Sprawozdanie

-Uczenie Maszynowe-

Autor: Mateusz Kucharz

Spis treści

1. O Projekcie

W ramach projektu, celem było przeprowadzenie analizy danych dotyczących **rotacji pracowników** (ang. **employee attrition**), które pochodzą z publicznego zbioru danych dostępnego na Kaggle (https://www.kaggle.com/datasets/stealthtechnologies/employee-attrition-dataset/data). Zbiór danych zawiera ponad 70 tysięcy obserwacji, które obejmują różnorodne cechy związane z pracownikami, takie jak wiek, dział, długość zatrudnienia, poziom wynagrodzenia oraz inne czynniki, które mogą wpływać na decyzję o odejściu z pracy.

W projekcie dokonano szczegółowej analizy eksploracyjnej danych, której celem było zrozumienie rozkładów poszczególnych cech, identyfikacja wartości odstających oraz wykrycie ewentualnych nieprawidłowości w danych. Następnie, dane zostały podzielone na zbiory treningowy i testowy. W dalszej części projektu, zastosowano trzy modele klasyfikacyjne: Random Forest, Regresję Logistyczną oraz SVM. Na podstawie tych modeli wykonano predykcje, a wyniki zostały omówione i porównane. Na koniec przeprowadzono interpretację wyników za pomocą wykresów Ceteris Paribus dla modelu Random Forest, co umożliwiło lepsze zrozumienie wpływu poszczególnych cech na decyzję o rotacji pracowników.

2. EDA

Zmienne

Poniżej przedstawiono zmienne z zestawu danych:

- **Wiek**: Wiek pracownika, w zakresie od 18 do 60 lat.
- **Płeć**: Płeć pracownika.
- **Lata w Firmie**: Liczba lat, które pracownik przepracował w firmie.
- **Miesięczne Wynagrodzenie**: Miesięczna pensja pracownika, wyrażona w dolarach.
- **Stanowisko**: Dział lub rola, w której pracuje pracownik, zakodowane na kategorie takie jak Finanse, Opieka zdrowotna, Technologia, Edukacja, Media.

- **Równowaga Praca-Życie**: Oceniana przez pracownika równowaga między pracą a życiem prywatnym (Poor, Below Average, Good, Excellent).
- **Satysfakcja z Pracy**: Satysfakcja pracownika z pracy (Very Low, Low, Medium, High).
- **Ocena Wydajności**: Ocena wydajności pracownika (Low, Below Average, Average, High).
- **Liczba Awansów**: Łączna liczba awansów, które pracownik otrzymał.
- **Odległość od Domu**: Odległość między domem pracownika a miejscem pracy, wyrażona w milach.
- **Poziom Wykształcenia**: Najwyższy poziom wykształcenia uzyskany przez pracownika (High School, Associate Degree, Bachelor's Degree, Master's Degree, PhD).
- **Stan Cywilny**: Stan cywilny pracownika (Divorced, Married, Single).
- **Poziom Stanowiska**: Poziom stanowiska pracownika (Entry, Mid, Senior).
- **Wielkość Firmy**: Rozmiar firmy, w której pracuje pracownik (Small, Medium, Large).
- **Staż Firmy**: Całkowita liczba lat, które firma pracownika działa w branży.
- **Praca Zdalna**: Czy pracownik pracuje zdalnie (Yes or No).
- **Szanse na Kierownictwo**: Czy pracownik ma szanse na objęcie roli lidera (Yes or No).
- **Szanse na Innowacje**: Czy pracownik ma szanse na angażowanie się w innowacje (Yes or No).
- **Reputacja Firmy**: Postrzegana przez pracownika reputacja firmy (Very Poor, Poor, Good, Excellent).
- **Wydajność Pracownika**: Poziom wydajności, jaką pracownik prezentuje (Very Low, Low, Medium, High).
- **Rotacja**: Czy pracownik opuścił firmę, zakodowane jako 0 (stayed) lub 1 (Left).

Typy danych:

```
Age                          int64
Gender                      object
Years at Company             int64
Job Role                    object
Monthly Income               int64
Work-Life Balance           object
Job Satisfaction            object
Performance Rating          object
Number of Promotions         int64
Overtime                    object
Distance from Home           int64
Education Level             object
Marital Status              object
Number of Dependents         int64
Job Level                   object
Company Size                object
Company Tenure               int64
Remote Work                 object
Leadership Opportunities    object
Innovation Opportunities    object
Company Reputation          object
Employee Recognition        object
Attrition                   object
dtype: object
```

W zestawie jest 7 zmiennych ilościowych oraz 17 zmiennych jakościowych

.

Statystyki Opisowe

Dla zmiennych ilościowych

	count	mean	std	min	25%	50%	75%	max
Age	74498.0	38.529746	12.083456	18.0	28.0	39.0	49.0	59.0
Years at Company	74498.0	15.721603	11.223744	1.0	7.0	13.0	23.0	51.0
Monthly Income	74498.0	7299.379514	2152.508566	1226.0	5652.0	7348.0	8876.0	16149.0
Number of Promotions	74498.0	0.832935	0.995289	0.0	0.0	1.0	2.0	4.0
Distance from Home	74498.0	49.991584	28.513611	1.0	25.0	50.0	75.0	99.0
Number of Dependents	74498.0	1.650326	1.553633	0.0	0.0	1.0	3.0	6.0
Company Tenure	74498.0	55.727456	25.399349	2.0	36.0	56.0	76.0	128.0

Statystyki ilościowe wskazują, że średni wiek pracowników wynosi 38 lat, z odchyleniem standardowym 12 lat, a większość mieści się w przedziale 28–49 lat. Średni staż pracy w firmie to 15 lat, z dużym zróżnicowaniem – od 1 roku do 51 lat. Dochód miesięczny wynosi średnio 7299 jednostek, z dużym rozrzutem, podczas gdy większość zarabia między 5652 a 8876 jednostek. Pozostałe zmienne, takie jak liczba awansów czy odległość od miejsca pracy, również wykazują duże zróżnicowanie, co wskazuje na zróżnicowany profil pracowników.

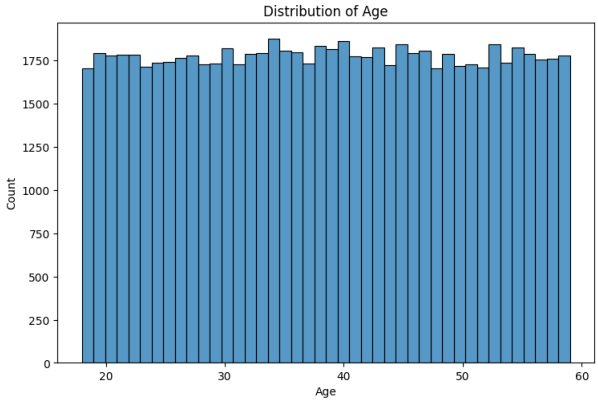
Dla zmiennych jakościowych

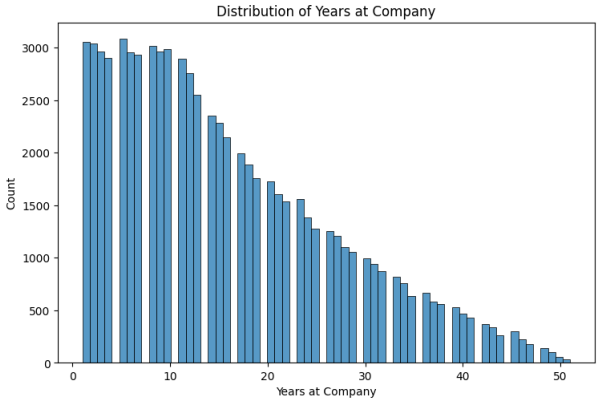
	count	unique	top	freq
Gender	74498	2	Male	40826
Job Role	74498	5	Technology	19322
Work-Life Balance	74498	4	Good	28158
Job Satisfaction	74498	4	High	37245
Performance Rating	74498	4	Average	44719
Overtime	74498	2	No	50157
Education Level	74498	5	Bachelor's Degree	22331
Marital Status	74498	3	Married	37419
Job Level	74498	3	Entry	29780
Company Size	74498	3	Medium	37231
Remote Work	74498	2	No	60300
Leadership Opportunities	74498	2	No	70845
Innovation Opportunities	74498	2	No	62394
Company Reputation	74498	4	Good	37182
Employee Recognition	74498	4	Low	29620
Attrition	74498	2	Stayed	39128

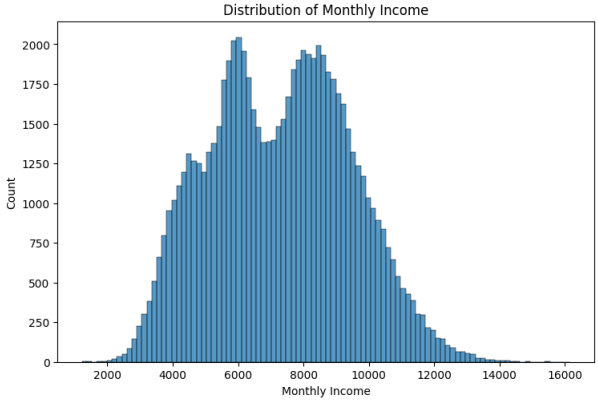
Statystyki pokazują, że większość pracowników to mężczyźni, dominującą rolą jest technologia, a najwyżej oceniana jest równowaga między pracą a życiem osobistym. Najczęściej spotykany poziom wykształcenia to licencjat, większość pracowników jest w związku małżeńskim, a dominujący poziom stanowisk to entry-level. Ponadto większość pracowników nie pracuje zdalnie i nie otrzymuje dodatkowych możliwości przywódczych czy innowacyjnych.

