**DOKUMENTACIJA O IZRADI PROJEKTA IZ PREDMETA SOFTVERSKI ALATI U SISTEMIMA AUTOMATSKOG UPRAVLJANJA**

**“Bank Marketing Classification”**

**Student:**

**Emilija Lazić RA 59/2022**

**SADRŽAJ**

1. Uvod
2. Analiza i priprema podataka
3. Modelovanje i Evaluacija
4. Zaključak

1. **Uvod**

1.1. Definicija problema

Projekat se bavi problemom optimizacije direktnih marketinških kampanja portugalske banke. Cilj je povećati efikasnost telefonskih poziva tako što će se unapred identifikovati klijenti sa najvećom verovatnoćom da će se pretplatiti na rok-depozit.

1.2. Cilj projekta

Glavni cilj ovog projekta je razvoj i evaluacija modela mašinskog učenja koji će klasifikovati klijente u dve grupe: one koji će se pretplatiti na depozit *(yes)* i one koji neće *(no)*.

1.3. Skup podataka (Dataset)

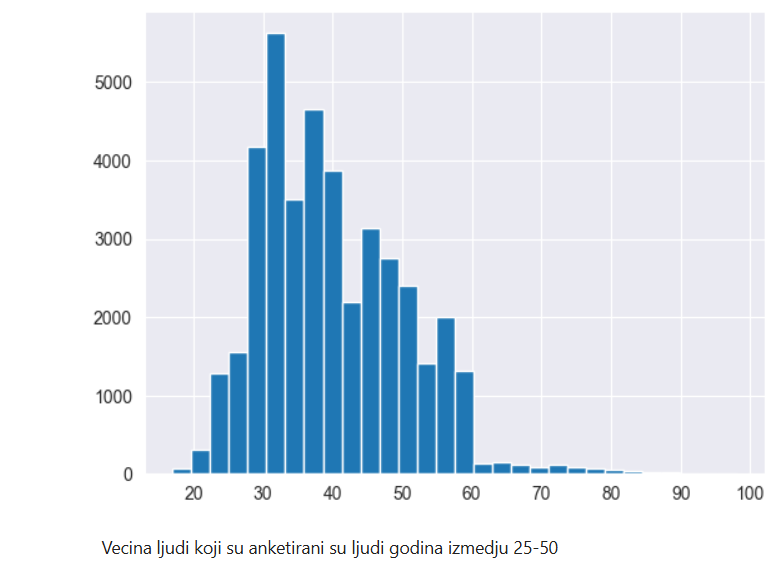
Korišćen je javno dostupan skup podataka "Bank Marketing" sa UCI Machine Learning repozitorijuma. Korišćena je kompletna verzija (bank-additional-full.csv), koja sadrži 41,188 redova i 21 kolonu, obuhvatajući demografske podatke o klijentima, informacije o prethodnim kampanjama i ekonomske indikatore.

1. **Analiza i priprema podataka**

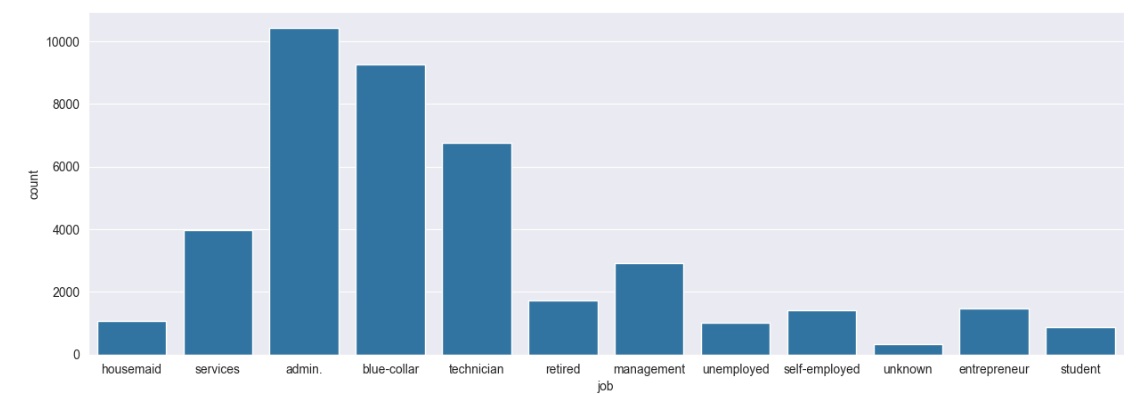
Proces pripreme podataka bio je ključan za uspeh modela i sastojao se iz nekoliko faza.

