A blue and white logo with a building in the background

Description automatically generated

Институт за математику и информатику

Природно-математички факултет

Универзитет у Крагујевцу

Семинарски рад

Представљање и тумачење скупа података „***HR Analytics***“

Ментор:  
др Бранко Арсић

Студенти:  
Јован Радовановић 85/2018  
Немања Тракић 130/2018

[1. Увод 4](#_Toc183906041)

[2. Припрема података 5](#_Toc183906042)

[2.1 Преглед података 7](#_Toc183906043)

[2.2 Уклањање дупликата и неважних колона 31](#_Toc183906044)

[2.3 Исправљање структурних грешака 32](#_Toc183906045)

[2.4 Уклањање екстремних вредности 32](#_Toc183906046)

[2.5 Руковање недостајућим вредностима 33](#_Toc183906047)

[2.6 Креирање нових променљивих 33](#_Toc183906048)

[DistanceFromHomeGroup 33](#_Toc183906049)

[3. Анализа података 34](#_Toc183906050)

[4. Креирање модела 49](#_Toc183906051)

[4.1 Логистичка регресија 50](#_Toc183906052)

[4.1.1 Логистички модел и унакрсна валидација 51](#_Toc183906053)

[4.2 GLM модел 52](#_Toc183906054)

[4.2.1 GLM модел и унакрсна валидација 54](#_Toc183906055)

[4.3 Random Forest модел 55](#_Toc183906056)

[4.3.1 Random Forest модел и унакрсна валидација 56](#_Toc183906057)

[4.4 Неуронске мреже 57](#_Toc183906058)

[4.5 Резултати модела 59](#_Toc183906059)

[4.6 Resampling методе 61](#_Toc183906060)

[4.6.1 Oversampling 61](#_Toc183906061)

[4.6.2 Undersampling метода 62](#_Toc183906062)

[4.6.3 Oversampling и Undersampling 62](#_Toc183906063)

[4.6.4 Резултати 62](#_Toc183906064)

[5. Закључак 63](#_Toc183906065)

[6. Литература 64](#_Toc183906066)

# Увод

Предвиђање напуштања запослених омогућава фирмама да на време препознају факторе који воде ка томе и омогућава им деловање у циљу спречавања одласка запослених. Ово може укључивати побољшање услова за рад, пружање додатне обуке, реорганизацију, проналазак улога који више одговарају свакоме појединачно, као и друге мере за повећање задовољства унутар организације.

Основни циљ ове анализе је да се идентификују кључни фактори који утичу на одлазак запослених. Као што су старосна доб, удаљеност од посла, задовољство на послу, учесталост пословних путовања и сл.

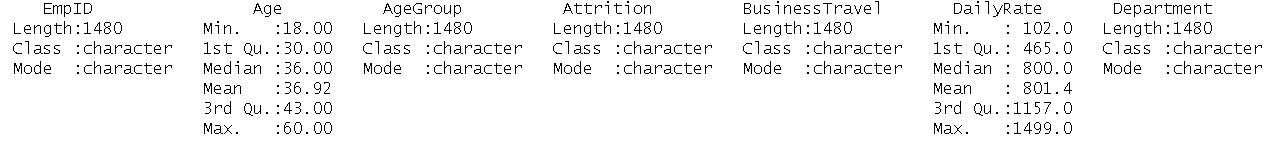
„**Attrition“** (у преводу осипање) или одлазак запослених подразумева када запослени својевољно или присилно напушта компанију. Висока стопа може довести до значајних трошкова за организацију, како новчаних тако и временских. Након одласка потребно је пронаћи, обучити и прилагодити нове запослене.

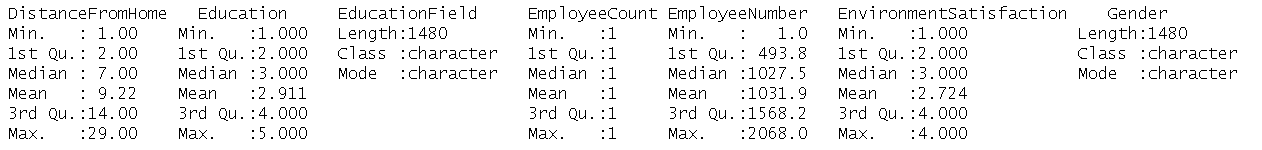
На крају, прецизно предвиђање омогућава запосленима у Људским ресурсима да доносе информисане и стратешке одлуке које директно и позитивно утичу на задовољство запослених.

Скуп података можете погледати [овде (HR Analytics Dataset (kaggle.com))](https://www.kaggle.com/datasets/saadharoon27/hr-analytics-dataset/data).

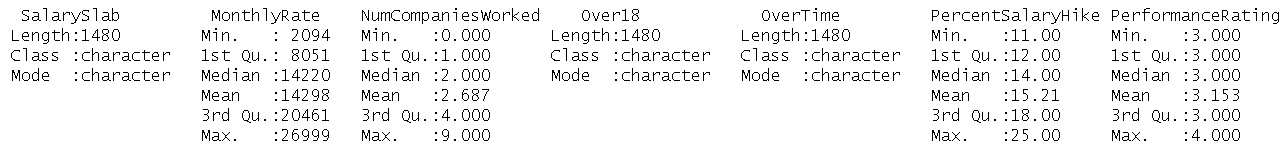
# Припрема података

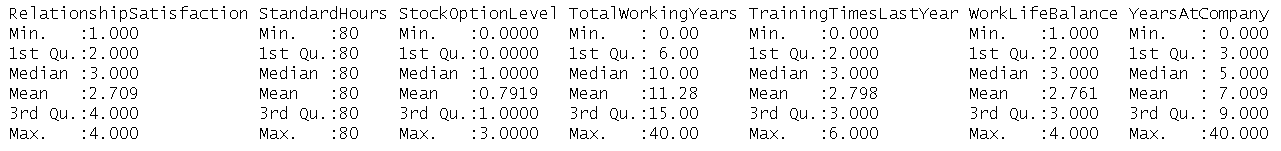
Скуп података ***HR\_Analytics*** обухвата 38 колона и садржи 1480 записа који пружају детаљан увид у различите аспекте живота запослених. Овај скуп података покрива како професионалне аспекте, попут задовољства послом, учешћа у обуци, и могућности за унапређење, тако и личне факторе као што су породични статус, баланс између пословног и приватног живота и задовољство радним окружењем. На тај начин, скуп података пружа комплетну слику о факторима који утичу на добробит и ангажованост запослених у компанији.



A close up of a white background

Description automatically generated





A black text on a white background

Description automatically generated

За анализу података креираћемо две функције („*plot\_histogram*“, „*plot\_bar*“ и „plot\_*boxplot*“) помоћу којих ћемо добити графички приказ поделе података по колонама:

A screenshot of a computer code

Description automatically generated

За категоријских променљивих користићемо *plot\_bar*,док за приказ нумеричких променљивих користимо *plot\_histogram*. Уколико буде потребе, увешћемо додатне начине приказа расподеле како бисмо лакше увидели недостатке.

## ****2.1 Преглед података****

**Колона „EmpId“**

Представља јединствени идентификатор запосленог. Уз помоћ следеће функције можемо проверити колико заиста јединствених вредности постоји унутар ове колоне. A close-up of a number

Description automatically generated



Видимо да постоји 1470 јединствених вредности док у се скупу података налази 1480 података. Закључујемо да постоје дупликати. Дупликате и ирелевантне податке ћемо исправити у следећим корацима.

**Колона „Age“**

Представља године запосленог. Најмлађи запослени има 18 година док најстарији 60 година. Средња вредност износи 36,92 године.

A close up of a number

Description automatically generated

A graph of age and age

Description automatically generated

Расподела година запослених указује на релативно нормалну дистрибуцију, без грешака или екстремних вредности које би могле негативно утицати на наш модел.

Видимо да су средња вредност као и медијана сличе што додатно потврђује да не постоје изузеци у подацима.