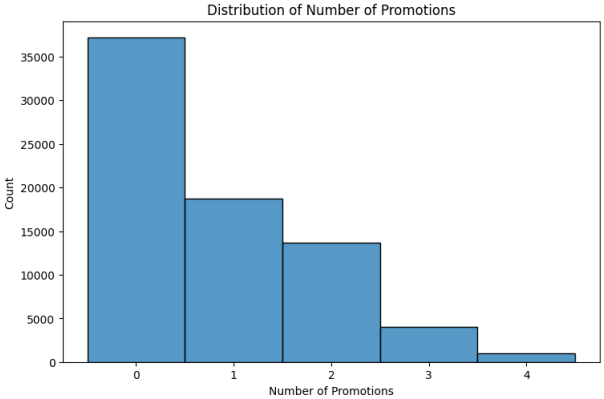
Rozkłady Zmiennych

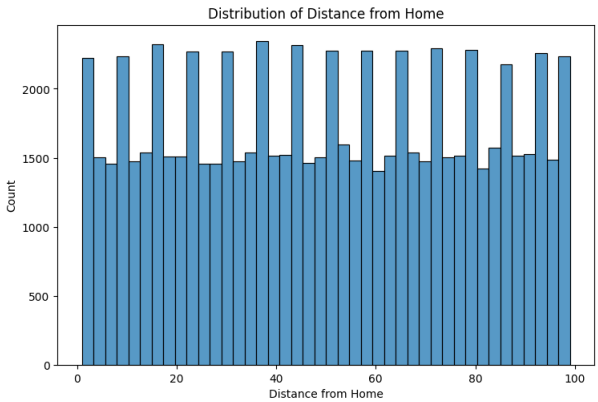
Zmienne ilościowe

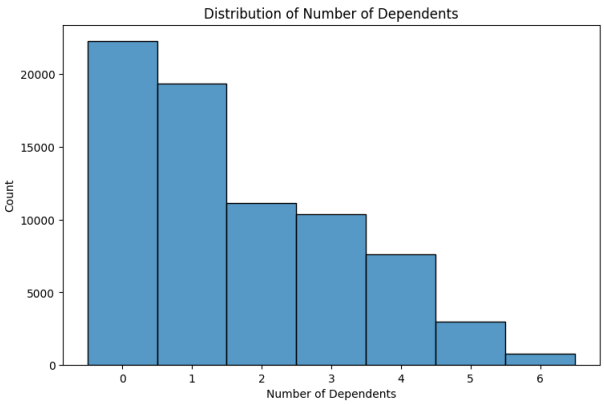


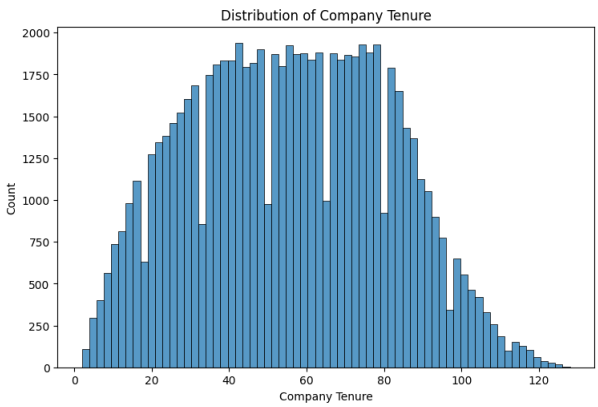






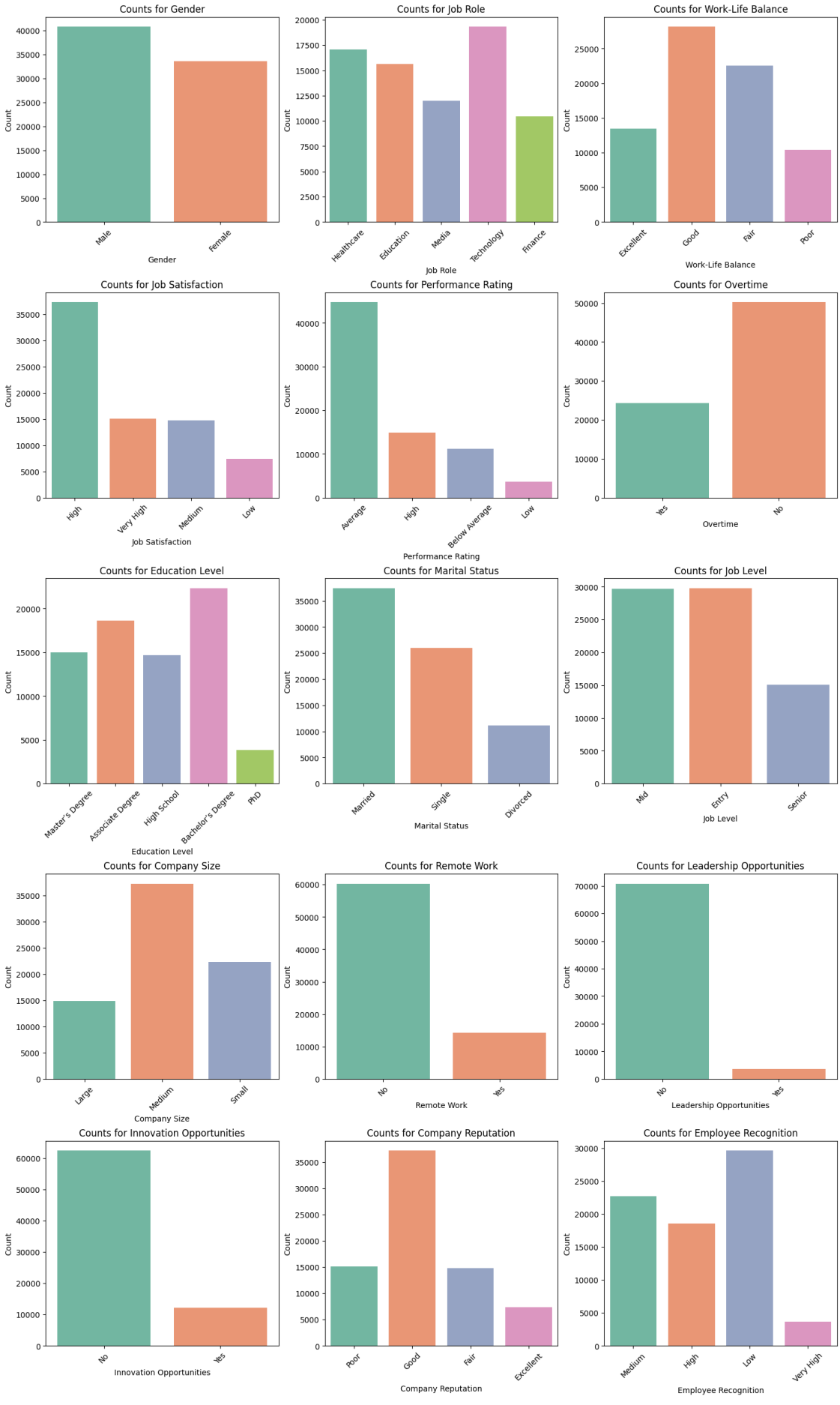






Żaden z rozkładów nie jest blisko rozkładu normalnego. Po **Age** i **Distance from Home** można zauważyć, że dane były przez kogoś wygenerowane. **Number of Promotions**, **Number of Dependends**, **Years at Company** są przechylone do lewej strony, podczas gdy **Monthly Income** i **Company Tenure** najbardziej trzymają się środka rozkładu.

Zmienne jakościowe

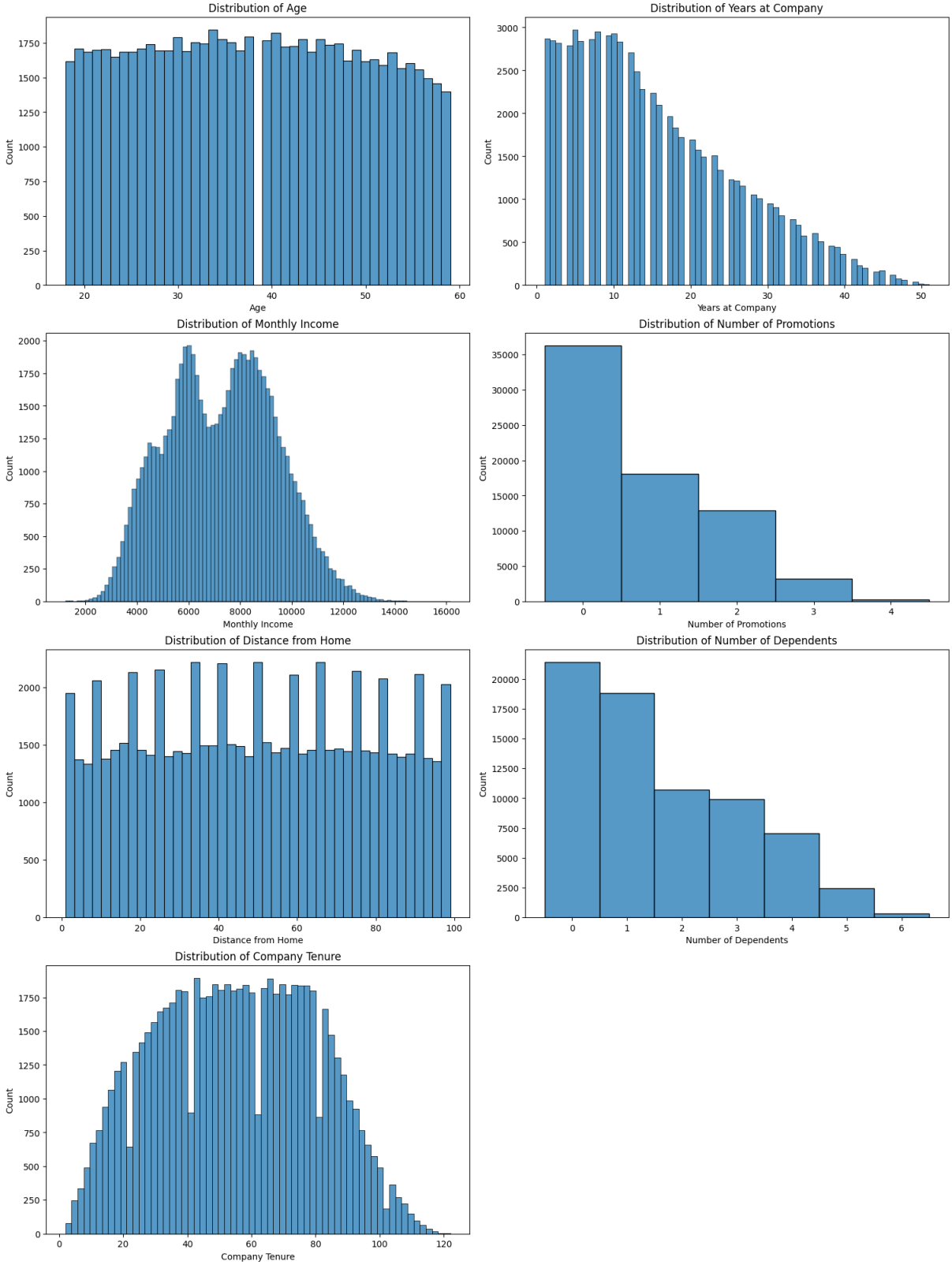

Counts for Gender
Count
Male
Female
Gender
Counts for Job Role
Healthcare
Education
Media
Technology
Finance
Job Role
Counts for Work-Life Balance
Excellent
Good
Fair
Poor
Work-Life Balance
Counts for Job Satisfaction
High
Very High
Medium
Low
Job Satisfaction
Counts for Performance Rating
Average
High
Below Average
Low
Performance Rating
Counts for Overtime
Yes
No
Overtime
Counts for Education Level
Master's Degree
Associate Degree
High School
Bachelor's Degree
PhD
Education Level
Counts for Marital Status
Married
Single
Divorced
Marital Status
Counts for Job Level
Mid
Entry
Senior
Job Level
Counts for Company Size
Large
Medium
Small
Company Size
Counts for Remote Work
No
Yes
Remote Work
Counts for Leadership Opportunities
No
Yes
Leadership Opportunities
Counts for Innovation Opportunities
No
Yes
Innovation Opportunities
Counts for Company Reputation
Poor
Good
Fair
Excellent
Company Reputation
Counts for Employee Recognition
Medium
High
Low
Very High
Employee Recognition

Większość rozkładów zmiennych kategorycznych prezentuje się jedną klasą o wysokiej częstotliwości i paru mniej licznych tak jak w **Overtime** czy **Performance Rating**. Najbardziej zrównoważone rozkłady wydaje się mieć **Gender** oraz **Job Role.**

Obserwacje Odstające

Zmienne ilościowe były analizowane pod kątem wartości odstających w kontekście wielu zmiennych za pomocą algorytmu Isolation Forest. Usuwanie wartości odstających jest istotne, ponieważ mogą one zniekształcać analizy statystyczne, wpływać na dokładność prognoz modeli oraz prowadzić do błędnych wniosków. Isolation Forest to algorytm uczenia maszynowego stosowany do wykrywania anomalii w sposób nienadzorowany. Działa on poprzez izolowanie wartości odstających, wykonując rekurencyjne podziały danych przy użyciu losowych cięć, co pozwala na skuteczną identyfikację nietypowych punktów w zbiorze danych.

Rozkłady po usunięciu wartości odstających

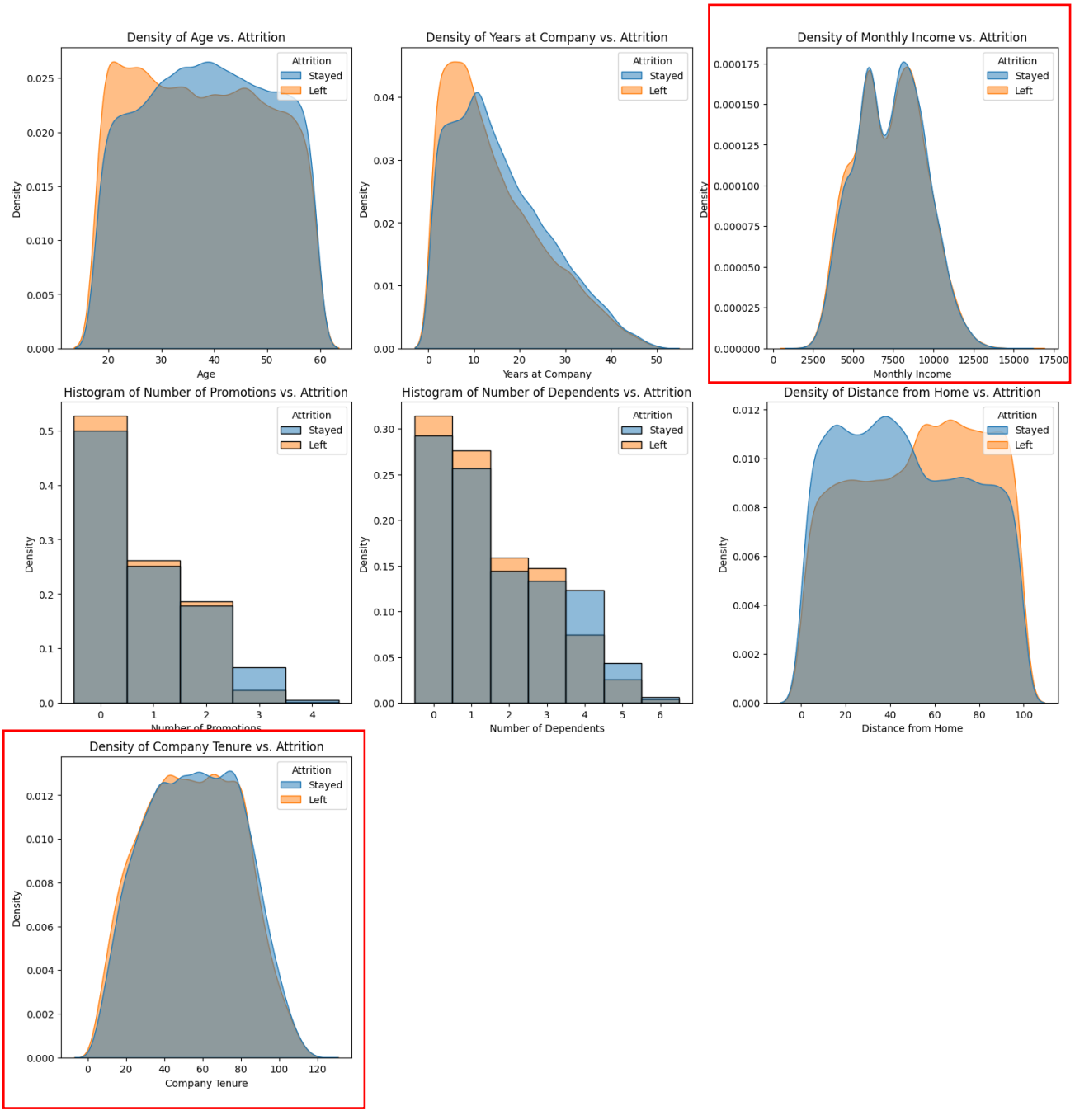


Rozkłady uległy niewielkim zmianom po usunięciu wartości odstających. Najbardziej widoczne jest to w przypadku zmiennych **Number of Dependents** oraz **Number of Promotions**, gdzie mniej liczniejsze wartości zostały znacząco zredukowane.

Wpływ Na Zmienną Y

Przeanalizowano wpływ zmiennych X na zmienną Y, a następnie zidentyfikowano te o najsłabszym wpływie i usunięto je w celu redukcji liczby zmiennych.

Zmienne ilościowe



Monthly Income oraz **Company Tenure** nie wydają się mieć dużego wpływu na **Attrition**, ponieważ ich wykresy gęstości dla różnych kategorii Y prawie się nakładają. Z tych powodów zostaną one usunięte. W reszcie wykresów jasno widać, że od pewnej wartości dominuje jedna kategoria.

Zmienne jakościowe

Job Role, **Leadership Opportunities**, **Innovation Opportunities**, **Company Size** oraz **Employee Recognition** nie wydają się mieć wpływu na **Attrition**, ponieważ

obie kategorie mieszczą się w zakresie 0.5 we wszystkich grupach na ich wykresach. Wybrane zmienne zostaną usunięte.

Balans Klas

Proporcja obserwacji z klasą 1 do klasy 0 wynosi 48%, co wskazuje na zbalansowane dane, które nie wymagają dalszego balansowania. W dalszej części projektu jako punkt odcięcia przyjęto 0.5.

3. Modele

Przygotowanie Danych

Dane zostały transformowane w zależności od ich typu by można było ich użyć do dalszych predykcji.

