RAZVOJ I USPOREDBA KOMPETITIVNIH MODELA ZA PREDIKCIJU PREKIDA UGOVORNIH ODNOSA S PRIMJENAMA TEHNIKA PROFILIRANJA

Lovrić-Matijević, Kristina

VISOKO UČILIŠTE ALGEBRA

ZAVRŠNI RAD

Razvoj i usporedba kompetitivnih modela za predikciju prekida ugovornih odnosa s primjenama tehnika profiliranja

Kristina Lovrić-Matijević

Zagreb, veljača 2023.

„Pod punom odgovornošću pismeno potvrđujem da je ovo moj autorski rad čiji niti jedan dio nije nastao kopiranjem ili plagiranjem tuđeg sadržaja. Prilikom izrade rada koristila sam tuđe materijale navedene u popisu literature, ali nisam kopirala niti jedan njihov dio, osim citata za koje sam navela autora i izvor te ih jasno označila znakovima navodnika. U slučaju da se u bilo kojem trenutku dokaže suprotno, spremna sam snositi sve posljedice, uključivo i poništenje javne isprave stečene dijelom i na temelju ovoga rada.“

U Zagrebu 28. veljače 2023.

Predgovor

Zahvaljujem mentoru prof. dr. sc. Goranu Klepcu na prenesenom znanju, trudu i motivaciji. Završni rad posvećujem svojoj iracionalnijoj polovici i vjetru u leđa – Miro, ovo je za tebe.

A Gradišćanska 24, HR-10000 Zagreb
T (01) 2222 182
E student@algebra.hr
www.algebra.hr

ALGEBRA

Temeljem članka 8. Pravilnika o završnom radu i završnom ispitu na preddiplomskom studiju Visokog učilišta Algebra sačinjena je ova

Potvrda o dodjeli završnog rada

kojom se potvrđuje da studentica Kristina Lovrić-Matijević, JMBAG 0130070847, OIB 13789446339 u šk. godini 2021./2022., studij: Primjenjeno računarstvo - Preddiplomski studij, smjer: Programsko inženjerstvo, od strane povjerenstva za provedbu završnog ispita, dana 21.02.2022. godine, ima odobrenu izradu završnog rada

s temom: **Razvoj i usporedba kompetitivnih modela za predikciju prekida ugovornih odnosa s primjenama tehnika profiliranja**

i sažetkom rada: Prekid ugovornih odnosa (engl. churn) vrlo je važna metrika u djelatnosti telekomunikacija i u njegovo sprječavanje ulažu se veliki napori. Ovaj završni rad opisuje te napore u vidu konstrukcije izvedenih varijabli na temelju inputa domenskih stručnjaka, pomnog odabira varijabli s pomoću radi izrade prediktivnih modela i mjerenja prediktivne snage tih modela. Na primjeru javno dostupnog skupa podataka iz područja telekomunikacija preuzetog s interneta napravljena je priprema podataka, analiza relevantnosti atributa, izrađeni su prediktivni modeli utemeljeni na logističkoj regresiji, naivnom Bayesovom teoremu i neuronskoj mreži te je testirana njihova prediktivna snaga putem standardnih postupaka primjene Kolmogorov-Smirnovljeva testa i ROC krivulje. Na kraju je dan prijedlog sveobuhvatne strategije sprječavanja prekida ugovornih odnosa na temelju proučenog primjera.

Mentor je: Goran Klepac.

Odobrenjem završnog rada studentici je omogućen upis kolegija "Izrada završnog projekta/Praksa" te je sukladno članku 8. Pravilnika o završnom radu i završnom ispitu dužan najkasnije do početka nastave ljetnog semestra u sljedećoj školskoj godini, uspješno obraniti završni rad uspješnim polaganjem završnog ispita.

U protivnom studentica može zatražiti novog mentora/icu i temu te ponovo upisati kolegij "Izrada završnog projekta/Praksa" budući da rad koji nije predan i obranjen na završnom ispitu u roku određenom Pravilnikom završnom radu i završnom ispitu prestaje vrijediti. Izrada novog završnog rada se izvodi sukladno rokovima određenima za školsku godinu u kojoj je studentici određen novi mentor/ica i dodijeljen novi završni rad.

Potpis studentice:	Potpis mentora:	Potpis predsjednika povjerenstva:

Ova potvrda izdaje se u 4 (četiri) primjerka od kojih 3 (tri) idu kao prilog završnom radu.

Sažetak

Prekid ugovornih odnosa (engl. *churn*) vrlo je važna metrika u djelatnosti telekomunikacija i u njegovo sprječavanje ulažu se veliki napori. Ovaj završni rad opisuje te napore u vidu konstrukcije izvedenih varijabli na temelju inputa domenskih stručnjaka, pomnog odabira varijabli radi izrade prediktivnih modela i mjerenja prediktivne snage tih modela. Na primjeru javno dostupnog skupa podataka iz područja telekomunikacija preuzetog s interneta napravljena je priprema podataka, analiza relevantnosti atributa, izrađeni su prediktivni modeli utemeljeni na logističkoj regresiji, naivnom Bayesovom teoremu i neuronskoj mreži te je testirana njihova prediktivna snaga putem standardnih postupaka primjene Kolmogorov-Smirnovljeva testa i ROC krivulje. Na kraju je dan prijedlog sveobuhvatne strategije sprječavanja prekida ugovornih odnosa na temelju proučenog primjera.