**Колона „AgeGroup“**

Представља опсег година у којем се налази запослени.

A close-up of a text

Description automatically generated

Садржи 5 различитих група: 18-25, 26-35, 36-45, 46-55, 55+. Расподелу по групама можемо прикази на следећи начин.

A graph of blue rectangular bars

Description automatically generated with medium confidence

**Колона „Attrition“**

Представља да ли је запослени напустио организацију, уједно и колона коју ћемо предвиђати у овом раду.

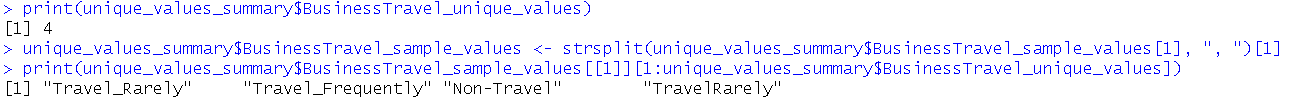
A graph with a red rectangle and blue rectangle

Description automatically generated

Из доступних података можемо закључити да је укупно 237 запослених напустило компанију. Овај податак указује на одређени ниво флуктуације радне снаге, што може бити резултат различитих фактора, као што су незадовољство послом, боље прилике у другим компанијама, или неусклађеност између личних и професионалних потреба запослених. Анализа разлога за одлазак ових запослених може пружити драгоцене увиде у области које треба побољшати у оквиру компаније, као и у стратегије задржавања радне снаге.

**Колона „BusinessTravel“**

Представља информације о учесталости путовања запосленог. Садржи 4 могуће вредности: „Travel\_Rarely“, „Travel\_Frequently“, „Non-Travel“, „TravelRarely“.



Након исправљања структурних грешака, које ћемо објаснити у даљем току рада, расподела за ову колону је:

A graph of a business travel

Description automatically generated

Већина запослених ретко има обавезу да иде на пословни пут, док нешто више од 250 запослених често путује због пословних обавеза. Најмањи број чине они који никада не иду на службена путовања, око 150 запослених.

**Колона „StandardHours“**

Ова колона указује на уговорени број радних сати на месечном нивоу. Сви запослени имају уговорено 80 сати, што значи да ова колона неће имати значаја у нашем даљем истраживању. Поступак уклањања објаснићемо у даљем тексту.

**Колона „HourlyRate“**

Представља уговорени износ запосленог по радном сату. Можемо проверити расподелу ове колоне као и да ли ова колона садржи екстремне вредности који би негативно утицали на наш модел.

A graph with a line

Description automatically generated

Видимо да ова колона не садржи изузетке и да задржава нормалну диструбуцију.

**Колона „DailyRate“**

Представља уговорени износ дневне накнаде запосленог. Као и за „HourlyRate“ колону, можемо проверити екстремне вредности и њену расподелу.

A screenshot of a graph

Description automatically generated

Добијамо сличну расподелу без изузетака као код „HourlyRate“, што је и логично, с обзиром на то да сви запослени имају исти уговорени број сати.

**Колона „MonthlyRate“**

Представља уговорену месечну зараду запосленог.

A graph with a line

Description automatically generated

Као и код осталих типова ове колоне тако ни ова колона не садржи екстремне вредности.

**Колона „MonthlyIncome“**

Представља укупне месечне приходе запосленог.

Можемо проверити расподелу ове колоне на боксплоту и проверити екстремне вредности.

A graph with a line

Description automatically generated

Видимо да постоје вредности које се налазе изван граница (Q3+1.5×IQR) проверићемо у даљем раду да ли је потребно да избацимо ове вредности.

**Колона „SalarySlab“**

Представља категорију у којој се налази запослени на основу укупног месечног прихода.

A graph with blue squares

Description automatically generated

**Колона „OverTime“**

Представља увид да ли запослени ради прековремено или не.

A graph with a rectangle

Description automatically generated

Можемо закључити да већина запослених не ради прековремено али такође да нам ово можда може помоћи при креирању модела.

**Колона „Department“**

Представља ком одељењу припада запослени.

A graph of a bar chart

Description automatically generated

**Колона „DistanceFromHome“**

Представља удаљеност места пребивалишта од фирме. Можемо проверити да ли ова колона садржи екстремне вредности који би негативно утицали на наш модел. У наставку рада ћемо додатно обрадити ову колону.

A graph with a line

Description automatically generated

**Колона „Education“**

Представља ниво образовања запосленог.

A graph of a graph showing different levels of education

Description automatically generated with medium confidence

**Видимо да и ова колона има униформну расподелу. Највише запослених има завршене факултете, потом мастер студије и средњу школу. Такође видимо да 15% запослених има завршену само основну школу, а веома мали број запослених има докторску дисертацију.**

**Колона „EducationField“**

Представља поље образовања запосленог.

A graph of a graph showing a number of blue rectangular bars

Description automatically generated with medium confidence

Највише запослених долази из сфера природних наука и медицине.

**Колона „EnvironmentSatisfaction“**

Представља задовољство пословним окружењем. Оцене су приказане на скали од 1 до 4, где оцена 1 означава да запослени уопште није задовољан док оцена 4 означава да је веома задовољан.

A graph with blue rectangular bars

Description automatically generated with medium confidence

Запослени су углавном задовољни својим пословним окружењем. Може бити добар предиктор.

**Колона „Gender“**

Представља пол запосленог.

A graph with blue squares

Description automatically generated

**Видимо да трећину запослених чине особе женског пола.**

**Колона „JobInvolvement“**

Ова колона одражава степен укључености запосленог у активности и одлуке у компанији.

A graph with blue squares

Description automatically generated

**Колона „JobLevel“**

Представља ниво позиције запослених у компанији. Ова колона обично категоризује запослене у различите нивое, који могу указивати на степен одговорности, сложеност посла и положај у организацији. Подела скупа података на основу ове колоне изгледа:

A graph of a graph with text

Description automatically generated with medium confidence

Добијамо очекивану расподелу, где се највише људи налази на нижим позицијама, а како позиције расту тако и број запослених опада.

**Колона „JobRole“**

Представља звање запосленог унутар компаније.

A graph with blue squares

Description automatically generated

**Колона „JobSatisfaction“**

Представља задовољство запосленог на пословном плану.

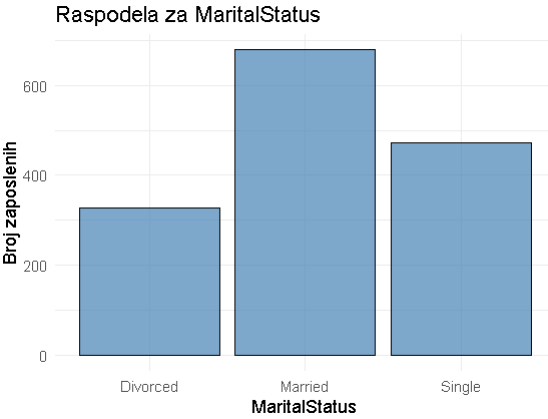
A graph showing a number of blue rectangular bars

Description automatically generated

Више запослених је задовољно самим послом, може бити један од важнијих предиктора.

**Колона „MartialStatus“**

Представља брачно стање запосленог.



Највише има запослених који су у браку. Можда у комбинацији са „*WorkLifeBalance“* подацима можемо добити корисне информације за наш модел.

**Колона „NumCompaniesWorked“**

Ова колона указује на број компанија у којима је запослени раније радио. Не укључујући тренутну компанију.

A graph with a line

Description automatically generated

Видимо да се у скупу података налази запослени који је чак 8 пута променио организацију. Можемо га избацити из овог скупа.

**Колона „PercentSalaryHike“**

Представља проценат последње повишице. Проценат је израчунат на основу уговорене месечне зараде.

A close up of a date

Description automatically generated

Видимо да је у последњој години најмања повишица износила 11% док је максимална износила 25% уговорене месечне зараде. Изгледа да немамо екстремне вредности али вредело би погледати и расподелу ове колоне.