- **Zmienne ilościowe**: zmienne te nie były zmienianie, w dalszej części projektu przeprowadzono na nich standaryzację.
- **Zmienne jakościowe z 2 kategoriami**: zmienione na zmienne binarne.
- **Zmienne jakościowe, w których kategorie mają określoną kolejność, ale odległości między nimi nie są jednoznaczne**: Wykonano OrdinalEncoding czyli przypisano im liczby by na potrzeby modelu wyznaczyć odległości między nimi.
- **Zmienne jakościowe z 2+ kategoriami**: Wykonano OneHotEncoding, odrzucono 1 kolumnę w celu uniknięcia współliniowości (nie wszystkie modele tego wymagają, ale też nie koniecznie im to przeszkadza).

Dane zostały podzielone na zbiór treningowy oraz testowy w proporcjach: 75% oraz 25%.

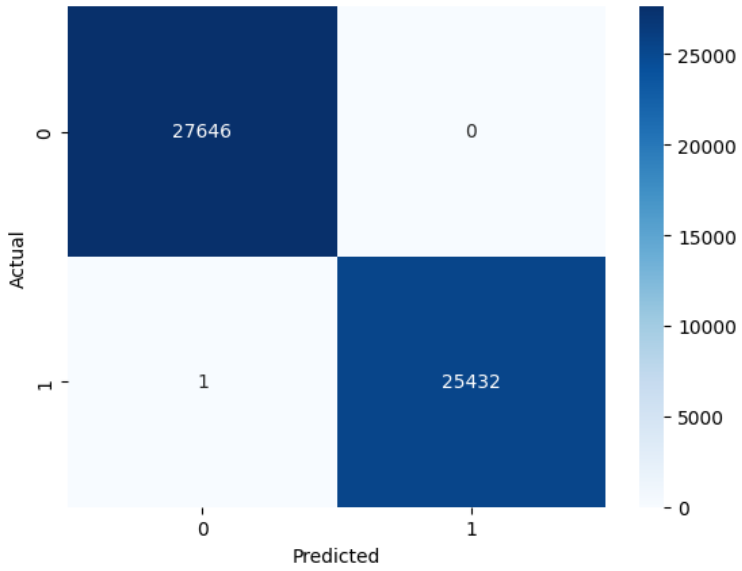
Random Forest

Random Forest to model ensemble w uczeniu maszynowym, który łączy wiele drzew decyzyjnych, aby poprawić dokładność i zredukować przeuczenie. Poprzez agregowanie prognoz poszczególnych drzew, zapewnia wydajność zarówno w zadaniach klasyfikacji, jak i regresji.

Najważniejsze hiperparametry:

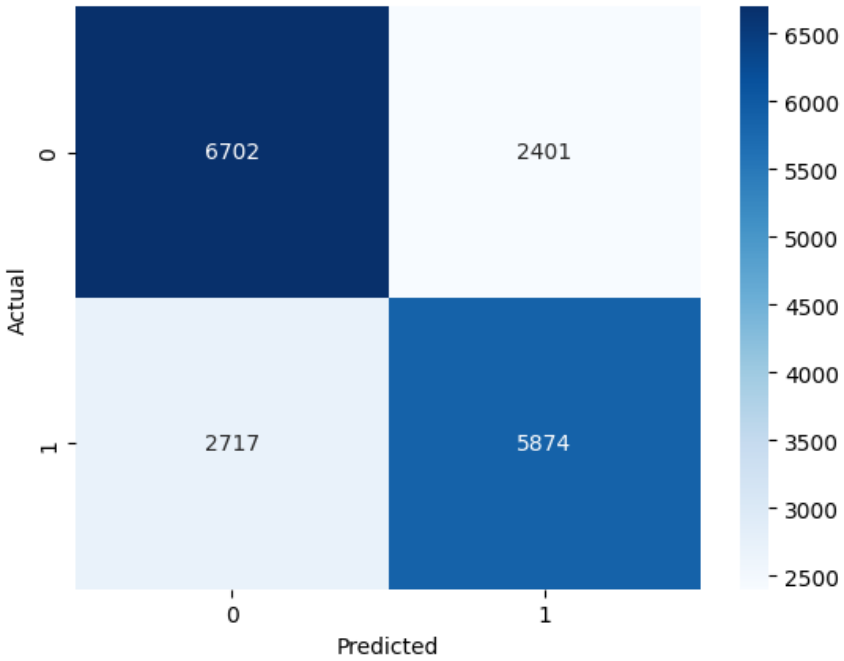
- **n_estimators**: Liczba drzew w lesie.
- **max_depth**: Maksymalna głębokość każdego drzewa. Kontroluje przeuczenie, ograniczając wzrost drzewa.
- **min_samples_split**: Minimalna liczba próbek wymagana do podziału węzła wewnętrznego.
- **min_samples_leaf**: Minimalna liczba próbek wymagana w węźle liściowym.
- **max_leaf_nodes**: Maksymalna liczba węzłów liściowych w pojedynczym drzewie.

Model z bazowymi parametrami – zbiór treningowy



Accuracy	Specificity	Recall
1	1	1

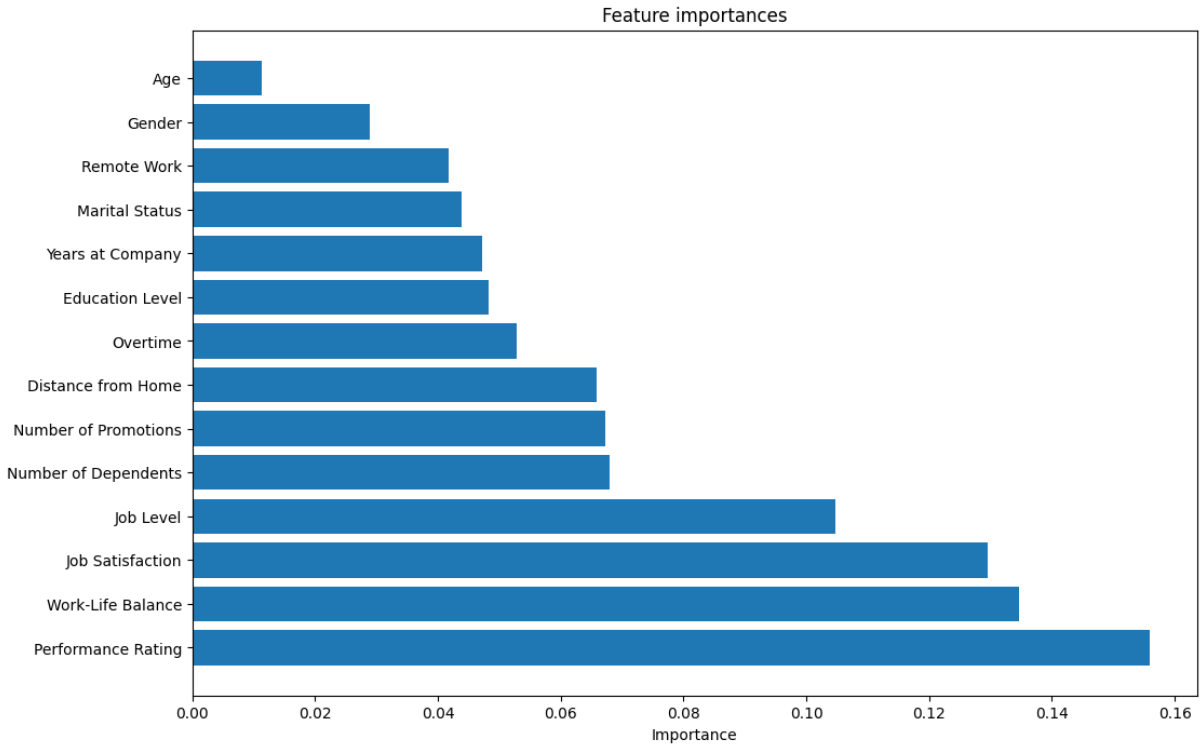
Model z bazowymi parametrami – zbiór testowy



Accuracy	Specificity	Recall
0.71	0.74	0.68

Wyniki modelu na zbiorze treningowym wskazują na idealne dopasowanie, co może sugerować przeuczenie modelu, ponieważ osiąga on perfekcyjne wyniki na danych, na których był trenowany. Na zbiorze testowym wyniki są znacznie niższe: Accuracy wynosi 71%, Specificity 74%, a Recall 68%. Oznacza to, że model radzi sobie przeciętnie z danymi, których wcześniej nie widział. Szczególnie Recall (68%) wskazuje, że model ma problemy z wykrywaniem pozytywnych przypadków, a obniżone Specificity (74%) świadczy o umiarkowanej zdolności do unikania fałszywie pozytywnych wyników. Te różnice między wynikami na zbiorach sugerują, że model mógł zostać nadmiernie dopasowany do danych treningowych.

Model z wybranymi parametrami – istotność zmiennych



GridSearch

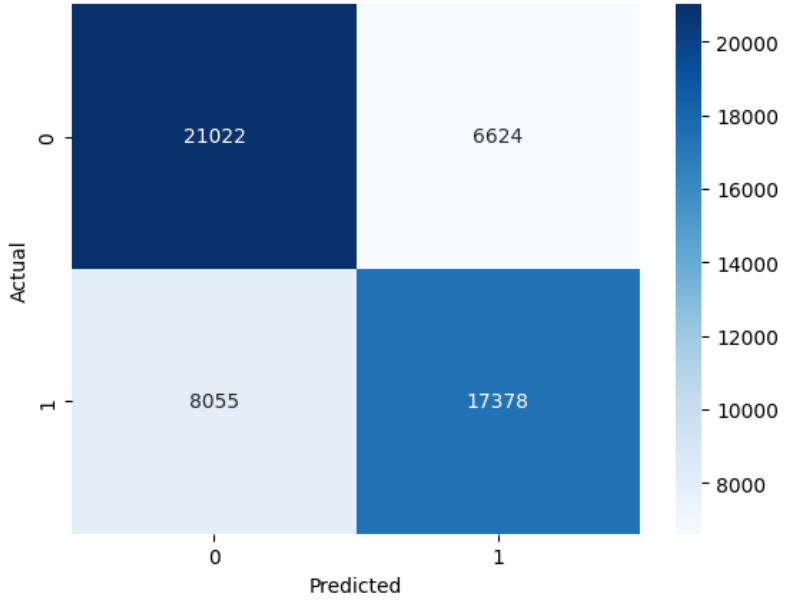
Dopasowanie hiperparametrów to proces wyboru optymalnego zestawu parametrów dla modelu uczenia maszynowego w celu poprawy jego wydajności i uogólnienia na nieznanych danych. W tym przypadku autor użył GridSearch z 5-krotną walidacją krzyżową, która jest techniką systematycznego przeszukiwania określonej przestrzeni hiperparametrów przez trenowanie i ocenianie modelu dla każdej możliwej kombinacji hiperparametrów. GridSearch pomaga zidentyfikować najlepsze hiperparametry poprzez ocenę wydajności modelu za pomocą walidacji krzyżowej.

Podczas używania GridSearch, ocenianie opiera się na funkcji zdefiniowanej na początku projektu. Funkcja ta uwzględnia dokładność, specyficzność i czułość, przypisując im wagi (1, 0.5, 0.5), a wynik to suma tych metryk. Takie podejście zapewnia, że wszystkie trzy metryki są brane pod uwagę, przy czym ogólna dokładność jest najważniejsza, aby uniknąć sytuacji, w których jedna metryka jest znacznie niższa od pozostałych.

Najlepsze parametry

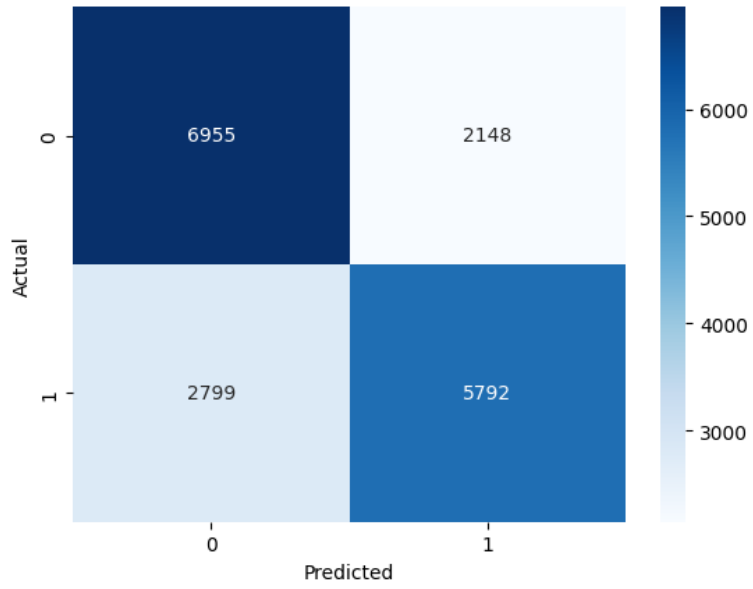
max_depth	max_features	min_samples_leaf	min_samples_split	n_estimators
None	sqrt	250	500	200

