Analiza bankovnih marketinških podataka

Studentski tim: Tina Bakić
Mateo Elez
Josip Prpić

Nastavnik: Damir Pintar
Mihaela Vranić

Tehnička dokumentacija

1. Uvod

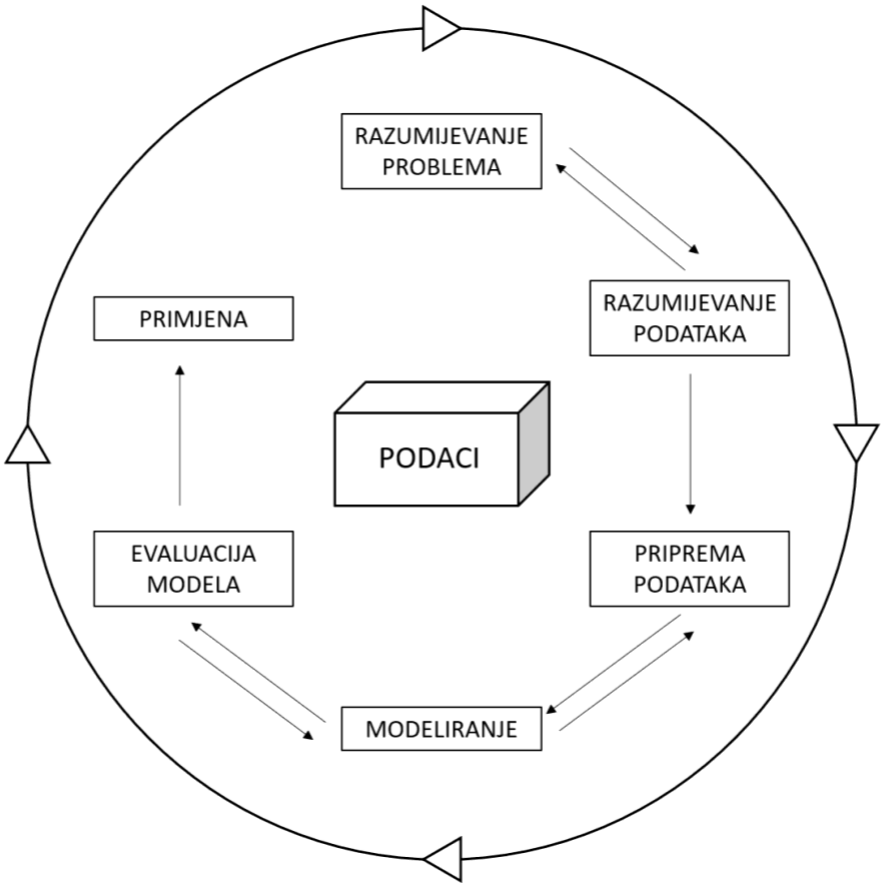
Danas, u svakoj sferi života postoji velika količina podataka. Područje za koje smo se odlučili u sklopu Projekt R je analiza poslovnih podataka. Ono što smo htjeli je ovisno o karakteristikama osobe kao što je dob, obrazovanje, bračni status i sl., odrediti kreditne usluge koje bi pojedina osoba mogla koristiti te ujedno provjeriti što nam još može biti dobar pokazatelj uz naše početne pretpostavke.

Skup podataka korišten u sklopu projekta preuzet je s platforme www.kaggle.com. Radi se o podacima prikupljenim telefonskim anketiranjem za potrebe Portugalske banke. Anketiranje je provedeno u periodu od svibnja 2008. godine do studenog 2011. godine. Cijeli skup podataka te njegove značajke moguće je proučiti na sljedećoj poveznici Bank Marketing Data Set | Kaggle.

Uz važnost prikupljanja, važnija je kvalitetna obrada i tumačenje podataka jer tek time imamo mogućnost razumijevanja i shvaćanja što nam prikupljeni podaci ustvari označavaju te je upravo to ono što smo se potrudili što bolje razviti u sklopu našeg projekta.

Najveću ulogu u tome imala je dubinska analiza podataka. To je proces otkrivanja globalnih pravilnosti i odnosa u velikim podatkovnim skupovima, u kojima su zbog prevelikog broja i nepreglednosti upravo te značajke skrivene. Za analizu podataka, dubinska analiza podataka služi se metodama iz računarske znanosti, statistike i strojnog učenja.

Slika 1.: Prikaz presjeka na kojem se nalazi dubinska analiza podataka



Slika 2.: Osnovne faze dubinske analize podataka

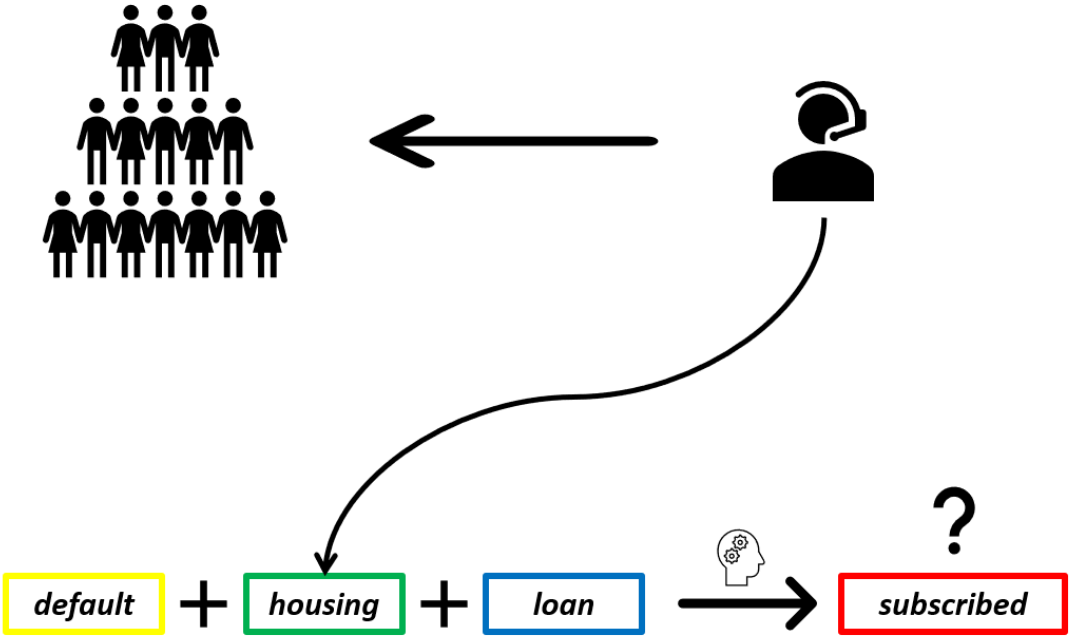
2. Cilj projekta

Cilj našeg projekta bilo je analizirati odabrani skup podataka, očistiti podatke, uredit ih i među njima pronaći regresore koji stvaraju dovoljno dobar model za kategoričke podatke s kojima raspolažemo, a vezani su za korištenje bankovnih usluga kredita.

Kategorički podaci u našem skupu podataka su: *loan, housing, default* i *subscribed*. Dok nam varijabla *default* ne govori puno obzirom da su se ispitanici izjašnjavali u 79% slučajeva s „no“, a u ostalih 29% s „unknown“, tj. nepoznatom vrijednošću, *loan* i *housing* imaju značajan utjecaj na varijablu *subscribed*. Upravo varijabla *subscribed* od najvećeg je značaja za banku i provedeno ispitivanje po kojem je formiran skup podataka koji koristimo.

Važnost te varijable i razlog zbog kojeg ju banke promatraju je ta što općenito u sklopu marketinških anketa provjerava se što je to što klijenti žele, a varijabla *subscribed* govori nam je li se ispitanik odlučio za oročeni depozit u ovoj banci, tj. za polaganje svog novca na određeno vrijeme upravo u ovoj banci, što bi trebao biti pokazatelj vjernosti klijenta.

U trenutku pronalaska povezanosti između varijabli o korištenju kreditnih usluga (*loan, housing*), dobivamo informaciju koja rješava značajni marketinški zadatak za što veću uspješnost banke. A ono što nam to pokazuje je kojem tipu ljudi treba usmjeriti marketinške pozive banke te ujedno na temelju podataka imamo mogućnost predviđanja hoće li osoba ostaviti depozit u toj banci ili neće, a to je vrijednost varijable *subscribed*.