A graph with a bar

Description automatically generated

Нема екстремних вредности. С обзиром да ова колона утиче позитивно на запослене може бити добар предиктор.

**Колона „PerformanceRating“**

Представља оцену перформансе запосленог. Оцене су приказане на скали 1 до 4.

A pie chart with a blue and red circle

Description automatically generated

Видимо да су дате оцене само 3 и 4. Као и да осмину скупа података чине запослени који имају изузетне перформансе. Можемо уклонити ову колону из даљих опсервација.

**Колона „RelationshipSatisfaction“**

Представља тренутни статус везе запослених на приватном плану. Ова информација може бити значајна за анализу различитих аспеката радне динамике и задовољства запослених.

**A graph showing a bar graph

Description automatically generated**

**Чини се да је расподела приближна и да ће нам можда користити у комбинацији са још неким предиктором.**

**Колона „StockOptionLevel“**

Показује да ли запослени има могућност да купује акције компаније. Највећи број запослених нема могућност поседовања акција компаније (0), опције 1 ,2 , 3 представљају ниво могућности поседовања акција компаније. Виши нивои представљају услове за привилеговане запослене на вишим пизицијама и дужим радним стажом, док нижи нивои представљају ограничену могућност куповине акција и мање повољне услове. Како бисмо лепше приказали односе, уместо нумеричких ознака користићемо текстуалне ознаке. Оцену 0 ћемо описати „Без опција“, 1 „Минимално опција“, 2 „Умерено“ и 3 „Значајно“.

A graph showing a number of blue rectangular bars

Description automatically generated with medium confidence

**Колона „TotalWorkingYears“**

Представља укупан радни стаж запосленог.

A graph with a line and a dot

Description automatically generated

Видимо да постоје екстремне вредности, њих ћемо обрадити у следећем делу рада.

**Колона „TrainingTimesLastYear“**

Представља колико је пута радник био на професионалном обучавању у претходној години.

A graph of a number of blue bars

Description automatically generated

Видимо да је расподела униформна и да не постоје екстремне вредности.

**Колона „WorkLifeBalance“**

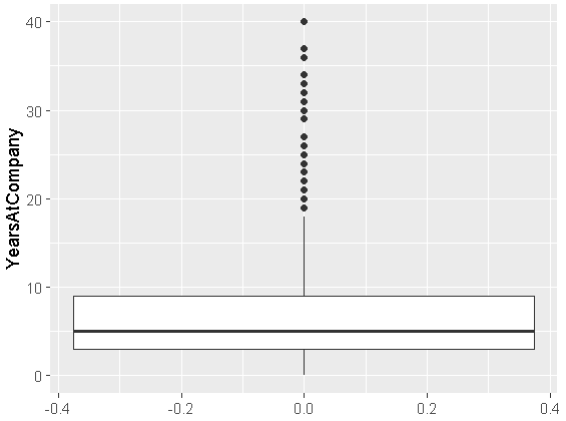
Представља ниво којим запослени успева да се посвети како пословном тако и приватном животу.

A graph of a bar chart

Description automatically generated

**Колона „YearsAtCompany“**

Представља колико година је запослени провео у фирми.



Видимо да и овде постоје екстремне вредности. То су запослени који су цео радни век провели у једној или две компаније.

**Колона „YearsInCurrentRole“**

Представља колико година је запослени провео на истој позицији.

A graph with blue squares

Description automatically generated

**Колона „YearsSinceLastPromotion“**

Представља колико година је прошло од последње промоције.

A graph with a bar

Description automatically generated

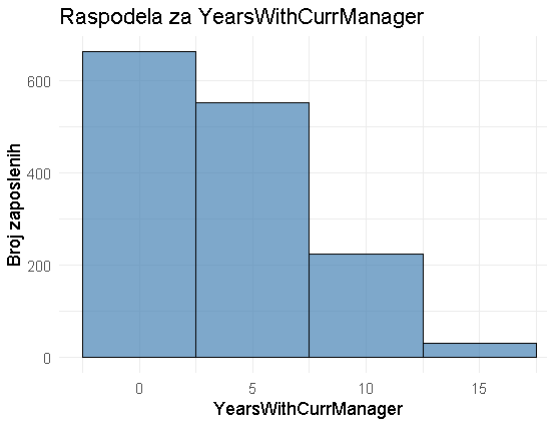
**Колона „YearsWithCurrentManager“**

Ова колона показује број година које запослени провео под тренутним менаџером.

A close-up of a number

Description automatically generated

Након попуњавања недостајућих вредности, где ћемо детаљне кораке представити у даљем раду, можемо погледати коначну расподелу.



Видимо да број запослених опада, што смо и очекивали.

## 2.2 Уклањање дупликата и неважних колона

С обзиром да смо раније уочили дуплиране уносе података колоне „EmpID“ и како нам ова колона није од превелике користи за наш модел, можемо проверити да ли су цели редови дупликати.

A close-up of a document

Description automatically generated

Пронашли смо укупно 7 дуплираних редова које ћемо уклонити из скупа података уз помоћ:

A close-up of a computer screen

Description automatically generated

Успешно смо уклонили 7 редова, сада се у скупу налази 1473 јединствених редова података.

Три преостала реда можемо сврстати под грешку при уносу колоне EmpID. С обзиром на то да ова колона не игра улогу у нашем моделу, можемо их оставити у скупу података без потребе за уклањањем.

За почетак уклонићемо податке везане за идентификацију запослених попут **EmployeeCount**, **EmpID**, **EmployeeNumber**:



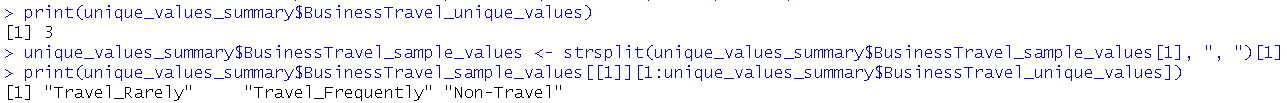
Даљом анализом приметили смо да су вредности колона **Over18** и **StandardHours** исте за све запослене, па ћемо их такође избацити:



## 2.3 Исправљање структурних грешака

Видели смо да је потребно извршити трансформацију података с обзиром да постоје уноси за TravelRarely и Travel\_Rarely који представљају исту учесталост. Уз помоћ функције „mutate“ претворићемо „TravelRarely“ у „Travel\_Rarely“ како би унос био конзистентан са осталим вредностима.





## 2.4 Уклањање екстремних вредности

С обзиром да се на графику налазе екстремне вредности можемо проверити да ли би нам ове вредности негативно утицале на модел. A graph showing a number of objects

Description automatically generated with medium confidence

Колико видимо са графика, постоји веза између прихода и колоне која описује напуштање запослених где је просечни месечни приход мањи код људи који су напустили организацију. Пошто се екстремне вредности налазе и код људи који су напустили фирму и оних који нису, нема потребе за избацивањем изузетака.

## 2.5 Руковање недостајућим вредностима

Ова колона садржи 57 Н/А вредности. Како бисмо што прецизније одредили средњу вредност, податке ћемо груписати на основу „*AgeGroup*“ колоне а затим за сваку групу извући њену просечну вредност којом ћемо заменити НА вредности.

A screenshot of a computer

Description automatically generated

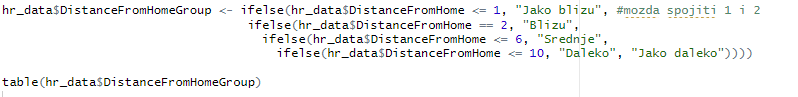
2.6 Креирање нових променљивих

DistanceFromHomeGroup

Примећујемо да највећи број људи живи на удаљености мањој од 1, док је удаљеност 10+ јако ретка. Због тога ћемо направити нову променљиву DistanceFromHomeGroup која ће садржати 5 категорија удаљености ради лакше анализе податка удаљености запосленог од куће.