GridSearch – zbiór treningowy



Accuracy	Specificity	Recall
0.72	0.76	0.68

GridSearch – zbiór testowy

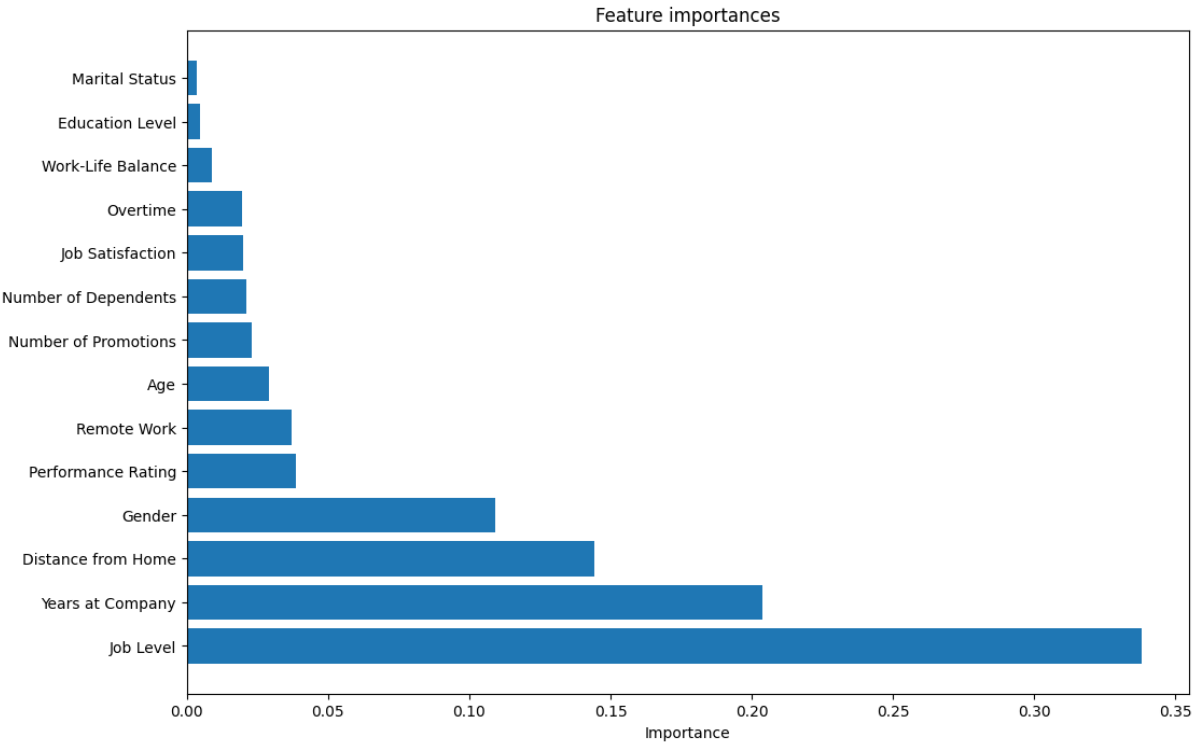


Accuracy	Specificity	Recall
0.72	0.76	0.67

Wyniki modelu wskazują na zbliżoną jakość predykcji zarówno na zbiorze treningowym, jak i testowym, co świadczy o dobrym ogólnym dopasowaniu modelu. Na obu zbiorach Accuracy wynosi około 72%, co oznacza, że model poprawnie klasyfikuje 72% przypadków. Specificity na poziomie 76% wskazuje, że model dobrze identyfikuje przypadki negatywne. Recall wynosi 67%–68%, co oznacza, że model nieco gorzej radzi sobie z wykrywaniem przypadków pozytywnych. Ogólnie model jest stabilny w działaniu, ale umiarkowana wartość Recall sugeruje, że jego skuteczność w identyfikacji pozytywnych przypadków mogłaby zostać poprawiona, np. przez dalsze strojenie hiperparametrów.

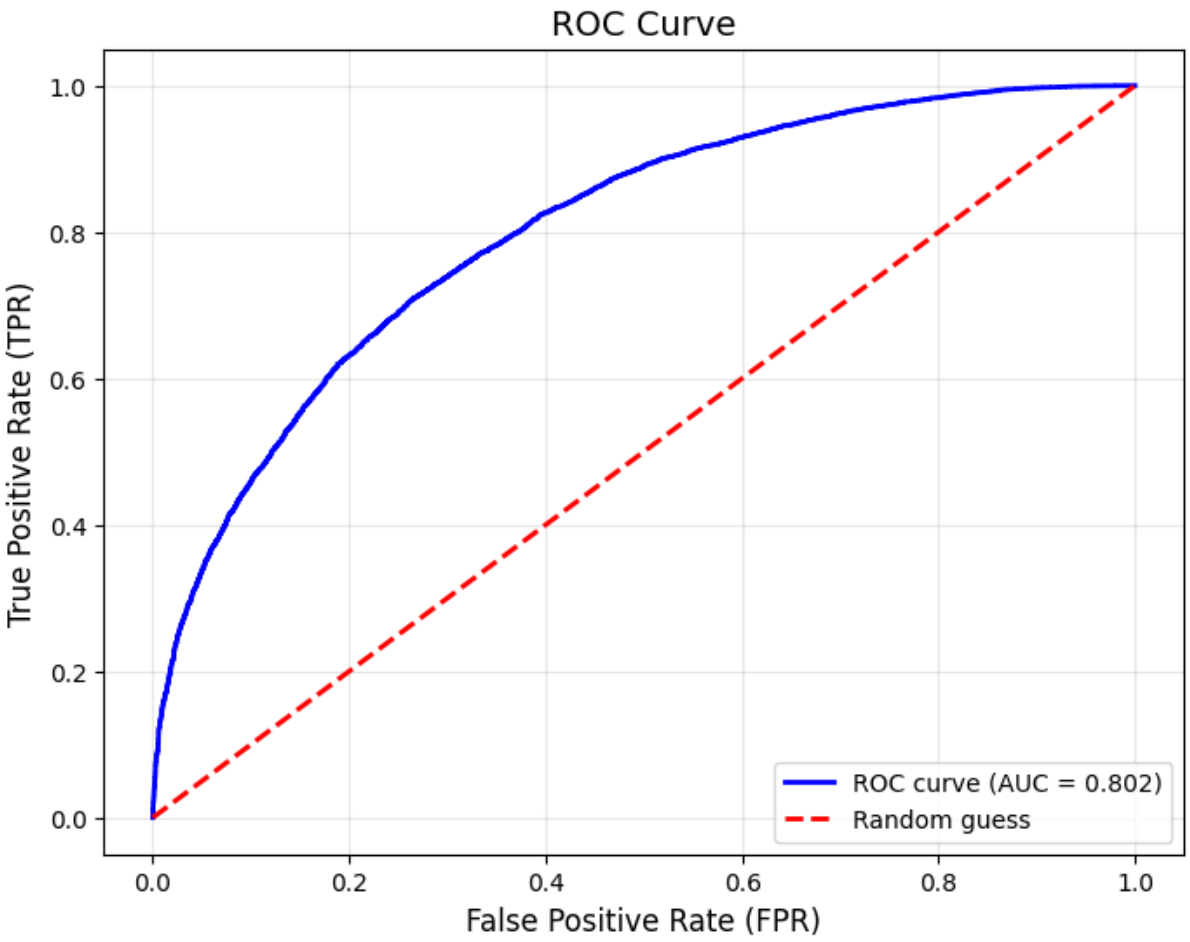
Wyniki na zbiorach treningowych sugerują, że po GridSearch model radzi sobie z przetrenowaniem, a wyniki na zbiorach testowych pokazują, że znalazł on lepszą kombinację hiperparametrów.

Model z wybranymi parametrami – istotność zmiennych



Po GridSearch zmieniła się istotność zmiennych. **Job Level** oraz **Years at Company** stały się o wiele bardziej istotne w porównaniu z modelem bazowym. Tak samo jak wcześniej spora część zmiennych jest mało istotna ~0.05.

ROC Curve



Wartość AUC wynosząca 0.802 oznacza, że model ma 80.2% prawdopodobieństwo poprawnego rozróżnienia między losowo wybraną pozytywną a losowo wybraną negatywną obserwacją.

Regresja Logistyczna

Logistic Regression to model statystyczny wykorzystywany do zadań klasyfikacji binarnej, w których wynik jest prawdopodobieństwem, które można odwzorować na wynik binarny. Model opisuje relację między cechami wejściowymi a

prawdopodobieństwem wystąpienia klasy docelowej, wykorzystując funkcję logistyczną.

Najważniejsze hiperparametry (**statsmodel Logit**):

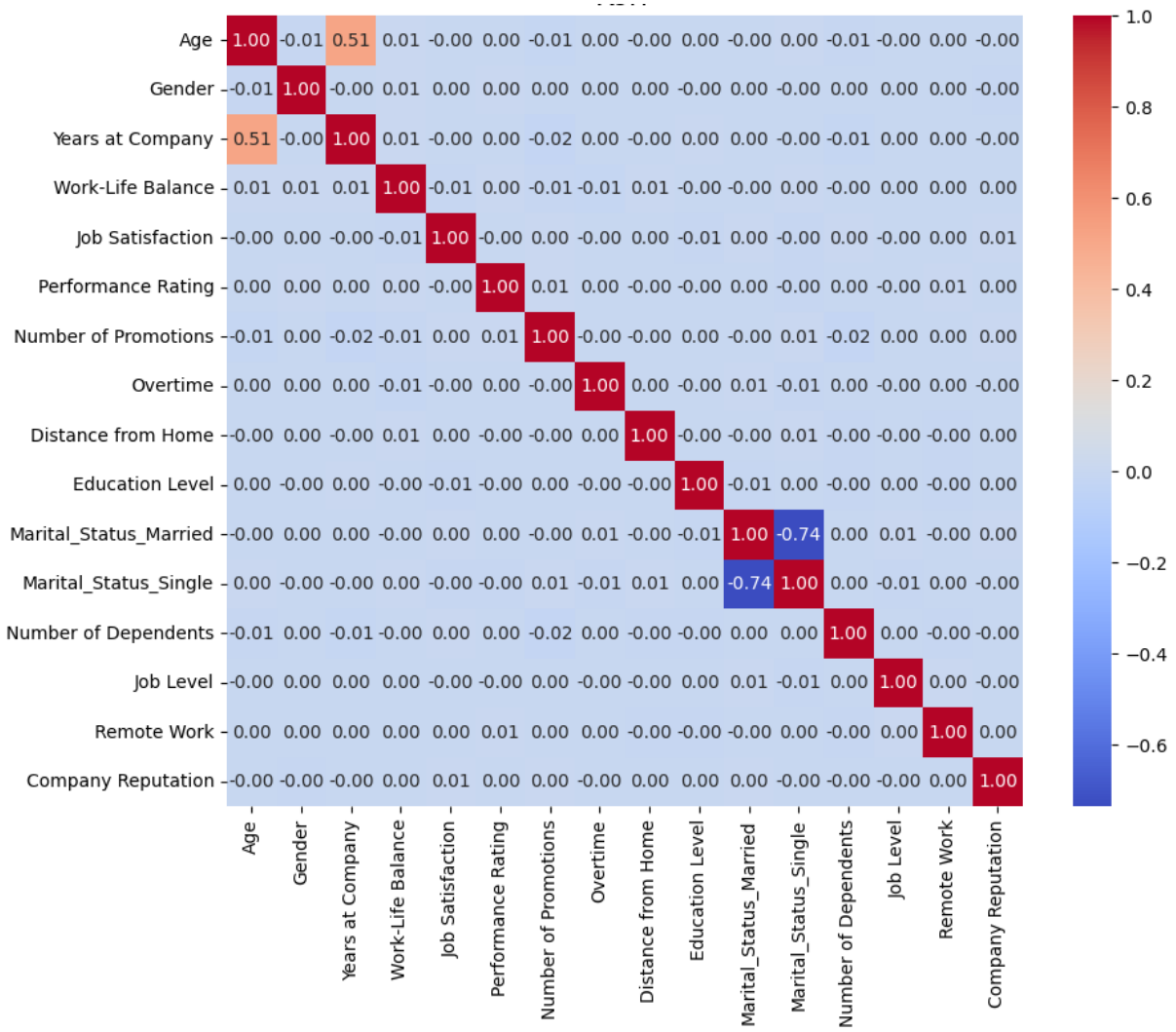
- **method**: Metoda optymalizacji używana do dopasowania modelu.
- **maxiter**: Maksymalna liczba iteracji wykonywanych podczas optymalizacji.
- **tol**: Tolerancja dla zbieżności.

Do modelu dodano stałą oraz przeprowadzono standaryzację zmiennych ilościowych, a także jednej zmiennej zakodowanej za pomocą OrdinalEncoding. Standaryzację zastosowano dla tej zmiennej, aby zmniejszyć jej wpływ na model z uwagi na dużą liczbę kategorii (5). Pozostałe zmienne nie zostały standaryzowane, ponieważ odległości między ich kategoriami nie są dokładnie określone, a liczba kategorii jest nieco mniejsza.

Równanie logarytmu szans

$$\log(p/(1-p)) = \beta_0 + \beta_1 * Age + \beta_2 * Gender + \beta_3 * Years\ At\ Company + \beta_4 * WorkLifeBalance + \beta_5 * JobSatisfaction + \beta_6 * PerformanceRating + \beta_7 * NumberOfPromotions + \beta_8 * Overtime + \beta_9 * DistanceFromHome + \beta_{10} * EducationLevel + \beta_{11} * MaritalStatusMarried + \beta_{12} * MaritalStatusSingle + \beta_{13} * NumberOfDependents + \beta_{14} * JobLevel + \beta_{15} * RemoteWork + \beta_{16} * CompanyReputation$$

Macierz korelacji X



Większość zmiennych X nie jest ze sobą skorelowana. Jednak zmienne **Years at Company** i **Age** wykazują umiarkowaną korelację, co jest zrozumiałe, ponieważ obie wartości mają tendencję do wzrostu w czasie, jeśli pracownik pozostaje w firmie. Dwie zmienne typu dummy pochodzące ze zmiennej **Marital Status** również wykazują korelację, co jest oczekiwane, ponieważ zostały wyprowadzone z tej samej zmiennej oryginalnej.