Ključne riječi: prekid ugovornog odnosa, telekomunikacije, analiza relevantnosti atributa, *churn*, prospektivna vrijednost korisnika, izvedene varijable, logistička regresija, Bayesov teorem, neuronske mreže, K-S test, ROC krivulja, ROC AUC.

Summary

Churn is a very important metric in the telecommunications industry and great efforts are being made to prevent it. This final paper describes these efforts in terms of the construction of derived variables based on the inputs of domain experts, careful selection of variables as a prerequisite for creating predictive models and measuring the predictive power of these models. On the example of a publicly available dataset from the telecommunication industry downloaded from the Internet, data was prepared and attribute relevance analysis was calculated, followed by the construction of predictive models based on logistic regression, naïve Bayesian theorem and neural network. Predictive power of the models was tested through standard procedures such as Kolmogorov-Smirnov test and ROC curve. At the end, a proposal for a comprehensive strategy for preventing the termination of contractual relations was given on the basis of the example studied.

Keywords: *churn, telecommunications, attribute relevance analysis, prospective customer value, derived variables, logistic regression, Naive Bayes, neural networks, K-S test, ROC curve, ROC AUC.*

Sadržaj

1. Uvod

Pružanje telekomunikacijskih usluga nesumnjivo je jedna od djelatnosti s najoštrijom konkurencijom. Operateri se bore da pridobiju i zadrže korisnike, tim više što pridobiti novog korisnika košta više nego zadržati postojećeg. Upravo je zato prekid ugovornih odnosa jedna od najvažnijih metrika koje telekomi prate u području odnosa s korisnicima.

Telekomi uz pomoć podatkovne znanosti razvijaju modele kojima nastoje predvidjeti koji bi korisnici mogli prijeći drugom operateru – odnosno prekinuti ugovorni odnos – prije nego što se to doista dogodi kako bi ih pokušali zadržati. Za tu se svrhu koriste brojni alati iz instrumentarija podatkovne znanosti, poput raznih klasifikatora, neuronskih mreža itd. Razvijeni prediktivni modeli postaju dio sveobuhvatnih rješenja ili sustava za sprječavanje prekida ugovornih odnosa.

U ovom će se radu prikazati postupak izrade jednostavnih analitičkih modela za predviđanje prekida ugovornih odnosa na primjeru javno dostupnog skupa podataka iz telekomunikacijskog sektora. Tom će postupku, baš kao i u stvarnom životu, prethoditi priprema i čišćenje podataka, podjela uzorka na dio za razvoj i dio za testiranje te zatim primjena tehnike analize relevantnosti atributa namijenjene odabiru varijabli. Bit će pokazano da se, osim za odabir varijabli, analiza relevantnosti atributa može koristiti i za profiliranje korisnika.

Nakon tih pripremnih faza uslijedit će konstrukcija triju prediktivnih modela: za to će biti upotrijebljena logistička regresija, naivni Bayesov algoritam i neuronska mreža. Sve će to biti izvedeno s pomoću koda u Pythonu i podatkovnoznanstvenih alata koje nude biblioteke pandas, numpy, scikitlearn, matplotlib, seaborn i druge.

Na posljetku će se razvijeni modeli usporediti s pomoću dviju često korištenih metoda za dokazivanje prediktivne snage – K-S testa i ROC krivulje, te će se odabrati najbolji za dani primjer uz prijedlog strategije zadržavanja korisnika.

2. Općenito o prekidu ugovornih odnosa

Najjednostavnije rečeno, prekid ugovornih odnosa (engl. *churn* ili *customer attrition / defection / turnover*) pojam je iz svijeta poslovanja kojim se opisuje gubitak klijenata ili korisnikâ (Klepac et al., 2015). Moglo bi se reći da je *churn* suprotan konceptu zadržavanja korisnika (engl. *customer retention*) – oni su dvije suprotstavljene sile. Prekid ugovornih odnosa važna je metrika u djelatnostima bankarstva, osiguranja, telekomunikacija, maloprodaje, izdavaštva periodike (novine, časopisi)… i općenito svugdje gdje se između korisnika i davatelja proizvoda/usluge sklapa pretplatnički ili ugovorni odnos ili pak uspostavlja vjernost u upotrebi proizvoda/usluge. Prekid ugovornih odnosa ima takvu poslovnu važnost jer gubitak korisnika dovodi do smanjenja prodaje i povećanja potrebe za privlačenjem novih korisnika/kupaca, što je čak pet do šest puta skuplje od troškova potrebnih za zadržavanje postojećih korisnika (Lazarov et al., 2010).