Подела на групе:

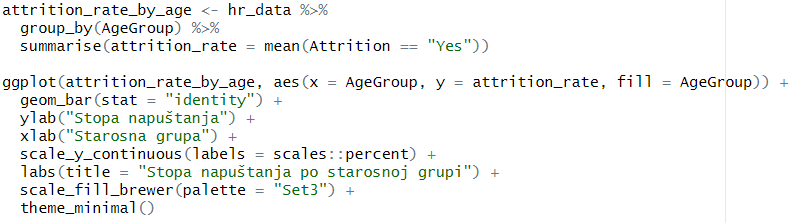
* Јако близу (0 до 1)
* Близу (1 до 2)
* Средње (2 до 6)
* Далеко (6 до 10)
* Јако далеко (10 и више)

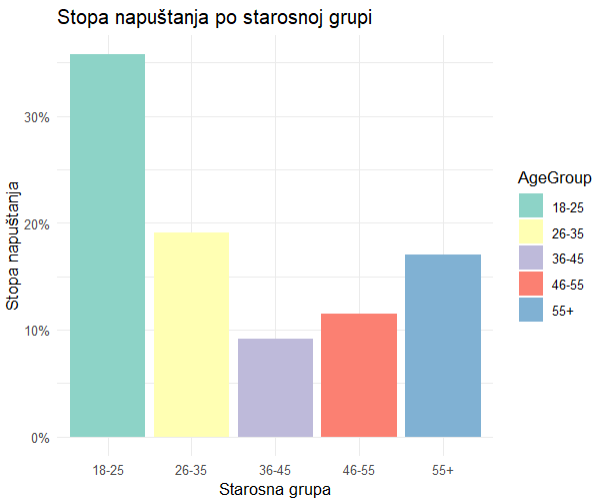


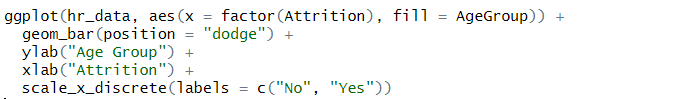
# Анализа података

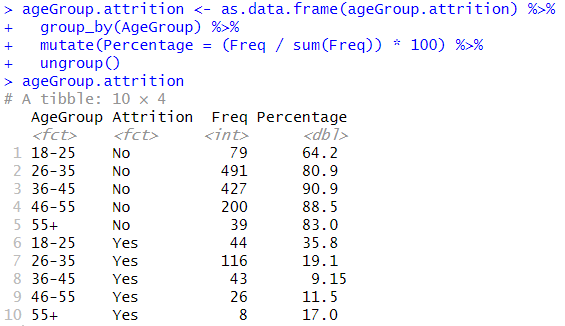
У овом поглављу ћемо покшати да уочимо повезаност између колона на основу досадашње анализе.

* **Age group и Attrition**









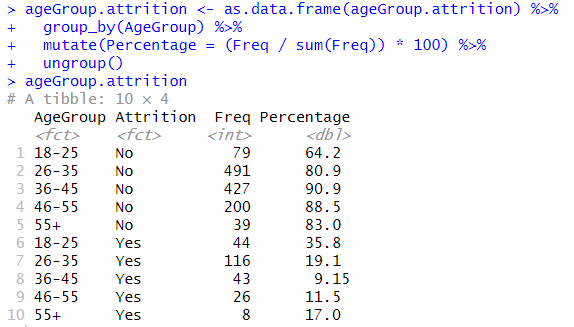
На основу графика можемо да закључимо да запослени који процентуално више напуштају компанију припадају млађим старосним групама. Тај тренд је најизраженији (35.8%) у размаку од 18 до 25 година и од 26 до 35 у мањој мери (19.1%), након тога је процентуално мања одлазност из компаније. Истиче се и група најстаријих од 55+ са степеном одлазности од 17% али њих има знатно мање, па нам овај податак није превише важан јер се вероватно ради о особама које одлазе у пензију или напуштају посао због година.



* **Age и Monthly Income**







У овом примеру смо упоредили године запослених и месечна примања и примећујемо да повезаност постоји. Најизраженија је за млађе запослене и можемо приметити како са порастом година расту и приманја уз доста изузезака након 30-те године.

* **Martial status и Bussines travel factor**

A graph of different colored squares

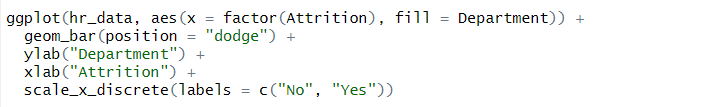
Description automatically generated with medium confidence

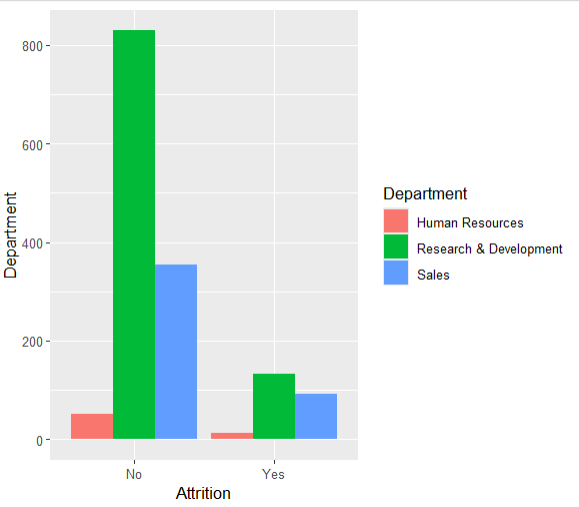
На овом графику можемо видети скалиран однос брачног статуса и учесталости пословних путовања запослених као и удео запослених који су напустили компанију.

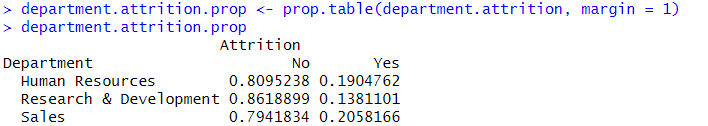
Видимо да су највећа удела оних који су отишли управо они који најчешће путују.

Процентуално најчешће компанију напуштају људи који (често) путују а нису у браку.

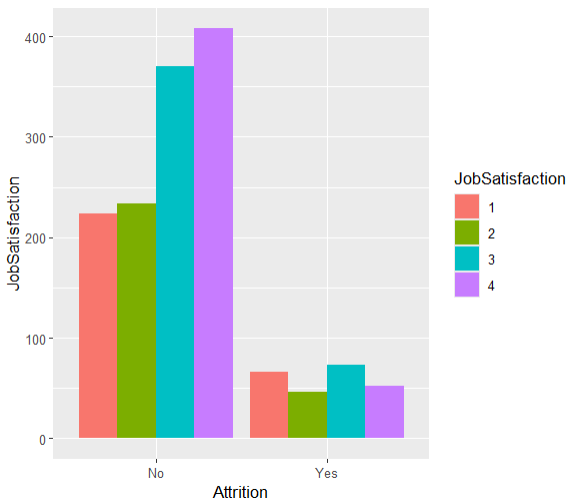
* **Attrition и Department**

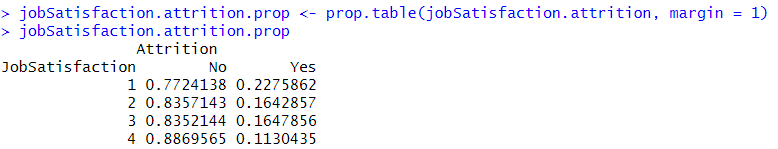




Примећујемо да је мала разлика у процентима међу одељењима, али да ”Sales” има највећи проценат људи који напуштају фирму.