Trenowanie modelu

Logit Regression Results

Dep. Variable:	Attrition	No. Observations:	53079
Model:	Logit	Df Residuals:	53062
Method:	MLE	Df Model:	16
Date:	Wed, 15 Jan 2025	Pseudo R-squ.:	0.2728
Time:	20:27:11	Log-Likelihood:	-26721.
converged:	True	LL-Null:	-36745.
Covariance Type:	nonrobust	LLR p-value:	0.000

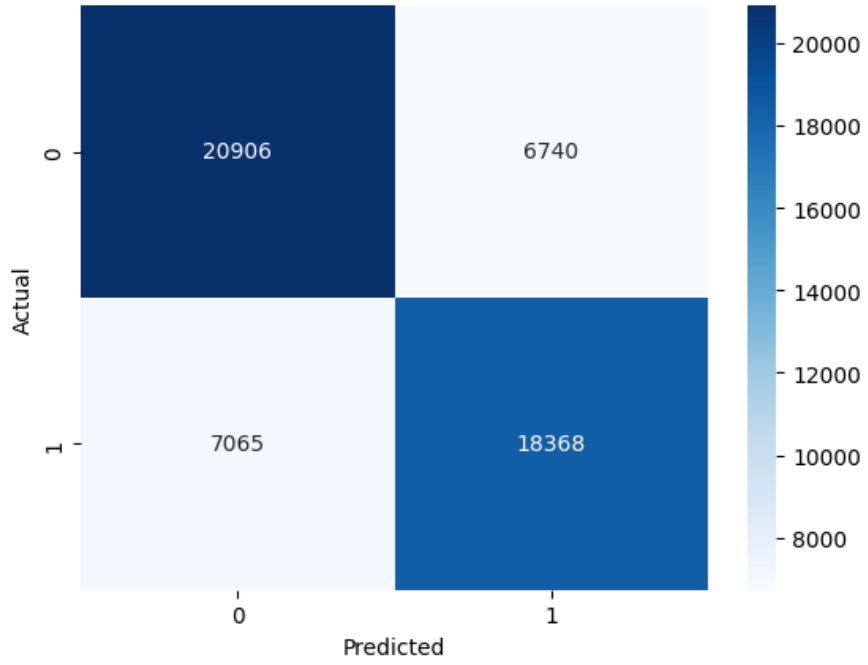
	coef	std err	z	P>|z|	[0.025	0.975]
const	2.6915	0.057	47.478	0.000	2.580	2.803
Age	-0.0779	0.012	-6.302	0.000	-0.102	-0.054
Gender	-0.5502	0.021	-25.613	0.000	-0.592	-0.508
Years at Company	-0.1547	0.012	-12.460	0.000	-0.179	-0.130
Work-Life Balance	-0.5961	0.012	-50.204	0.000	-0.619	-0.573
Job Satisfaction	0.0371	0.012	3.046	0.002	0.013	0.061
Performance Rating	-0.1781	0.014	-12.370	0.000	-0.206	-0.150
Number of Promotions	-0.2000	0.011	-18.681	0.000	-0.221	-0.179
Overtime	0.3522	0.023	15.523	0.000	0.308	0.397
Distance from Home	0.2701	0.011	25.198	0.000	0.249	0.291
Education Level	-0.1369	0.011	-12.848	0.000	-0.158	-0.116
Marital_Status_Married	-0.2689	0.031	-8.780	0.000	-0.329	-0.209
Marital_Status_Single	1.4847	0.033	44.729	0.000	1.420	1.550
Number of Dependents	-0.2171	0.011	-20.263	0.000	-0.238	-0.196
Job Level	-1.2069	0.016	-76.937	0.000	-1.238	-1.176
Remote Work	-1.7306	0.030	-57.278	0.000	-1.790	-1.671
Company Reputation	-0.3304	0.012	-28.375	0.000	-0.353	-0.308

Wyniki pokazują, że wszystkie zmienne są istotne, ponieważ wszystkie wartości p (**P>|z|**) są mniejsze niż 0.05. Dodatkowo, tabela przedstawia zmienne wraz z odpowiadającymi im współczynnikami.

Interpretacja wybranych współczynników:

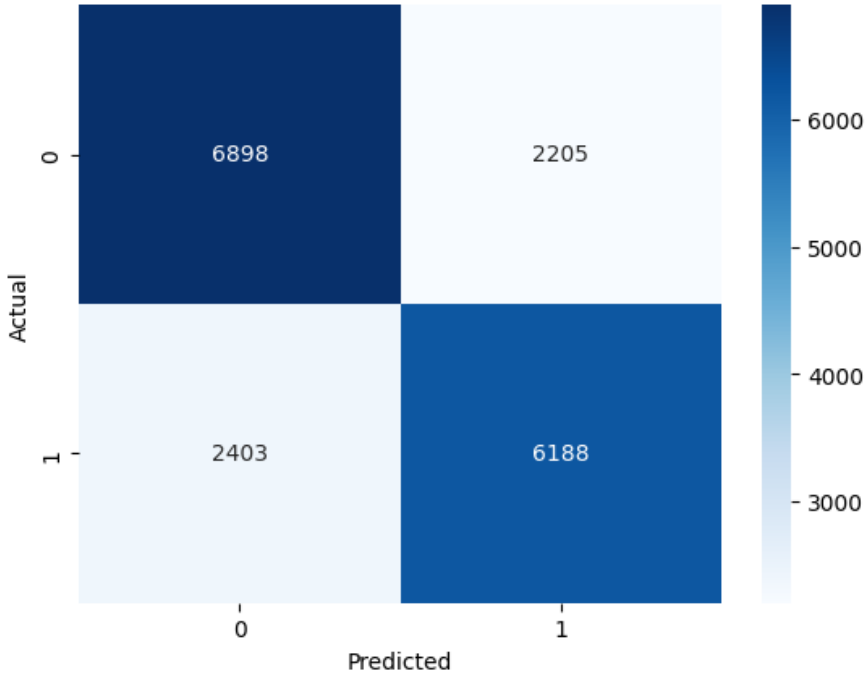
- Współczynnik dla **Remote Work** wynosi **-1.7306**. Oznacza to, że przy założeniu stałości pozostałych zmiennych, pracownicy pracujący zdalnie są mniej skłonni do opuszczenia firmy w porównaniu z tymi, którzy nie pracują zdalnie. Szanse na odejście pracownika zmniejszają się o czynnik **exp(-1.7306) ≈ 0.176** dla osób pracujących zdalnie. Pracownicy zdalni mają o około **82.4% mniejsze szanse** na odejście z firmy w porównaniu z pracownikami, którzy nie pracują zdalnie.
- Współczynnik dla **Overtime** wynosi **0.3522**. Wskazuje to, że przy założeniu stałości pozostałych zmiennych, pracownicy wykonujący nadgodziny są bardziej skłonni do opuszczenia firmy w porównaniu z tymi, którzy nie pracują

w nadgodzinach. Szanse na odejście wzrastają o czynnik **exp(0.3522) ≈ 1.423** dla pracowników pracujących nadgodzinach. Pracownicy wykonujący nadgodziny mają o około **42.3% większe szanse** na odejście z firmy w porównaniu z tymi, którzy nie pracują w nadgodzinach.



Accuracy	Specificity	Recall
0.74	0.76	0.72

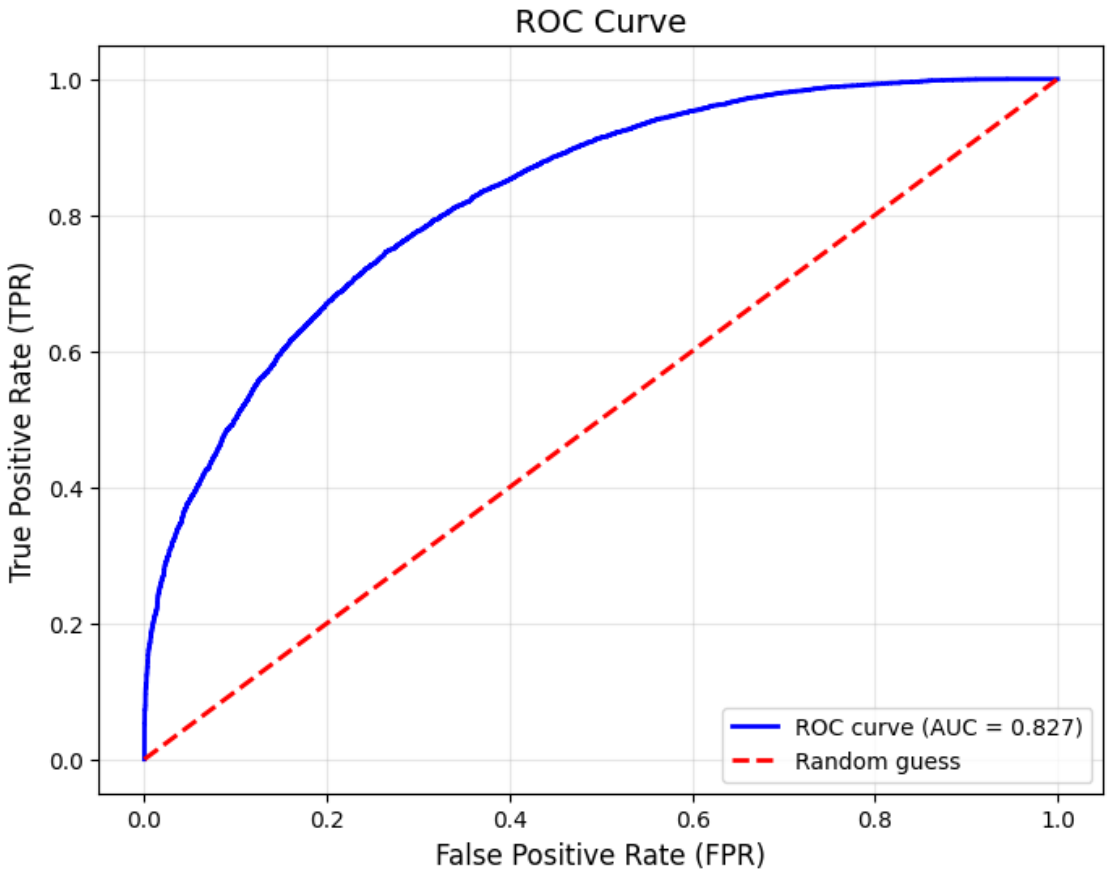
Testowanie modelu



Accuracy	Specificity	Recall
0.74	0.76	0.72

Nie zaobserwowano istotnych zmian w wartościach metryk pomiędzy zbiorami treningowym a testowym. Podobnie jak w poprzednim modelu, obserwacje należące do klasy 0 są nieco częściej poprawnie rozpoznawane przez model. Ogólnie jednak wszystkie metryki mają zbliżone wartości. Zarówno liczba fałszywie pozytywnych, jak i fałszywie negatywnych klasyfikacji jest podobna, co sugeruje, że model nie wykazuje stronniczości w żadnym kierunku.

ROC Curve



Wartość AUC wynosząca 0.827 oznacza, że model ma 82.7% prawdopodobieństwo poprawnego rozróżnienia między losowo wybraną pozytywną a losowo wybraną negatywną obserwacją.

SVM

Support Vector Machine to nadzorowany model uczenia maszynowego, często stosowany do zadań klasyfikacji, choć może być również używany do regresji. SVM działa poprzez znalezienie hiperprzestrzeni, która najlepiej oddziela klasy w przestrzeni wyższego wymiaru.

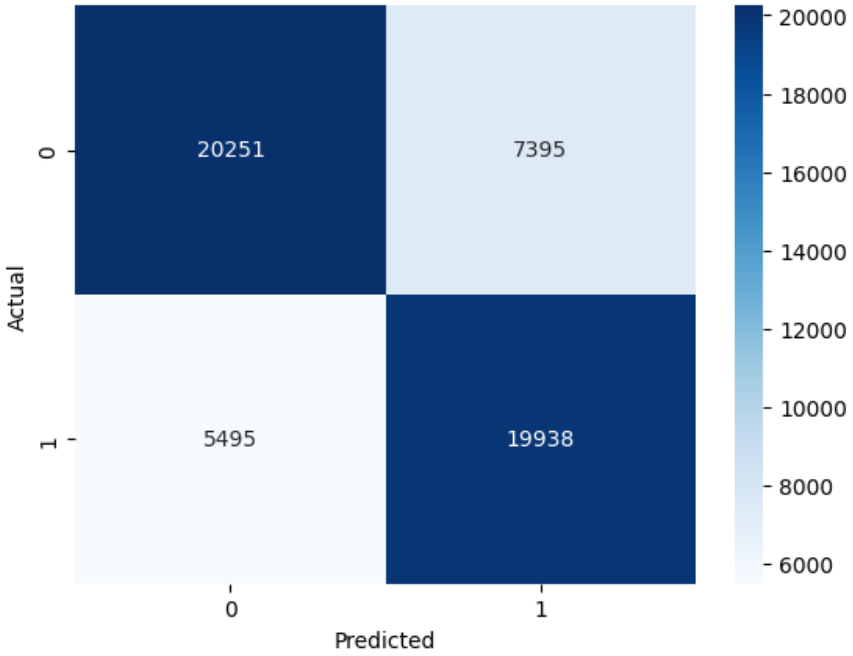
Najważniejsze hiperparametry:

- **C**: Parametr regularyzacji. Kontroluje kompromis między maksymalizacją marginesu a minimalizacją błędów klasyfikacji.
- **kernel**: Określa rodzaj jądra używanego w algorytmie.

- **degree**: Stopień wielomianu w funkcji jądra (dla jądra 'poly'). Ignorowany, gdy używane jest inne jądro.
- **gamma**: Współczynnik jądra dla jąder 'rbf', 'poly' i 'sigmoid'. Kontroluje, jak duży wpływ ma pojedynczy przykład treningowy. Niskie wartości rozpraszają wpływ każdego punktu, podczas gdy wysokie wartości lokalizują wpływ na mniejsze obszary.
- **coef0**: Niezależny składnik w funkcji jądra. Używany wyłącznie dla jąder 'poly' i 'sigmoid'. Pomaga kontrolować elastyczność granicy decyzyjnej.
- **tol**: Tolerancja dla kryterium zatrzymania. Jeśli zmiana w funkcji celu optymalizacji jest mniejsza niż ta wartość, optymalizacja zostanie przerwana.