Točna metrika koja se prati zapravo je količina prekinutih ugovornih odnosa (engl. *churn rate*, CR) koja se definira ovako (Klepac et al., 2015) i obično se izražava postotkom:

$$CR = \frac{\text{broj korisnika koji su prekinuli ugovorni odnos tijekom promatranog razdoblja}}{\text{ukupan broj korisnika prvog dana promatranog razdoblja}} \qquad (1)$$

Praćenje prekida ugovornih odnosa uvijek podrazumijeva stanovitu dozu nepreciznosti jer poslovni subjekt ne može biti siguran da njegov korisnik razmišlja o prelasku konkurenciji sve dok ne bude prekasno (Klepac et al., 2015). U tom je svjetlu najbolje što tvrtka može učiniti da na temelju korisnikova ponašanja i drugih podataka čim bolje – sa što većom vjerojatnošću – predvidi tko će otići. Za to se koriste prediktivni modeli utemeljeni na rudarenju podataka (engl. *data mining*).

Nakon predviđanja treba uslijediti ublažavanje (engl. *mitigation*) i donošenje mjera za smanjenje *churna*, koji zajedno čine rješenje za upravljanje prekidima ugovornih odnosa.

No, dobar prediktivni model nije jamac dobre strategije smanjenja prekida ugovornih odnosa, pa se čak govori o stavu da je za razvoj rješenja za *churn* dovoljan analitičar za rudarenje podataka kao o mitu (Klepac et al., 2015). Naime, rješenje za *churn* bi osim odgovora na pitanje **tko** će prekinuti ugovorni odnos, na koje odgovara prediktivni model,

trebalo ponuditi i odgovor na pitanje **zašto**, a iskustvo iz stvarnog poslovanja pokazalo je da je za to prijeko potreban input domenskih stručnjaka u vidu:

- konstrukcije izvedenih (bihevioralnih, virtualnih, engl. *derived*) varijabli
- profiliranja korisnika s pomoću analize relevantnosti atributa
- definiranja pravila za *fuzzy* ekspertne sustave kojima se računa prospektivna vrijednost korisnika kako bi se prvenstveno probali zadržati takvi korisnici.

Neke od tih tehnika bit će prikazane u ovom radu kao i nužnost kombiniranja tehnika rudarenja podataka s domenskim znanjem pri razvoju rješenja za smanjenje prekida ugovornih odnosa.

Prepoznavanje korisnikâ s najvećom vjerojatnosti prekida ugovornog odnosa, a zatim i s najvećom prospektivnom vrijednošću omogućit će da se korisnicima pristupi s ponudom za zadržavanje bez koje bi takvi korisnici otišli (Berry at al., 2014). U konačnici, glavni je cilj predviđanja *churna* zadržavanje visokovrijednih korisnika radi maksimiziranja dobiti.

2.1. Vrste prekida ugovornih odnosa

Nisu svi prekidi ugovornog odnosa isti. Primjerice, u području klasične maloprodaje *churn* izgleda nešto drugačije nego u izdavaštvu ili telekomunikacijama pa kupci ne sklapaju pretplatničke odnose s trgovačkim lancima, ali postoje kartice vjernosti (engl. *loyalty cards*) putem kojih se često nagrađuju kupci koji se vraćaju ili redovito kupuju u nekom trgovačkom lancu. Ili pak osoba može biti korisnik jednog servisa za *streaming* TV sadržaja i istovremeno se pretplatiti na konkurentski servis.

U tom svjetlu, razlikuju se sljedeće vrste prekida ugovornih odnosa (Klepac at al., 2015):

Osnovna je podjela na tzv.

1. ***Hard churn*** **–** podrazumijeva konkretan prekid odnosa, odnosno postoji precizno definiran događaj koji označava *churn* - primjerice zatvaranje računa ili raskid ugovora.
2. ***Soft churn*** **–** podrazumijeva razdoblje latentnosti kada korisnik nije aktivan u upotrebi proizvoda/usluga, ali istovremeno nije ni otišao niti se odredio kao korisnik koji je prekinuo odnos. To se primjerice događa kad korisnik neko vrijeme nema transakcija s tvrtkom.

Hard i *soft churn* trebaju se analizirati zasebno, pri čemu se za *soft* varijantu prekida ugovornih odnosa najčešće koriste tehnike segmentacije i analize ponašanja te *fuzzy* ekspertni sustavi (Klepac at al., 2015).