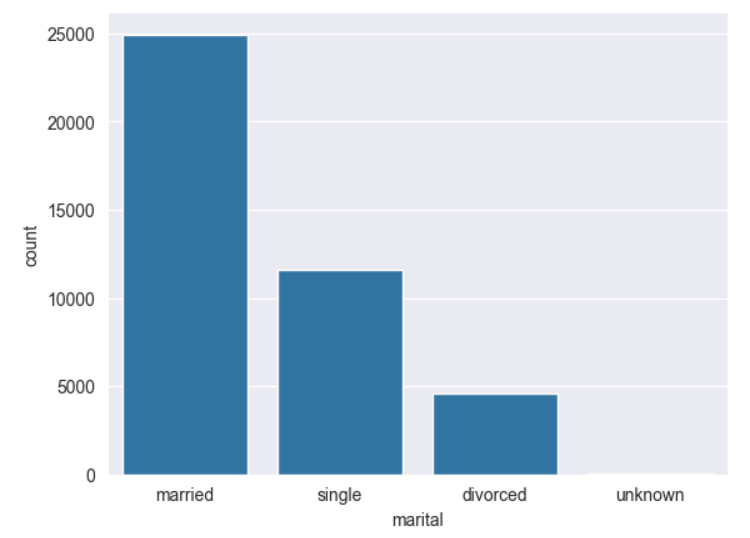
2.1. Pregled podataka

“Age”

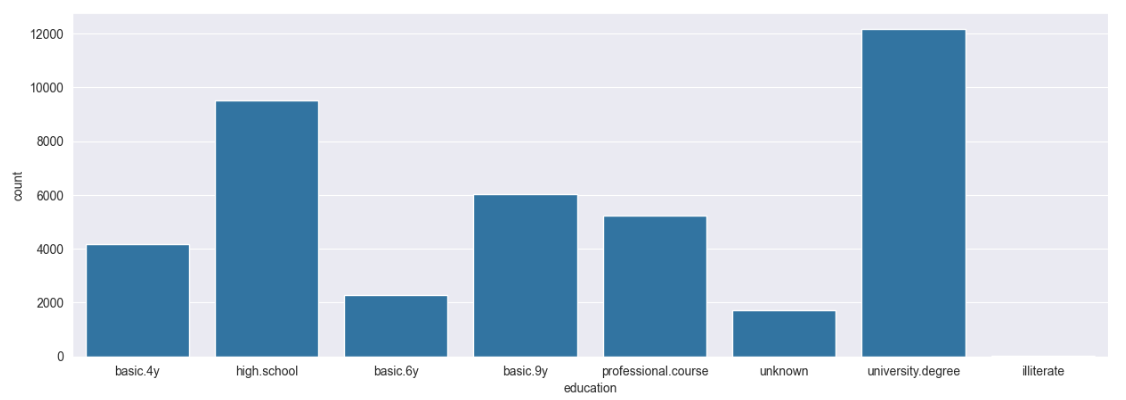


“Job”

“Marital”