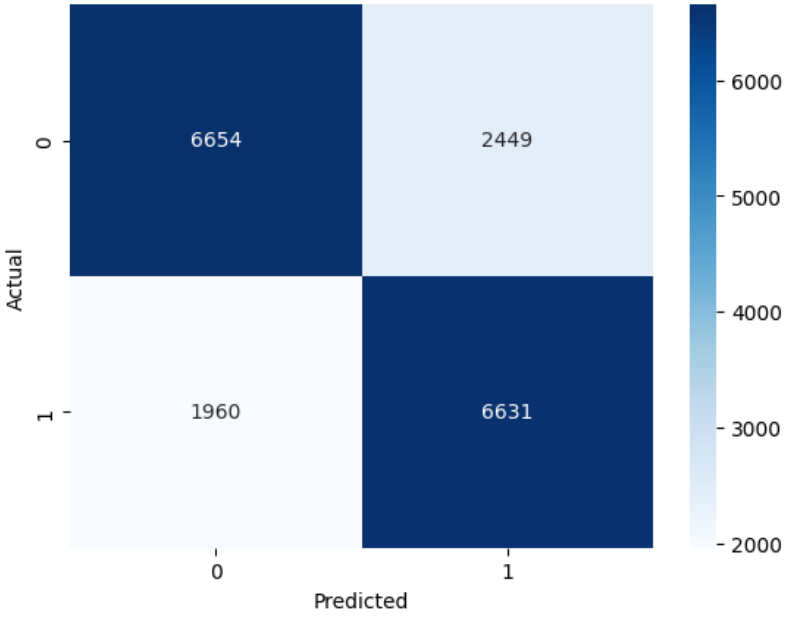
W tym przypadku użyto jądra **'poly'**, co oznacza, że algorytm zastosuje funkcję wielomianową do przekształcenia danych wejściowych. Dzięki temu model będzie w stanie lepiej oddzielać punkty danych, których nie da się oddzielić liniowo w oryginalnej przestrzeni cech.

Model z bazowymi parametrami – zbiór treningowy



Accuracy	Specificity	Recall
0.76	0.73	0.78

Model z wybranymi parametrami – zbiór testowy



Accuracy	Specificity	Recall
0.75	0.73	0.77

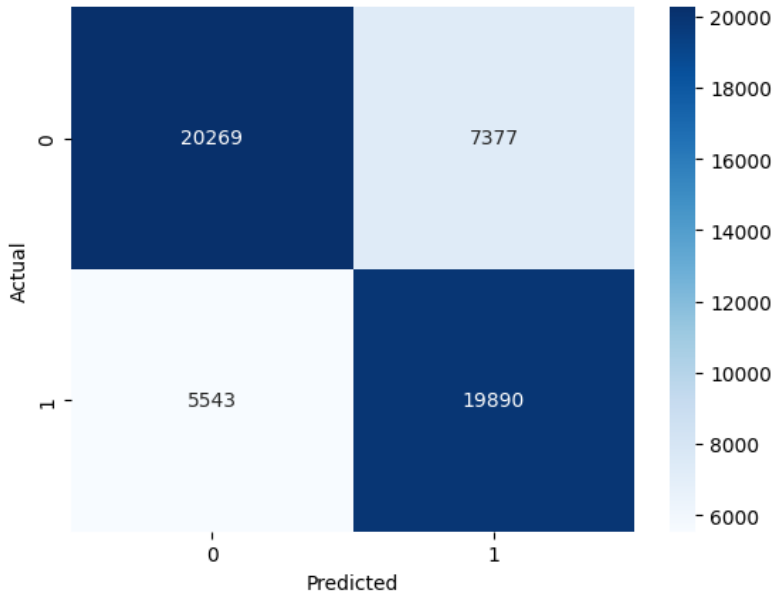
Również jak w poprzednich przypadkach różnice między zestawami danych nie są znaczące. Accuracy wynosi 0.75, co oznacza, że model poprawnie sklasyfikował 75% wszystkich obserwacji. Specificity to 0.73, co wskazuje, że model poprawnie zidentyfikował 73% obserwacji należących do klasy 0, ignorując False Positives. Recall wynosi 0.77, co oznacza, że model poprawnie sklasyfikował 77% obserwacji klasy 1, co wskazuje na wysoką zdolność modelu do wykrywania pozytywnych przypadków.

GridSearch - Najlepsze parametry

W przypadku SVM użyto mniejszej ilości kombinacji oraz 3 krotną walidację krzyżową z uwagi na dużą ilość obserwacji.

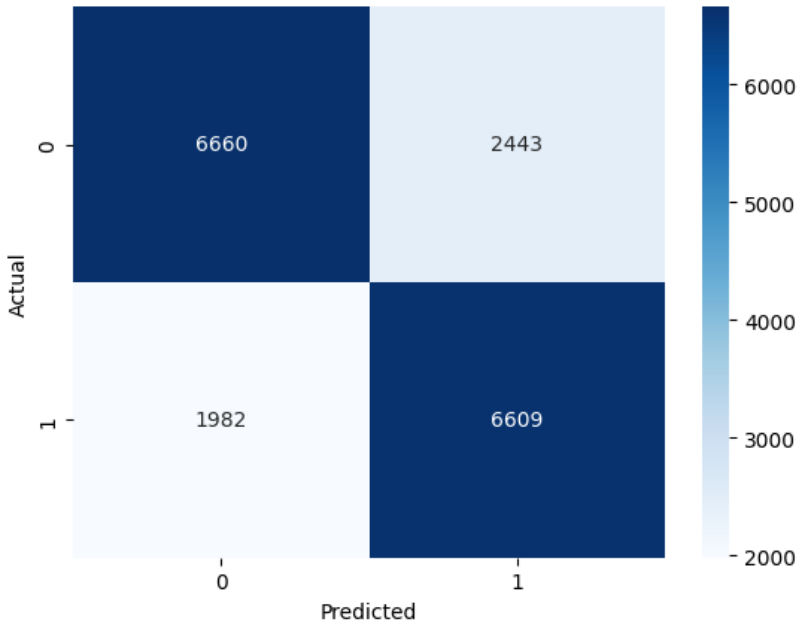
C	degree	gamma
0.1	3	0.1

GridSearch – zbiór treningowy



Accuracy	Specificity	Recall
0.76	0.73	0.78

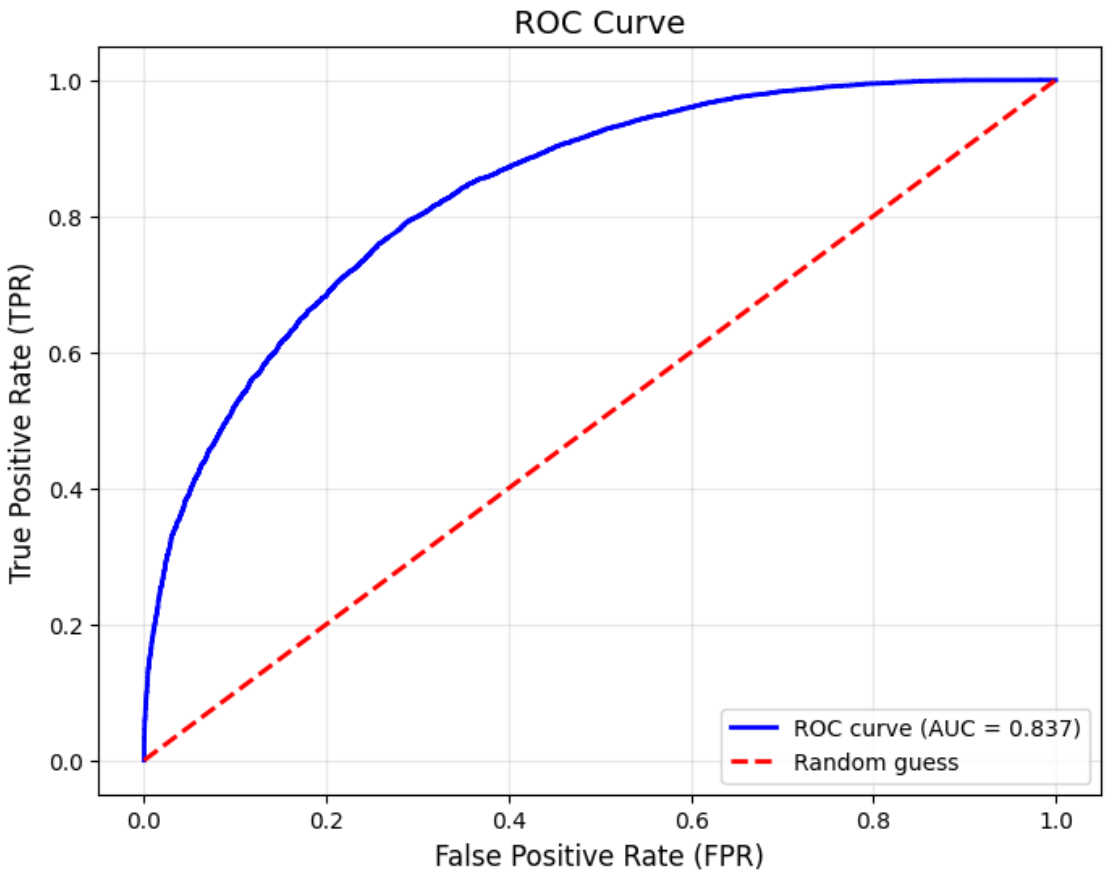
GridSearch – zbiór testowy



Accuracy	Specificity	Recall
0.75	0.73	0.77

Różnice między zestawami danych nie są znaczące. Wyniki modelu dla zbioru testowego są identyczne, co może wynikać z mniejszej liczby sprawdzonych kombinacji, ale są jak najbardziej zadowalające. Wysoki Recall (0.77) pokazuje, że model skutecznie identyfikuje obserwacje z klasą 1, natomiast Accuracy wynoszące 0.75 wskazuje na ogólną poprawność klasyfikacji. Specificity (0.73) sugeruje, że model poprawnie klasyfikuje większość obserwacji klasy 0, minimalizując liczbę fałszywych pozytywów.

ROC Curve



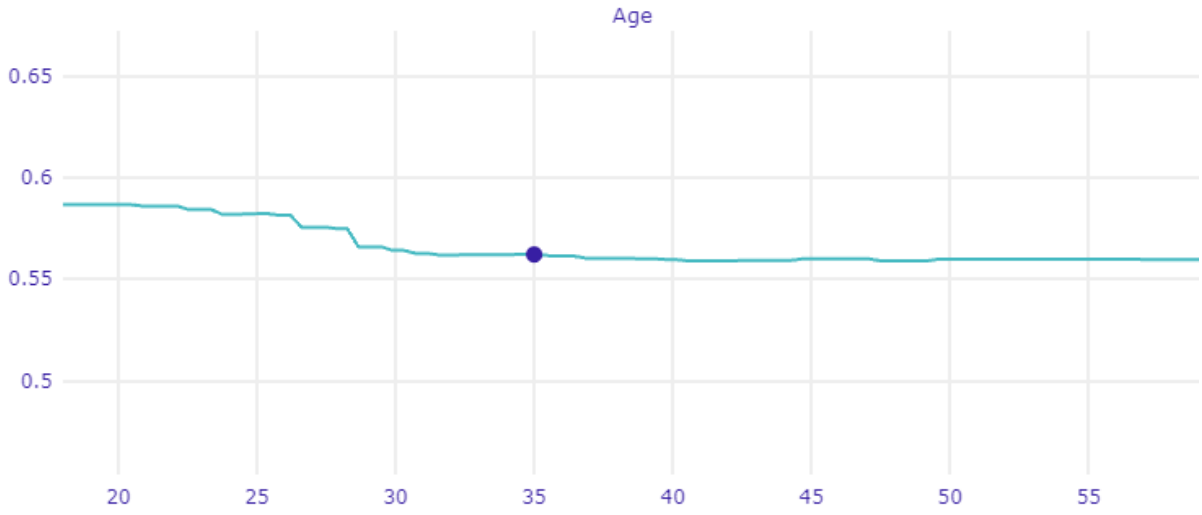
Wartość AUC wynosząca 0.837 oznacza, że model ma 83.7% prawdopodobieństwo poprawnego rozróżnienia między losowo wybraną pozytywną a losowo wybraną negatywną obserwacją.

4. Interpretacja Oraz Wyniki

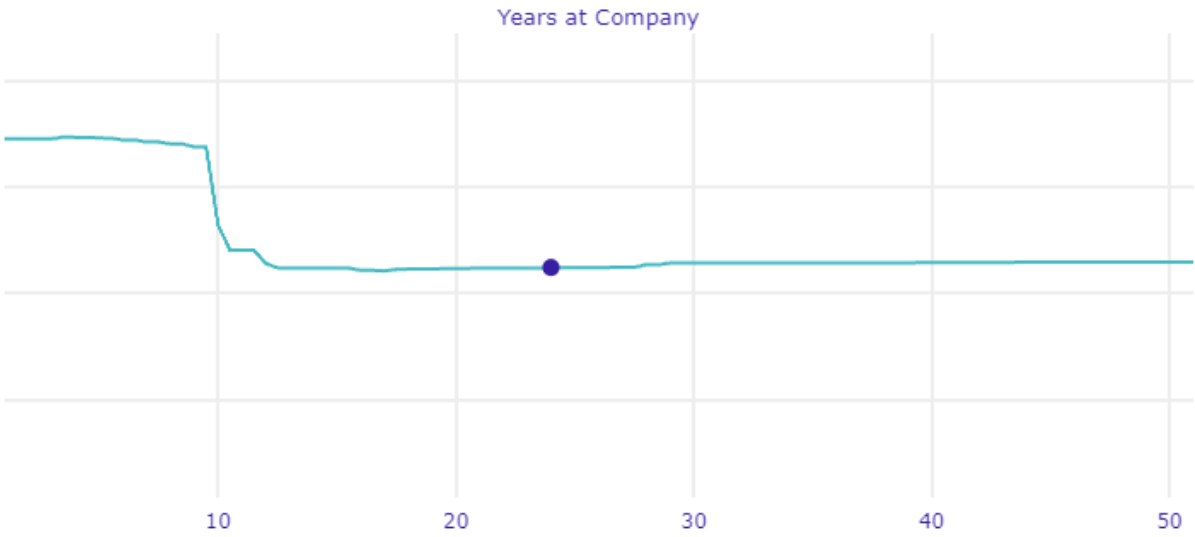
Wykresy CP

Wykresy Ceteris Paribus służą do analizy wpływu pojedynczej zmiennej na wynik modelu, przy założeniu, że pozostałe zmienne są stałe. Pomagają one w zwiększeniu interpretowalności i transparentności modeli predykcyjnych, umożliwiając zrozumienie, jak zmiana konkretnej cechy wpływa na prognozowane wyniki.