Daljnja je podjela detaljnija, pa se razlikuje:

1. **Dobrovoljan *churn*** (engl. *voluntary*) – prekid ugovora od strane korisnika. Ta se vrsta prekida ugovornog odnosa dalje dijeli na:
    a. **Slučajan** (engl. *incidental* ili *rotational*) – javlja se u slučaju preseljenja korisnika ili kad si korisnik više ne može priuštiti proizvod/uslugu. Treba uočiti da u slučaju te vrste *churna* korisnik NE prelazi kod izravne konkurencije (jer se seli na lokaciju gdje ni tvrtka ni konkurenti nisu dostupni i jer si ne može priuštiti ničiji proizvod/uslugu).
    b. **Namjeran** ili aktivan (engl. *deliberate*) – „pravi" dobrovoljni *churn,* kad korisnik sam pokreće prekid ugovora.
2. **Nedobrovoljan** ili pasivan (engl. *involuntary*) – kod te vrste *churna* kompanija je ta koja korisniku uskraćuje uslugu, najčešće zbog duga, zloupotrebe ili neaktivnosti.

Uz dobrovoljan i nedobrovoljan prekid ugovornog odnosa Berry i Linoff razlikuju i **očekivani** (engl. *expected*) *churn*. Definiraju ga kao prekid koji nastaje kad korisnik više nije dio ciljnog tržišta za proizvod, primjerice kad korisnik telekoma seli u drugu državu i stoga više ne može koristiti usluge tog telekoma (Berry at al., 2014).

Lazarov i Capota rade još jednu distinkciju prekida ugovornih odnosa prema statusu ugovora i odnosu s konkurencijom, pa tako razlikuju:

1. **ukupan** *churn* (engl. *total*) – kada se događa službeni raskid ugovora
2. **sakriven** (engl. *hidden*) – kad ugovor nije raskinut, ali je korisnik duže vrijeme neaktivan
3. **djelomičan** (engl. *partial*) – kad ugovor nije službeno raskinut, ali korisnik djelomično koristi usluge konkurencije

U slučajevima sakrivenog i djelomičnog *churna* korisnik primjerice plaća pretplatu, ali ima malu ili vrlo malu potrošnju unutar ponude i nepostojeću dodatnu potrošnju (Lazarov at al., 2010). Različite vrste prekida ugovornih odnosa različito se tretiraju te im se pristupa drugačijim setom analitičkih tehnika.

U ovom će se radu, budući da će se temeljiti na primjeru iz područja telekomunikacija, prikazati upravljanje prekidima ugovornih odnosa koji bi se svrstali u kategoriju *hard churna* dobrovoljnog, namjernog i ukupnog tipa.

2.2. Značenje prekida ugovornih odnosa u svakodnevnom poslovanju telekom-kompanije

Churn je najlakše definirati u djelatnostima utemeljenima na pretplati ili zasnivanju ugovornog odnosa (Klepac et al., 2015), a upravo su takve telekomunikacije – bilo da se radi o usluzi fiksne ili mobilne telefonije, davanju pristupa internetu ili ponudi televizijskih paketa. Ima u telekomunikacijama i *soft churna,* primjerice u području pružanja *prepaid* usluga, ali od korisnika s kojima se sklapa klasičan ugovorni odnos dolazi glavnina prihoda, pa su stoga oni u fokusu.

No, prevlast *hard churna* nije jedini razlog iznimne „popularnosti" analize prekida ugovornih odnosa u telekom-svijetu. Tržište telekomunikacija jedan je od najeklatantnijih primjera **zasićenog tržišta**, što znači da je dosegnuta točka u životnom vijeku telekomunikacijskih usluga u kojoj su one tako široko dostupne korisnicima da je teško plasirati novu uslugu ili proizvod. Od te točke nadalje nijedna tvrtka ne može osvojiti novi udio na tržištu (na istom tržištu za istu uslugu), a da pritom ne preuzme tržišni udio druge tvrtke (CFI Team, 2023).

Drugim riječima, telekom ne može osvojiti novog korisnika a da ga ne preotme konkurenciji ili, gledano iz druge perspektive, ne može izgubiti korisnika a da ga ne „pošalje" konkurenciji. Zasićenost tržišta znači i veći trošak akvizicije novih korisnika, pa je praćenje prekida ugovornih odnosa u području telekomunikacija jedan od ključnih problema koje valja adresirati da bi se dobro poslovalo (Mahajan et al., 2015).

Telekomunikacije su djelatnost u kojoj ima puno prilika za interakciju s korisnicima/pretplatnicima, a time i puno prilika za konstruiranje izvedenih varijabli. Neki autori to nazivaju **privilegijom telekomâ**: „S jedne strane imaju jasnu sliku o tome kad je pretplatnik počeo i prestao koristiti njihove proizvode i usluge, a s druge strane korisnici su u svakodnevnoj interakciji s kompanijom. To je pravo bogatstvo za konstrukciju izvedenih varijabli." (Klepac et al., 2015)

