

|   |    |
|---|----|
| Wstęp .....                                   | 2  |
| Opis i wstępna analiza danych .....           | 2  |
| Przygotowanie danych .....                    | 8  |
| Podział na zbiór uczący i testowy .....       | 8  |
| Zastosowanie modeli uczenia maszynowego ..... | 8  |
| Analiza interpretowalności.....               | 13 |
| Podsumowanie .....                            | 24 |

## Wstęp

W dzisiejszym dynamicznym środowisku biznesowym, skomplikowane decyzje personalne, takie jak zwolnienie pracownika, stają się nieodłącznym elementem zarządzania zasobami ludzkimi. Organizacje zmagają się z wyzwaniem efektywnej oceny potencjalnego ryzyka zwolnienia pracownika, przy jednoczesnym minimalizowaniu zakłóceń w strukturze zespołów. W tym kontekście, zastosowanie modeli uczenia maszynowego jako narzędzi wsparcia podejmowania decyzji zyskuje na znaczeniu.

Projekt, skupia się na zastosowaniu modeli uczenia maszynowego w analizie szans na zwolnienie pracownika w organizacji. Celem tego badania jest stworzenie precyzyjnego i skutecznego narzędzia prognostycznego, które wspomoże kadry zarządzające w identyfikacji pracowników narażonych na ryzyko zwolnienia. Analiza opiera się na różnorodnych zmiennych, takich jak poziom edukacji, staż pracy, miejsce zamieszkania, poziom wynagrodzenia, wiek, płeć oraz doświadczenie w bieżącej dziedzinie.

Przez zastosowanie modeli uczenia maszynowego, takich jak SVM czy drzewa decyzyjne, planowane jest dostarczyć narzędzie, które nie tylko zidentyfikuje potencjalne ryzyko zwolnienia pracownika, ale również umożliwi zrozumienie kluczowych czynników wpływających na tę decyzję. W efekcie, organizacje będą miały możliwość skoncentrowania swoich działań na zindywidualizowanych strategiach retencyjnych oraz lepszym zarządzaniu zasobami ludzkimi, co przyczyni się do zwiększenia stabilności kadrowej oraz efektywności organizacyjnej.

## Opis i wstępna analiza danych

W projekcie został użyty zestaw danych Employee pobrany ze strony kaggle.com. Zawarte w nim są następujące dane:

Education – opisuje stopień naukowy pracownika

Joining Year – mówi o tym, w którym roku pracownik dołączył do firmy

City – miasto, w którym pracuje pracownik

Payment Tier – poziom wynagrodzenia pracownika, gdzie 3 to najmniejsza płaca a 1 najwyższa

Age – wiek pracownika

Gender – płeć pracownika

Ever Benched – informuje, czy pracownik był kiedyś bez przypisanej mu pracy

Experience in Current Domain – informuje o ilości lat, które pracownik spędził w branży

Leave or Not – jest to podjęta decyzja czy pracownik został zwolniony (1) czy pozostał w firmie (0)

Większość danych jest kategoryczna, więc zostanie opisany ich rozkład. Jedyną daną numeryczną jest wiek.

|            | Bachelors | Masters | PHD |
|------------|-----------|---------|-----|
| Education: | 3601      | 873     | 179 |

zdecydowaną większość stanowią pracownicy z wykształceniem licencjat/inżynier. Intuicyjnie jest mniej ludzi z magistrem, a jeszcze mniej ludzi z wykształceniem doktorskim.

|               | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 |
|---------------|------|------|------|------|------|------|------|
| Joining Year: | 504  | 669  | 699  | 781  | 525  | 1108 | 367  |

w bazie są pracownicy, którzy zostali przyjmowani do pracy w formie w latach 2012-2018. Najwięcej pracowników zostało przyjętych w 2017, a najmniej w 2018.

Age

|         |        |
|---------|--------|
| Min.    | :22.00 |
| 1st Qu. | :26.00 |
| Median  | :28.00 |
| Mean    | :29.39 |
| 3rd Qu. | :32.00 |
| Max.    | :41.00 |

Age: najmłodszy pracownik ma 22 lata, a najstarszy 41.

| Female | Male |
|--------|------|
| 1875   | 2778 |

Gender: w firmie jest więcej mężczyzn od kobiet.

| No   | Yes |
|------|-----|
| 4175 | 478 |

Ever Benched: zdecydowanej mniejszości bo tylko 478 pracownikom zdarzyło się być kiedyś bez przypisanej im pracy.

| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
|---|---|---|---|---|---|---|---|

|                               |     |     |      |     |     |     |   |   |
|-------------------------------|-----|-----|------|-----|-----|-----|---|---|
| Experience in Current Domain: | 355 | 558 | 1087 | 786 | 931 | 919 | 8 | 9 |
|-------------------------------|-----|-----|------|-----|-----|-----|---|---|

lata doświadczeń w branży rozciągają się od 0 do 7. Najwięcej osób ma 2 lata doświadczenia, a najmniej ci którzy mają go 6 i 7 lat.

| 0 | 1 |
|---|---|
|---|---|

|               |      |      |
|---------------|------|------|
| Leave or Not: | 3053 | 1600 |
|---------------|------|------|

1600 osób zostało zwolnionych a 3053 pozostało na swoim stanowisku. Jest to zmienna prognozowana, więc można od razu zauważać, że jest zbalansowana, ponieważ żadna z wartości w znaczący sposób nie przyćmiewa drugiej.

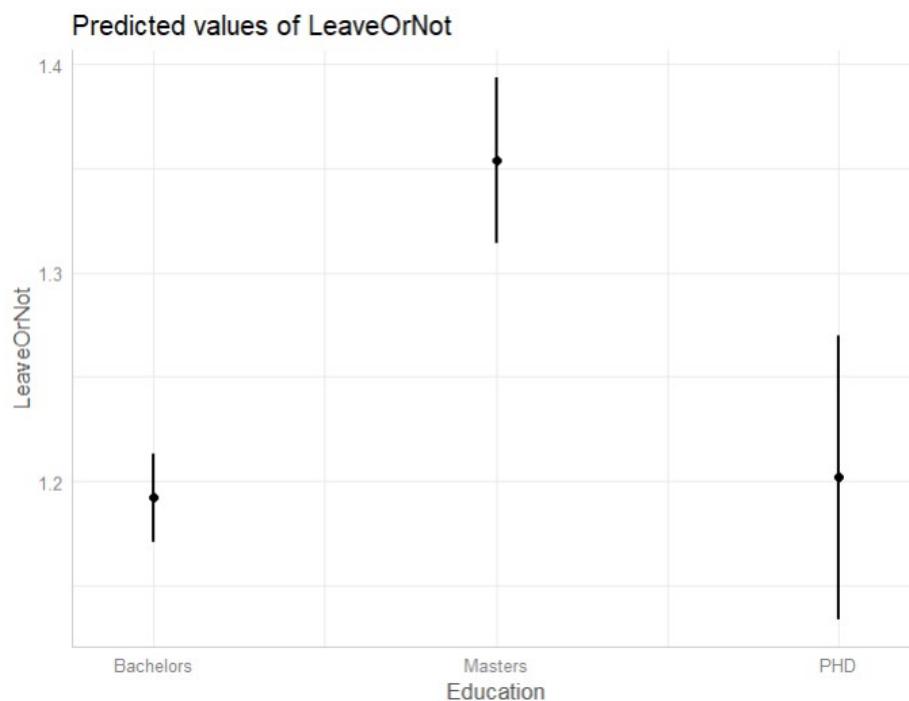
Teraz jest pokazany wpływ poszczególnych zmiennych na zmienną prognozowaną poprzez użycie modelu regresji liniowej.

Coefficients:

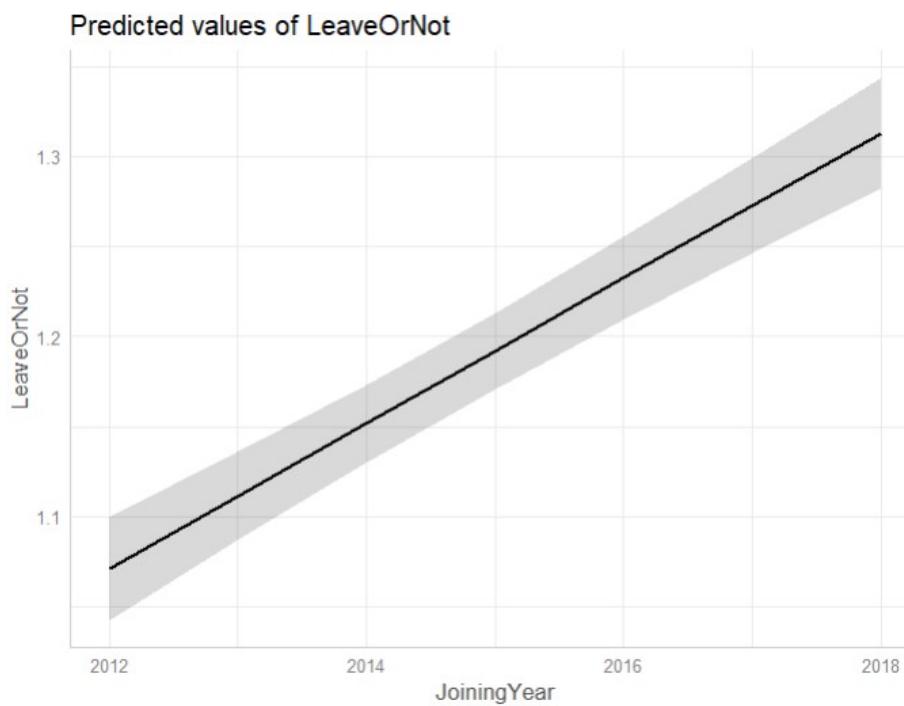
|                           | Estimate   | Std. Error | t value | Pr(> t )     |
|---------------------------|------------|------------|---------|--------------|
| (Intercept)               | -79.472684 | 7.130738   | -11.145 | < 2e-16 ***  |
| EducationMasters          | 0.161925   | 0.018706   | 8.656   | < 2e-16 ***  |
| EducationPHD              | 0.009939   | 0.034371   | 0.289   | 0.7725       |
| JoiningYear               | 0.040328   | 0.003538   | 11.398  | < 2e-16 ***  |
| CityNew Delhi             | -0.108793  | 0.018250   | -5.961  | 2.69e-09 *** |
| CityPune                  | 0.146567   | 0.016268   | 9.009   | < 2e-16 ***  |
| PaymentTier               | -0.074283  | 0.012404   | -5.988  | 2.28e-09 *** |
| Age                       | -0.005505  | 0.001345   | -4.092  | 4.35e-05 *** |
| GenderMale                | -0.194297  | 0.013740   | -14.141 | < 2e-16 ***  |
| EverBenchedYes            | 0.115068   | 0.021244   | 5.417   | 6.38e-08 *** |
| ExperienceInCurrentDomain | -0.008565  | 0.004165   | -2.057  | 0.0398 *     |

Po zbudowaniu modelu można zauważać, że wszystkie wypisane wartości są znaczące dla prognozowania wartości docelowej poza wykształceniem doktorskim i doświadczeniem w danej branży (może się to wiązać z tym, że są w nich małe ilości danych). Każda ze zmiennych zostanie jeszcze osobna sprawdzona pod względem wpływów jej konkretnych wartości na Leave or Not.