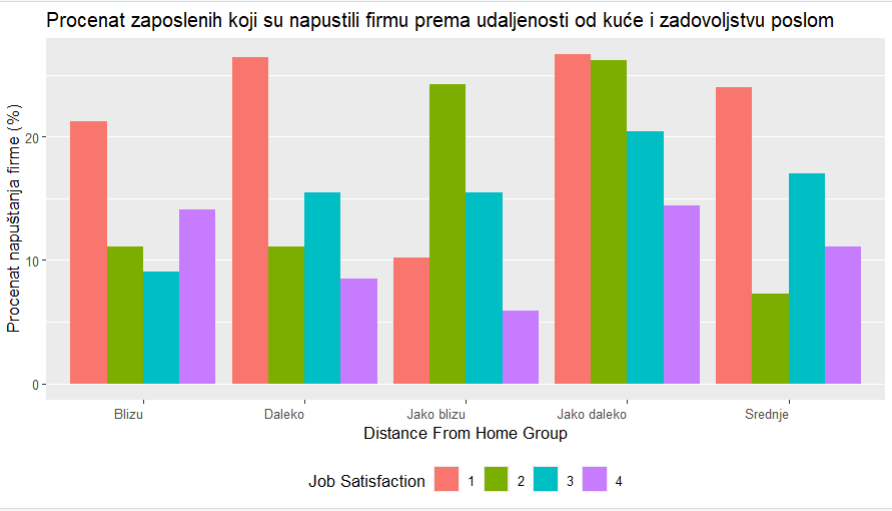
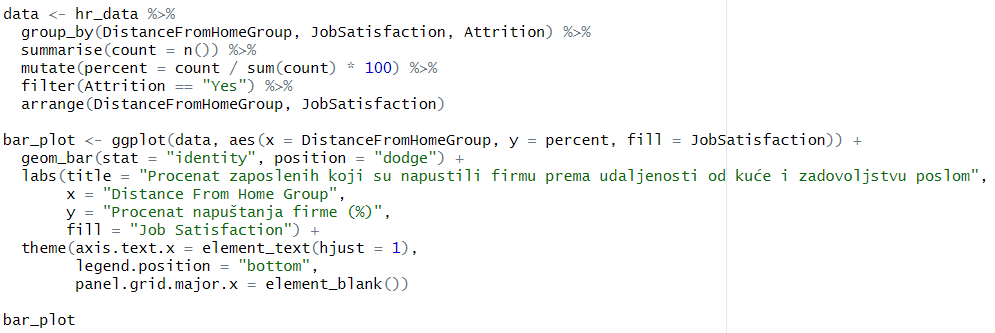
* **Job Satisfaction и Attrition**





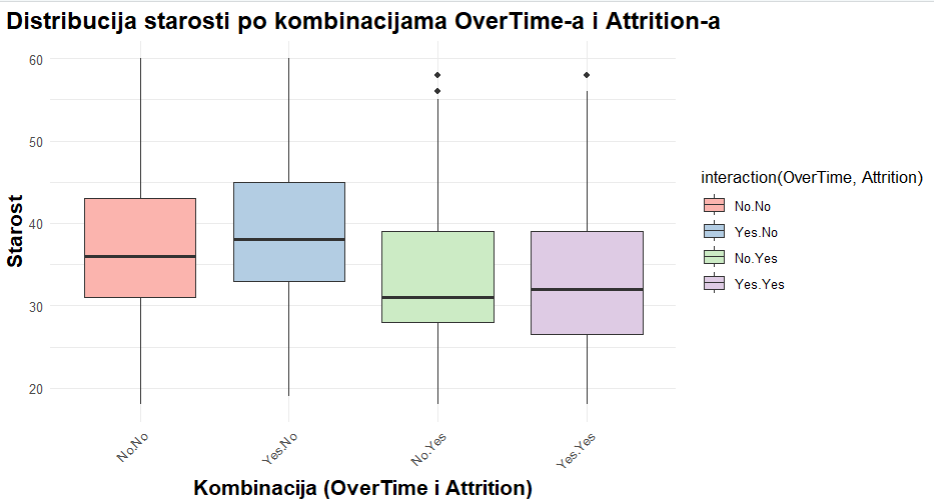
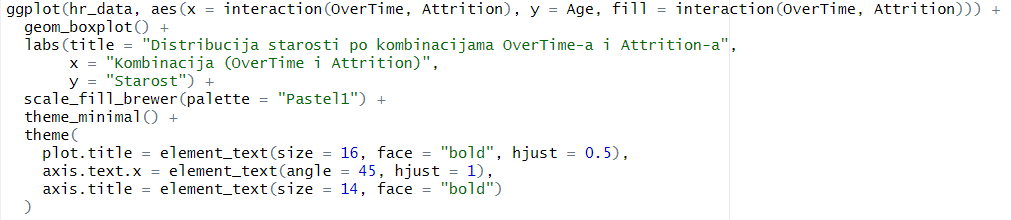
Запослени који су незадовољни послом чешће напуштају фирму, али то није превише изражено.

* **Distance from home и Job satisfaction за запослене који напуштају фирму (Attrition-Yes)**



Из овог графика можемо да закључимо да запослени који живе јако далеко од фирме чешће напуштају исту. Такође су то у највећем проценту запослени који нису задовољни послом (ниво 1 и 2). Удаљеност од посла од куће може бити јако добар предиктор.

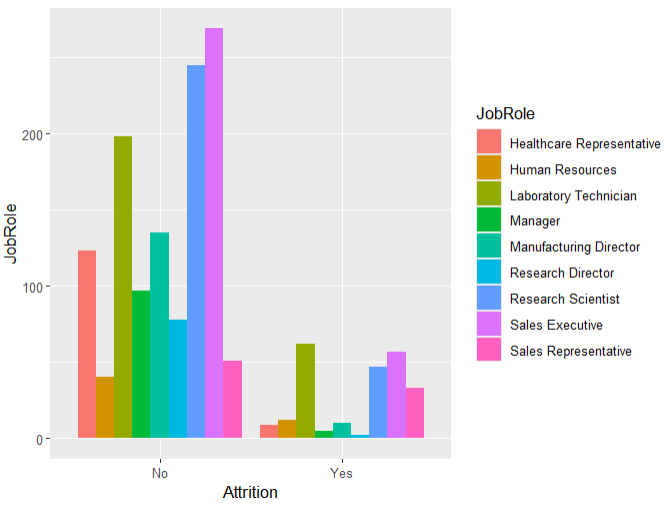
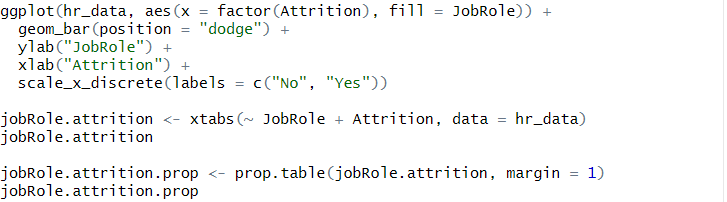
* **Age, OverTime и Attrition**

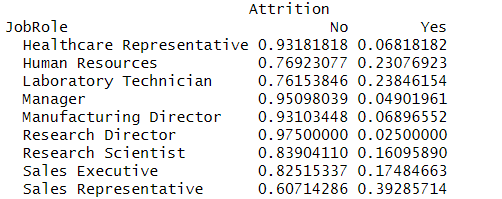


Примећујемо да комбинација предиктора прековремени рад и напуштање посла у односу на старост запосленог може бити корисна за креирање модела. Старије особе које раде прековремено не напуштају фирму, док млађе особе напуштају фирму без обзира да ли раде прековремено или не.

Постоји неколико аутлајер-а у категоријама запослених који напоштају фирму, што указује да постоје и случајеви старијих у тим категоријама.