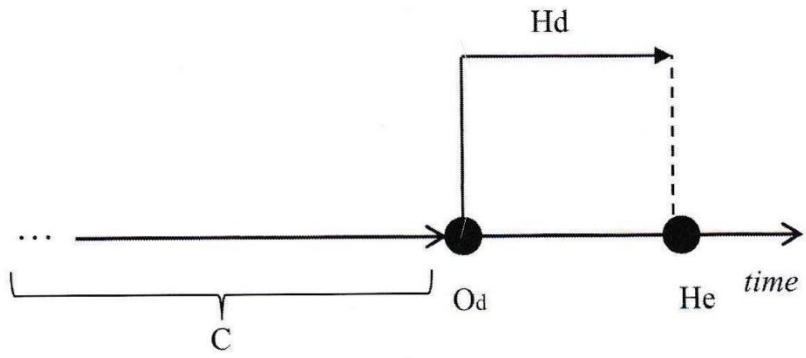
Često se događa da su telekomovo vrijeme i resursi za nastojanja oko zadržavanja korisnika ograničeni, pa se tada među vjerojatnim *churnerima* – korisnicima za koje je velika vjerojatnost da će prekinuti ugovorni odnos – dodatno izdvajaju oni najvjerojatniji. To je u slučaju telekoma, koji imaju velik broj pretplatnika, često samo gornjih 1 %.

Uz to, da bi dodatno smanjio troškove zadržavanja (engl. *retention*), telekom će za korisnike s najvećom vjerojatnošću prekida ugovora pokušati utvrditi i koliko će prihoda donijeti tijekom razdoblja ostanka kako bi targetirao one s najvećom potrošnjom, a time i vrijednošću za telekom. To je postupak identifikacije **vrijednih korisnika** (engl. *valuable customers*) i za njega je predviđanju prekida ugovornih odnosa potrebno dodati tehnike određivanja vrijednosti korisnika tijekom njihova životnog vijeka kao što su evolucijski pristup ili samoorganizirajuće mape (Lazarov at al., 2010).

3. Priprema skupa podataka

Početni korak u razvoju svakog modela utemeljenog na rudarenju podataka jest upoznavanje sa skupom podataka, njegovo čišćenje i priprema. Sva literatura navodi da se na taj pripremni korak troši oko 80 % ukupnog vremena.

Pritom je prvi zadatak s kojim se susreće analitičar za rudarenje podataka u stvarnom scenariju modeliranja rješenja za sprječavanje prekida ugovornih odnosa planiranje konstrukcije uzorka za izradu prediktivnog modela. Shematski prikazano, to planiranje konstrukcije uzorka izgleda ovako (Klepac et al., 2015):



Slika 3.1 Shema konstrukcije uzorka podataka za prediktivno modeliranje

C = razdoblje promatranja aktivnih ugovora; Od = točka promatranja i početka praćenja ishoda; He = točka završetka razdoblja praćenja ishoda; Hd = razdoblje praćenja ishoda za uzorak

U razdoblju promatranja C prati se kako se kupac ponaša u neposrednoj okolini točke promatranja i stvaraju se bihevioralne ili izvedene varijable. U razdoblju praćenja ishoda Hd prati se reakcija kupaca, odnosno ponašanje u vezi s ciljnom varijablom. Tako je posložena konstrukcija uzorka za prediktivno modeliranje prekida ugovornih odnosa, no valja imati na umu da pristup kreiranju uzorka ovisi o vrsti problema i prilagođava se specifičnoj namjeni.

Nadalje, u postupak pripreme podataka potrebno je uključiti **poslovnu logiku**, npr. izbaciti neprofitabilne korisnike i neke druge kategorije iz uzorka. Upravo je konstrukcija uzorka jedan od načina integriranja poslovnih zahtjeva u modele (Klepac et al., 2015).

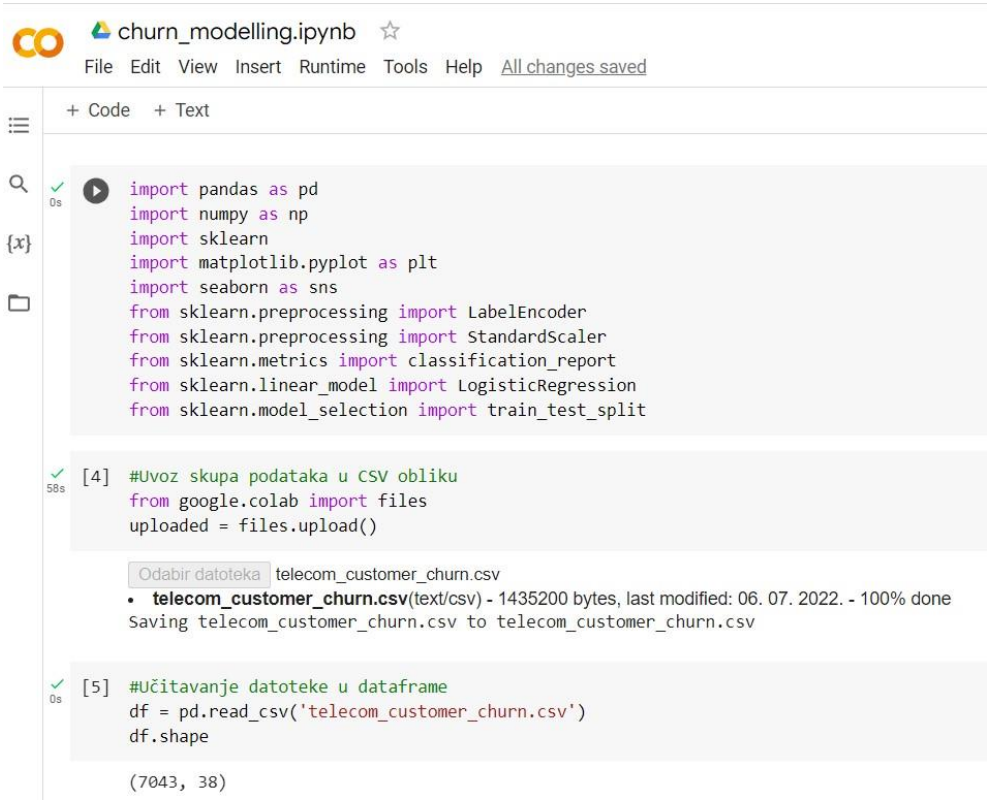
U nastavku će biti pokazan postupak pripreme unaprijed uzorkovanog skupa podataka *telecom_customer_churn.csv* iz područja telekomunikacija, javno dostupnog na web-mjestu