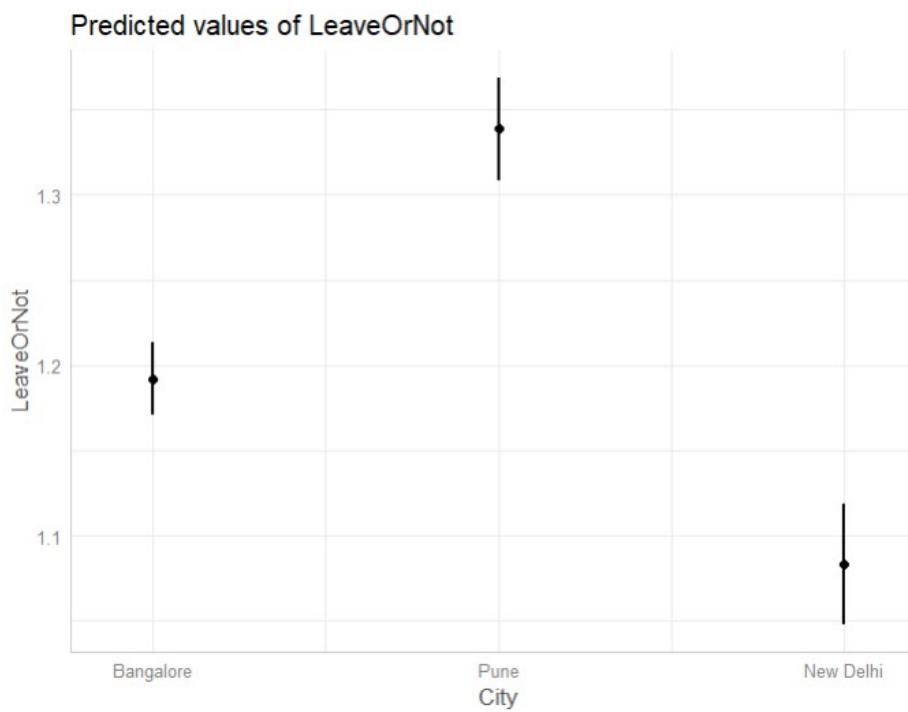
Education: Najwięcej ludzi zostaje zwolnionych z tytułem magistra.



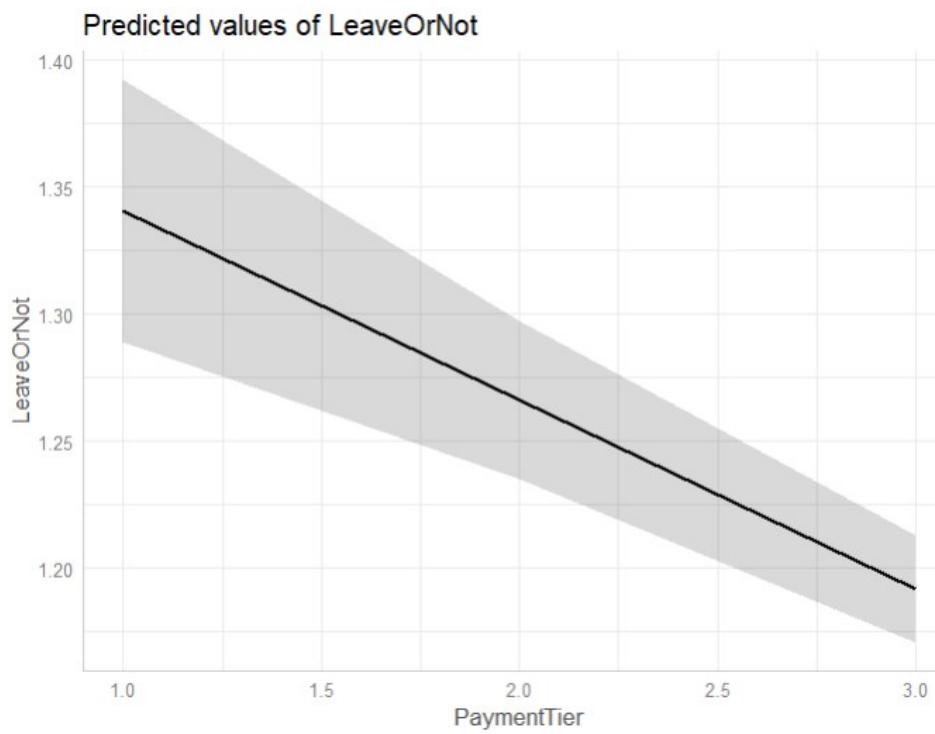
Joining Year: im później ktoś zaczął pracę w formie tym większa szansa, że zostanie zwolniony.



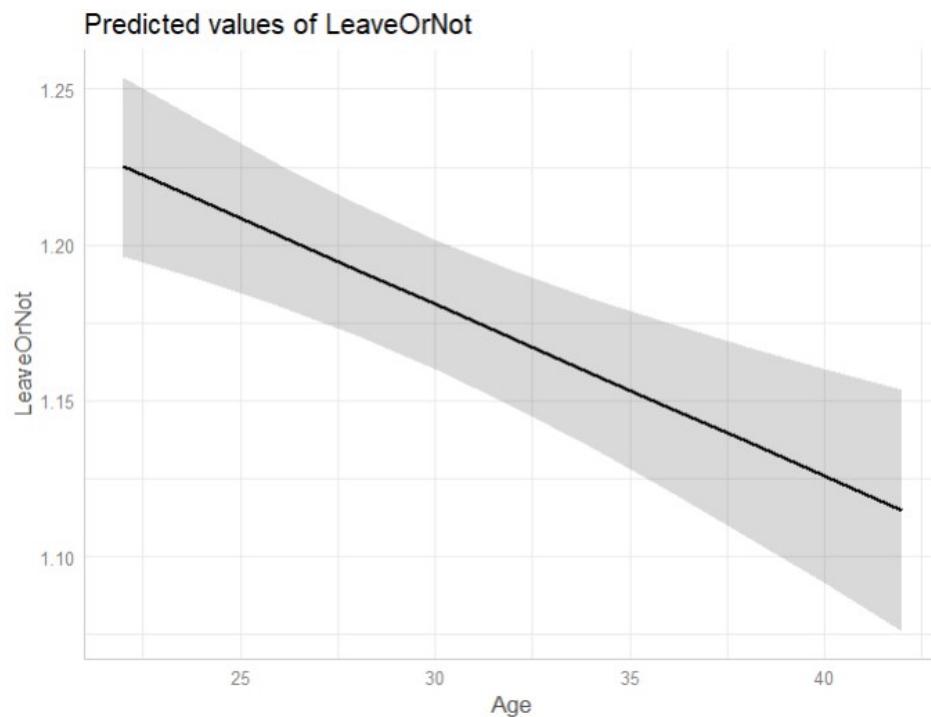
City: pracownicy z Pune mają największą szansę na bycie zwolnionymi.



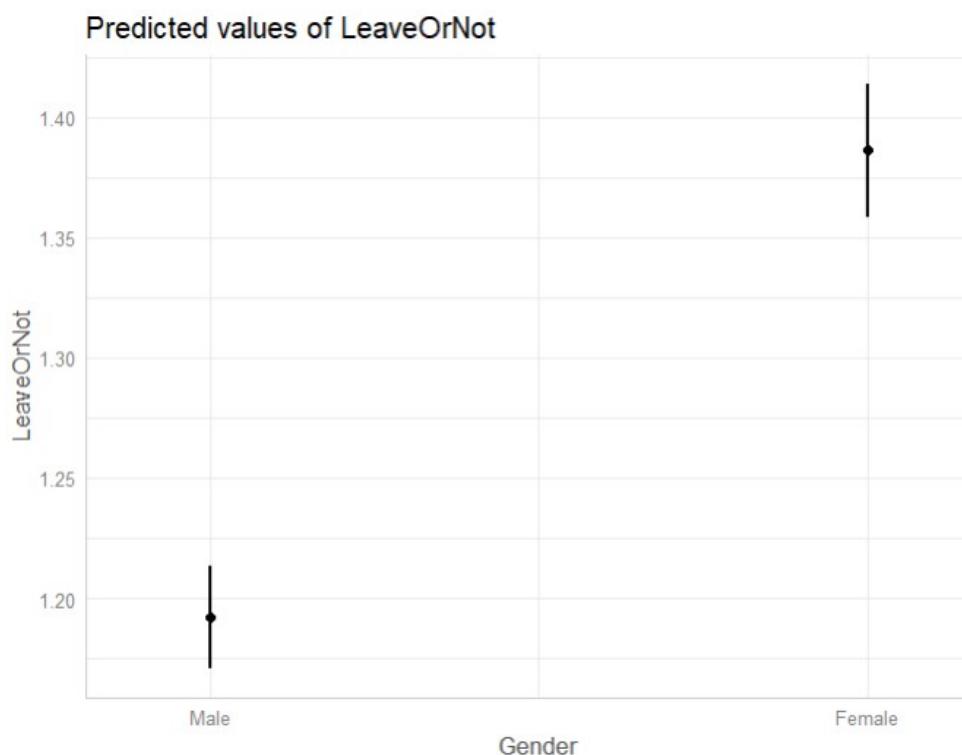
Payment Tier: pracownicy z wyższym wynagrodzeniem mają większą szansę na zostanie zwolnionymi niż ci z mniejszym.



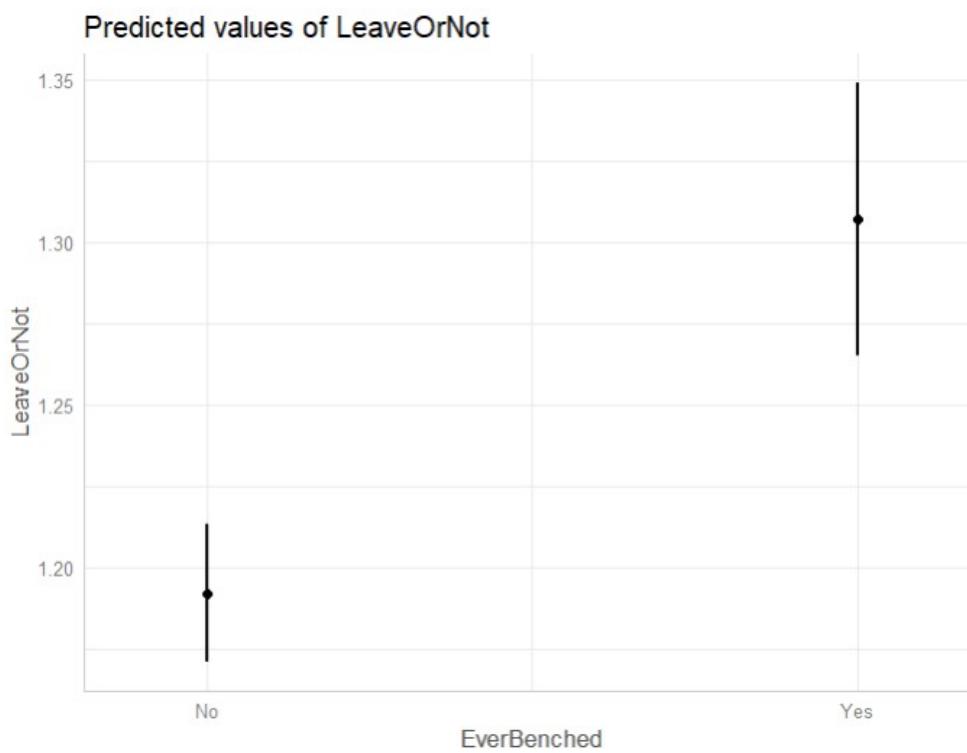
Age: młodzi pracownicy mają większą szansę na bycie zwolnionym niż starsi.



