożonych odejściem z wyprzedzeniem 3-6 miesięcy
* **Personalizowane interwencje** - dostosowanie działań retention do profilu ryzyka każdego pracownika
* **Continuous monitoring** - stały monitoring wskaźników engagement i satisfaction
* **ROI optimization** - cost-sensitive approach pozwalający na optymalizację zwrotu z inwestycji w retention

**1.1.4 Cost-sensitive learning jako innowacja w HR Analytics**

Tradycyjne podejścia do machine learning w HR koncentrują się na maksymalizacji standardowych metryk jak accuracy czy AUC-ROC. Jednak w kontekście biznesowym, różne typy błędów predykcji mają różne konsekwencje finansowe.

**W przypadku predykcji attrition:**

* **False Negative (FN)** - przeoczenie pracownika, który rzeczywiście odejdzie: **koszt 80,000 PLN**
* **False Positive (FP)** - niepotrzebna interwencja retention: **koszt 3,500 PLN**

Stosunek kosztów FN:FP wynosi około **23:1**, co oznacza, że znacznie bardziej kosztowne jest przeoczenie odejścia niż niepotrzebna interwencja. **Cost-sensitive optimization** pozwala na uwzględnienie tej asymetrii i optymalizację modelu pod kątem minimalizacji całkowitych kosztów biznesowych, a nie tylko poprawy standardowych metryk ML.

Ta innowacyjna metodologia stanowi bridge między world of data science a business requirements, umożliwiając stvarzenie modeli, które nie tylko dobrze przewidują, ale również generują maksymalną wartość biznesową.

**1.2 Problem badawczy i pytania badawcze (2-3 strony)**

**1.2.1 Sformułowanie problemu głównego**

W obliczu rosnących kosztów rotacji pracowników i zwiększającej się konkurencji o talenty, organizacje potrzebują skutecznych narzędzi do predykcji i prewencji odejść pracowników. Tradycyjne metody zarządzania retencją, oparte na periodic surveys i reactive approaches, okazują się niewystarczające w dynamicznym środowisku biznesowym.

**Problem główny niniejszego badania brzmi:**

*"Jak opracować skuteczny model predykcyjny rotacji pracowników wykorzystujący algoritmy uczenia maszynowego, który jednocześnie optymalizuje koszty biznesowe i umożliwia proactive talent retention w organizacji?"*

Problem ten ma charakter wielowymiarowy i obejmuje:

1. **Wymiar techniczny** - wybór i optymalizacja algorytmów ML dla specyfiki danych HR
2. **Wymiar biznesowy** - uwzględnienie rzeczywistych kosztów różnych typów błędów predykcji
3. **Wymiar operacyjny** - zapewnienie implementowalności i użyteczności rozwiązania w praktyce
4. **Wymiar etyczny** - zachowanie fairness i privacy w automatyzacji decyzji HR

**1.2.2 Pytania badawcze szczegółowe**

Dla kompleksowego rozwiązania problemu głównego, sformułowano następujące pytania badawcze szczegółowe:

**PB1: Które algorytmy uczenia maszynowego są najskuteczniejsze w predykcji rotacji pracowników?**

* Jak porównują się modele liniowe vs nieparametryczne w kontekście danych HR?
* Czy modele ensemble przewyższają pojedyncze algorytmy?
* Jaki wpływ ma hyperparameter tuning na skuteczność predykcji?
* Które metryki najlepiej odzwierciedlają jakość modelu w kontekście biznesowym?

**PB2: Jakie czynniki najsilniej wpływają na decyzje o odejściu pracowników?**

* Które zmienne mają największą siłę predykcyjną?
* Jak interpretować wyniki feature importance w kontekście teorii HR?
* Czy istnieją interakcje między zmiennymi zwiększające ryzyko attrition?
* Jakie nowe zmienne można skonstruować przez feature engineering?

**PB3: Jak zoptymalizować model pod kątem kosztów biznesowych?**

* Jaki jest optymalny próg decyzyjny minimalizujący całkowite koszty?
* Jak wpływa cost-sensitive optimization na standardowe metryki ML?
* Jaka jest trade-off między precision a recall w kontekście biznesowym?
* Jak oszacować ROI z implementacji predykcyjnego modelu retention?

**PB4: Jakie są praktyczne implikacje implementacji takiego rozwiązania?**

* Jakie są requirements techniczne i organizacyjne dla wdrożenia?
* Jak zapewnić continuous monitoring i model maintenance?
* Jakie są potencjalne ryzyka i ograniczenia rozwiązania?
* Jak zmierzyć success i business impact po implementacji?