platforme za podatkovnu znanost Maven Analytics https://www.mavenanalytics.io/blog/maven-churn-challenge. Taj skup čine podaci o 7043 korisnika jedne kalifornijske telekomunikacijske tvrtke iz drugog tromjesečja 2022. godine, što znači da je razdoblje promatranja i stvaranja bihevioralnih varijabli bilo 90 dana (travanj, svibanj i lipanj 2022).

Za pisanje koda u Pythonu korištena je bilježnica (engl. *notebook*) na platformi Google Colab i ona je spremana na Google Drive.

3.1. Upoznavanje s podacima

Javno dostupan skup podataka preuzet je u CSV obliku i učitan u Google Colab bilježnicu. Korištenjem osnovnih pandas naredbi utvrđeno je da se sastoji od 7043 retka i 38 stupaca. Kako smo već utvrdili, radi se o podacima o 7043 korisnika jedne kalifornijske telekom-kompanije prikupljenim tijekom drugog tromjesečja 2022.



Slika 3.2 Učitavanje datoteke s podacima u Google Colab bilježnicu

Skup podataka čine sljedeći stupci:

- **Customer ID** – nasumce generiran jedinstveni ID za svakog korisnika, npr. „0002-ORFBO“
- **Gender** – spol korisnika: M/F
- **Age** – dob korisnika u godinama u trenutku Q2 2022
- **Married** – je li korisnik u braku: Yes/No
- **Number of Dependents** – broj uzdržavanih osoba koje žive s korisnikom u kućanstvu
- **City** – grad u Kaliforniji u kojem korisnik živi, npr. Sunnyvale
- **Zip Code** – poštanski broj grada
- **Latitude** – zemljopisna širina lokacije korisnika
- **Longitude** – zemljopisna dužina lokacije korisnika
- **Number of Referrals** – koliko je puta do točke promatranja korisnik preporučio kompaniju prijateljima ili obitelji
- **Tenure in Months** – koliko je mjeseci korisnik klijent kompanije do točke promatranja
- **Offer** – zadnja ponuda/paket koji je korisnik aktivirao: None, Offer A, B, C…
- **Phone Service** – je li korisnik pretplaćen na uslugu fiksnog telefona: Yes/No
- **Avg Monthly Long Distance Charges** – prosječan mjesečni trošak međugradskih poziva; ako korisnik nema fiksnu telefonsku liniju, podatak je „NaN“
- **Multiple Lines** – ima li korisnik više telefonskih linija: Yes/No; ako korisnik nema fiksnu telefonsku liniju, podatak je „NaN“
- **Internet Service** – je li korisnik pretplaćen na uslugu interneta: Yes/No
- **Internet Type** – koju vrstu internetske usluge korisnik ima: Fiber Optic/DSL/Cable; ako korisnik nema uslugu interneta, podatak je „NaN“
- **Avg Monthly GB Download** – prosječna mjesečna količina preuzetih GB; ako korisnik nema uslugu interneta, podatak je „NaN“
- **Online Security** – je li korisnik pretplaćen na dodatnu uslugu sigurnosti na internetu: Yes/No; ako korisnik nema uslugu interneta, podatak je „No“
- **Online Backup** - je li korisnik pretplaćen na dodatnu uslugu mrežnog sigurnosnog kopiranja: Yes/No; ako korisnik nema uslugu interneta, podatak je „No“
- **Device Protection Plan** - je li korisnik pretplaćen na dodatnu uslugu zaštite uređaja: Yes/No; ako korisnik nema uslugu interneta, podatak je „No“