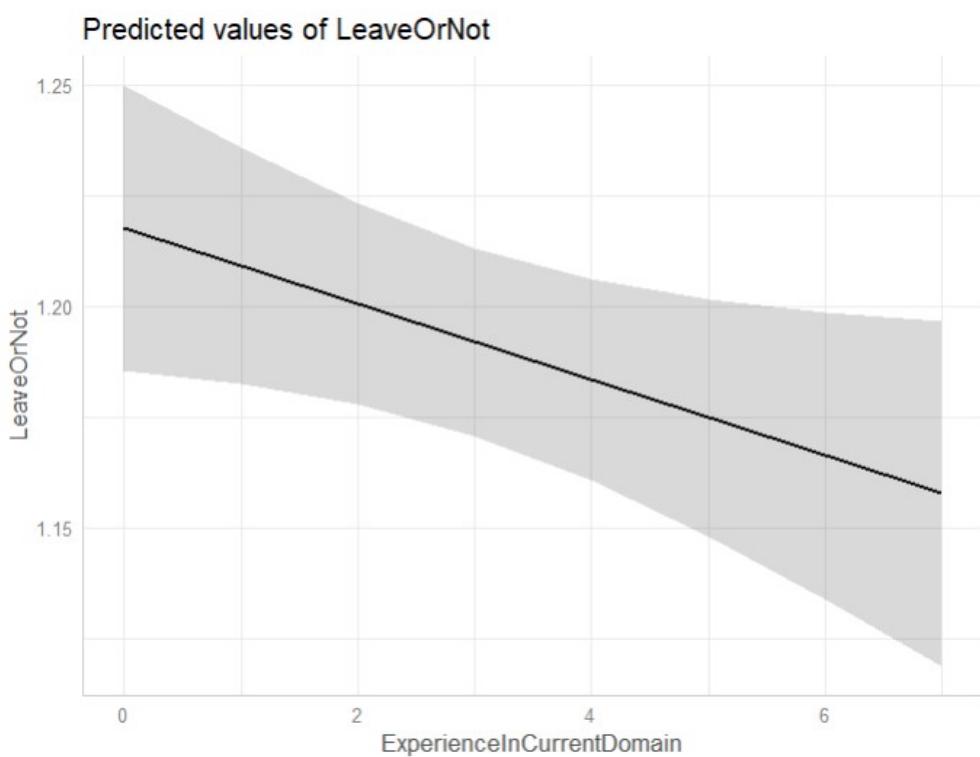
Gender: kobiety mają większą szansę na zostanie zwolnionymi od mężczyzn.



Ever Benched: jeśli ktoś nie miał kiedyś przypisanej pracy to ma większe szanse na zostanie zwolnionym.



Experience In Current Domain: im ktoś dłużej pracuje w branży to ma mniejsze szanse na to, że zostanie zwolniony.



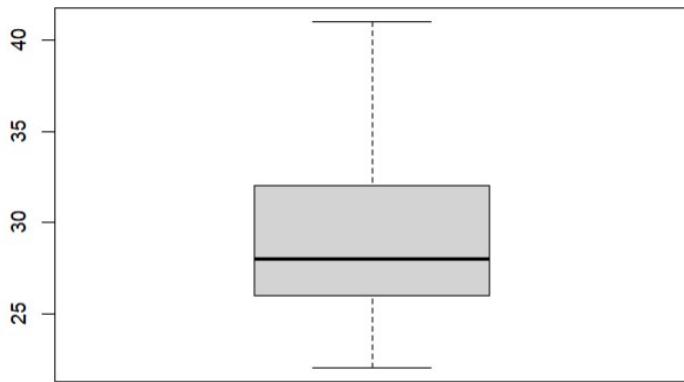
## Przygotowanie danych

Zostanie sprawdzone czy w danych występują braki.

```
> sum(is.na(data))  
[1] 0
```

W danych nie ma żadnych braków.

Jedyną wartością w jakiej mogą występować wartości odstające, ponieważ reszta zmiennych to dane kategoryczne, jest zmienna Age. Zostanie więc stworzony dla niej wykres pudelkowy.



Można zauważyć, że nie ma na nim żadnych wartości odstających. Z tego powodu nie ma potrzeby modyfikacji tej zmiennej.

## Podział na zbiór uczący i testowy

Poniższym kodem zbiór danych Employee zostanie podzielony na zbiór uczący i testowy.

```
set.seed(123)  
prop <- 0.7  
n <- nrow(data)  
train_indices <- sample(1:n, prop * n)  
train_data <- data[train_indices, ]  
test_data <- data[-train_indices, ]
```

Po podzieleniu zostanie sprawdzone czy te zbiory nie są niebalansowane

```
> table(train_data$LeaveOrNot)
```

|      |      |
|------|------|
| 0    | 1    |
| 2146 | 1111 |

zbiór uczący jest zbalansowany

```
> table(test_data$LeaveOrNot)
```

|     |     |
|-----|-----|
| 0   | 1   |
| 907 | 489 |

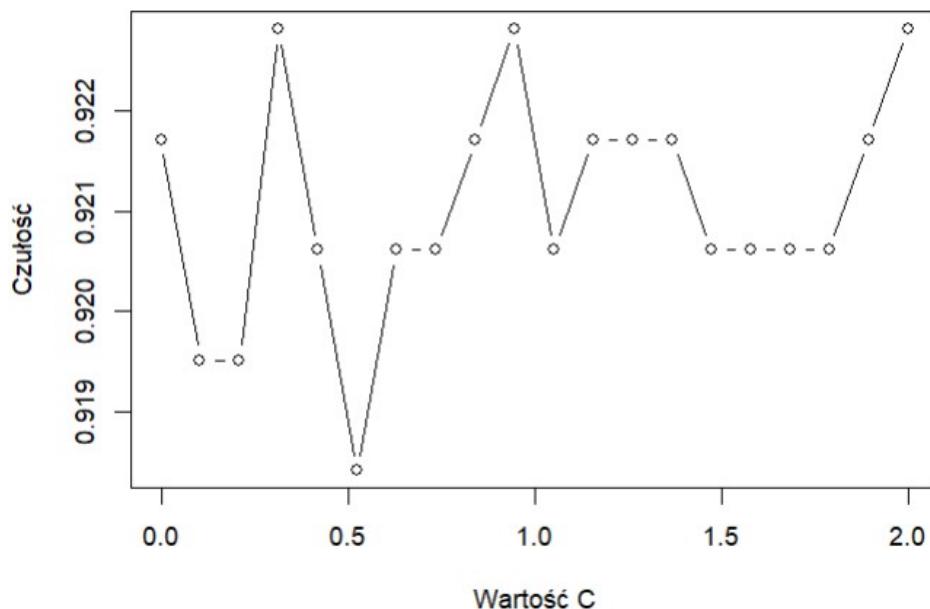
zbiór testowy również jest zbalansowany

## Zastosowanie modeli uczenia maszynowego

W tym projekcie zostaną użyte trzy metody uczenie maszynowego: SVM liniowy, lasy losowe oraz metoda KNN.

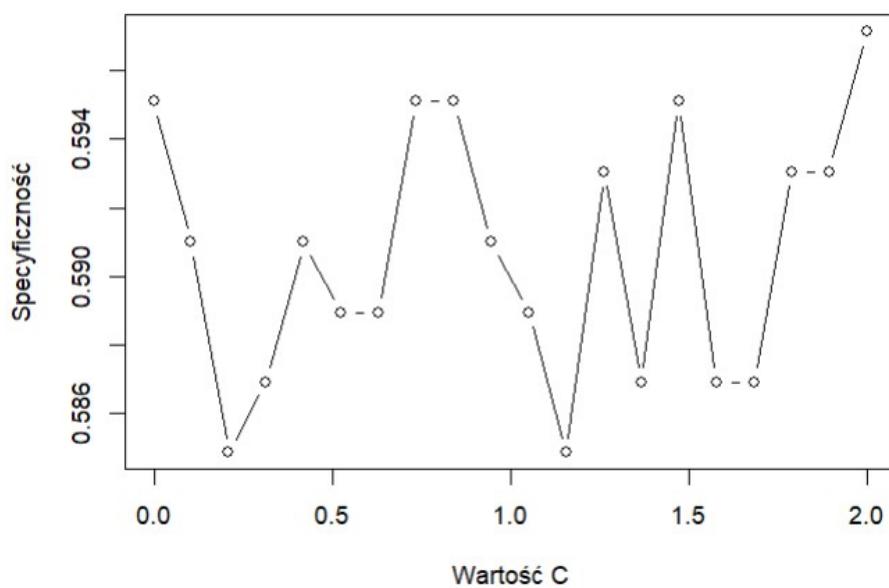
## SVM liniowy

Zależność czułości od wartości C dla SVM liniowego



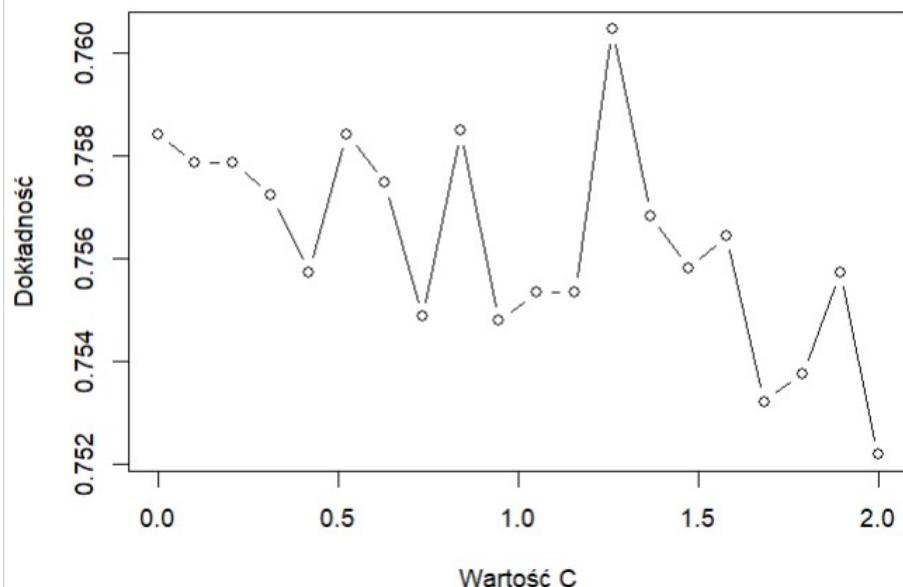
Dla wartości C w zakresie 0-2 czułość waha się od 0,918-0,922, gdzie najwyższe wyniki uzyskuje dla wartości C równych 0,3 0,9 i 2.

Zależność specyficzności od wartości C dla SVM liniowego



Dla wartości C w zakresie 0-2 specyficzność waha się od około 0,582 do około 0,598. Najwyższą wartość przyjmuje dla C równego 2

### Zależność dokładności od wartości C dla SVM liniowego



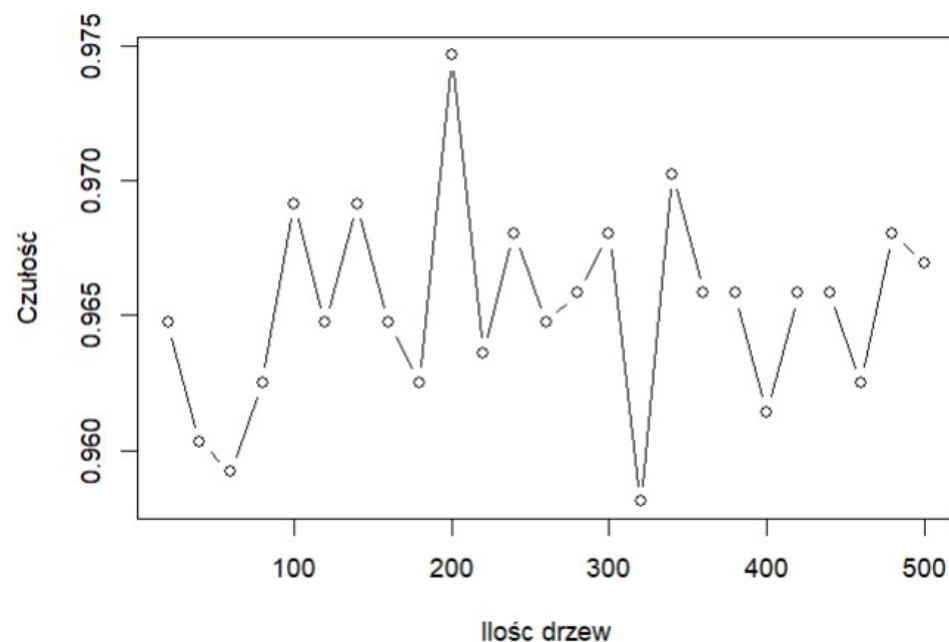
Dla wartości C w zakresie 0-2 dokładność waha się od 0,752 do około 0,762. Najwyższą wartość przyjmuje dla C równego 1,3.