* **Job role и Attrition**

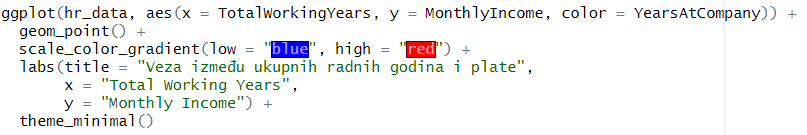


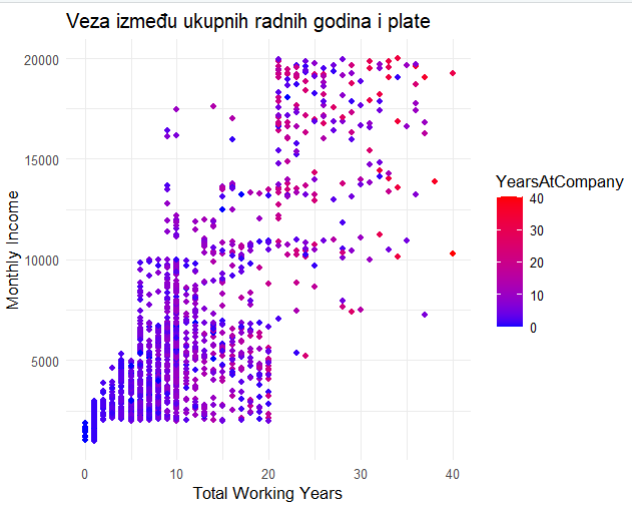


Позиције које у највећем проценту напуштају фирму јесу Sales Representative (39.28%), Laboratory Technician (23.84%) и Human Resources (23.07%).

Са друге стране позиције попут Research Director (2.5%), Manager (4.9%) и Manufacturing Director (6.89%) запослени скоро па не напуштају, што је и логично јер су то највише позиције.

* **Total working years и Years at company према Monthly Income**



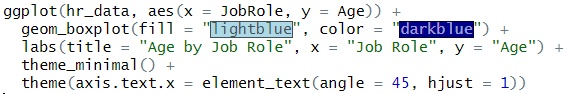


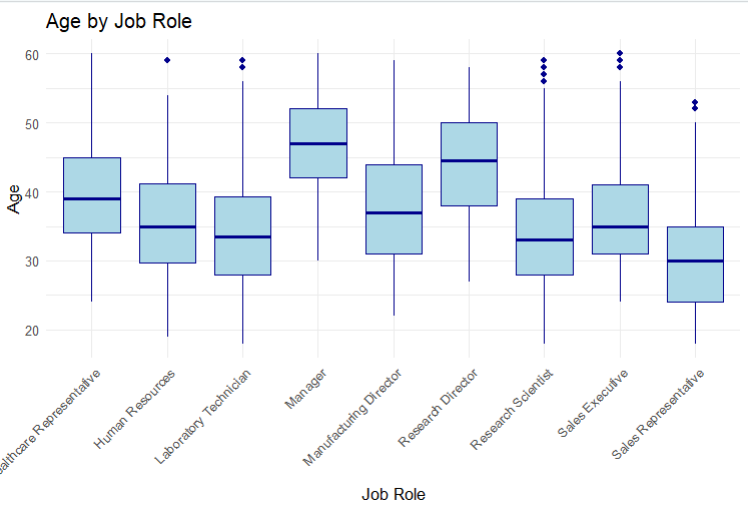
Са следећег графика закључујемо да посотји веза између радног стажа и плате. Пораст година радног стажа прати и пораст плате. Закључујемо и да број година у тренутној фирми не гарантује већу плату, већ је битније укупно радно искуство.

Након 20-те године радног стажа плате углавном достижу горњу границу (15000-20000).

Још један закључак са графика јесте да запослених са мањим радним искуством (0-10) има више и да имају ниже плате.

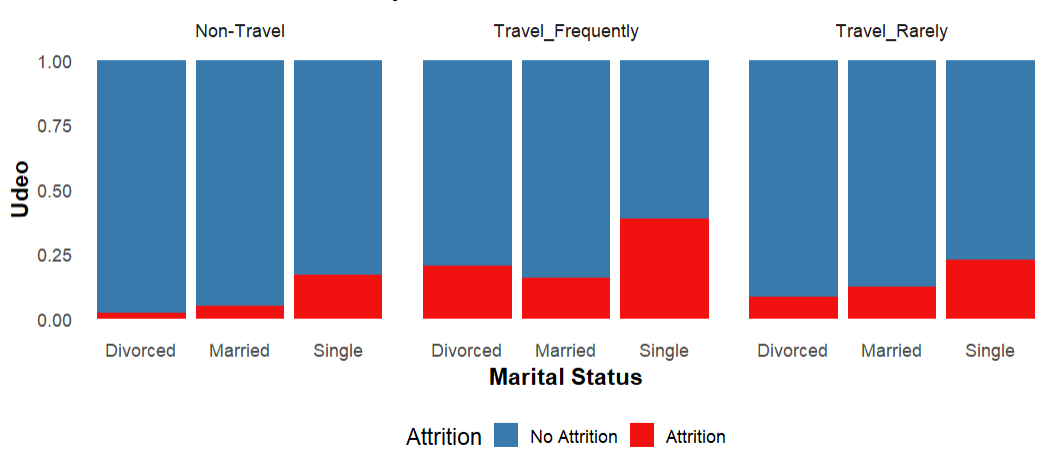
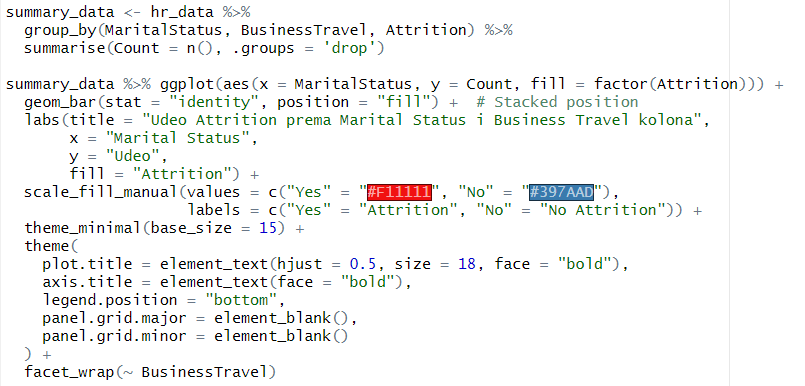
* **Job role и Age**





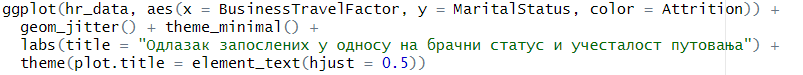
За више позиције у оквиру фирме, потребне су старије особе, млађе особе најчешће раде на нижим позивијама које се чешће напуштају.

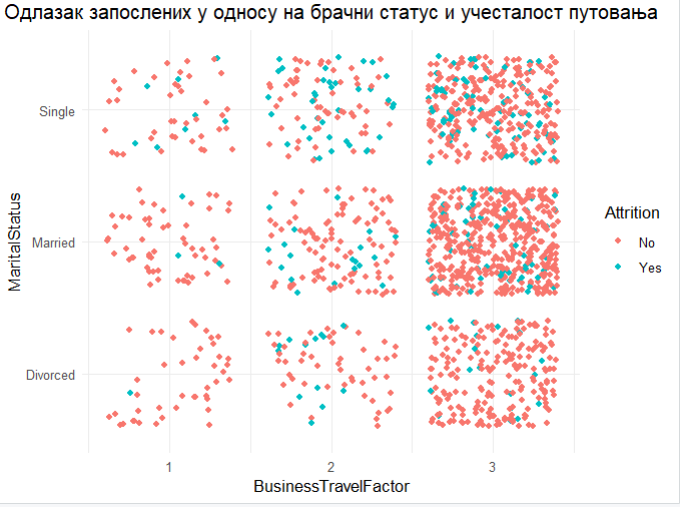
* **Martial status, Business Travel и Attrition**



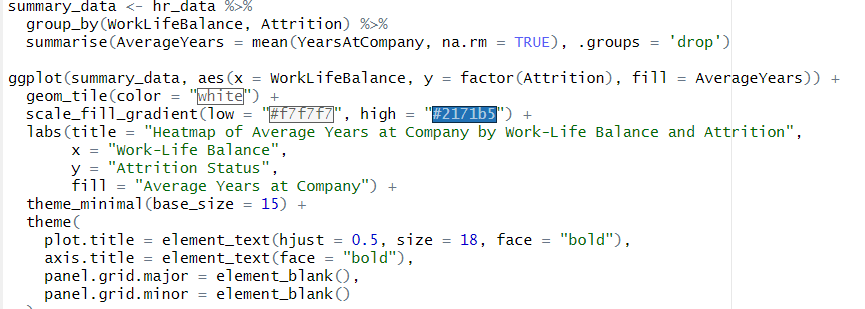
Са графика можемо закључити да посао најчешће напуштају запослени који често путују, а најчешће они који су слободни. Можемо и закључити и да невезано за учесталост путовања, слободни запослени најчешће напуштају, што је и логично.