Slika 3.: Prikaz procesa

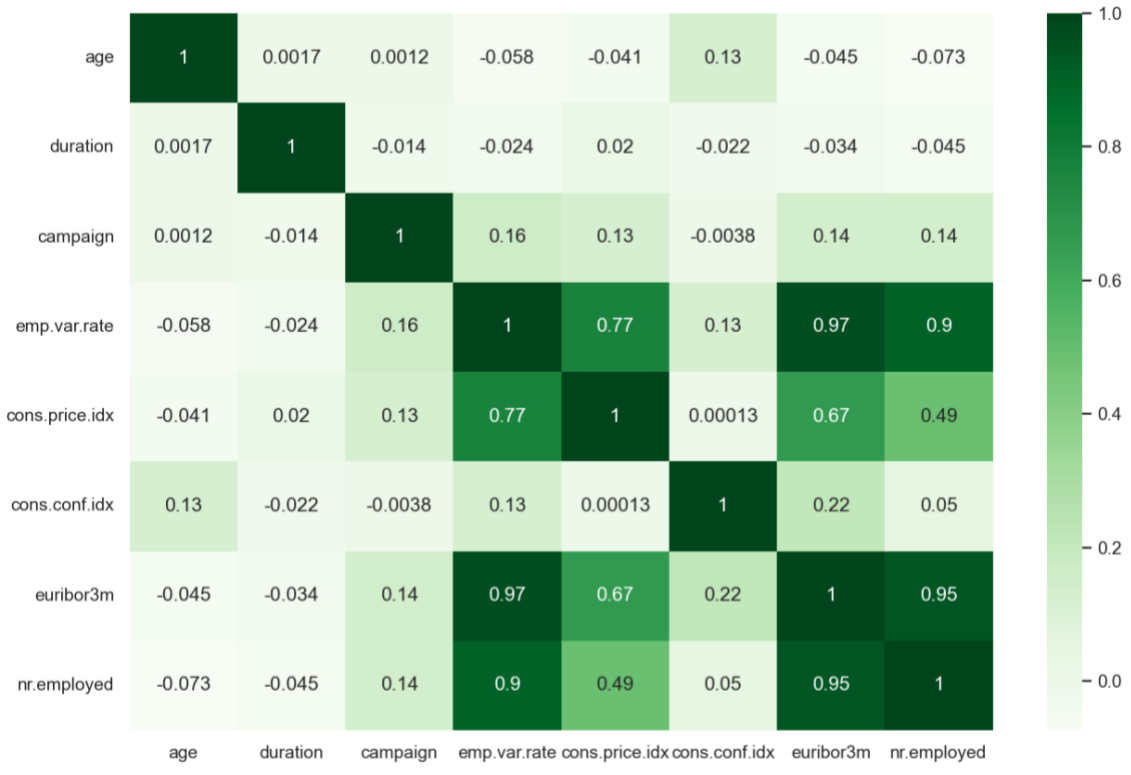
3. Eksploratorna analiza i pretprocesiranje

3.1 Pročišćavanje podataka

Početni skup podataka sastojao se od 21 (dvadeset i jednog) atributa te s preko 41000 (četrdeset i jedne tisuće) uzoraka. Nakon ukupne obrade koja će detaljnije biti opisana u nastavku sveli smo podatke na 15 (petnaest) atributa i nešto malo više od 30000 (trideset tisuća) uzoraka, s kojima smo nastavili raditi projekt.

Najprije smo izbacili atribute koji se odnose na ranije provedene ankete, čiji podaci nisu trenutni podaci koje koristimo. Zatim uzorke koji ukazuju na neuspješnost provedbe ankete kao što je trajanje poziva manje od 35 sekundi za koje vrijeme anketa ne može biti ispunjena te ostale nepoznate vrijednosti označili smo i sve uzorke koje sadrže za barem jedan atribut nepoznatu vrijednost izbacili smo iz skupa podataka. Za podatke koji su za bračni status stavili nepoznatu vrijednost postavili smo da su slobodni, za nepoznatu vrijednost zaposlenja postavili smo da je osoba nezaposlena, a za nepoznatu vrijednost kod obrazovanja postavili smo da osoba nije išla u školu.

Nakon uređivanja unesenih podataka, provjerili smo koji su atributi korelirani. Obzirom da bi nas prekorelirani prediktori mogli navesti na krivi zaključak, izračunom korelacije odlučili smo određene atribute izbaciti te ostaviti samo jedan od tih koreliranih.



Slika 4.: Matrica koreliraonosti

Po dobivenim rezultatima prikazanim na Slika 4., izbacili smo *emp.var.rate, euribor3m* i *nr.employed* te ostavili *cons.price.idx*.

Za jednostavniji rad s podacima kategorijske podatke kao što je vrsta posla, razina obrazovanja, dan u tjednu i ostale vrijednosti unesene u obliku stringa, prebacili smo u brojčane vrijednosti što konkretno možete vidjeti dijelom koda prikazanom na Slika 5.

```
job_types = {'admin.': 1, 'blue-collar': 2, 'entrepreneur': 3, 'housemaid': 4,
             'management': 5, 'retired': 6, 'self-employed': 7,
             'services': 8, 'student': 9, 'technician': 10, 'unemployed': 11}

marital_status = {'divorced': 1, 'married': 2, 'single': 3}

edu = {'basic.4y': 1, 'basic.6y': 2, 'basic.9y': 3, 'high.school': 4,
       'illiterate': 5, 'professional.course': 6, 'university.degree': 7}

contact_type = {'cellular': 1, 'telephone': 2}

months = {'jan': 1, 'feb': 2, 'mar': 3, 'apr': 4, 'may': 5, 'jun': 6,
          'jul': 7, 'aug': 8, 'sep': 9, 'oct': 10, 'nov': 11, 'dec': 12}

days = {'mon': 1, 'tue': 2, 'wed': 3, 'thu': 4, 'fri': 5}
```

Slika 5.: Isječak programskog koda koji prikazuje koje numeričke vrijednosti će zamijeniti koje kategoričke vrijednosti

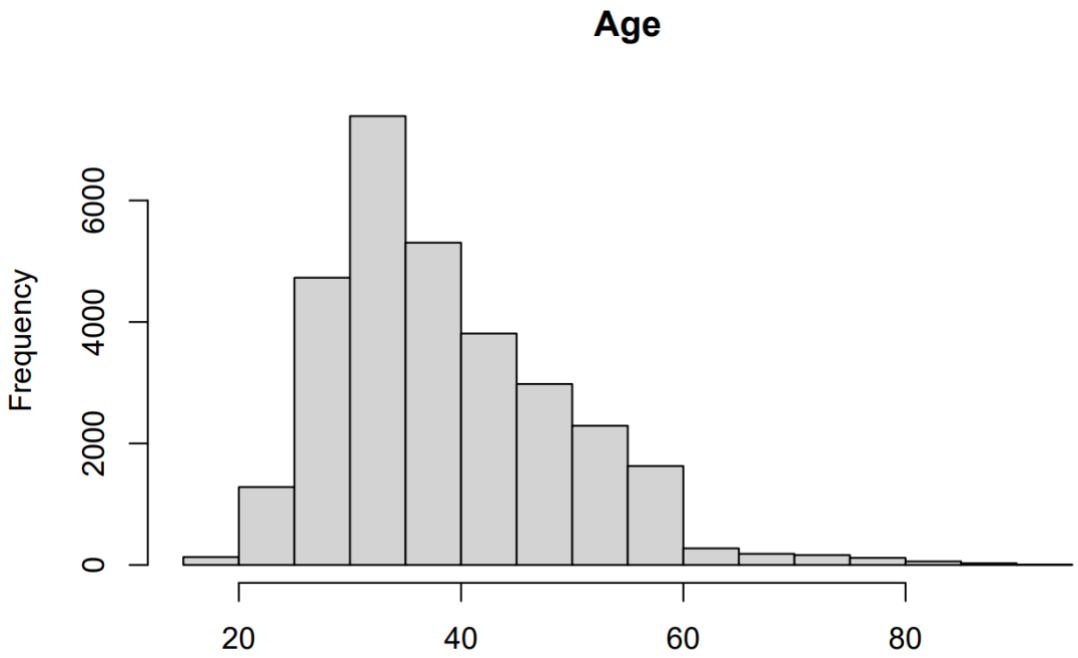
Istu stvar bilo je potrebno napraviti s binarnim podacima o korištenju kreditnih usluga, koji mogu poprimiti samo vrijednost „yes“ ili vrijednost „no“, stoga smo sve vrijednosti „yes“ pretvorili u numeričku vrijednost 1, a sve „no“ u numeričku vrijednost 0.