Tak prezentuje się macierz pomyłek dla przykładowego modelu SVM liniowego:

|            |     | Reference |  |
|------------|-----|-----------|--|
| Prediction | 0   | 1         |  |
| 0          | 834 | 203       |  |
| 1          | 73  | 286       |  |

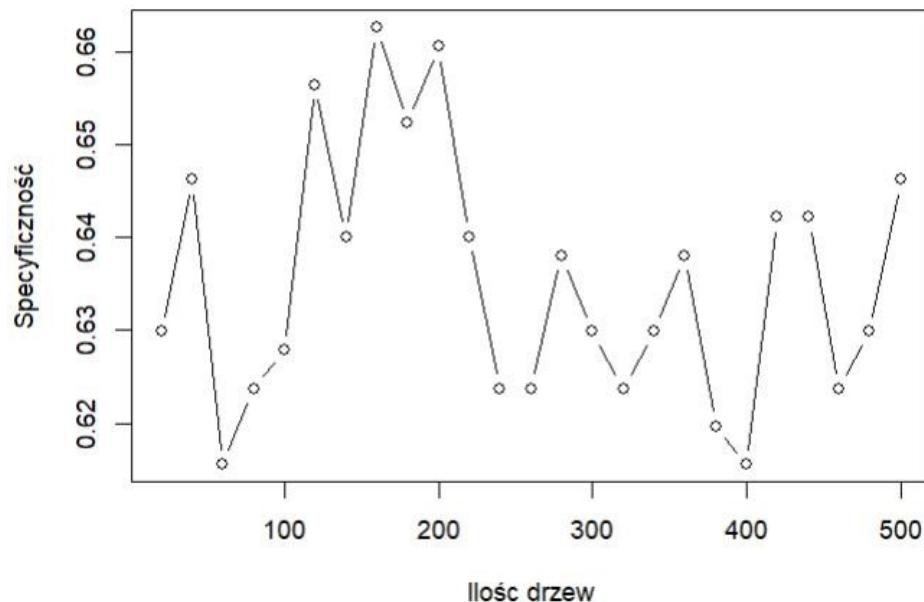
### Lasy losowe

### Zależność czułości od ilości drzew dla lasów losowych



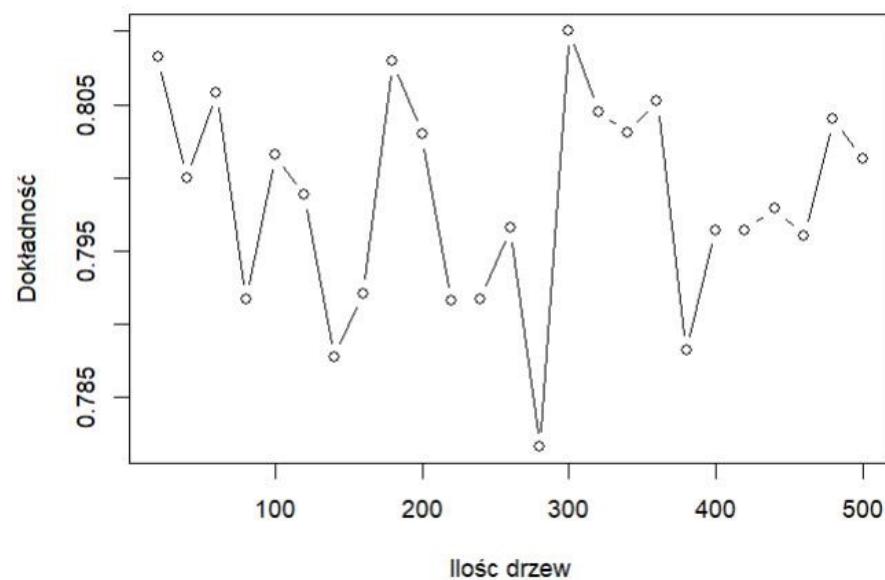
Dla liczby drzew od 20 do 500 czułość waha się od około 0,955 do 0,975. Najwyższa wartość przyjmuje dla 200 drzew

### Zależność specyficzności od ilości drzew dla lasów losowych



Dla liczby drzew od 20 do 500 specyficzność waha się od około 0,62 do 0,66. Najwyższą wartość przyjmuje dla 160 drzew.

### Zależność dokładności od ilości drzew dla lasów losowych



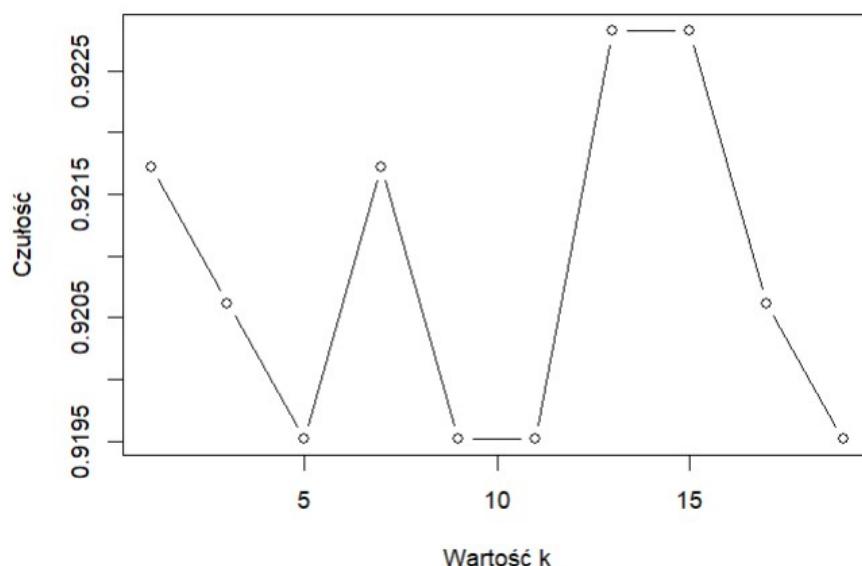
Dla liczby drzew od 20 do 500 dokładność waha się od około 0,775 do około 0,815. Najwyższą wartość przyjmuje dla 300 drzew.

Tak prezentuje się macierz pomyłek dla przykładowego modelu lasów losowych:

| Reference  |     |     |
|------------|-----|-----|
| Prediction | 0   | 1   |
| 0          | 873 | 176 |
| 1          | 34  | 313 |

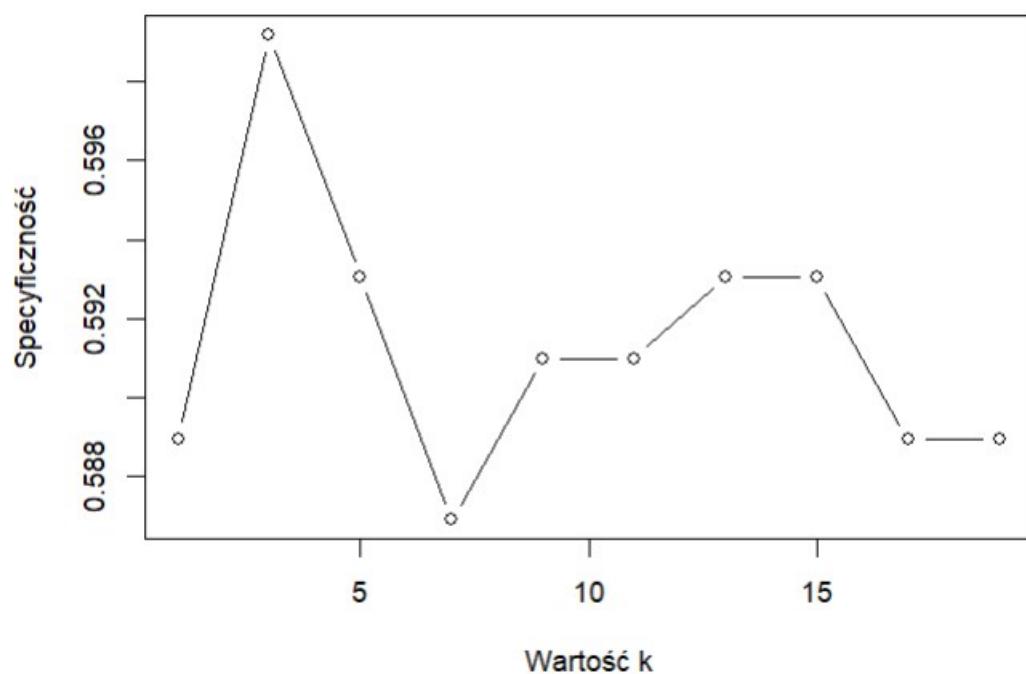
## Metoda KNN

Zależność czułości od wartości k dla metody KNN



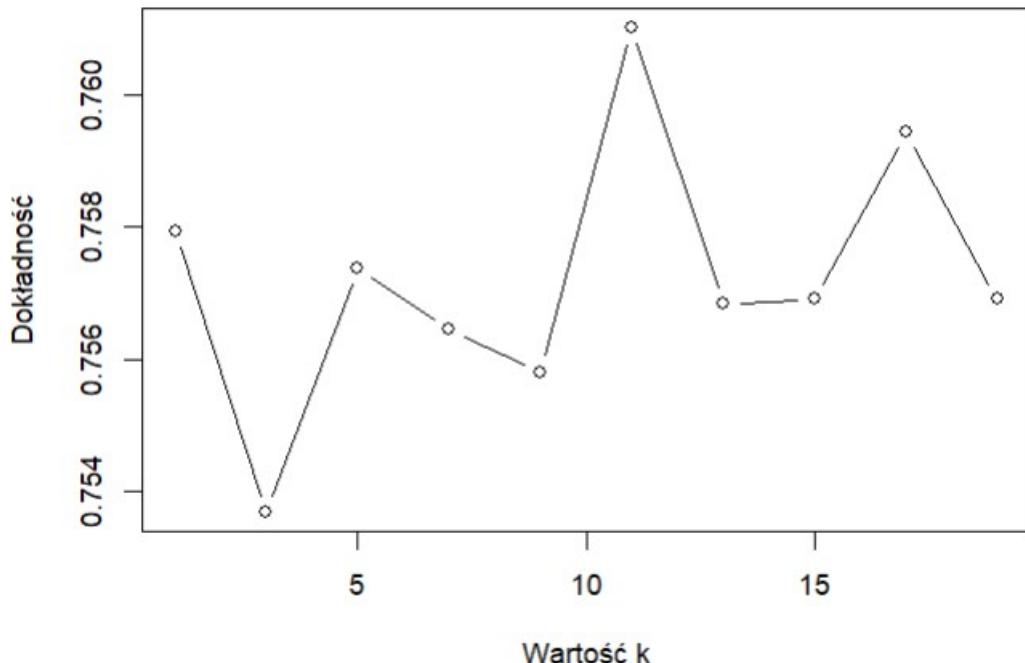
Dla K w zakresie od 1 do 20 czułość waha się od 0,9195 do około 0,9235. Najwyższe wartości przyjmuje dla K równych 13 i 15.

Zależność specyficzności od wartości k dla metody KNN



Dla K w zakresie od 1 do 20 specyficzność waha się od około 0,584 do około 0,6. Najwyższą wartość przyjmuje dla K równego 3.

### Zależność dokładności od wartości k dla metody KNN



Dla K w zakresie od 1 do 20 dokładność waha się od około 0,754 do około 0,762. Najwyższą wartość przyjmuje dla K równego 11.