- **Premium Tech Support** - je li korisnik pretplaćen na dodatnu uslugu premium tehničke podrške: Yes/No; ako korisnik nema uslugu interneta, podatak je „No“
- **Streaming TV** – koristi li korisnik uslugu besplatnog streaminga televizijskih sadržaja od treće strane: Yes/No; ako korisnik nema uslugu interneta, podatak je „No“
- **Streaming Movies** - koristi li korisnik uslugu besplatnog streaminga filmova od treće strane: Yes/No; ako korisnik nema uslugu interneta, podatak je „No“
- **Streaming Music** - koristi li korisnik uslugu besplatnog streaminga glazbe od treće strane: Yes/No; ako korisnik nema uslugu interneta, podatak je „No“
- **Unlimited Data** – je li korisnik kupio neki dodatni mjesečni paket za flat internet: Yes/No; ako korisnik nema uslugu interneta, podatak je „No“
- **Contract** – vrsta ugovora: Month-to-Month/One Year/Two Years
- **Paperless Billing** – ima li korisnik e-račun: Yes/No
- **Payment Method** – način plaćanja: Bank Withdrawal/Credit Card/Mailed Check
- **Monthly Charge** – ukupan mjesečni trošak korisnika za sve usluge kompanije
- **Total Charges** – koliko je ukupno naplaćeno od tog korisnika do točke promatranja
- **Total Refunds** - koliko je ukupno taj korisnik imao povrata do točke promatranja
- **Total Extra Data Charges** – koliko je korisnik ukupno platio troškove interneta izvan tarife/paketa do točke promatranja
- **Total Long Distance Charges** – koliki je korisnikov ukupan trošak za međugradske pozive izvan tarife/paketa do točke promatranja
- **Total Revenue** – ukupan prihod od tog korisnika do točke promatranja; računa se prema formuli *Total Charges - Total Refunds + Total Extra Data Charges + Total Lond Distance Charges*
- **Customer Status** – status korisnika na kraju Q2 2022: Stayed/Churned/Joined
- **Churn Category** – ako je status korisnika bio *Churned*, kategorizacija razloga: Attitude, Competitor, Dissatisfaction, Other, Price
- **Churn Reason** - ako je status korisnika bio *Churned*, opis razloga (slobodan unos)

U moru podataka primjećujemo da se podaci prikupljeni tijekom tromjesečnog razdoblja promatranja sastoje od **sociodemografskih varijabli** kao što su spol, dob, bračni status, mjesto boravišta itd. i od **izvedenih ili bihevioralnih varijabli** koje nam govore o ponašanju korisnika, npr. koliko je puta preporučio kompaniju, plaća li dodatni podatkovni promet ili razgovore izvan paketa, koliko korisnik mjesečno troši itd. Potonje su varijable vrlo važne

jer je iskustvo pokazalo da su upravo bihevioralni atributi najvažniji za konstrukciju modela za predviđanje prekida ugovornih odnosa, dok su sociodemografski atributi poput začina koji tim modelima daje okus (Klepac et al., 2015). Uostalom, ponašanje je ono koje može odražavati namjeru.

Izvedene, bihevioralne ili virtualne (engl. *derived*) varijable izravan su odraz domenskog znanja upotrijebljenog u modelima za predviđanje prekida ugovornih odnosa i alat za artikulaciju bihevioralnih obilježja. „Izvedene varijable nose *data science* projekte jer reprezentiraju ekspertno znanje artikulirano kroz varijable.“ (Klepac, 2019)

3.2. Čišćenje podataka

Nakon osnovnog upoznavanja s podacima slijedi faza njihova čišćenja i pripreme prvo za primjenu tehnike analize relevantnosti atributa, a zatim i modeliranje. U toj fazi i dalje se koriste jednostavni alati koje nude pandas, numpy i ostale korištene Python biblioteke namijenjene analizi podataka.

3.2.1. Uklanjanje stupaca koji ne nose informaciju

Za početak, bit će izbačeni stupci za koje je procijenjeno da ne nose korisnu informaciju.

```
[ ]  #Pregled svih stupaca/varijabli
     df.columns.values
```

```
array(['Customer ID', 'Gender', 'Age', 'Married', 'Number of Dependents',
       'City', 'Zip Code', 'Latitude', 'Longitude', 'Number of Referrals',
       'Tenure in Months', 'Offer', 'Phone Service',
       'Avg Monthly Long Distance Charges', 'Multiple Lines',
       'Internet Service', 'Internet Type', 'Avg Monthly GB Download',
       'Online Security', 'Online Backup', 'Device Protection Plan',
       'Premium Tech Support', 'Streaming TV', 'Streaming Movies',
       'Streaming Music', 'Unlimited Data', 'Contract',
       'Paperless Billing', 'Payment Method', 'Monthly Charge',
       'Total Charges', 'Total Refunds', 'Total Extra Data Charges',
       'Total Long Distance Charges', 'Total Revenue', 'Customer Status',
       'Churn Category', 'Churn Reason'], dtype=object)
```

Slika 3.3 Pregled svih početnih stupaca u *dataframeu*

Takav je sigurno stupac *Customer ID*. Stupac *City* sadrži čak 1106 jedinstvenih (engl. *unique*) vrijednosti i premda nije isključeno da bi taj stupac - možda u kombinaciji s podatkom o poštanskom broju da dobijemo precizniju lokaciju (za neke gradove ima više poštanskih brojeva, pa možemo pretpostaviti da su to gradske četvrti) - bio relevantan za

predviđanje *churnera*, rad s tim stupcem u kasnijem koraku analize relevantnosti atributa dao bi velik broj *dummy* varijabli te tako otežao i usporio analizu. Stoga će biti uklonjen.