3.2 Vizualizacija varijabli

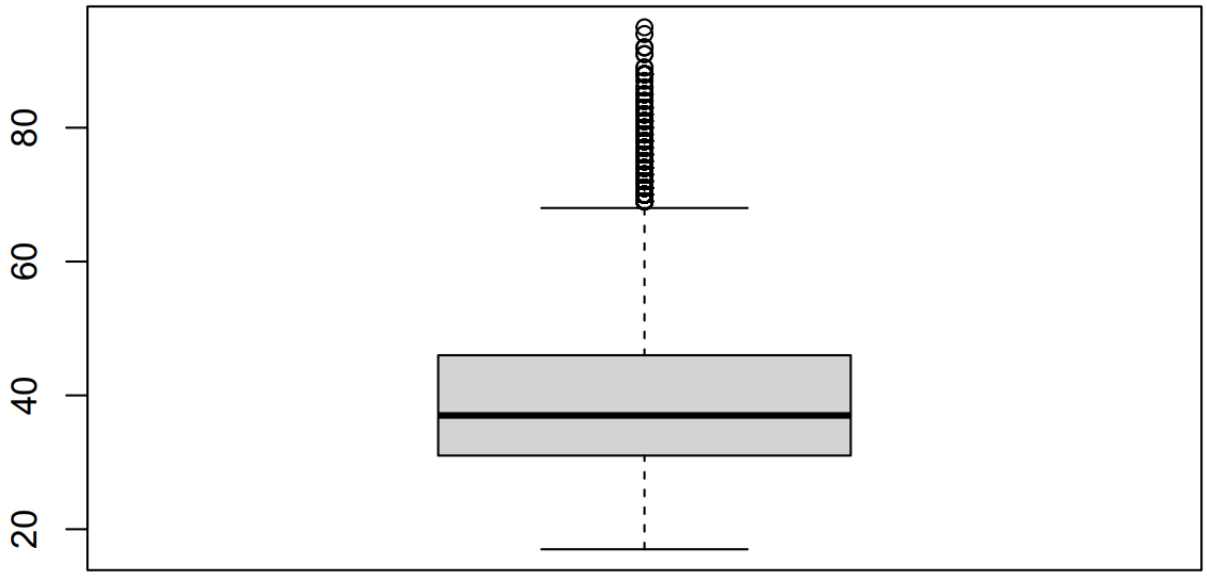
Zbog lakšeg razumijevanja i iščitavanja skupa podataka, u programskom jeziku R, vizualizirali smo najbitnije varijable s vrijednostima koje su unesene anketiranjem korisnika. Podaci su prikazani pomoću histograma, a funkcijom *boxplot*() dobili smo prikaz prosjeka varijabli te postojanih outliera, tj. stršećih vrijednosti.

Navedena vizualizacija nam je bila od velikog značaja za što konkretnije upoznavanje s podacima s kojima raspolažemo, u što kraćem vremenskom periodu te za jasniju sliku o tipu ispitanih korisnika, kao i za provjeru koje su varijable važne i ukazuju na nešto, tj. koje to nisu. Bez toga, početak rada i stvaranje konkretno projekta trajalo bi puno duže.

U nastavku možete vidjeti neke od varijabli koje smo odlučili izdvojiti te prikazat podatke i ukratko opisati što smo iščitali iz podataka.

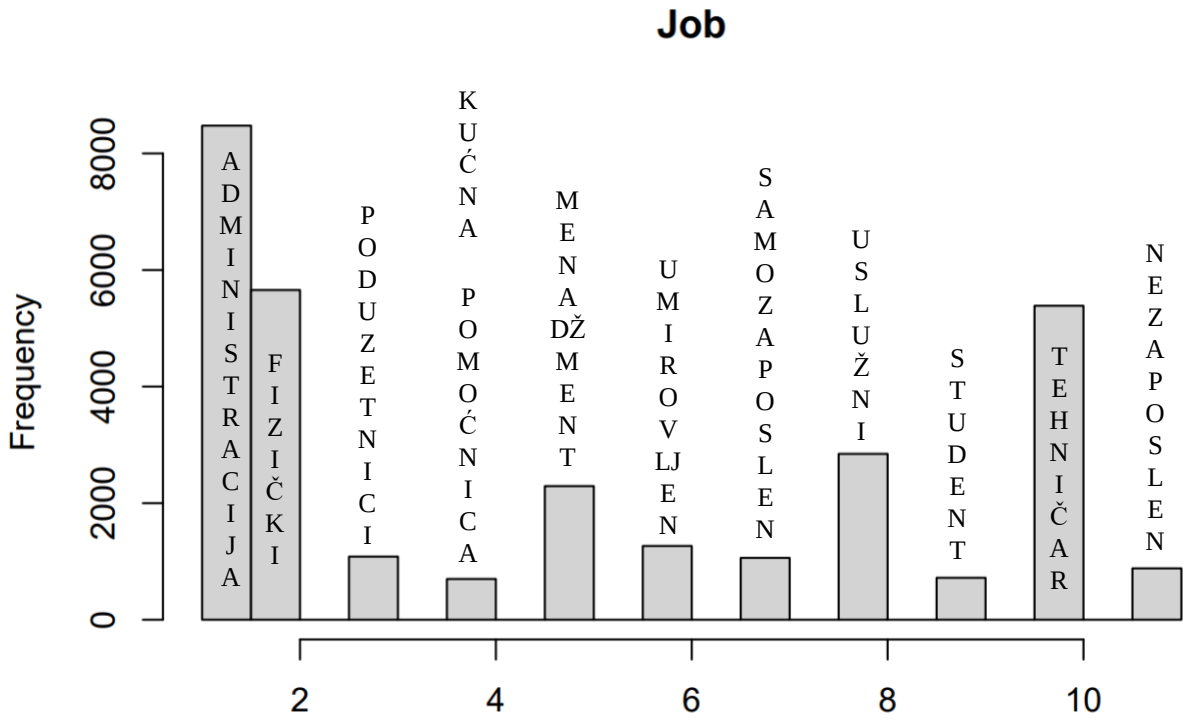


Slika 6.: Histogram varijable *age* (godine ispitanika)



Slika 7.: Boxplot varijable *age* (godine ispitanika)

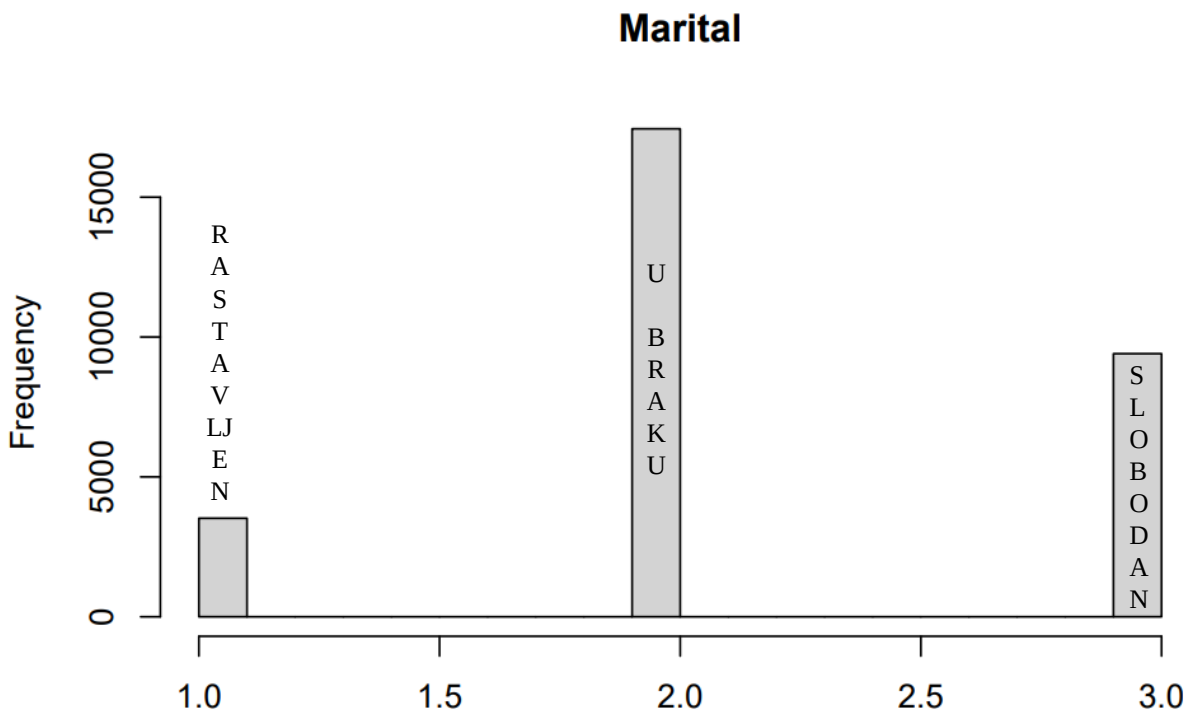
Slika 6. i Slika 7. vizualizacija su varijable godine (*age*) ispitanika čija je dob na intervalu od 18 do 70 godina. Iz Slika 7. možemo iščitati da je prosjek godina ispitanika malo manji od 40 godina, a stršeće vrijednosti uočljive su iznad 70 godina.