Tak prezentuje się macierz pomyłek dla przykładowego modelu metody KNN:

|            |     | Reference |   |
|------------|-----|-----------|---|
| Prediction |     | 0         | 1 |
| 0          | 837 | 200       |   |
| 1          | 70  | 289       |   |

## Wnioski

We wszystkich metodach czułość miała najwyższe wyniki sięgające ponad 0,9 natomiast specyficzność miała najmniejsze. Wynosiły one około 0,6. Dokładność zazwyczaj wynosiła około 0,78. Najwyższą czułość wyliczyła metoda lasów losowych. Tak zdarzyło także się dla specyficzności oraz dla dokładności. Można to zauważyć po samych macierzach błędów. W przypadku lasów losowych ilość 0(pracownik zostaje w firmie), które zostały przewidziane jako 1(pracownik zostaje zwolniony) jest dwukrotnie mniejsza niż w pozostałych metodach.

## Analiza interpretowalności

Interpretowana będzie analiza modelu KNN

## Profile ceteris-paribus (PCP)

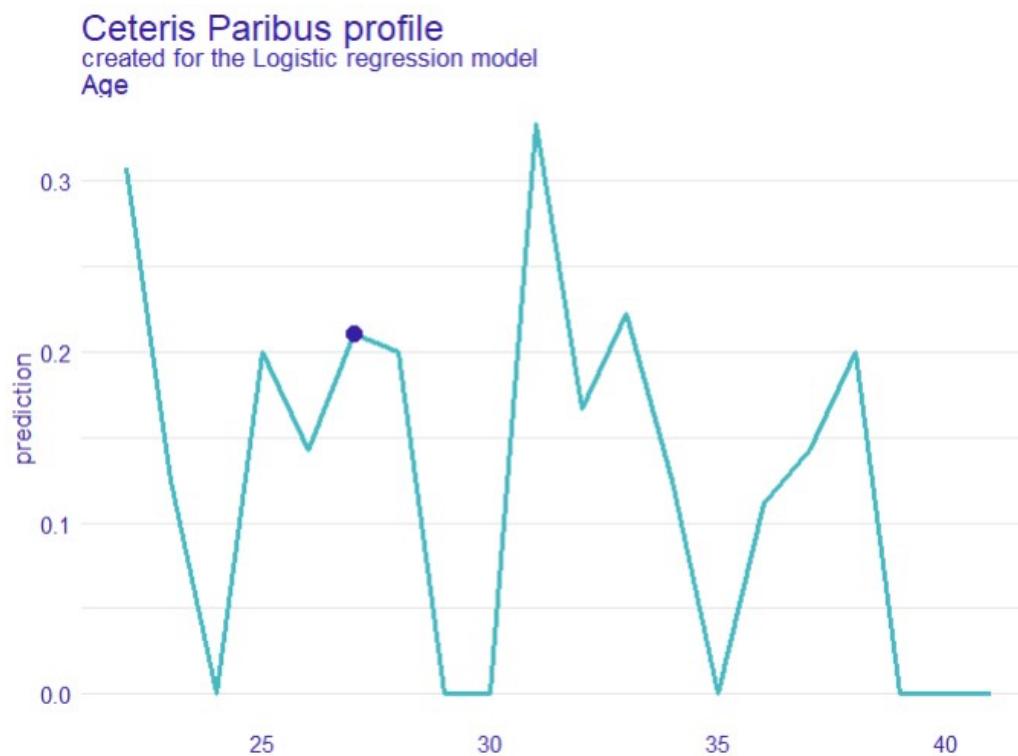
Został stworzony explainer

```
Preparation of a new explainer is initiated
-> model label      : Logistic regression
-> data              : 3257 rows 8 cols
-> target variable   : 3257 values
-> predict function  : yhat.train will be used ( default )
-> predicted values  : No value for predict function target column. ( default )
-> model_info         : package caret , ver. 6.0.94 , task classification ( default )
-> model_info         : type set to classification
-> predicted values  : numerical, min = 0 , mean = 0.3195357 , max = 1
-> residual function : difference between y and yhat ( default )
-> residuals          : numerical, min = 0.1 , mean = 1.021576 , max = 1.944444
A new explainer has been created!
```

Do wykorzystania należy wybrać losową obserwację. Niech będzie to wiersz numer 47:

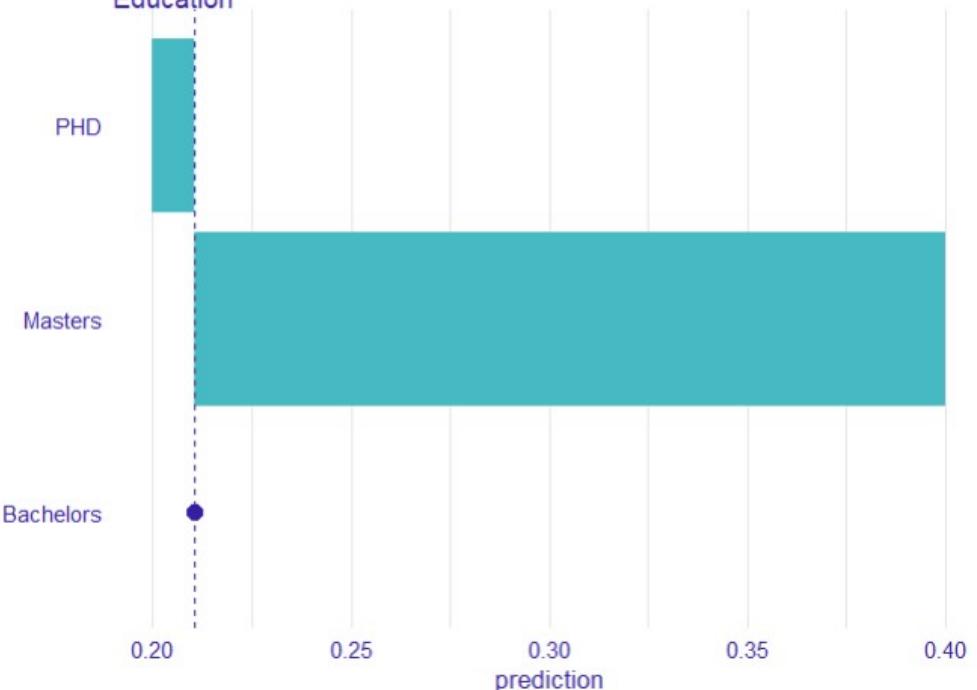
| Education | JoiningYear | City      | PaymentTier | Age | Gender | EverBenchched | ExperienceInCurrentDomain | LeaveOrNot |
|-----------|-------------|-----------|-------------|-----|--------|---------------|---------------------------|------------|
| Bachelors | 2012        | New Delhi | 3           | 27  | Female | No            | 5                         | 1          |

Najpierw zostaną zbadane zmienne współczynników i to jak ich zmiana wpłynęłaby na szansę zostania zwolnionym.



Jak widać sam wiek nie jest zbyt dobrym współczynnikiem sprawdzenia szans na zwolnienie, ponieważ jego przekrój jest w większości losowy.

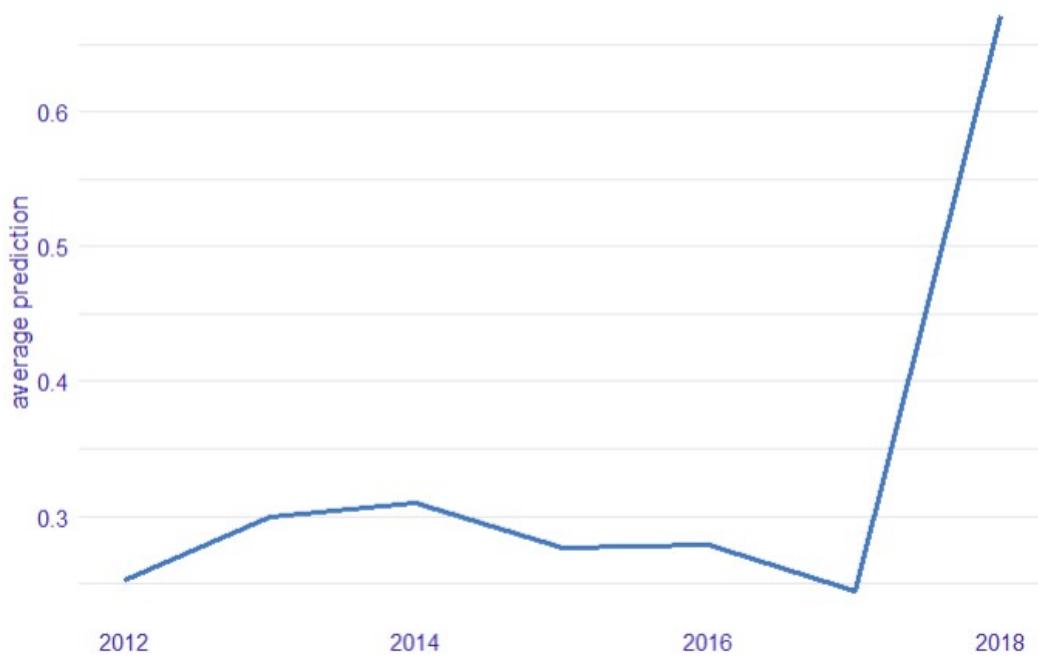
### Ceteris Paribus profile created for the Logistic regression model Education



Jeśli obserwacja posiadałaby magistra szansa na jej zwolnienie znacznie by wzrosły, a jeśli posiadałaby doktora szanse na zwolnienie nieznacznie by zmalały.