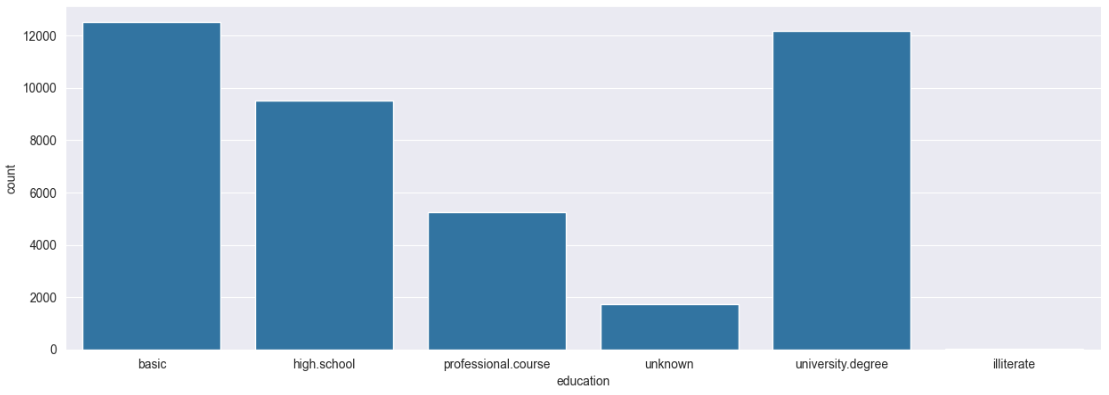


“Education”



* **Grupisanje kategorija:** Vrednosti u koloni education ('basic.4y', 'basic.6y', 'basic.9y') su grupisane u jednu zajedničku kategoriju 'basic' radi smanjenja kompleksnosti.

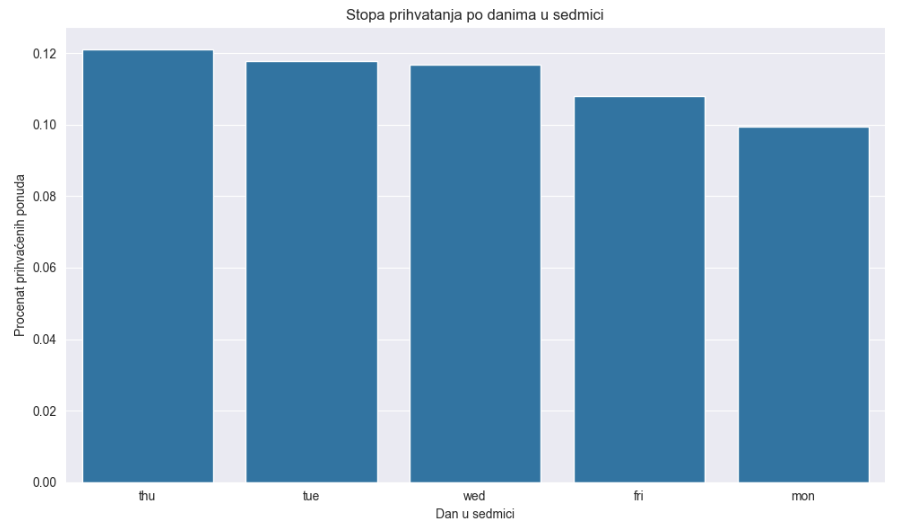
Nakon grupisanja:



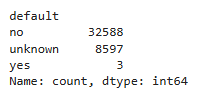
“Month”



“Day of week”



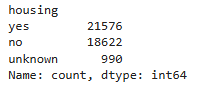
“Default” (da li poseduje kreditno zaduženje da/ne/nepoznato)



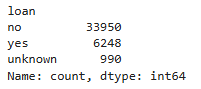
“Duration” (trajanje poziva), koristili smo za benchmark analizu.

* **Uklanjanje kolona:** Kolona default je izbačena zbog izrazite neuravnoteženosti i niskog prediktivnog značaja. Kolona duration je izbačena kako bi se izbeglo curenje podataka (data leakage), jer trajanje poziva nije poznato pre samog poziva.

“Housing”



“Loan”



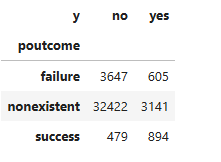
* **Rukovanje nepoznatim vrednostima:** 'Unknown' vrednosti u kolonama housing i loan su zamenjene najčešćom vrednošću u tim kolonama, kako bi se sačuvali podaci.