Slika 8.: Histogram varijable *job* (posao ispitanika)

Na Slika 5., prikazana je transformacija kategoričkih podatak u numeričke, na kojoj možemo vidjeti koje zanimanje je poprimilo koju brojčanu vrijednost. No, zbog lakšeg razumijevanja Slika 8. na histogramu za svaki stupac napisana je kategorička vrijednost.

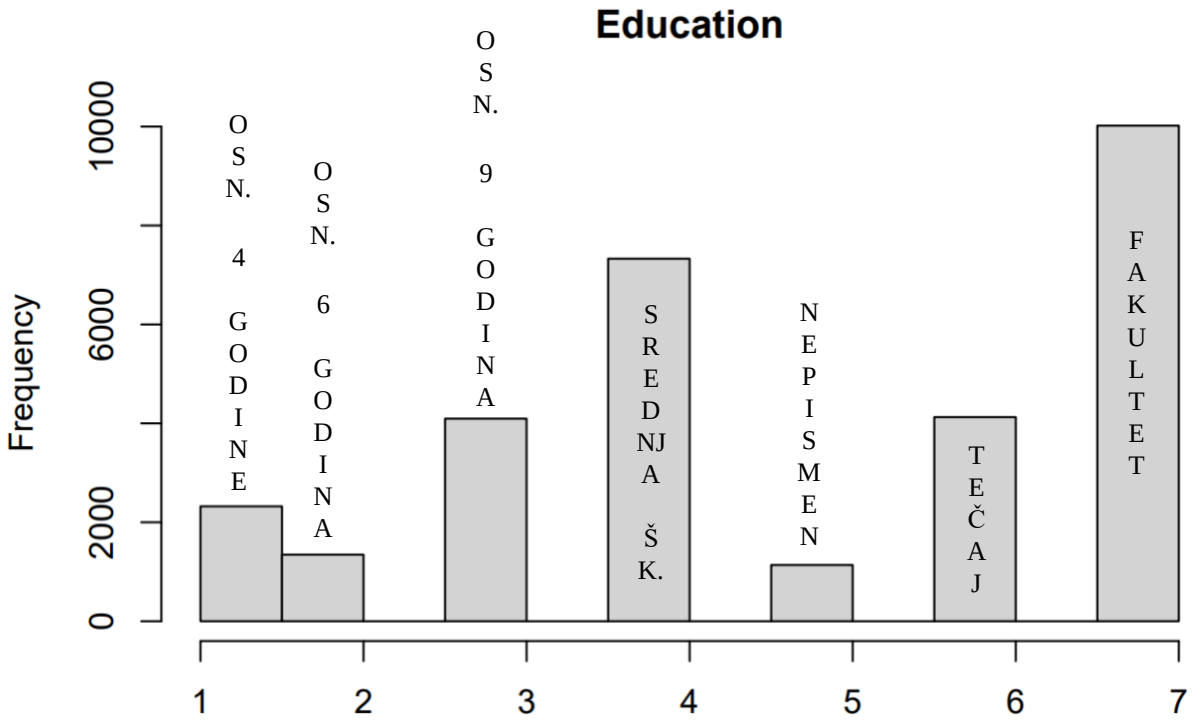
Iz histograma sa Slika 8. možemo zaključiti da najveći broj ispitanika radi u administraciji, zatim fizički posao i tehničke poslove. Najmanji broj ispitanika su kućne pomoćnice, samozaposleni, studenti te nezaposleni.



Slika 9.: Histogram varijable *marital* (bračni status ispitanika)

Marital, tj. bračni status, također je varijabla s kategoričkim podacima koje je za potrebe jednostavnijeg korištenja bilo potrebno pretvoriti u numeričke podatke te su zbog bolje preglednosti na histogramu dodani opisni nazivi stupaca.

Iz histograma sa Slika 9. možemo zaključiti da je najviše ispitanika u braku, a najmanje onih rastavljenih.



Slika 10.: Histogram varijable *education* (obrazovanje ispitanika)

Obrazovanje ispitanika, također je jedna od varijabla koja sadrži kategoričke podatke koje smo kao što je prikazano na Slika 5. morali transformirati. Na histogramu, podjela je po numeričkoj vrijednosti te su zbog lakšeg razumijevanja dodani opisni nazivi.

Kod najvećeg broja ispitanika, najviši stupanj obrazovanja je fakultet te srednja škola, a najmanji je broj onih koji su nepismeni te sa završenih samo 4 ili 6 godina osnovne škole što možemo vidjeti po histogramu sa Slika 10.

3.3 Vizualizacija odnosa s varijablama kredita

Nakon prikaza vizualizacije određenih varijabli, ono bitnije nama za projekt je odnos koji te varijable imaju s varijablama vezanim za kredit, a to su:

- *housing* – ima li ispitanik stambeni kredit
- *loan* – ima li ispitanik privatnu pozajmicu
- *default* – ima li ispitanik kredit u neplaćanju

U nastavku prikazane su vrijednosti podudaranja varijabli godine ispitanika, bračni status, obrazovanje te posao s gore navedenim varijablama za kredit.

3.3.1 Godine i krediti ispitanika

Ispitanike smo podijelili na 3 starosne skupine:

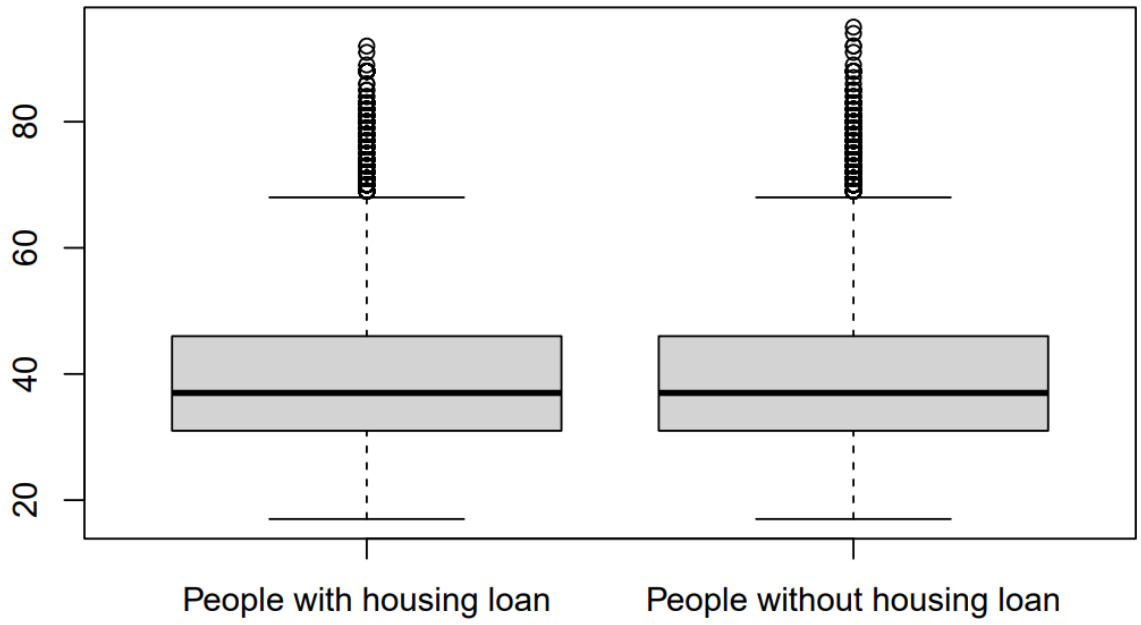
I. mladi: do 25 godina
II. srednje mladi: između 25 i 45 godina
III. stariji: od 45 godina

Zatim smo ih podijelili i po tome koji oblik kredita imaju, a koji nemaju.

```
## Prosječni postotak mladih ljudi koji imaju stambeni kredit  55.02833

## Prosječni postotak srednje starih ljudi koji imaju stambeni kredit  53.73711

## Prosječni postotak starijih ljudi koji imaju stambeni kredit  54.45839
```

Slika 11.: Isječak rezultata odnosa godina i stambenog kredita

Slika 11. prikazuje slične rezultate odnosa stambenog kredita sa starosnim skupinama, no najveći postotak ljudi u skupini sa stambenim kreditom je kod mladih ljudi.