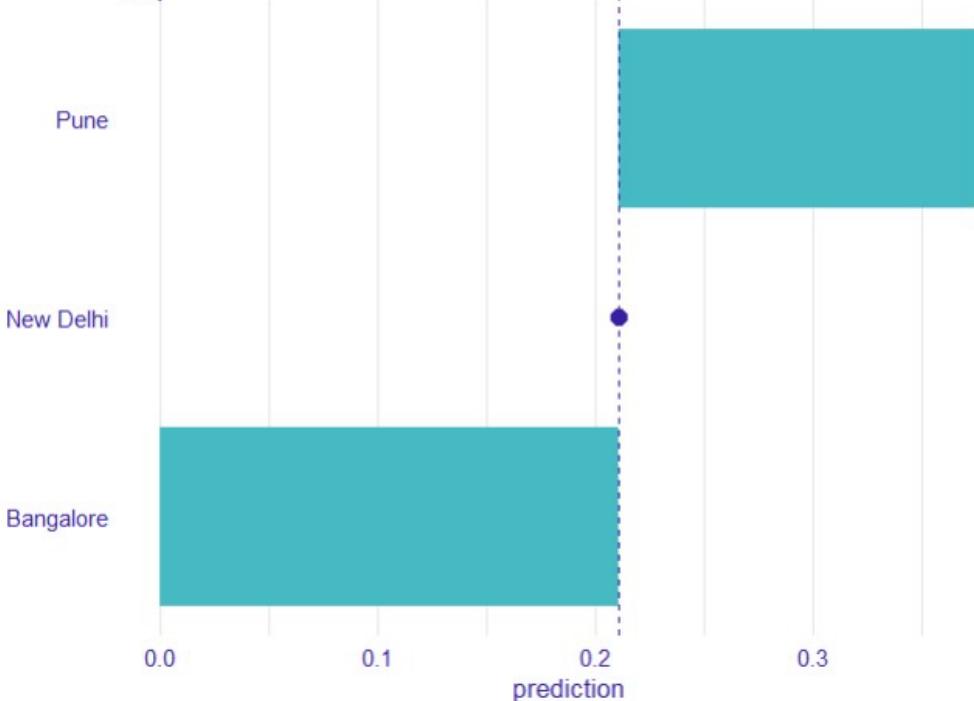
### Partial Dependence profile

Created for the Logistic regression model  
JoiningYear



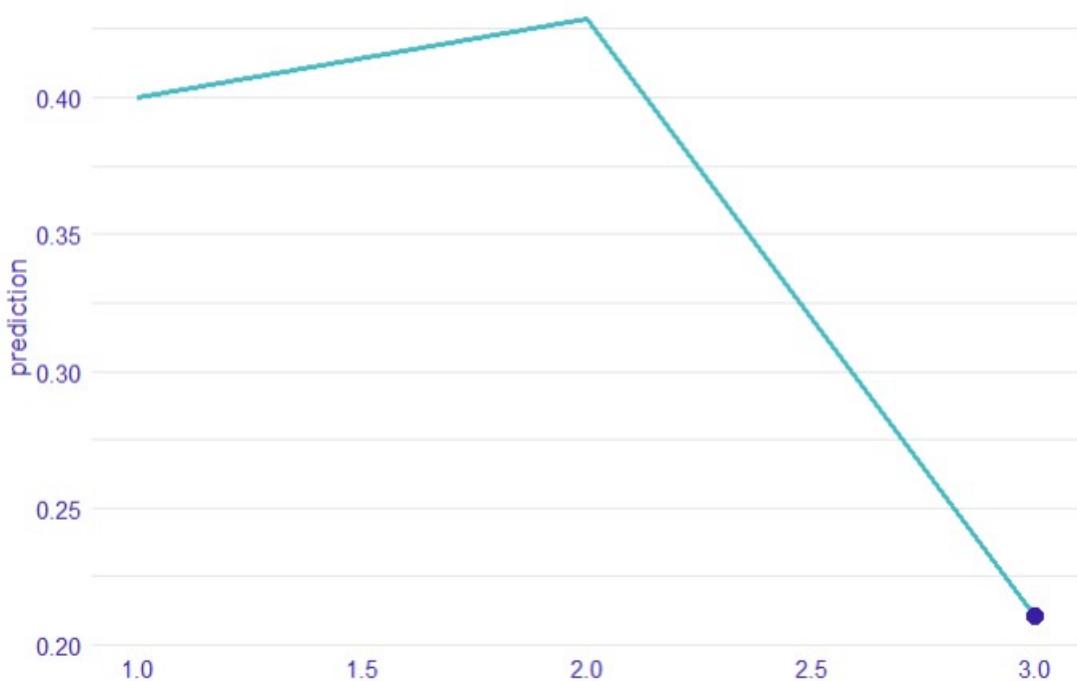
Im byłby późniejszy rok rozpoczęcia pracy w firmie tym statystycznie byłaby większa szansa na zwolnienie.

**Ceteris Paribus profile**  
created for the Logistic regression model  
City



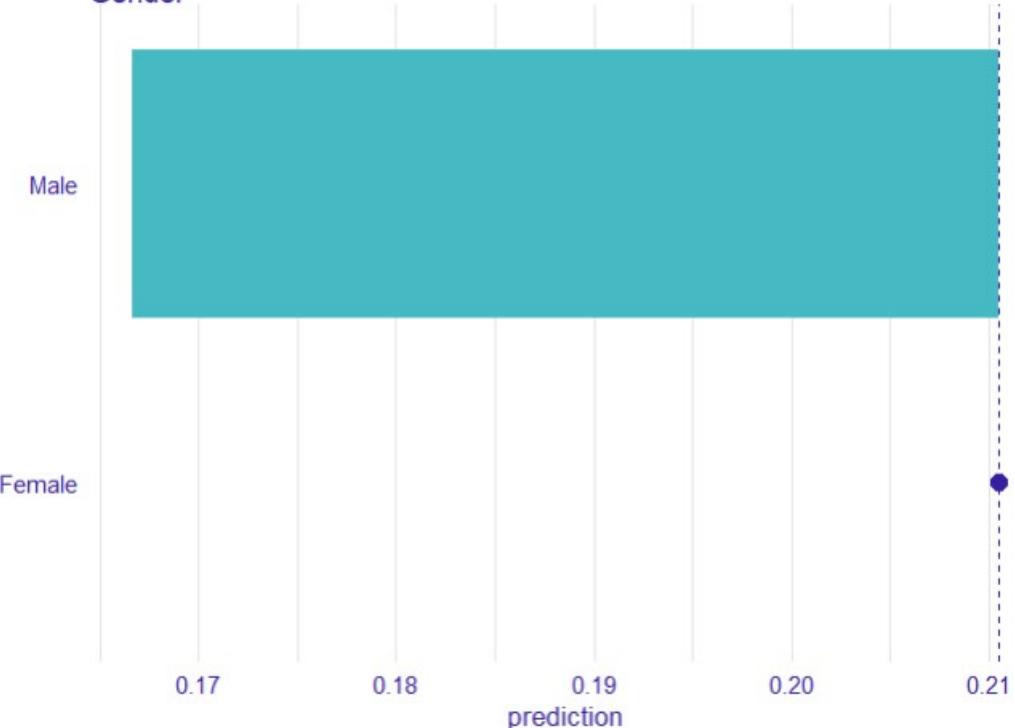
Gdyby obserwacja mieszkała w Pune szanse na zwolnienie wzrosłyby, a gdyby mieszkała w Bangalore zmalałyby.

**Ceteris Paribus profile**  
created for the Logistic regression model  
PaymentTier



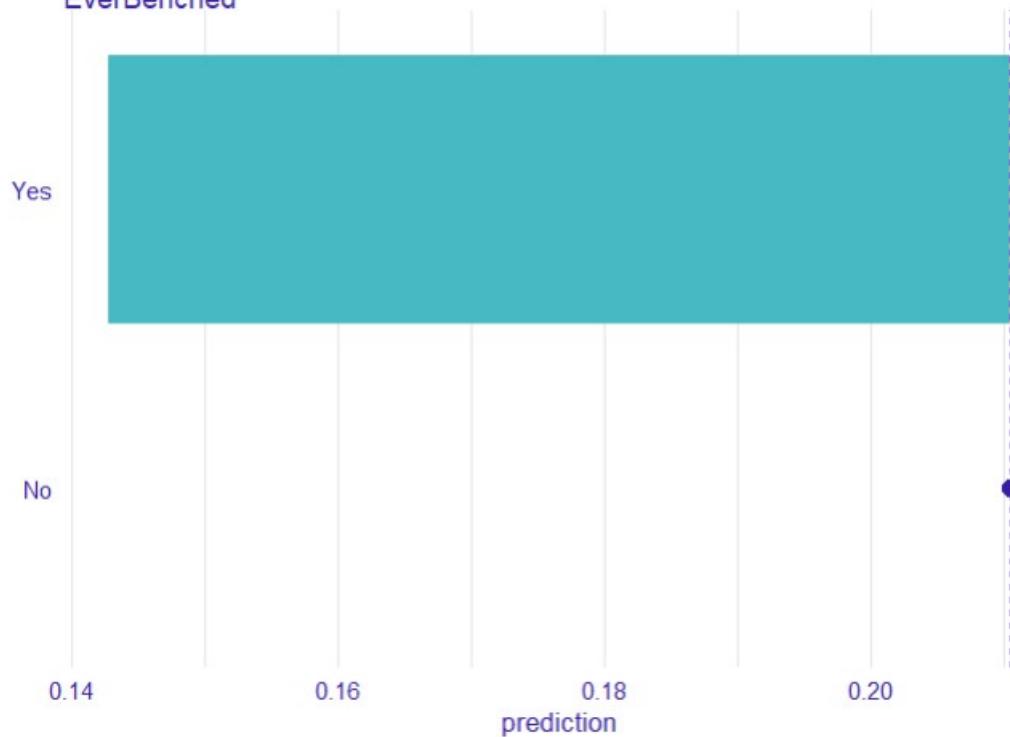
Gdyby obserwacja zarabiała więcej miałaby większe szanse na zwolnienie.

**Ceteris Paribus profile**  
created for the Logistic regression model  
**Gender**



Gdyby obserwacja była mężczyzną miałaby mniejsze szanse na zwolnienie.

**Ceteris Paribus profile**  
created for the Logistic regression model  
**EverBench**



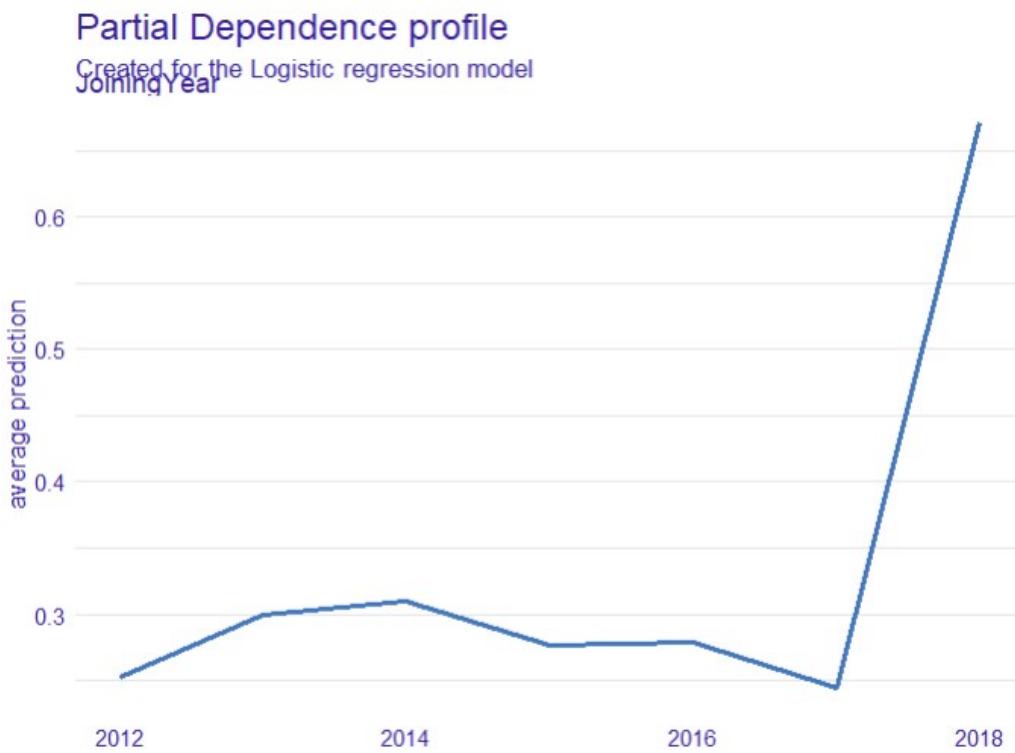
Gdyby obserwacja miała kiedyś okres bez przydzielonej pracy to szansa na jej zwolnienie spadłaby. Nie jest to intuicyjne.

**Ceteris Paribus profile**  
created for the Logistic regression model  
`ExperiencelnCurrentDomain`



Obserwacja ma największe szanse na zwolnienie przy tej ilości przepracowanych lat w branży.

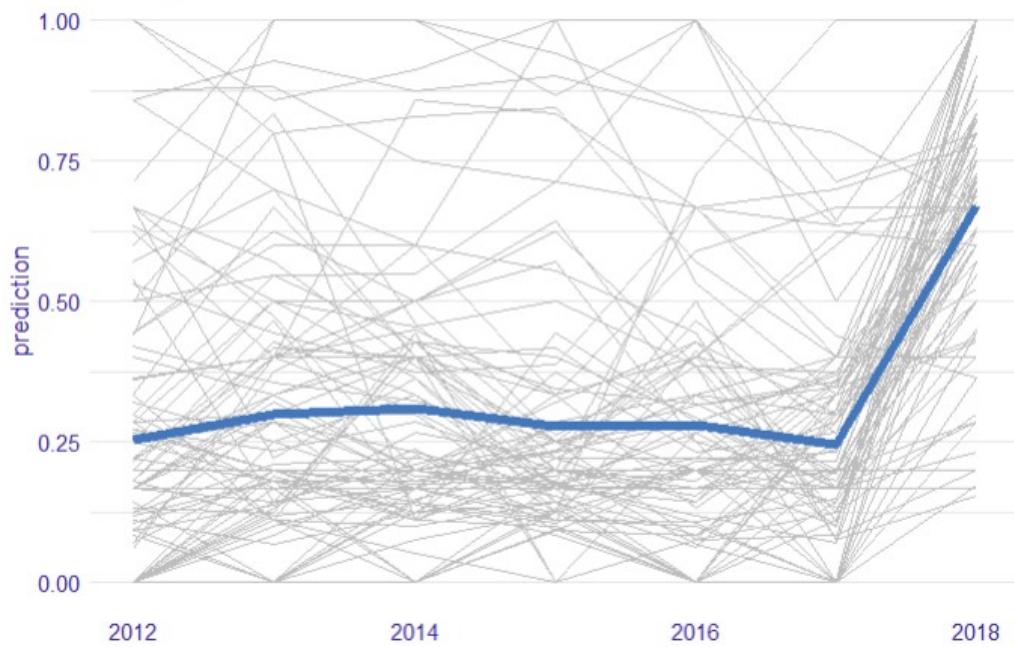
## Wykresy częściowej zależności (PDP)



Im mniejszy staż pracy w firmie tym większa szansa na zwolnienie.