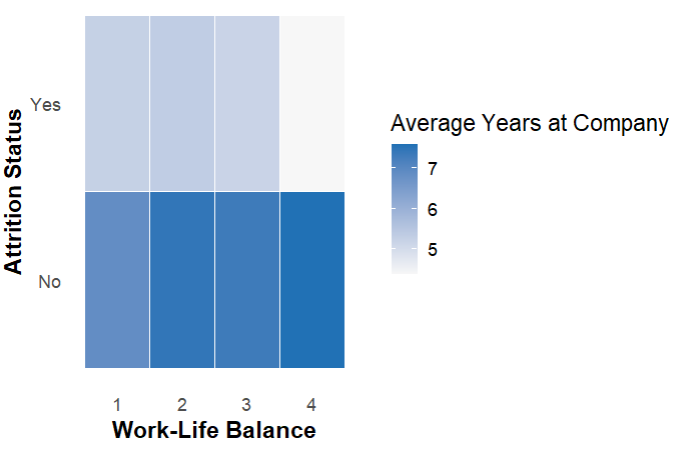
У наставку се налази још један график који нам је помогао да донесемо ове закључке.





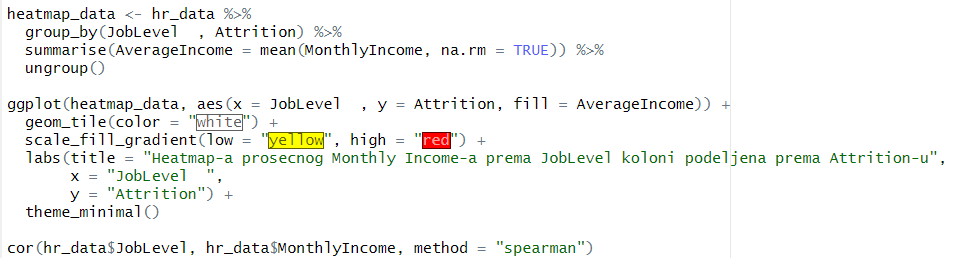
* **Work life balance, Average years и Attrition**

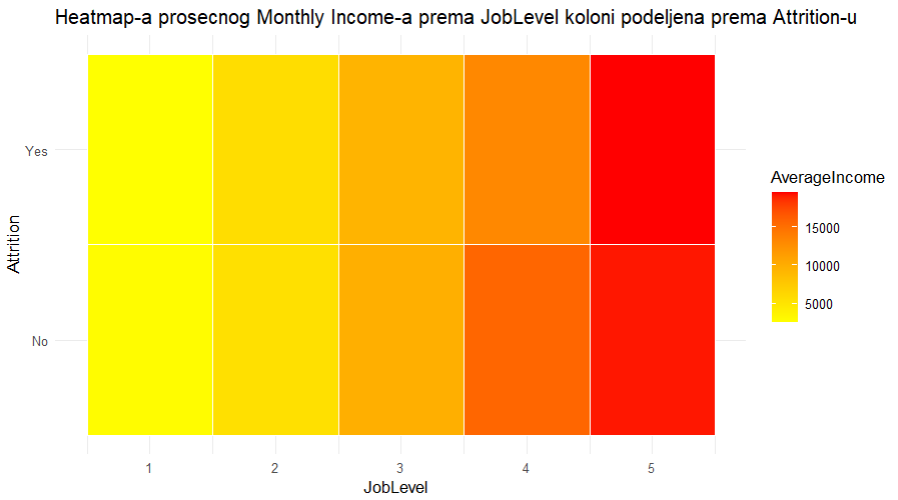




Закључујемо да запослени који су пронашли добар баланс дуже остају у фирми, што значи и да је мања вероватноћа да ће је напустити.

* Job level и Monthly income





На основу графика закључујемо да више позиције имају већа примања, али и такође да пораст месечних примања јесте пропорцијалан расту висини позиције и да ове две колоне имају велику корелацију која износи 0.919878, због чега колону JobLevel нећемо узимати у обзир у даљој анализи.

# Креирање модела

Пре креирања модела, потребно је извршити поделу скупа података на податке за тренинг као и податке за валидацију модела.

A close-up of a code

Description automatically generated

Као први корак користимо команду „set.seed(int)“ која нам помаже да реплицирамо случајни избор тестних и тренинг података.

Податке делимо уз помоћ функције „*createDataPartition“* из библиотеке „*caret*”. Уз помоћ ове функције поделили смо, случајним одабиром, овај скуп на два дела у односу 80:20. Где 80 процената скупа припада валидационом скупу док 20 тестном. Овај проценат смо одабрали због ограниченог обима података у нашем скупу. Потребно је обезбедити више података како би модел постао функционалан и пружио прецизније резултате.

Развијаћемо 4 модела: модел логистичке регресије, стабло одлучивања, „*Random Forest*“ модел и модел Неуронских мрежа.

Користићемо различите предикторе за сва три модела како бисмо постигли оптималне резултате. Као праг сигурности модела поставићемо вредност од 0,5, а за приказ резултата модела користићемо матрицу конфузије.

## 4.1 Логистичка регресија

За предикторе овог модела изабрали смо колоне „Age“, „OverTime“ и „YearsInCurrentRole“.

A screenshot of a computer code

Description automatically generatedA screenshot of a computer

Description automatically generated

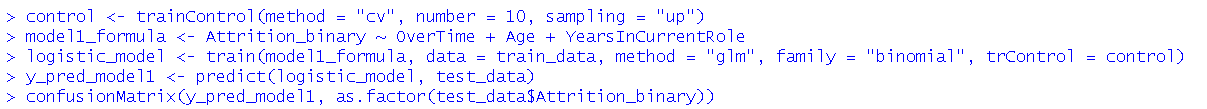
1Резултати логистичког модела

Према резултатима модела видимо да овај модел има високу прецизност као и сензитивност док је специфичност веома ниска. Балансирана прецизност модела износи 54% што је поприлично ниско у односу на пријављену прецизност.

Један од разлога за ове резултате представља то да је у тренинг скуп ушло само 5 особа које су напустиле компанију. Ово је веома мали узорак података на коме модел није успео да уочи довољно карактеристика.

### 4.1.1 Логистички модел и унакрсна валидација

Како бисмо побољшали резултате претходног модела можемо искористи десетоструку унакрсну валидацију. Као још један параметар овој функцији можемо проследи и параметар „*sampling*“ с вредношћу „*up*“ која нам помаже да избалансирамо податке.

A screenshot of a computer

Description automatically generated

У поређењу са првим моделом, модел са унакрсном валидацијом пружа бољу осетљивост и специфичност. Специфичност је знатно побољшана самим тим модел пруижа прецизније резултате.

Иако је укупна тачност опала, *Kappa* вредност је значајно боља што се осликава у томе да модел боље рефлектује стварне податке.

Следећи корак може бити смањење прага (енг. „treshold“) како бисмо повећали осетљивост модела.

## 4.2 GLM модел

Како бисмо постигли боље резултате, можемо размотрити коришћење GLM модела који нуди већу флексибилност у односу на логистички модел.