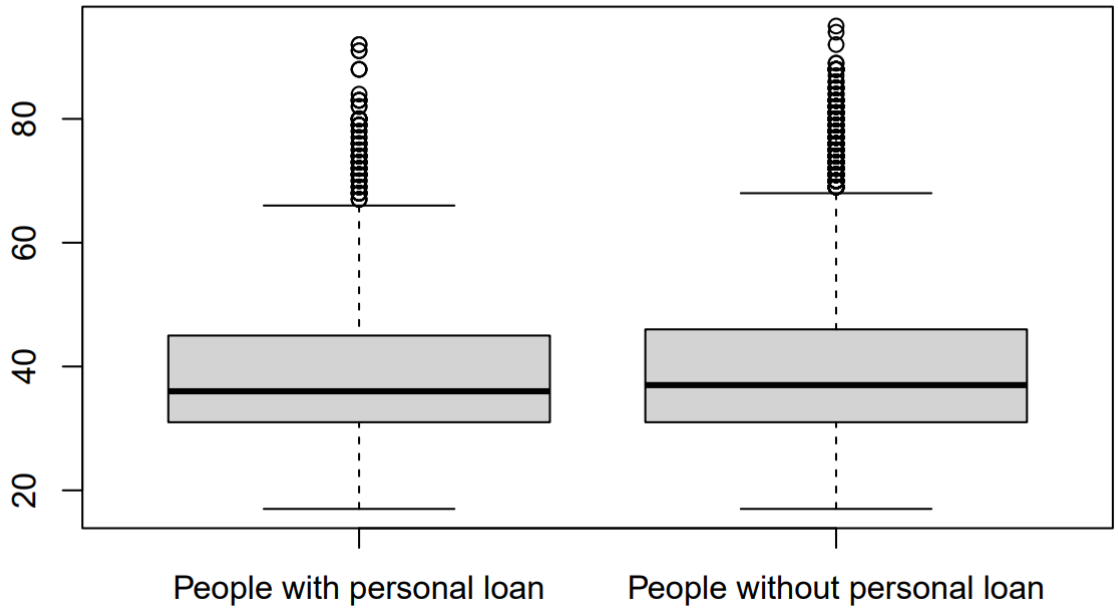
Slika 12.: Boxplot varijable godina ljudi koji imaju (lijevo) i koji nemaju (desno) stambeni kredit

Obzirom da je postotak onih s kreditom malo veći od 50% kod sve tri starosne skupine, boxplotovi sa Slika 12. su međusobno veoma slični, kao što su slični i s boxplotom prikazanim na Slika 7.

```
## Prosječni postotak mladih ljudi koji imaju privatni kredit  15.72238

## Prosječni postotak srednje starih ljudi koji imaju privatni kredit  15.72081

## Prosječni postotak starijih ljudi koji imaju privatni kredit  15.077
```

Slika 13.: Isječak rezultata odnosa godina i privatnog kredita

Kao i kod stambenog kredita, i kod privatnog nema velike razlike u postotku ljudi iz skupine s navedenim kreditom. Najmanji postotak je među starijim ljudima, a između postotka kod mladih i srednje mladih ljudi je veoma mala, što je prikazano na Slika 13.



Slika 14.: Boxplot varijable godina ljudi koji imaju (lijevo) i koji nemaju (desno) privatni kredit

Na Slika 14. možemo vidjeti da je prosjek godina ljudi bez privatnog kredita malo veća od prosjeka godina ljudi s privatnim kreditom tj. da je veći broj mladih ljudi koji koriste uslugu privatnog kredita.

3.3.2 Bračni status i krediti ispitanika

Ovisno o bračnom statusu ispitanike smo podijelili na dvije skupine:

I. Oženjeni

II. Neoženjeni (slobodni ili rastavljeni)

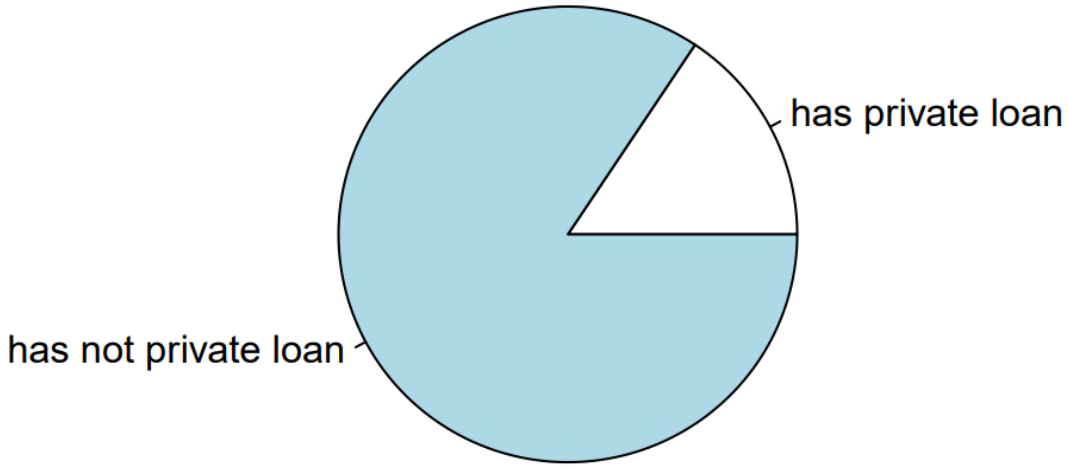
```
## Prosjek oženjenih ljudi koji imaju privatni kredit  15.66065

## Prosjek neoženjenih ljudi koji imaju privatni kredit  15.41734
```

Slika 15: Isječak rezultata odnosa bračnog statusa i privatnog kredita

Slika 15. prikazuje nam da je veći postotak ljudi koji imaju privatni kredit u skupini ljudi koji su u braku, s 15,66065%. Slika 17. je kružni graf koji prikazuje omjer oženjenih ljudi koji imaju i koji nemaju privatni kredit, a isječak programskog koda kojim smo to ostvarili je Slika 16.

```
x <- c(15.66, 84.44)
labels <- c("has private loan", "has not private loan")

pie(x, labels, main = "Pie chart of married people")
```

Slika 16.: Isječak programskog koda izrade kružnog grafa za posjedovanje privatnog kredita oženjenih ljudi



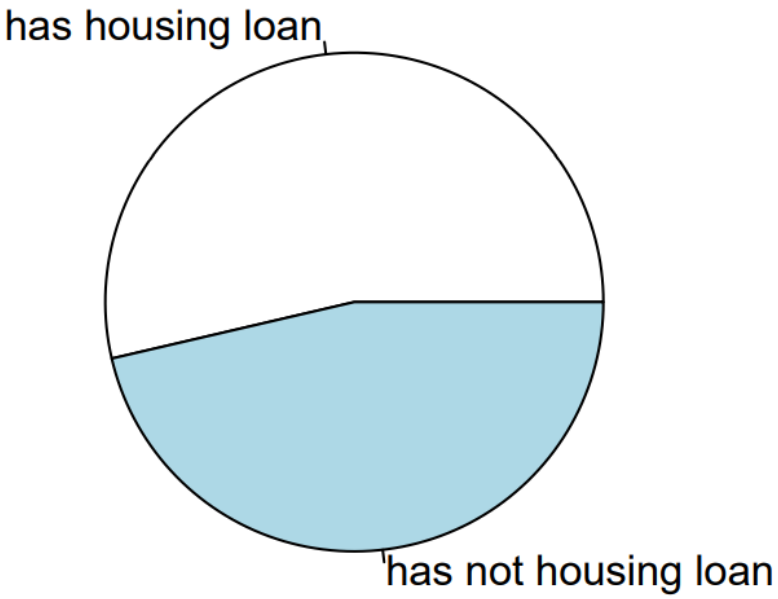
Slika 17.: Kružni graf posjedovanja privatnog kredita oženjenih ljudi

Isto je napravljeno i za stambeni kredit.

```
## Prosjek oženjenih ljudi koji imaju stambeni kredit  53.63715

## Prosjek neoženjenih ljudi koji imaju stambeni kredit  54.44419
```

Slika 18.: Isječak rezultata odnosa bračnog statusa i stambenog kredita

Po dobivenim rezultatima sa Slika 18., veći postotak ljudi sa stambenim kreditom je iz skupine neoženjenih ljudi. Slika 19. kružni je graf koji prikazuje omjer oženjenih ljudi koji imaju i koji nemaju stambeni kredit.