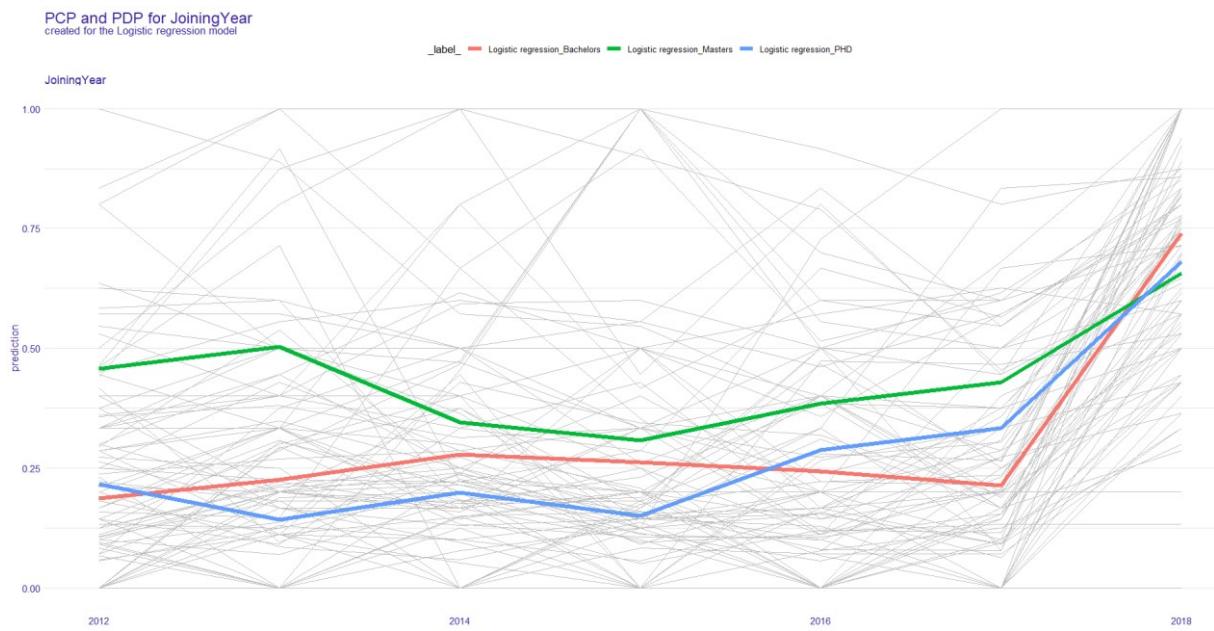
## PCP and PDP for JoiningYear

created for the Logistic regression model  
JoiningYear

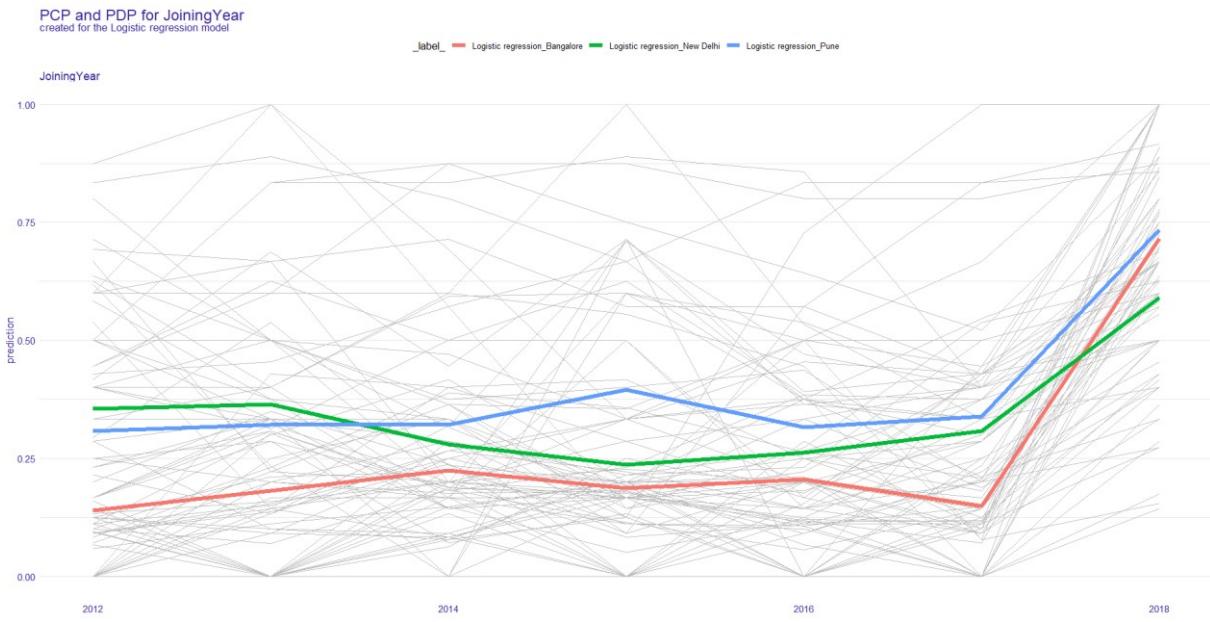


Wszystkich czynników jest za dużo i są w większości zbyt losowe by je na raz zinterpretować.  
Widać jednak że spora część z nich wzrasta przy zmiennej rozpoczęcia pracy w 2018.

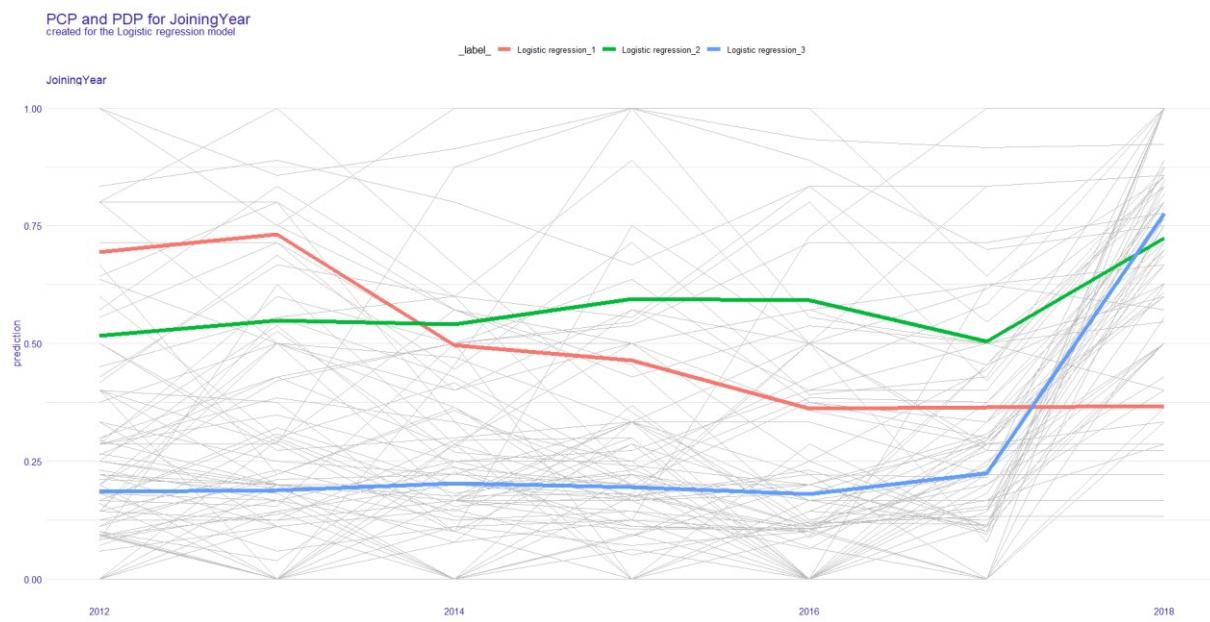
Zostaną teraz zbadane wszystkie zmienne po kolej.



Widać że największą szansę przez prawie cały czas mają osoby z magistrem.

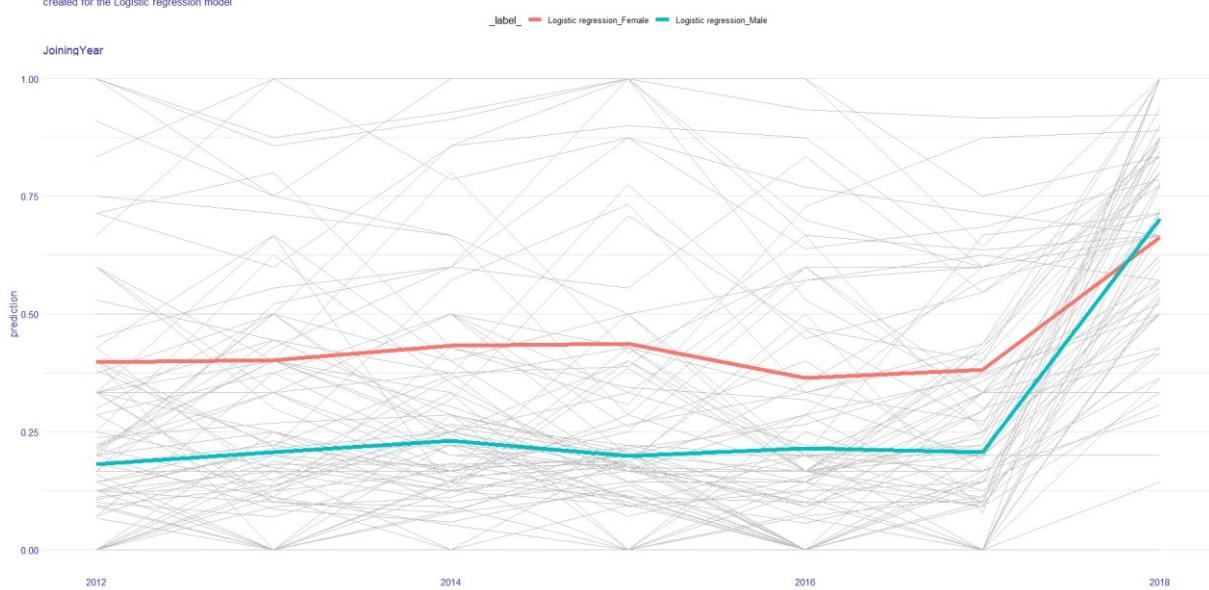


Przez większość czasu największe szanse na zwolnienie mają osoby z Pune a najmniejsze Bangalore.