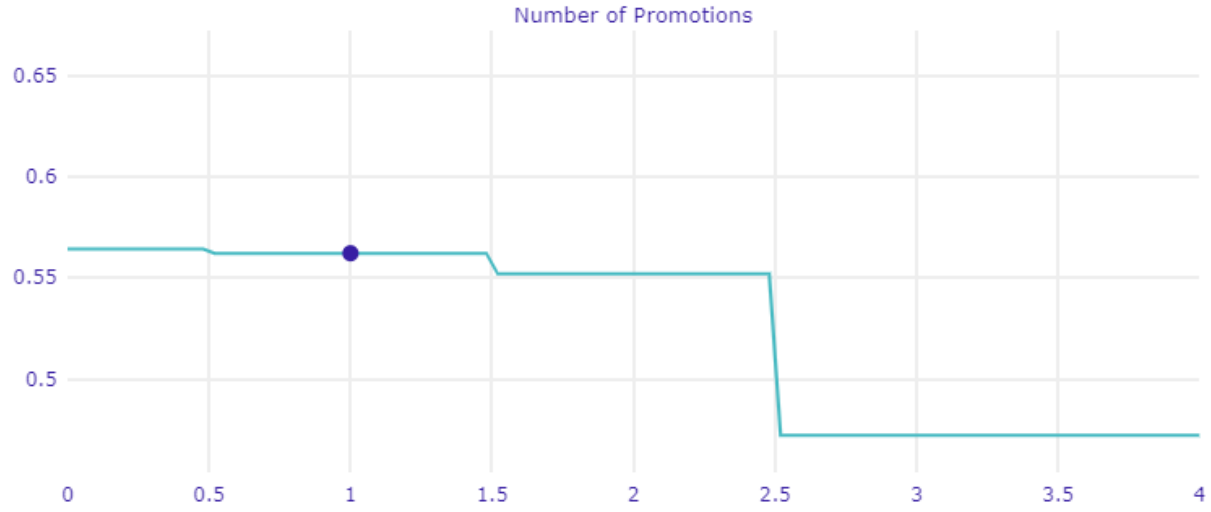
Zmienne ilościowe



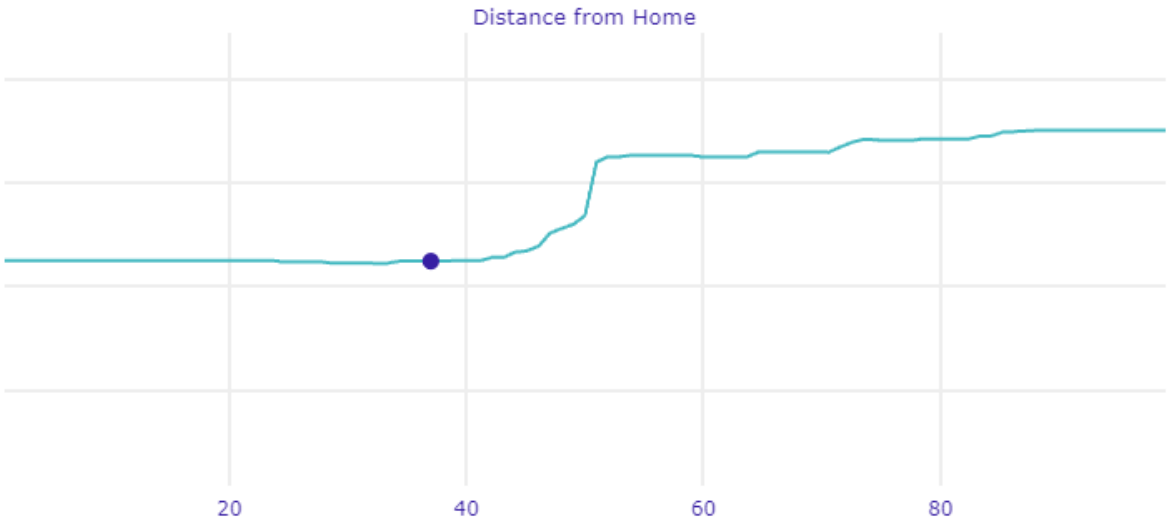
Przy założeniu stałości pozostałych zmiennych, jeśli wybrana osoba byłaby młodsza, jej prawdopodobieństwo opuszczenia firmy rośnie. W przypadku gdy była by starsza, szansa na odejście jest bardzo podobna. Jest to logiczne, ponieważ starsze osoby są zazwyczaj mniej skłonne do dokonywania zmian zawodowych.



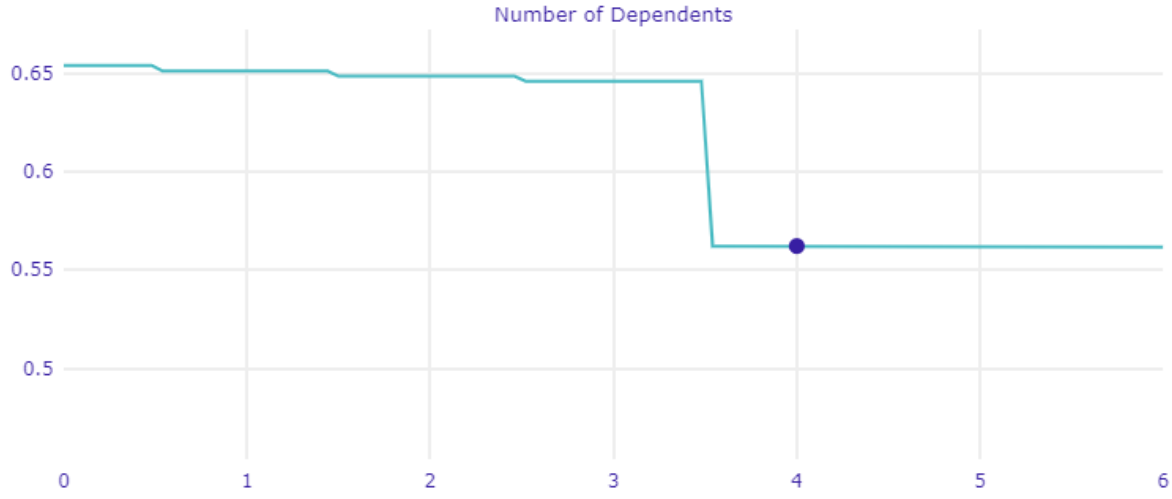
Przy założeniu stałości pozostałych zmiennych, jeśli wybrana osoba miałaby więcej lat spędzonych w firmie, jej prawdopodobieństwo opuszczenia organizacji pozostałoby relatywnie stałe. Gdyby miała staż mniejszy niż 10 lat, szansa na odejście wrosłaby o około 10 p.p.. Po 10 latach pracownik zazwyczaj ma ugruntowaną pozycję w firmie, korzysta z benefitów, a odejście wiązałoby się z większym ryzykiem i utratą przywilejów.



Przy założeniu stałości pozostałych zmiennych, jeśli wybrana osoba miałaby większą liczbę awansów, jej prawdopodobieństwo opuszczenia firmy malało by wraz z nimi. Po osiągnięciu 3 lub więcej awansów prawdopodobieństwo to wyraźnie maleje. Podobnie jak w poprzednich przypadkach, większa liczba awansów oznacza, że osoba ma silną pozycję w firmie, a jej odejście wiązałoby się z większym ryzykiem.

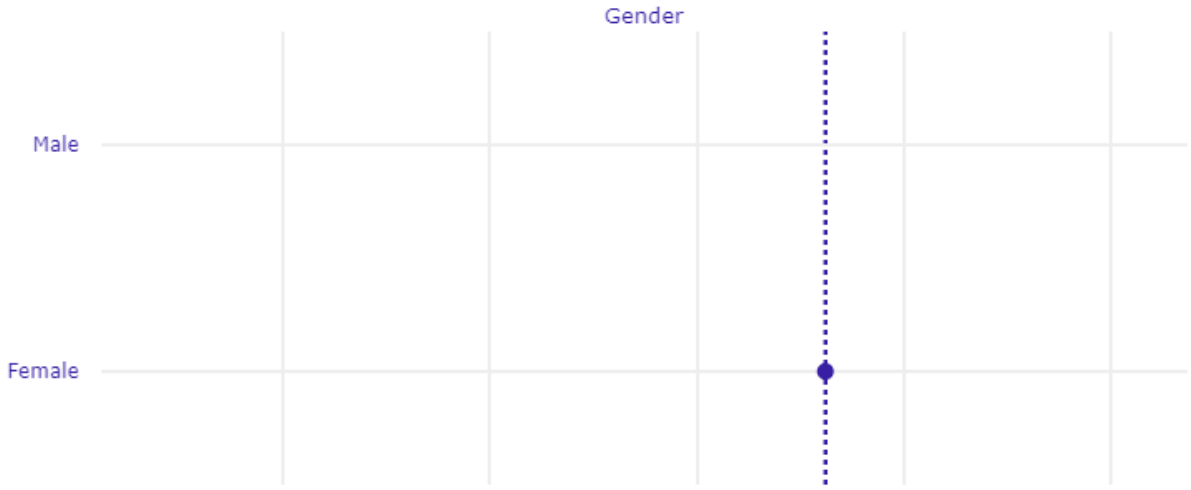


Przy założeniu stałości pozostałych zmiennych, jeśli wybrana osoba miałaby większą odległość od miejsca zamieszkania do pracy, jej prawdopodobieństwo opuszczenia firmy rosło by wraz z nią. Dla mniejszych odległości jest podobne.

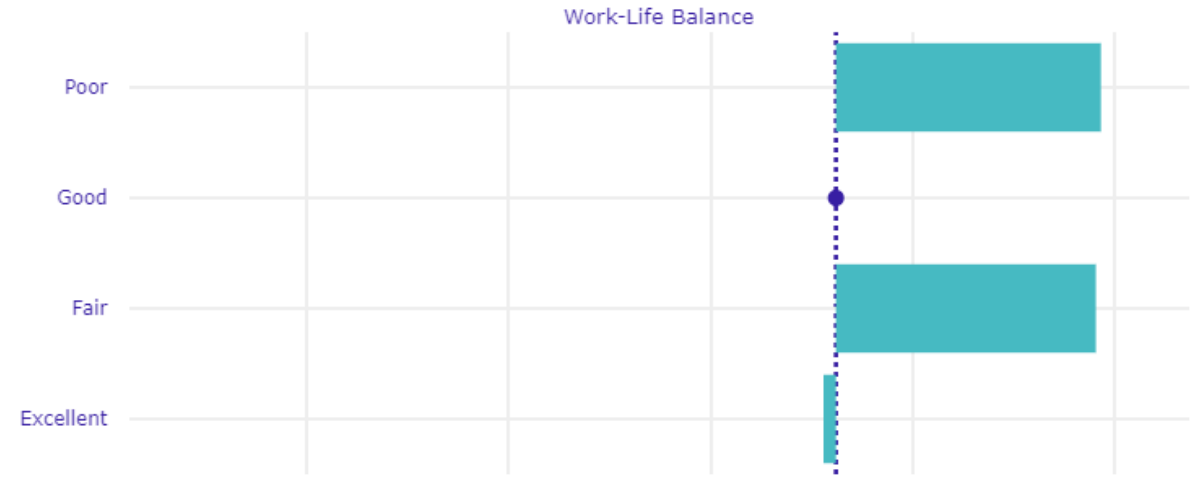


Przy założeniu stałości pozostałych zmiennych, jeśli wybrana osoba miałaby większą liczbę osób, za które odpowiada, jej prawdopodobieństwo opuszczenia firmy raczej by się nie zmieniło znacząco. Dla mniejszych wartości wzrosło by o około 10p.p..

Zmienne jakościowe



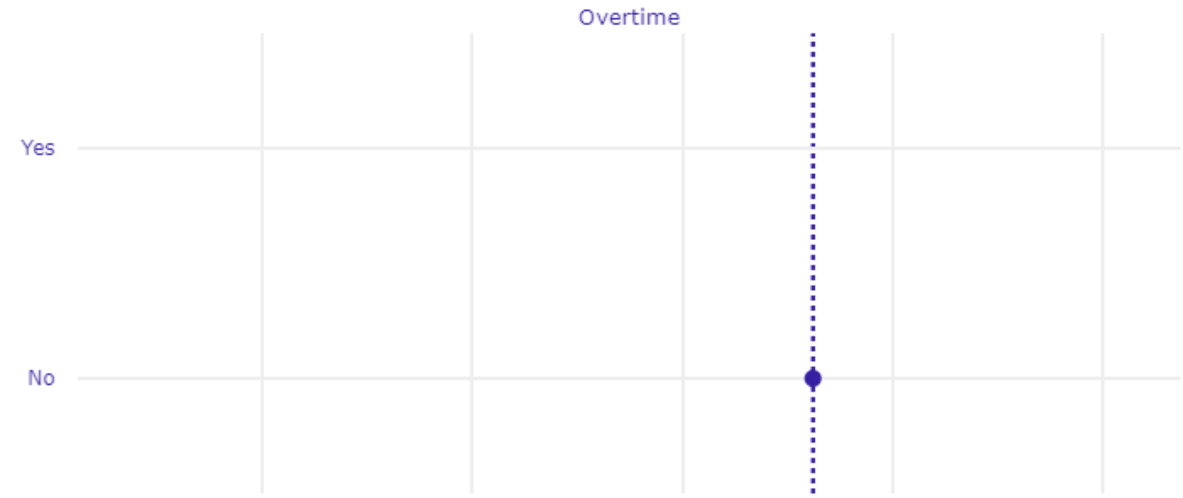
Przy założeniu stałości pozostałych zmiennych, płeć osoby nie ma znaczenia dla prawdopodobieństwa opuszczenia firmy.