Slika 19.: Kružni graf posjedovanja stambenog kredita oženjenih ljudi

3.3.3 Obrazovanje i krediti ispitanika

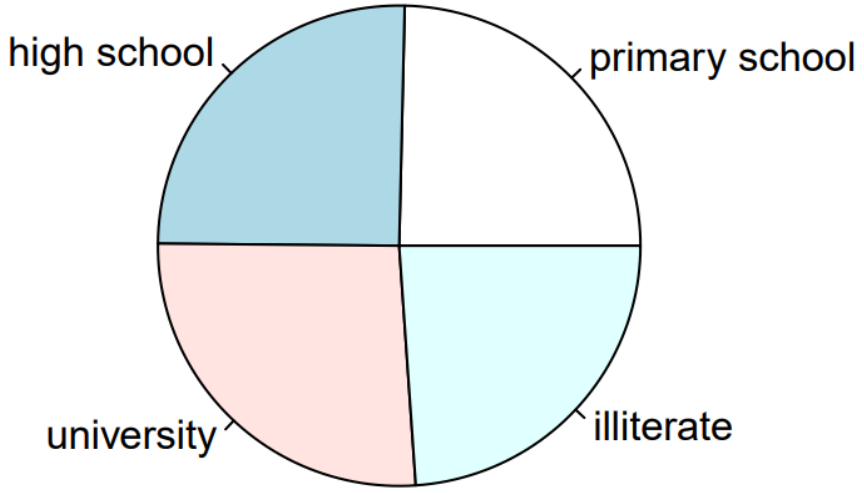
Ovisno o stupnju obrazovanja, ispitanike smo podijelili na:

I. Završena osnovna škola (obuhvaćeni su korisnici koji su završili 4, 6 ili 9 razreda osnovne škole)
II. Završena srednja škola
III. Završen fakultet
IV. Nepismeni

Zbog manje i jasnije podjele, zanemarili smo ispitanike sa završenim tečajem.

```
## Prosjek ljudi sa završenom samo osnovnom školom koji imaju privatni kredit  15.15073

## Prosjek ljudi sa završenom srednjom školom koji imaju privatni kredit  15.51371

## Prosjek ljudi sa završenim fakulteteom koji imaju privatni kredit  16.13934

## Prosjek nepismenih ljudi koji imaju privatni kredit  14.7007
```

Slika 20.: Isječak rezultata odnosa stupnja obrazovanja i privatnog kredita



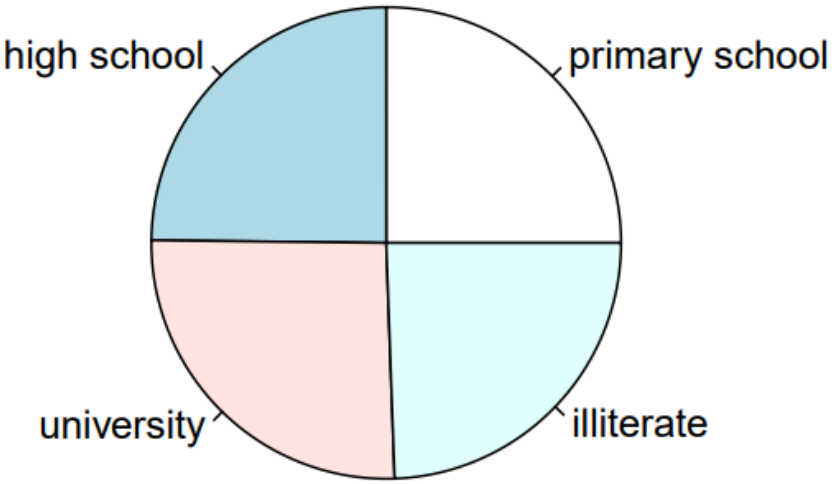
Slika 21.: Kružni graf podjele stupnja obrazovanja ljudi s privatnim kreditom

Iz Slika 20. možemo iščitati da što je veći stupanj obrazovanja to je veći broj ljudi koji iz te skupine imaju privatni kredit. Najveći postotak ljudi iz skupine koji imaju privatni kredit su oni sa završenim fakultetom, tj. oni najobrazovaniji, a najmanji postotak ljudi iz skupine s privatnim kreditom su nepismeni ljudi. Također, u obzir treba uzeti da skupina nepismenih broji značajno manje ljudi od ostalih skupina kod podjele po stupnju obrazovanja.

Identična stvar napravljena je za stambeni kredit te je u nastavku prikazan odnos stupnja obrazovanja sa imanjem stambenog kredita.

```
## Prosjek ljudi sa završenom samo osnovnom školom koji imaju stambeni kredit  53.24659

## Prosjek ljudi sa završenom srednjom školom koji imaju stambeni kredit  52.84486

## Prosjek ljudi sa završenim fakulteteom koji imaju stambeni kredit  54.83581

## Prosjek nepismenih ljudi koji imaju stambeni kredit  52.02465
```

Slika 22.: Isječak rezultata odnosa stupnja obrazovanja i stambenog kredita



Slika 23.: Kružni graf podjele stupnja obrazovanja ljudi sa stambenim kreditom

Kao i kod privatnog kredita, najveći prosjek ljudi sa stambenim kreditom su iz skupine sa završenim fakultetom. Dobivene rezultate možemo potvrditi situacijom iz stvarnog života, po kojoj je sigurnost i visina plaće ključna kod dobivanja kredita, pogotovo stambenog kredita, a najbolje uvjete za kredit većinom imaju upravo oni koji su fakultetski obrazovani.

Na kružnom grafu sa Slika 23. vidljiva je mala razlika, ali ne veća od one prikazane na kružnom grafu sa Slika 21., za privatni kredit.

3.4 Vizualizacija odnosa s varijablom *subscribed*

Prethodne usporedbe varijabli *age, education* i *marital* s posjedovanjem privatnog i stambenog kredita, zbog dosta sličnih rezultata nisu nam značajno podijelili skupine. Varijablu *job* smo izostavili zbog toga što je u sklopu nje podjela vrsta poslova na 11 skupina i rezultati ne bi mogli biti iskoristivi.

Nakon toga, promatrali smo odnos ranije navedenih varijabli *age, education* i *marital* s ključnom varijablom *subscribed*. Varijabla *subscribed* u ulozi je output varijable, koja nam govori je li ispitanik odlučio oročiti depozit u banci.

Na početku morali smo podijeliti ispitanike na one koji su odlučili ostaviti depozit s vrijednošću „1“ te one koji nisu s vrijednošću „0“.

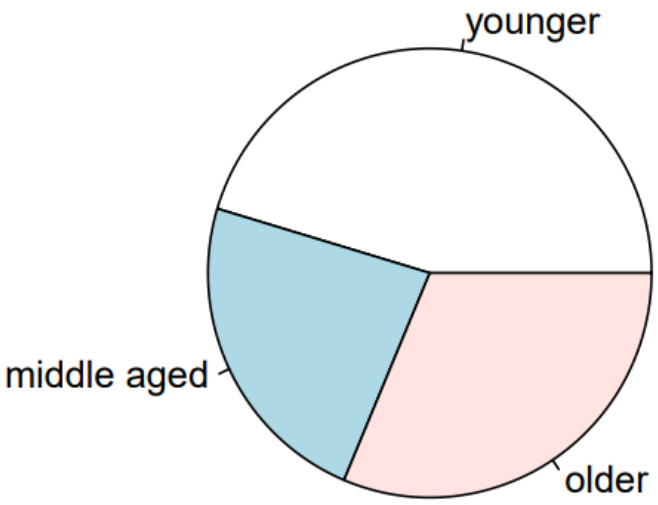
3.4.1 Godine ispitanika i varijabla *subscribed*

```
## Prosjek pretplaćenih mladih ljudi  23.22946
## Prosjek pretplaćenih srednje starih ljudi  11.93425
## Prosjek pretplaćenih starijih ljudi  16.03468
```

Slika 24.: Isječak rezultata odnosa godina i vrijednosti varijable *subscribed*

Iz rezultata prikazanih na Slika 24. jasan je utjecaj dobi na to odlučuju li se ljudi oročiti depozit u banci. To da se većinom mlađi ljudi odlučuju za to, može se obrazložiti time da oni do sada nisu koristili tu uslugu banke, dok je za starije ljude vjerojatnije da navedenu uslugu već koriste.