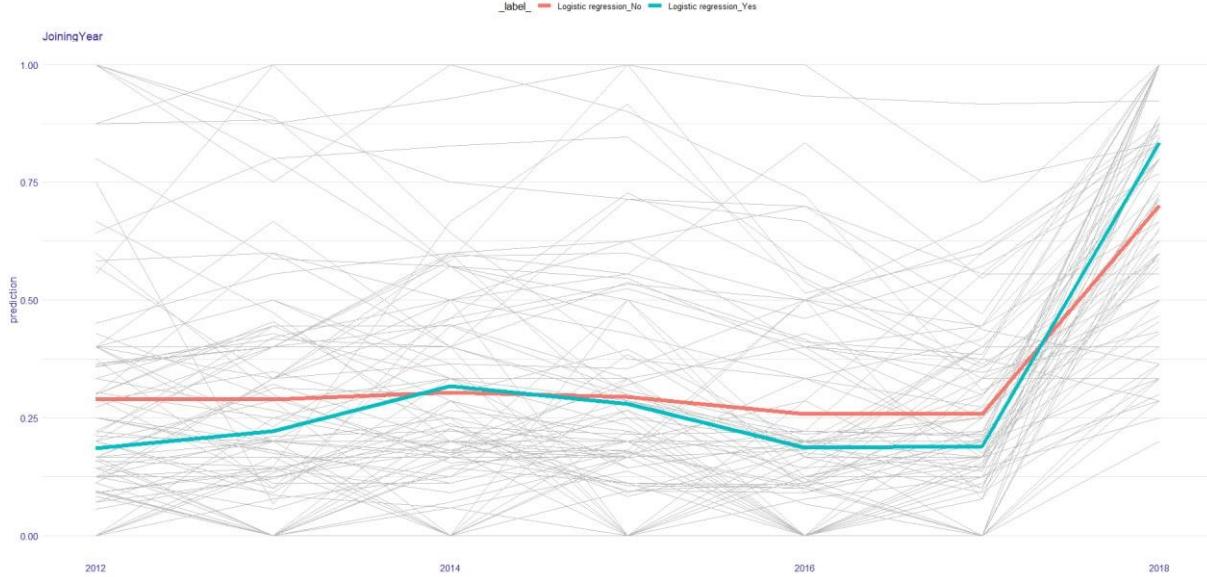
Można zauważyć, że osoby z najmniejszymi zarobkami przez większość czasu mają najmniejsze szanse za zwolnienie, lecz gdy wykres zbliża się do roku 2018 staje się ono najwyższe. Osoby z największymi zarobkami mają coraz mniejsze szanse na zwolnienie wraz z przesuwaniem się wykresu.

PCP and PDP for JoiningYear  
created for the Logistic regression model



Kobiety mają większe szanse na zwolnienie od mężczyzn. Zrównuje się ono jednak przy zmiennej 2018.

PCP and PDP for JoiningYear  
created for the Logistic regression model



Wyniki zmiennej czy ktoś był kiedyś bez przydzielonej pracy przeplatają się przez cały wykres. Jednak przy zmiennej 2018 większą szansę na zwolnienie ma osoba, której zdarzyło się nie mieć przypisanej pracy.

Dane dla wieku i ilości doświadczenia w branży nie zostały użyte ponieważ jest ich na tyle dużo, że są nieczytelne.

# Wartości SHAP

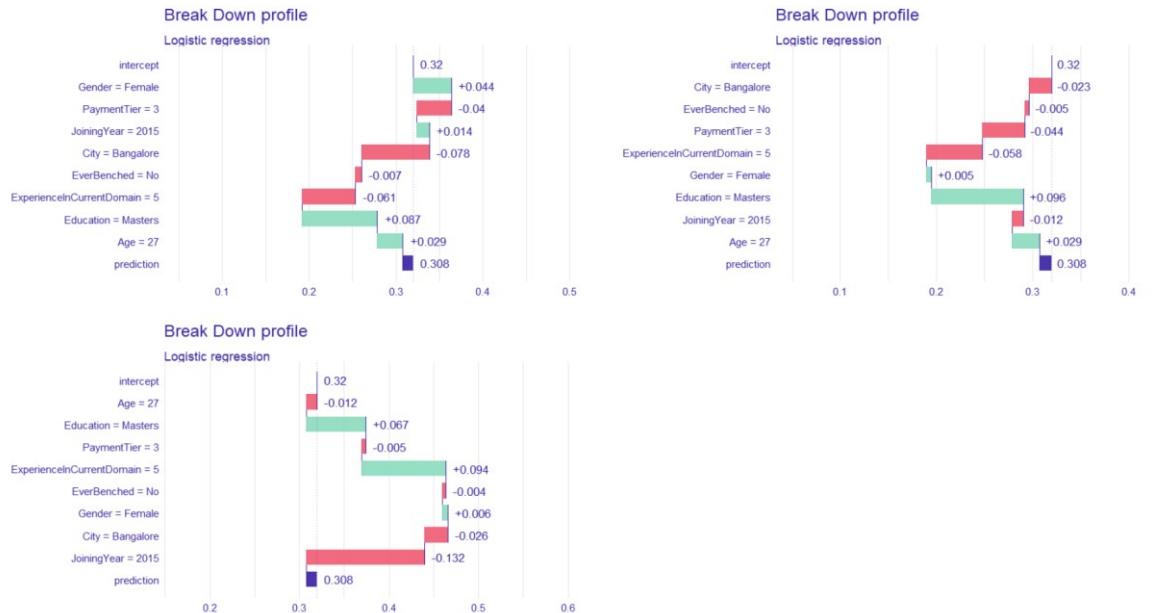
Wykresy BD:

Badaną obserwacją niech będzie obserwacja numer 29:

| Education | JoiningYear | City      | PaymentTier | Age | Gender | EverBenchend | ExperienceInCurrentDomain | LeaveOrNot |
|-----------|-------------|-----------|-------------|-----|--------|--------------|---------------------------|------------|
| Masters   | 2015        | Bangalore | 3           | 27  | Female | No           | 5                         | 1          |

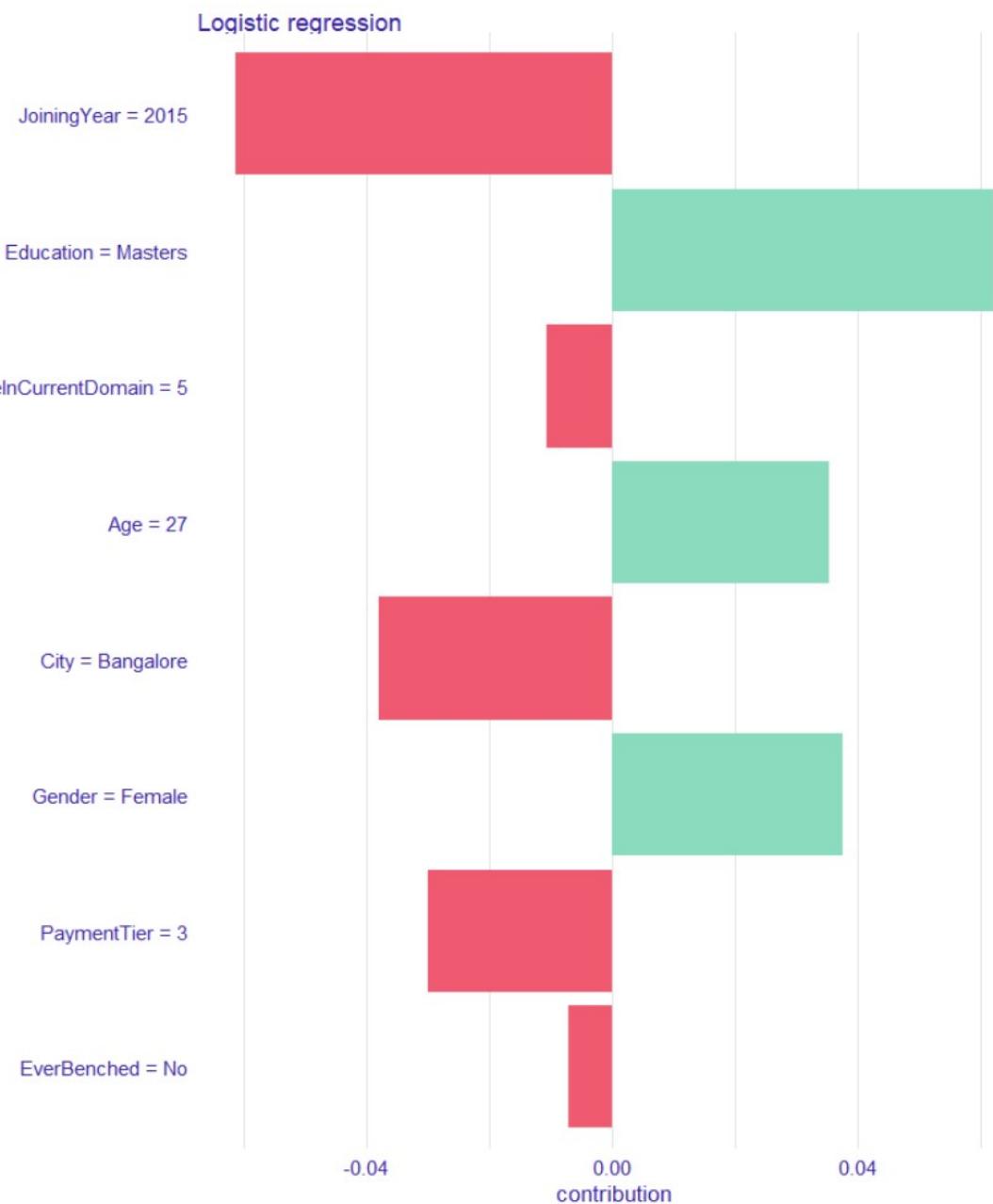
Jak widać po samych zmiennych model przewiduje, że obserwacja na 69% nie powinna zostać

zwolniona  $0.6923077 - 0.3076923 = 0.3846154$  jak wiadomo rzeczywistość jest jednak inna.



Bycie kobietą zawsze sprzyja większej szansie na zwolnienie, lecz im wcześniej ona wprowadzona tym większy ma wpływ. Zarabianie mniej zawsze wpływa na zmniejszenie szans na zwolnienie lecz może zarówno być to duży wpływ jak i prawie niezauważalny. Rok dołączenia do firmy 2015 może być zarówno czynnikiem wpływającym na zwolnienie jak i nie. Mieszkanie w Bangalore zmniejsza szansę na zwolnienie. Brak posiadania braku kiedykolwiek zadań w pracy zmniejsza szansę na zwolnienie. 5 lat doświadczenia w branży może zarówno pozytywnie jak i negatywnie wpłynąć na zwolnienie. Posiadanie magistra zwiększa szansę na zwolnienie. Wiek 27 lat może zarówno wpłynąć na zwolnienie jak i nie nie wpłynąć.

Wartości SHAP:



Jak widać wykres jest podzielony na wyniki które wpływają zarówno na zwolnienie jak i wpływające na to że tego zwolnienia nie będzie. Na zwolnienie w dużym stopniu wpływa wykształcenie magisterkie, wiek 27 lat i to że obserwacja jest kobietą. Natomiast na brak zwolnienia najbardziej wpływa to, że obserwacja pracuje w firmie od 2015 roku, pracuje w

Banglore oraz mało zarabia. 5 letnie doświadczenie w branży oraz to że nigdy nie miała okresu bez wykonywania zadań w pracy wpływa na zmniejszenie szans na zwolnienie, ale w mniejszym stopniu.

## Podsumowanie

W ramach tego projektu badawczego skoncentrowaliśmy się na zastosowaniu zaawansowanych modeli uczenia maszynowego w analizie szans na zwolnienie pracownika w kontekście zarządzania zasobami ludzkimi. Naszym celem było stworzenie efektywnego narzędzia wspierającego proces podejmowania decyzji personalnych oraz minimalizującego potencjalne zakłócenia w strukturze organizacyjnej.

Przeprowadzona analiza opierała się na różnorodnych czynnikach, takich jak poziom edukacji, staż pracy, miejsce zamieszkania, poziom wynagrodzenia, wiek, płeć, doświadczenie w bieżącej dziedzinie. Wykorzystaliśmy modele uczenia maszynowego, takie jak SVM liniowy, drzewa losowe i metody KNN, aby opracować narzędzie prognozujące potencjalne ryzyko zwolnienia pracownika.

Finalnie okazało się że najlepszą metodą w tym przypadku jest metoda drzew losowych. Dawała ona największe wyniki zarówno dla dokładności, czułości i specyficzności.

Podczas interpretowania modelu okazało się, że dużo zależy od obserwacji. Baza ma na tyle dużo pracowników, że wybór danego ma spore znaczenie w dalszej interpretacji wyników.