**1.2.3 Kontekst teoretyczny i praktyczny**

Problem rotacji pracowników ma silne osadzenie w teoriach organizational behavior i human resource management. Kluczowe teorie stanowiące foundation dla niniejszego badania to:

**1. Teoria dwuczynnikowa Herzberga (1959)**

* Rozróżnienie między czynnikami higieny (zapobiegającymi niezadowoleniu) a motywatorami (zwiększającymi satysfakcję)
* Implikacje dla identyfikacji key drivers of attrition

**2. Model Job Demands-Resources (JD-R) Bakkera i Demerouti (2007)**

* Równowaga między wymaganiami pracy a dostępnymi zasobami
* Work-life balance jako kluczowy element modelu

**3. Teoria wymiany społecznej (Social Exchange Theory)**

* Reciprocity w relacjach pracodawca-pracownik
* Psychological contract i jego wpływ na commitment

**Praktyczny kontekst** badania obejmuje:

* Rosnące koszty acquisition nowych talentów
* Shortage of skilled workers w branży IT
* Increasing employee expectations dotyczące work-life balance
* Digital transformation w HR processes

**1.3 Cele pracy i hipotezy badawcze (2-3 strony)**

**1.3.1 Cel główny**

**Celem głównym** niniejszego badania jest **opracowanie i walidacja kompleksowego modelu predykcji rotacji pracowników wykorzystującego algoritmy uczenia maszynowego z cost-sensitive optimization, zapewniającego maksymalizację wartości biznesowej przy zachowaniu wysokiej skuteczności predykcyjnej.**

Model ten ma stanowić practical solution dla organizacji dążących do proactive talent management i data-driven HR decisions, oferując konkretne business value w postaci:

* Redukcji kosztów rotacji o minimum 50%
* ROI przekraczający 300% w pierwszym roku
* Skrócenia czasu reakcji na ryzyko attrition z miesięcy do tygodni

**1.3.2 Cele szczegółowe**

**CS1: Analiza porównawcza algorytmów uczenia maszynowego**

* Implementacja i ewaluacja 5 różnych algorytmów ML (Logistic Regression, SVM, Random Forest, XGBoost, Extra Trees)
* Comprehensive hyperparameter tuning dla każdego algorytmu
* Statistical comparison wyników z wykorzystaniem cross-validation
* Identyfikacja najskuteczniejszego algorytmu dla danych HR

**CS2: Identyfikacja i interpretacja kluczowych predyktorów attrition**

* Feature importance analysis z wykorzystaniem multiple methods
* Business interpretation wyników w kontekście teorii HR
* Categorization predyktorów według obszarów wpływu (compensation, work-life balance, career development, etc.)
* Development nowych zmiennych przez advanced feature engineering

**CS3: Implementacja cost-sensitive optimization**

* Quantification rzeczywistych kosztów biznesowych błędów predykcji
* Development cost-sensitive threshold optimization
* ROI analysis i business case development
* Comparison standardowych metryk ML vs business-oriented metrics

**CS4: Validacja praktycznej użyteczności rozwiązania**

* Assessment implementability w real-world environment
* Development practical recommendations dla HR professionals
* Risk assessment i mitigation strategies
* Framework dla continuous model monitoring i improvement

**1.3.3 Hipotezy badawcze**

Na podstawie przeglądu literatury i preliminary analysis danych, sformułowano następujące hipotezy badawcze:

**H1: Algorytmy ensemble przewyższają modele liniowe w predykcji attrition**

*Uzasadnienie:* Dane HR charakteryzują się complex non-linear relationships i interactions między zmiennymi. Modele ensemble jak Random Forest czy XGBoost powinny lepiej capture te zależności niż linear models.

*Operacjonalizacja:* AUC-ROC modeli ensemble > AUC-ROC Logistic Regression o minimum 0.05

**H2: Feature engineering znacząco poprawia skuteczność predykcji**

*Uzasadnienie:* Raw variables w datasecie HR często nie capture pełnego business context. Engineering nowych zmiennych opartych na domain knowledge powinien improve model performance.

*Operacjonalizacja:* AUC-ROC modelu z engineered features > AUC-ROC modelu z raw features o minimum 0.08

**H3: Cost-sensitive optimization generuje ROI przekraczający 500%**

*Uzasadnienie:* Traditional ML metrics nie uwzględniają business costs różnych typów błędów. Cost-sensitive approach powinien significantly improve business value przy marginalnym spadku technical metrics.

*Operacjonalizacja:* ROI z cost-optimized model ≥ 500% w pierwszym roku operacyjnym

**H4: Work-life balance stanowi najważniejszy predyktor attrition**

*Uzasadnienie:* Literature review i preliminary data analysis wskazują na overriding importance czynników related do work-life balance, szczególnie overtime i job satisfaction.

*Operacjonalizacja:* Variables related do work-life balance (OverTime, WorkLifeBalance, JobSatisfaction) stanowią ≥40% total feature importance

**1.3.4 Oczekiwany wkład do wiedzy**

**Wkład metodologiczny:**

* Development integrated framework łączącego ML techniques z business optimization
* Novel approach do cost-sensitive learning w HR context
* Practical methodology dla ROI assessment ML solutions w HR

**Wkład teoretyczny:**

* Empirical validation key theories of employee retention
* Quantitative analysis względnej ważności różnych retention factors
* Evidence-based insights dla HR theory development

**Wkład praktyczny:**

* Ready-to-implement solution dla HR professionals
* Actionable recommendations oparte na data-driven insights
* Template dla scaling podobnych solutions w różnych organizacjach