Dobiveni rezultati vrlo su lijepo prikazano na kružnom dijagramu na Slika 25. te se značajno veći prosjek s vrijednošću „1“ varijable *subscribed*, vidi kod skupine mlađih ljudi.



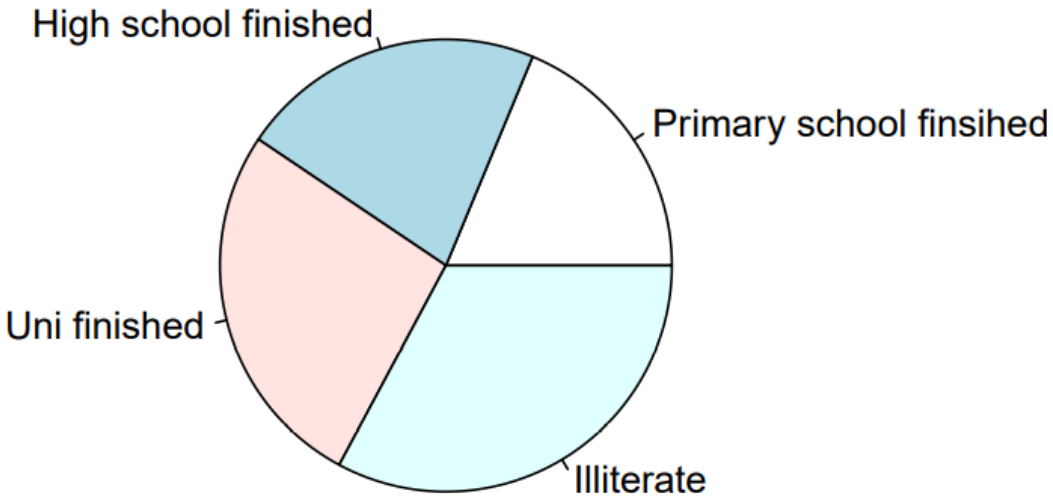
Slika 25.: Kružni graf podjele ljudi po starosti koji su se odlučili pretplatiti

3.4.2 Stupanj obrazovanja ispitanika i varijabla *subscribed*

```
## Prosječno pretplaćenih ljudi sa završenom osnovnom školom  10.95079

## Prosječno pretplaćenih ljudi sa završenom srednjom školom  12.77118

## Prosječno pretplaćenih ljudi sa završenim fakultetom  15.55045

## Prosječno pretplaćenih nepismenih ljudi  19.19014
```

Slika 26.: Isječak rezultata odnosa stupnja obrazovanja i vrijednosti varijable *subscribed*

Iz rezultata prikazanih na Slika 26. najveći prosjek onih koji su se odlučili za uslugu oročenja depozita su iz skupine nepismenih ljudi. Za razumijevanje ovih rezultata, treba znati da je broj ispitanika koji spadaju u skupinu nepismenih ljudi značajno manji od ostalih skupina. No, ono što nam pokazuje ostatak rezultata je to da porastom stupnja obrazovanja, raste i prosjek onih koji su se iz pojedine skupine odlučili za ranije spomenutu usluge banke. Isto je prikazano i na kružnom grafu na Slika 27.

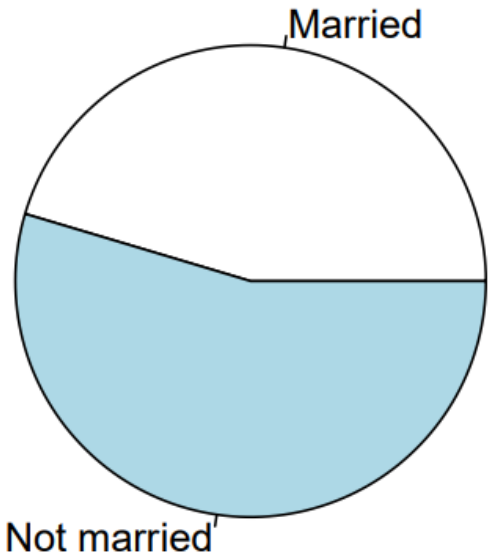


Slika 27.: Kružni graf podjele ljudi po stupnju obrazovanja koji su se odlučili pretplatiti

3.4.3 Bračni status ispitanika i varijabla *subscribed*

```
## Prosjek preplaćenih ljudi koji su oženjeni  12.43336

## Prosjek preplaćenih ljudi koji su neoženjeni ili rastavljeni  14.94546
```

Slika 28.: Isječak rezultata odnosa bračnog statusa i vrijednosti varijable *subscribed*



Slika 29.: Kružni graf podjele ljudi po bračnom statusu koji su se odlučili pretplatiti

Iz Slika 28. i Slika 29. vidljivo je da je veći prosjek onih koji su se odlučili za oročenje depozita unutar skupine neoženjenih, točnije slobodnih i rastavljenih.

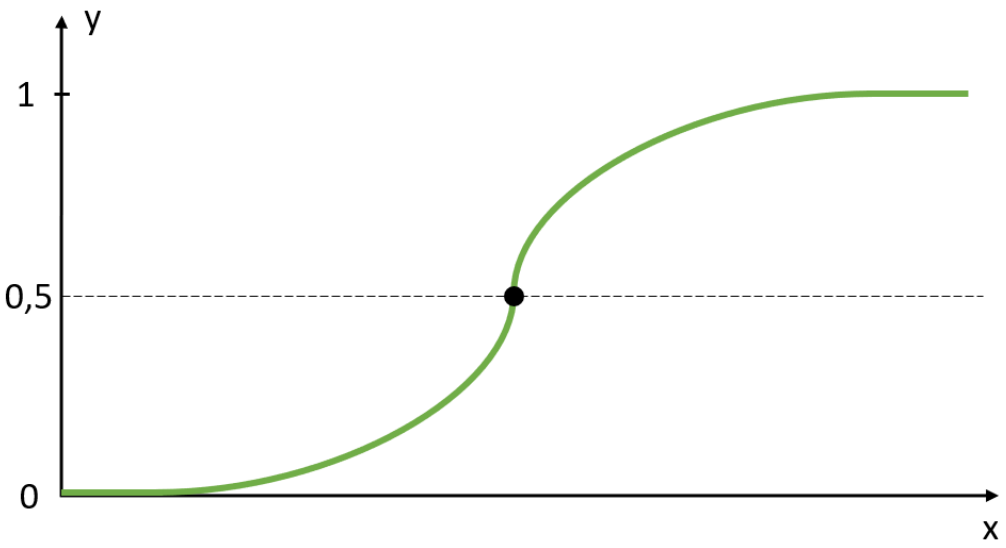
4. Logistička regresija

Nakon ostvarene vizualizacije, zaključke je potrebno potkrijepiti i učvrstiti rezultatima logističke regresije.

Logistička regresija jedan je od važnijih algoritama strojnog učenja. Tip učenja koji se koristi u sklopu logističke regresije je nadzirano učenje što znači da kao izlaz možemo dobiti vrijednosti *da* ili *ne*, a sami cilj je ostvariti predikciju kategoričkih ovisnih varijabli pomoću skupa neovisnih varijabli.

Navedena regresija koristi se za potrebu klasifikacije te je važna jer je upravo pomoću nje moguće odrediti najučinkovitiju varijablu za potrebu klasifikacije.

Funkcija koju logistička regresija čini je u obliku slova S, na osi y nalazi se kategorička varijabla u intervalu [0, 1]. Upravo ona nam govori koja je vjerojatnost nekog događaja. Primjer jedne logističke funkcije možete vidjeti na Slika 5.



Slika 30.: Primjer logističke funkcije → S krivulja

U sklopu projekta proveli smo četiri različite logističke regresije. Za svaku, podatke smo podijelili na dva dijela, na trenirane i testne. Podijelili smo ih tako da 70% podataka su za treniranje modela, a 30% za testiranje modela.

Slika 31.: Prikaz omjera podjele podataka

Zatim vrlo važno bilo je provesti provjeru je li otprilike jednak broj onih kod kojih je vrijednost „1“ varijable *subscribed* kod podataka za treniranje modela i kod podataka za testiranje modela, kako ne bi došlo do *overfittinga*. *Overfittig* je kada izrada analize preblizu ili potpuno odgovara određenom skupu podataka, zbog čega nije moguće uklopiti dodatne podatke i pouzdano napraviti predviđanje. Dio programskog koda koji nam služi za to nalazi se na Slika 32.

```
print('Trenirani model: ')
print(y_train["subscribed"].value_counts()/len(y_train))
print('Testni model: ')
print(y_test["subscribed"].value_counts()/len(y_test))
```

Slika 32.: Isječak programskog koda koji vrši provjeru ravnomjerne raspodjele podataka

Također za sve logističke regresije bilo je potrebno postaviti da vrijednosti varijabli budu „true“ ili „false“, to smo napravili pomoću funkcije *LabelEncoder()*.

```
classifier = LogisticRegression()
classifier.fit(X_train, y_train)
```

Slika 33.: Isječak programskog koda pozivanja logističke regresije i stvaranje modela naredbom *fit()*

```
y_pred = classifier.predict(X_test)
```

Slika 34.: Linija programskog koda pomoću koje predviđamo rezultate

Dio programskog koda prikazan i opisan pod Slika 33. i Slika 34. isto smo proveli kod logističkih regresija u sklopu projekta.