“Pdays”

* Kolona pdays (dani od poslednjeg kontakta), gde je vrednost 999 označavala da klijent nikada nije bio kontaktiran, transformisana je u novu, binarnu kolonu bio\_kontaktiran\_ranije. Ovo je modelu dalo jasniji signal o istoriji klijenta.

“Poutcome”

Tabela pokazuje drastične razlike u ponašanju klijenata u zavisnosti od njihovog prethodnog iskustva sa bankom.



* **Grupa nonexistent (Novi klijenti):**

Ovo je najveća grupa klijenata. Od njih 35,563, samo 3,141 je pristalo na ponudu, što daje stopu uspeha od **8.8%**. Ova grupa formira osnovnu (baseline) verovatnoću uspeha za "hladne pozive", odnosno za klijente bez prethodne interakcije.

* **Grupa failure (Prethodni neuspeh):**

Od 4,252 klijenta koji su u prošlosti odbili ponudu, njih 605 je ovog puta pristalo. Stopa uspeha za ovu grupu je **14.2%**. Zanimljivo je da čak i prethodni neuspeh ukazuje na veću šansu za uspeh u poređenju sa potpuno novim klijentima.

* **Grupa success (Prethodni uspeh):**

Ova grupa, iako najmanja, nosi najjači signal. Od 1,373 klijenta koji su u prošlosti već prihvatili ponudu, čak 894 je to uradilo ponovo. Stopa uspeha za ovu grupu je **65.1%**.

2.2. Enkodiranje kategoričkih varijabli

Inicijalni pristup sa LabelEncoder-om je odbačen jer unosi lažni redosled u podatke, što negativno utiče na linearne modele.

Finalni pristup je korišćenje **One-Hot Encoding** (pd.get\_dummies) tehnike za sve nominalne kategoričke varijable (job, marital, poutcome, itd.). Ovo je stvorilo robustan set podataka pogodan za sve tipove algoritama.

2.3. Skaliranje numeričkih podataka

Sve numeričke kolone (npr. age, campaign i ekonomski indikatori) su skalirane pomoću StandardScaler-a. Ovim je svakoj koloni data srednja vrednost 0 i standardna devijacija 1, što je rešilo problem ConvergenceWarning-a i omogućilo fer tretman svih atributa od strane modela.

2.4. Rešavanje neuravnoteženosti klasa

Analizom je utvrđena značajna neuravnoteženost ciljne promenljive y. Za rešavanje ovog problema, primenjena je tehnika - parametar **class\_weight='balanced'** unutar modela.

1. **Modelovanje i evaluacija**

3.1. Od Benchmarka do praktičnog modela

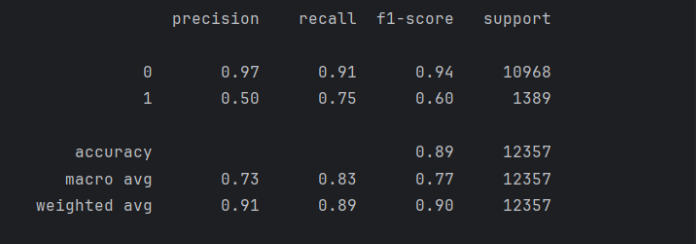
Proces razvoja modela bio je iterativan, sa ključnim odlukama donetim na osnovu analize performansi i ispravnosti.

#### **Inicijalni pristup i perealno visoke performanse**

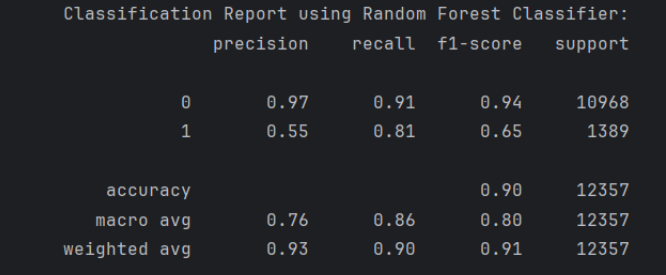
U prvoj fazi projekta, model je treniran na podacima koji su uključivali sve dostupne atribute, uključujući i *duration* (trajanje poslednjeg poziva). Za enkodiranje kategoričkih varijabli je prvobitno korišćen *LabelEncoder*.