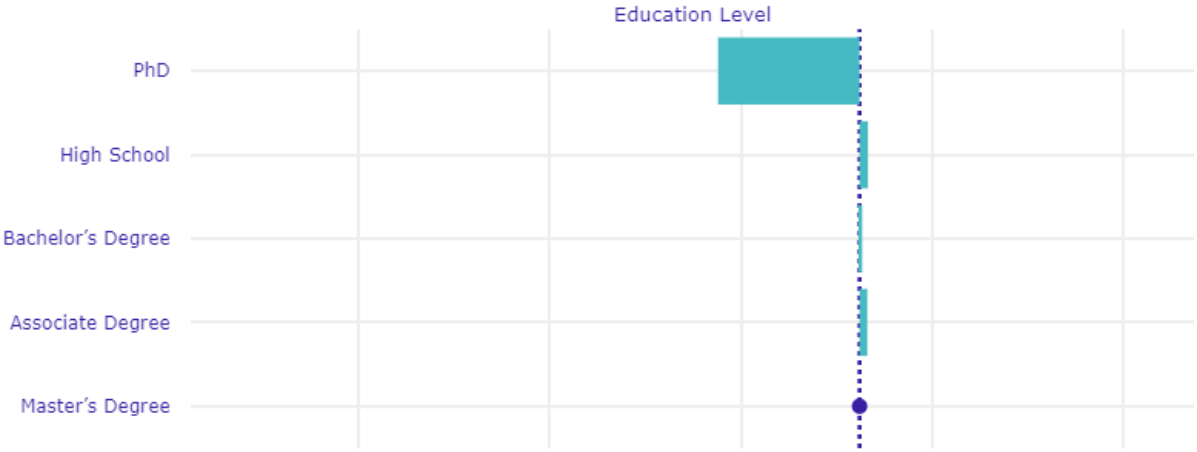
Przy założeniu stałości pozostałych zmiennych, jeśli wybrana osoba oceniłaby równowagę między życiem zawodowym a prywatnym jako "Poor" lub "Fair", jej prawdopodobieństwo opuszczenia firmy znacząco by rosło. Przy ocenie "Excellent" prawdopodobieństwo to delikatnie maleje. Im lepsze warunki w firmie, tym pracownik ma więcej argumentów, by nie odchodzić z organizacji.

Przy założeniu stałości pozostałych zmiennych, jeśli wybrana osoba miałaby ocenę satysfakcji z pracy "Low", jej prawdopodobieństwo opuszczenia firmy subtelnie by rosło. Dla oceny "Very High" prawdopodobieństwo to delikatnie rośnie, a przy ocenie "High" subtelnie maleje. Ciężko jest wytłumaczyć, dlaczego osoba bardzo zadowolona miała by chętniej opuścić firmę, być może jest to kwestia tej konkretnej obserwacji.

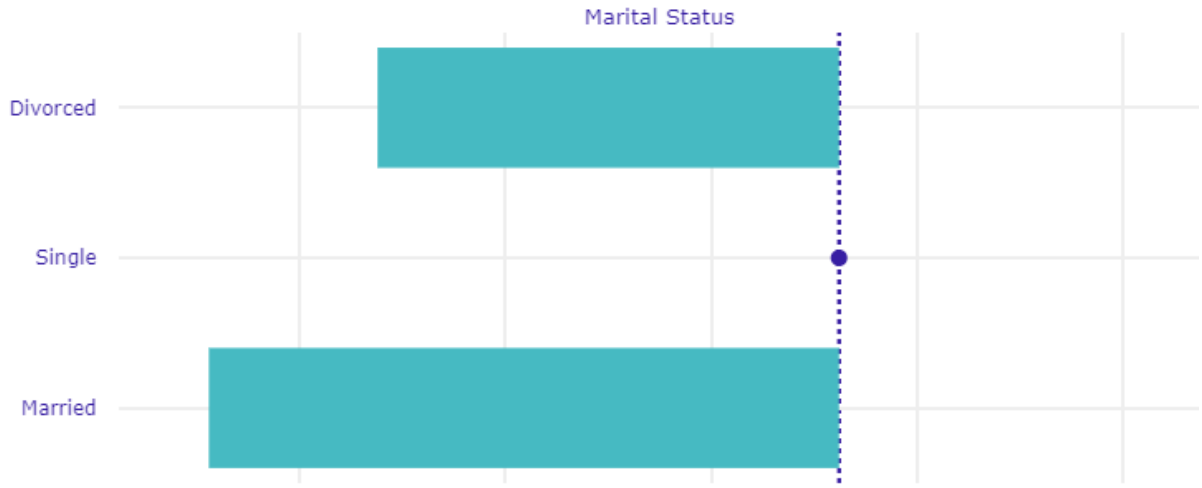
Przy założeniu stałości pozostałych zmiennych, dla oceny wydajności, jeśli osoba miała by niższe oceny to szansa na to, że odejdzie delikatnie rośnie.



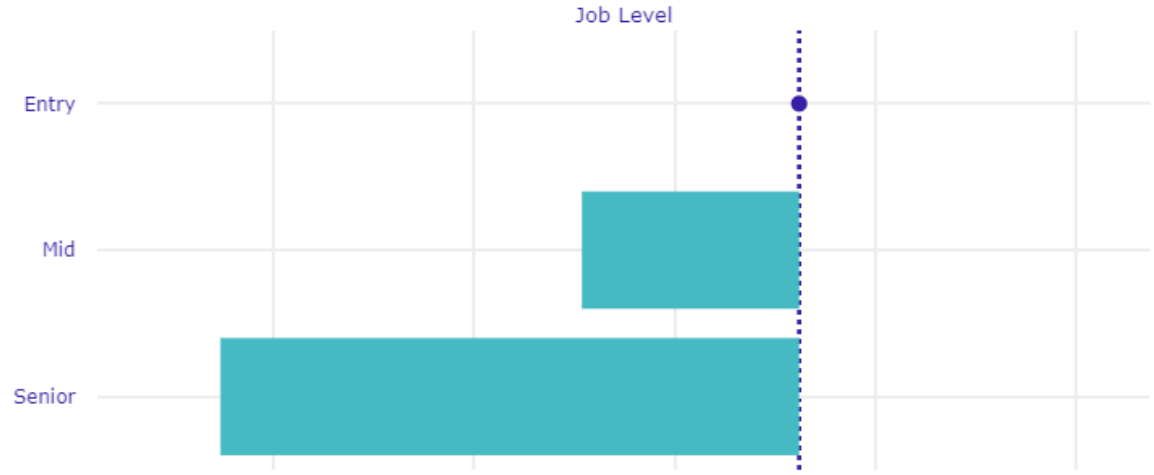
Przy założeniu stałości pozostałych zmiennych, nie ma znaczenia czy wybrana osoba robi nadgodziny.



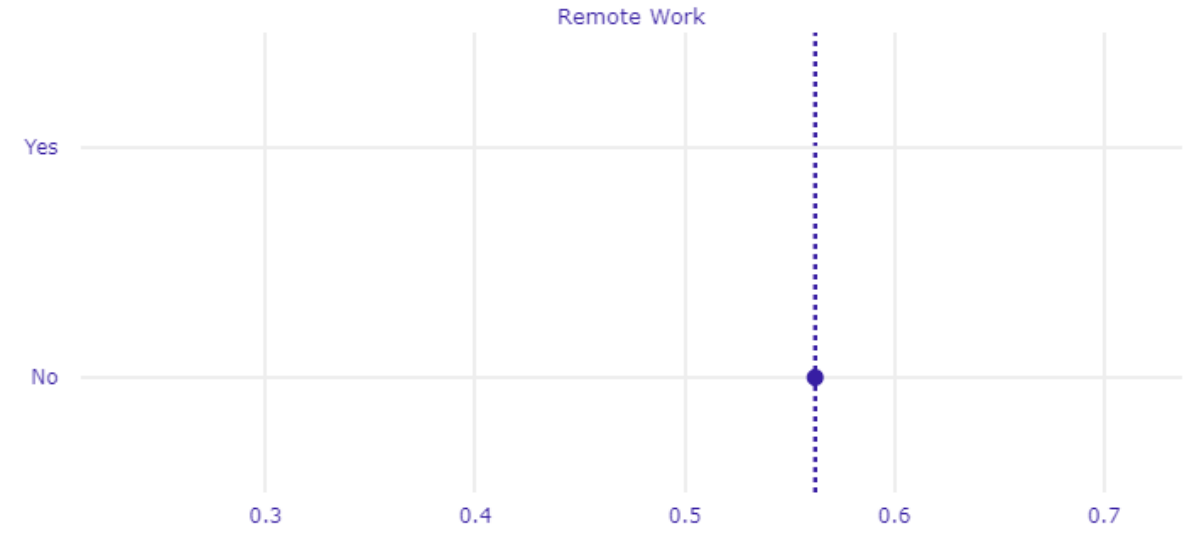
Przy założeniu stałości pozostałych zmiennych, jeśli wybrana osoba miałaby wykształcenie doktorskie, jej prawdopodobieństwo opuszczenia firmy znacząco by zmalało, podczas gdy w przypadku niższego poziomu wykształcenia prawdopodobieństwo to by delikatnie rosło.



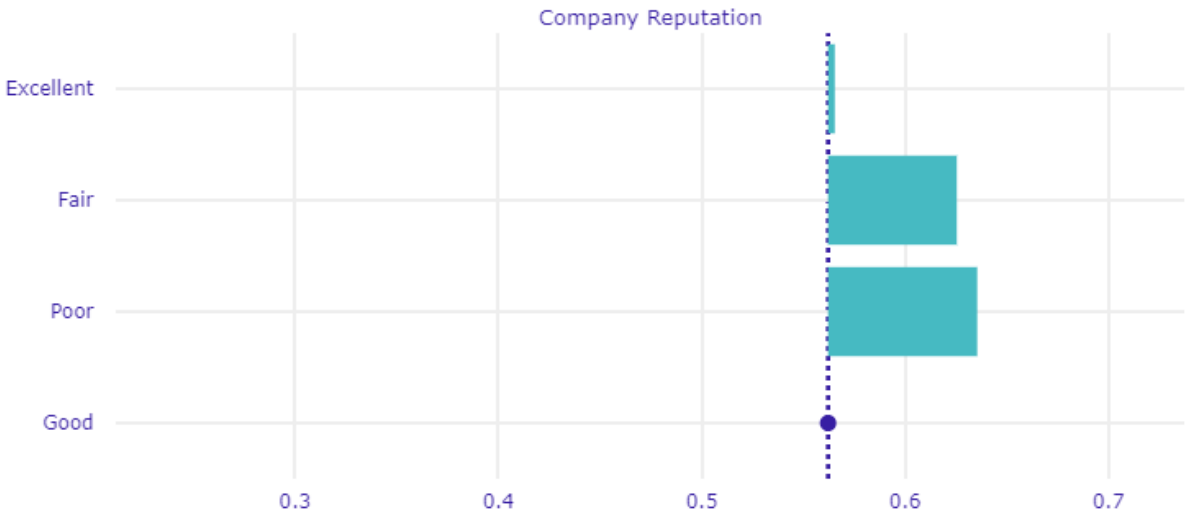
Przy założeniu stałości pozostałych zmiennych, stan cywilny ma wpływ na prawdopodobieństwo opuszczenia firmy. Jeśli wybrana osoba była by rozwiedziona lub zamężna/żonata prawdopodobieństwo opuszczenia firmy znacząco maleje.



Przy założeniu stałości pozostałych zmiennych, gdyby wybrana osoba była na stanowisku średnim lub wyższym, jej prawdopodobieństwo opuszczenia firmy znacznie by zmalało, szczególnie w przypadku stanowiska Senior. Wyższe stanowisko to lepsza pozycja w firmie.



Przy założeniu stałości pozostałych zmiennych, nie ma znaczenia czy wybrana osoba pracowałaby zdalnie.



Przy założeniu stałości pozostałych zmiennych, jeśli wybrana osoba pracowała by w firmie z średnią oraz słabą reputacją, szansa na to, że odejdzie się zwiększa. Co ciekawe, w przypadku bardzo dobrej reputacji, prawdopodobieństwo odejścia również delikatnie rośnie.

Porównanie Wyników

Trening – najlepsze wyniki

Model	Accuracy	Specificity	Recall
Random Forest	0.723450	0.760399	0.683285
Logistic Regression	0.739916	0.756203	0.722211
SVM	0.756589	0.733162	0.782055

Dla zbioru treningowego najlepsze wyniki przypadły modelowi SVM, który ma najwyższą Accuracy oraz Recall. Pomimo najniższego Specificity nie odbiega znacząco od reszty. Następna w kolejności wydaje się być Logistic Regression, z wynikami delikatnie lepszymi od ostatniego Random Forest.

Różnice między najlepszym a najgorszym modelem:

- Accuracy: ~ 3%
- Specificity: ~ -3%
- Recall: ~ 10%

Test - najlepsze wyniki

Model	Accuracy	Specificity	Recall	AUC
Random Forest	0.720414	0.764034	0.674194	0.802
Logistic Regression	0.739573	0.757772	0.720289	0.827
SVM	0.749915	0.731627	0.769293	0.837

Model SVM osiągnął najwyższe wartości w 3/4 metrykach, co oznacza, że jest on najbardziej skuteczny w rozróżnianiu różnych klas w danych. Wysoka wartość AUC wskazuje na dobrą zdolność SVM do odróżniania obserwacji należących do różnych klas, a wysoka dokładność, specyficzność i czułość potwierdzają jego ogólną skuteczność. Drugi w kolejności tak samo jest Logistic Regression a na końcu Random Forest.

Różnice między najlepszym a najgorszym modelem:

- Accuracy: ~ 3%
- Specificity: ~ -3%
- Recall: ~ 10%
- AUC: ~3.5%

5. Wnioski

W ramach realizacji projektu udało się w pełni spełnić założenia oraz wymagania określone na początku. Na etapie wstępnym dokładnie opisano dane i zmienne, które były przedmiotem analizy, oraz dokonano ich wizualizacji. Następnie przygotowano dane do analizy, co obejmowało oczyszczanie danych oraz kodowanie zmiennych kategorycznych. Dzięki tym krokom dane stały się gotowe do dalszego przetwarzania i wykorzystania w modelach predykcyjnych.

W projekcie wykonano predykcje z użyciem trzech różnych modeli predykcyjnych, z uwzględnieniem optymalizacji hiperparametrów w celu uzyskania jak najlepszych wyników. Po przeprowadzeniu optymalizacji porównano wyniki tych modeli, analizując ich skuteczność na podstawie wybranych metryk oceny.

Na końcowym etapie podjęto próbę interpretacji wyników modelu Random Forest za pomocą narzędzia Ceteris Paribus Profiles. Dzięki tej metodzie możliwe było dokładniejsze zrozumienie wpływu poszczególnych zmiennych na przewidywania modelu.

Podsumowując, wszystkie założenia projektowe zostały zrealizowane, a wyniki zostały dokładnie przeanalizowane i porównane, co pozwoliło na pełne zrozumienie mechanizmów predykcji oraz ich zastosowanie w analizowanej dziedzinie.