Rezultate smo zbog preglednosti prikazali u matrici pomoću funkcije *confusion_matrix()*, čiji je primjer vidljiv kasnije kod prikaza logističkih regresija pojedinačno.

Slika 35. sadrži kod kojim stvaramo brojčani pokazatelj toga koliko nam je dobar model.

```
print("Classification report table:")
print(classification_report(y_test, y_pred))
```

Slika 35.: Isječak programskog koda stvaranja izvješća klasifikacije

Važni dijelovi tablice izvješća klasifikacije su:

- *Accuracy*, tj. točnost omjer je točno predviđenog opažanja i ukupnog opažanja
- *Precision*, tj. preciznost omjer je točno predviđenih pozitivnih opažanja i ukupno predviđenih pozitivnih opažanja
- *Recall*, tj. prisjećanje omjer je točno predviđenih pozitivnih opažanja i svih opažanja u razredu „yes“; model je dobar ako je iznad 0.5

- *F1-score* ponderirana je srednja vrijednost preciznosti i prisjećanja

4.1 Logistička regresija I.

U prvoj logističkoj regresiji regresore X činile su sve varijable osim varijable *subscribed*, ona je varijabla y.

Nakon ranije napisanih nužnih dijelova logističke regresije, dobili smo rezultate koji su prikazani na Slika 36.

```
[[7732  144]
 [1021  215]]
Score for logistic regression model is:  0.8721466198419666
Classification report table:
              precision    recall  f1-score   support

           0       0.88      0.98      0.93      7876
           1       0.60      0.17      0.27      1236

    accuracy                           0.87      9112
   macro avg       0.74      0.58      0.60      9112
weighted avg       0.84      0.87      0.84      9112
```

Slika 36.: Rezultati prve logističke regresije

Rezultat modela logističke regresije iznosi više od 0.87. Dobiveni rezultat je odličan i model je potpuno sposoban predvidjeti hoće li se ispitanik odlučiti za oročenje depozita u banci što je i bio cilj samoga projekta.

Iz tablice izvješća klasifikacije uočljivo je da su i ostali rezultati dobri.

4.2 Logistička regresija II.

U sklopu druge logističke regresije regresori X bile su varijable *age*, *marital* i *education*, dok je kao izlazna varijabla, tj. y bila varijabla *housing*, koja označava ima li korisnik stambeni kredit.

Na Slika 37. možemo vidjeti rezultate dobivene drugom logističkom regresijom. Dobiveni rezultat modela iznosi 0.537 što nije katastrofalni rezultat, ali svakako nije zadovoljavajući model za potrebe banke.

```
Score for logistic regression model is:  0.5366549604916594
Classification report table:
              precision    recall  f1-score   support

    accuracy                           0.54      9112
   macro avg       0.27      0.50      0.35      9112
weighted avg       0.29      0.54      0.37      9112
```

Slika 37.: Rezultati druge logističke regresije

4.3 Logistička regresija III.

Ono po čemu se treća logistička regresija razlikuje je to što se kao izlazna varijabla y, umjesto stambenog kredita, gleda to ima li osoba privatni kredit. Regresori X i ovdje su varijable *age, marital* i *education*.

```
Score for logistic regression model is:  0.8454784899034 24
              precision    recall  f1-score   support

    accuracy                           0.85      9112
   macro avg       0.42      0.50      0.46      9112
weighted avg       0.71      0.85      0.77      9112
```

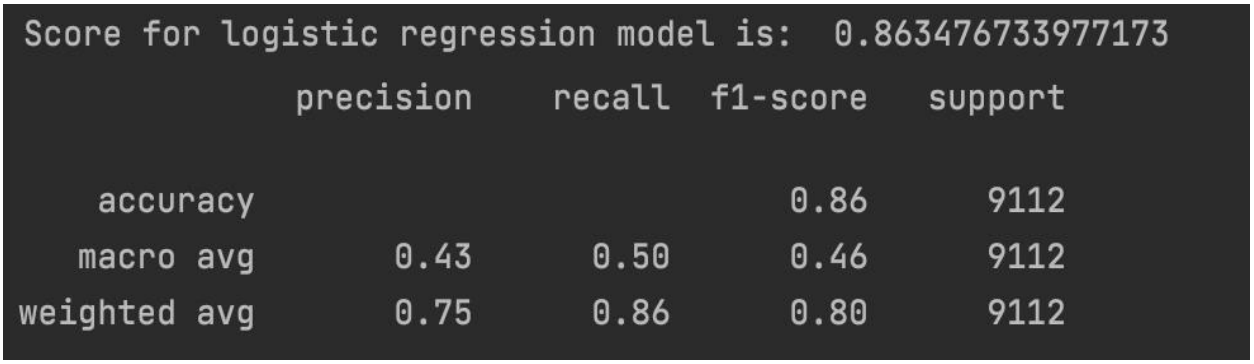
Slika 38.: Rezultati treće logističke regresije

Dobiveni rezultat možemo vidjeti na Slika 38. te je on za treću logističku regresiju jako dobar i iznosi 0.85 točnosti po čemu možemo zaključiti da je model sposoban dobro predvidjeti tko će se od ispitanika odlučiti oročiti depozit u banci.

4.4 Logistička regresija IV.

U četvrtoj logističkoj regresiji, y je opet varijabla *subscribed,* a regresori X su, kao i u drugoj i trećoj logističkoj regresiji, varijable *age, marital* i *education*.

Rezultat modela logističke regresije koji iznosi 86% točnosti, prikazan je na Slika 39. I ovaj model zbog visoke točnosti, potpuno bi bio sposoban predvidjeti koja će osoba staviti oročenje depozita, a koja ne.



Slika 39.: Rezultati četvrte logističke regresije

5. Zaključak

Ispravan rad s podacima od velike je važnosti za cjelokupno, što bolje funkcioniranje društva. Veliku količinu nepreglednih podataka, potrebno je razumjeti, a to je moguće samo onda kada su podaci smisleno prikazani i uređeni.

U sklopu projekta, upoznali smo se s do sada nama nepoznatim alatima te uz učenje uspjeli smo kao krajnji rezultat dobiti ostvarenje početnog cilja projekta.

Određene atribute našeg početnog skupa podataka bilo je potrebno maknuti, jer isti nam dolasku do našeg cilja nisu doprinosili. Određene varijable bile su ovisne o vremenu i promjenjive po mjesecima ili kvartalima, također, neke od varijabli bile su međusobno korelirane i to je ono čime smo se bavili u dijelu eksploratorne analize.

Nakon toga, veliku važnost razumijevanja podataka ostvarili smo pomoću vizualizacije, iz koje nismo mogli samo zaključiti krajnji rezultat, ali koja nam je pomogla u tome u kojem smjeru trebamo nastaviti rad na projektu. Sva opažanja dobivena iz vizualizacije, potkrijepili smo logističkom regresijom.
Iz napravljene četiri logističke regresije, najbolji rezultat dala je ona regresija kojoj je izlazna varijabla y varijabla subscribed, koja je ujedno i najznačajnija varijabla, dok su sve ostale varijable regresori X. Nakon toga, proveli smo još tri regresije gdje smo za kategoričke varijable, regresori X, gledali ovisnost drugih varijabli poput loan i housing, ali i dalje najbolji rezultat daje prva logistička regresija. Za vraijablu housing dobili smo rezultat 54% što nije loše, ali u našem slučaju, kada se radi o banci, to nije dovoljno dobro. Iz toga smo zaključili da upravo model gdje promatramo ovisnost varijable subscribed, o svim ostalim varijablama je najbolji model.

Postotak točnosti prve logističke regresije je 87%, čime smo dobili vrlo sposoban model za predviđanje toga hoće li se ispitanik odlučiti ostaviti oročeni depozit u banku. Upravo ovaj model, bio bi od velikog značaja u poslovanju banke, a mi smo upravo s njim ostvarili cilj našega projekta.