Ovaj inicijalni model je pokazao izuzetno visoke, ali nerealne metrike performansi.

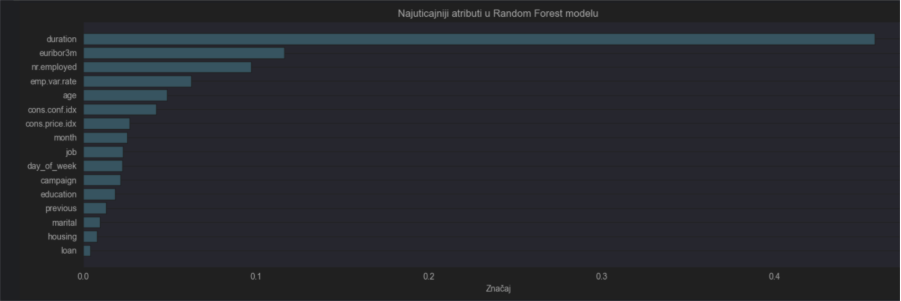
Logistic Regression:



Random Forest:



#### **Identifikacija curenja podataka (Data Leakage) i benchmark analiza**



Detaljnom analizom je utvrđeno da je atribut duration uzrok **curenja podataka (data leakage)**. Trajanje poziva (duration) je informacija koja nije poznata **pre** nego što se poziv obavi i klijent donese odluku. Korišćenje ove informacije za predviđanje ishoda istog tog poziva je metodološki neispravno, jer model koristi informaciju iz "budućnosti" koju u realnoj primeni (pre pozivanja klijenta) ne bi imao.

Iako je neupotrebljiv za praktičnu primenu, ovaj model sa duration kolonom služi kao **Benchmark Model**. Njegove visoke performanse definišu **teorijski maksimum** ili "zlatni standard" koji bi se mogao dostići kada bismo imali savršene, ali nedostupne informacije. Ovaj benchmark nam služi kao gornja granica za poređenje sa performansama našeg finalnog, realno primenjivog modela.

**Kreiranje Praktičnog Modela**

Da bi se kreirao model koji ima primenu u stvarnom svetu, doneta je ključna odluka da se kolona duration **potpuno ukloni** iz skupa podataka.

Ovaj korak je rezultirao očekivanim, ali realnim padom metrika. Performanse "Praktičnog Modela" su niže od Benchmarka, ali predstavljaju **istinit i realan prikaz** sposobnosti modela da predvidi ishod na osnovu informacija koje su zaista dostupne pre poziva.

Uporedo sa ovom odlukom, nova priprema podataka je takođe sadržala:

1. **LabelEncoder** je zamenjen sa **OneHotEncoder**-om kako bi se izbegla lažna interpretacija redosleda kategorija.
2. Uvedeno je **skaliranje numeričkih podataka** (StandardScaler) radi optimizacije performansi linearnih modela.
3. Zadržani su i pravilno obrađeni ključni prediktori kao što je **poutcome**.

Svi finalni rezultati i poređenja algoritama prikazani u ovom izveštaju odnose se na ovaj, metodološki ispravan i realno primenjiv, **Praktični Model**.

3.2. Strategija optimizacije modela

Nakon što je definisan metodološki ispravan pristup i pripremljen set podataka, fokus je prebačen na optimizaciju samih modela. Proces optimizacije se sastojao iz dva ključna dela: izbora metrike za optimizaciju i finog podešavanja hiperparametara.

#### **Izbor Metrike za Optimizaciju: Fokus na Odziv (Recall)**

Kod problema binarne klasifikacije sa neuravnoteženim klasama, kao što je ovaj, ukupna tačnost (accuracy) je nepouzdana metrika. Stoga je bilo neophodno doneti stratešku odluku o tome da li optimizovati model za Preciznost (Precision) ili za Odziv (Recall).