Такође, пробаћемо са другачијим предикторима, у овом случају „*OverTime*“, „*MonthlyIncome*“, „*JobInvolvement*“.

A black text on a white background

Description automatically generated A screenshot of a computer

Description automatically generated

Модел показује прецизност од 74% али Kappa вредност је ниска док је специфичност боља. Матрица конфузије потврђује да модел није успео да препозна ни један случај напуштања запослених, што указује на проблем у препознавању ове класе.

С обзиром да овај модел пружа знатно лошије резултате него претходни, можемо проверити да ли је проблем у моделу или је у предикторима. Покушаћемо да уз помоћ истих предиктора као у логистичком модела дођемо до бољих резултата.

A black text on a white background

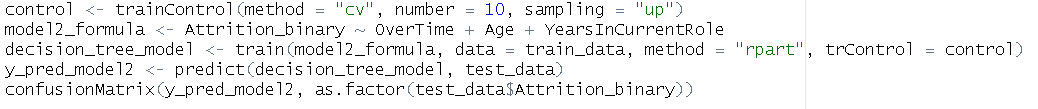
Description automatically generated

A screenshot of a computer

Description automatically generated

У овом случају добијамо мало боље резултате. Покушаћемо, исто као и са првим моделом, да уз помоћ десетоструке унакрсне валидације и *up-sampling-a* дођемо до бољих резултата.

### 4.2.1 GLM модел и унакрсна валидација



A screenshot of a computer

Description automatically generated

Као и код логистичко модела овде видимо знатно боље резултате у погледу специфичност и могућности модела да детектује запослене који су напустили компанију.

Прецизност овог модела износи 68.71%, док *Sensitivity* износи 70.85% а „*Specifity*“ је 57.45%. У овом случају *Kappa* вредност износ 0.195 што је ниже него побољшани први модел.

## 4.3 Random Forest модел

Знајући да *Random Forest* може боље обрадити сложеније податке и да већ користи унакрсну валидацију у сваком стаблу, што нам може помоћи да дођемо до боље генерализације, следећи корак ће нам бити управо креирање оваквог модела.

A black text on a white background

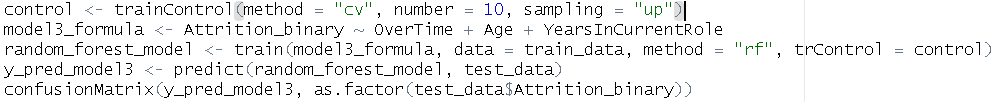
Description automatically generated

A screenshot of a computer

Description automatically generated

Видимо да поново долазимо до истих мана модела па ћемо онда одмах прећи на унакрсну валидацију уз примену *sampling-up* технике. Иако овај модел већ користи крос валидацију, она нам може помоћи у смањењу варијансе и пружити нам флексибилност у подацима, што је у нашем случају веома бито с обзиром на неуравнотеженост података.

### 4.3.1 Random Forest модел и унакрсна валидација



A screenshot of a computer

Description automatically generated

Као и у свим претходним случајевима, видимо опадање прецизности и осетљивости модела али пораст свих других вредности.

## 4.4 Неуронске мреже

Неуронске мреже су показале највећу прецизност у поређењу са осталим моделима. Резултати показују да мрежа најбоље препознаје запослене који напуштају компанију, са сензитивношћу од 81,78% и прецизношћу од 79,59%. Упркос добрим резултатима, специфичност је и даље нешто нижа (68,09%).

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

A black text on a white background

Description automatically generated A screenshot of a computer

Description automatically generated

A computer code with black text

Description automatically generated

A diagram of a network

Description automatically generated

На графику је приказана структура неуронске мреже која се користила у моделу. Мрежа се састоји од улазног слоја, једног скривеног слоја са 10 неурона и излазног слоја. Сваки чвор (неурон) у скривеном слоју повезан је са свим чворовима у претходном и следећем слоју.

Као најважнији предиктори показали су се „OverTime“ и „MonthlyIncome“, док су остали предиктори допринели мањем степену варијације у резултатима.

## 4.5 Резултати модела

Да бисмо упоредили ова три модела на једноставан начине, направићемо табелу у којој ће колоне садржати резултате појединачних модела а редови одговарајуће метрике.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Логистичка регресија | ГЛМ модел | Random forest | Неуронске мреже |
| Прецизност | 73,13% | 68,71% | 72,45% | 79,59% |
| Балансирана прецизност | 69,36% | 64,15% | 62,93% | 74,93% |
| Сензитивност | 74,9% | 70,85% | 76,92% | 81,78% |
| Специфичност | 63,83% | 57,45% | 48,94% | 68,09% |
| *Kappa* вредност | 0,2791 | 0,1954 | 0,2014 | 0,3963 |

Неуронске мреже се показују као најбољи модел према свим кључним метрикама.

## 4.6 Resampling методе

A graph with a bar and a rectangle

Description automatically generated

Имамо приближно 84% негативних случајева и 16% позитивних што указује на небалансиране податке колоне Attrition. Кључни проблем небалансираних података може бити пристрасност модела према већинској класи и имати потешкоћа у идентификацији мањинске.  
  
Покушаћемо да добијемо прецизније моделе коришћењем Resampling метода из ROSE библиотеке.

с

### 4.6.1 Oversampling

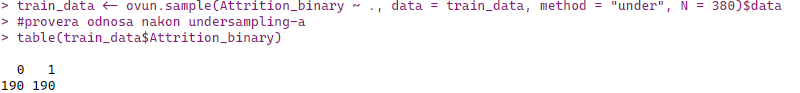
Коришћењем методе oversampling-а ћемо уједначити број негативних узорака са бројем позитивних.

A white background with black text

Description automatically generated

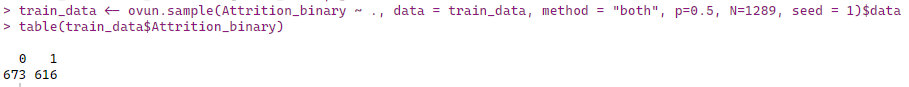
### 4.6.2 Undersampling метода

Коришћењем ове методе прилагодили смо број узорака класе са позитивним вредностима броју негативних.



### 4.6.3 Oversampling и Undersampling

Овом методом смо избалансирали вредности класе и сада имамо скоро идентичне бројеве негативних и позитивних вредности.



### 4.6.4 Резултати

Resampling методе коришћене су за решавање проблема небалансираности података. Најпре је примењен oversampling, где је број узорака мањинске класе повећан како би се изједначио са већинском. Затим је примењен undersampling, где је смањен број већинске класе. Комбиновањем обе методе, постигнута је равнотежа класа.

Иако је баланс постигнут, примена resampling метода није значајно побољшала параметре модела. Наши модели показали су сличне резултате пре и након примене ових метода. Ово указује да небалансираност није била доминантан фактор у прецизности модела, већ да је квалитет података и избор предиктора имао значајнији утицај.

# Закључак

Анализа је усмерена на предвиђање запослених код којих постоји највећа могућност да напусте компанију користећи податке Људских ресурса из претходно урађених анкета.

Израђени модели укључују логистичку регресију, *GLM* модел, модел *Random Forest* и модел неуронских мрежа*.* Ови модели пружили су нам увид у вероватноћу задржавања запослених са различитим процентима тачности.

Након упоређивања перформанси модела на основу кључних метрика, закључили смо да модел логистичке регресије даје најпрецизније резултате. Овај резултат сугерише да је однос између варијабли унутар скупа података линеаран.

# Литература

* *Увод у науку о подацима* - вежбе, предавања и материјали предмета
* *GGPlot2* - [документација](https://ggplot2.tidyverse.org/reference)
* *Caret* - [документација](https://topepo.github.io/caret/)
* *GLM* - [документација](https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/glm)
* *Neural Networks -* [документација](https://rpubs.com/chelseyhill/733053)