* Preciznost meri kvalitet pozitivnih predikcija (minimizira broj pogrešnih poziva, tj. lažnih alarma).
* Odziv meri sposobnost modela da pronađe sve pozitivne instance (minimizira broj propuštenih klijenata, tj. propuštenih prilika).

U kontekstu ovog projekta, doneta je odluka da se blaga prednost da Odzivu (Recall). Razlog za ovu odluku leži u pretpostavci da je trošak propuštenog klijenta (False Negative) značajno veći od troška pogrešnog poziva (False Positive). Gubitak potencijalnog klijenta koji bi ostvario oročeni depozit predstavlja direktan gubitak profita za banku. Sa druge strane, trošak pozivanja klijenta koji nije zainteresovan je relativno nizak (nekoliko minuta vremena agenta).

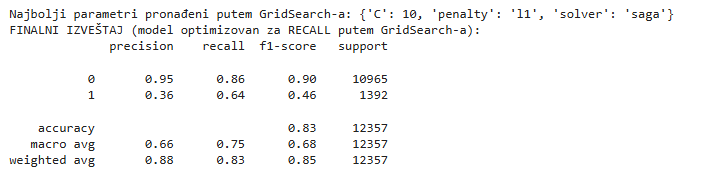
Stoga je cilj bio kreirati model koji će identifikovati što je moguće veći broj stvarnih potencijalnih klijenata. Iako je recall bio primarni cilj, tokom GridSearch optimizacije je korišćen i f1-score kao sekundarna metrika, kako bi se osigurao dobar balans i izbegla prevelika degradacija Preciznosti.

#### **Optimizacija Hiperparametara pomoću GridSearchCV**

Za fino podešavanje i pronalaženje najbolje verzije svakog algoritma, korišćena je GridSearchCV tehnika. Ova metoda sistematski testira različite kombinacije hiperparametara modela koristeći unakrsnu validaciju (cross-validation) kako bi se pronašla ona koja daje najbolje performanse prema zadatoj metrici (recall ili f1).

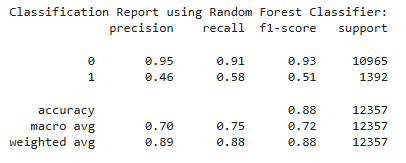
Ovim pristupom je osigurano da je iz svakog testiranog algoritma izvučen maksimum, omogućavajući fer i objektivno poređenje njihovih finalnih performansi.

Logistic Regression:



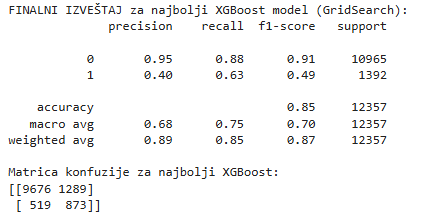


Random Forest:



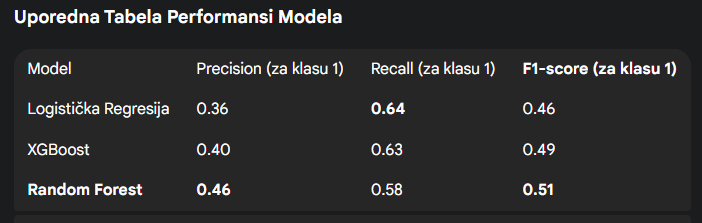


XGBoost:



**4. Zaključak**

4.1. Sumiranje rezultata i izbor najboljeg modela



Na osnovu evaluacionih metrika, **Random Forest se pokazao kao sveukupno najuspešniji model**. Sa F1-skorom od **0.51**, postigao je najbolji balans između Preciznosti (**0.46**) i Odziva (**0.58**). Iako su i XGBoost i Logistička Regresija pokazali konkurentne rezultate, Random Forest je ponudio superiornu kombinaciju performansi.