Nadalje, uklonit će se i stupci *Zip Code*, *Latitude* i *Longitude* – općenito, odlučeno je da se isključe lokacijski atributi.

I na posljetku, bit će uklonjeni i stupci *Churn Category* i *Churn Reason*. Iako su to vrlo zanimljivi podaci koji kategoriziraju *churnere* prema pet okvirnih razloga i navode konkretan razlog odlaska u formi slobodnog teksta, nisu korisni za konstruiranje prediktivnog modela. Ti bi podaci bili zanimljivi u nekoj postanalizi.

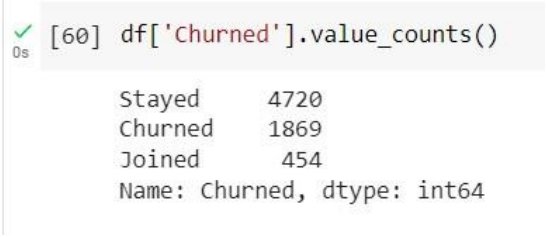
Na kraju će se preimenovati preostali stupci tako da ne sadrže razmak kako bi ih se moglo lakše referencirati u kodu.

Nakon prvog koraka čišćenja *dataframe* izgleda ovako:

Slika 3.4 *Dataframe* nakon uklanjanja stupaca koji ne nose informaciju

3.2.2. Ciljna varijabla

Ciljna varijabla nalazi se u stupcu *Churned* i - neočekivano - može poprimiti tri vrijednosti: *Stayed*, *Churned* i *Joined*.



Slika 3.5 Podaci u stupcu *Customer Status*

Uvidom u podatke saznaje se da status *Joined* imaju 454 korisnika koji su postali pretplatnici prije tri mjeseca ili manje. Među njima su zastupljene sve vrste ugovora: na godinu dana,

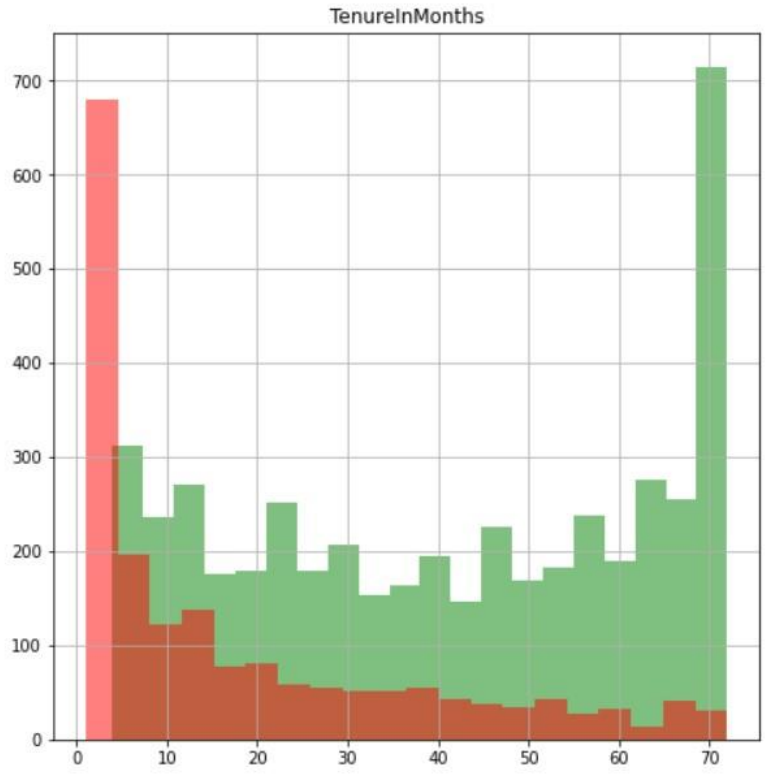
dvije godine, a najzastupljeniji su oni koji plaćaju uslugu iz mjeseca u mjesec, što znači da su slobodni otići svakog mjeseca.

```
[104] #Distribucija vrsta ugovora među korisnicima sa statusom "Joined"
      df[df.Churned == 'Joined'].Contract.value_counts()

      Month-to-Month    408
      One Year           24
      Two Year           22
      Name: Contract, dtype: int64
```

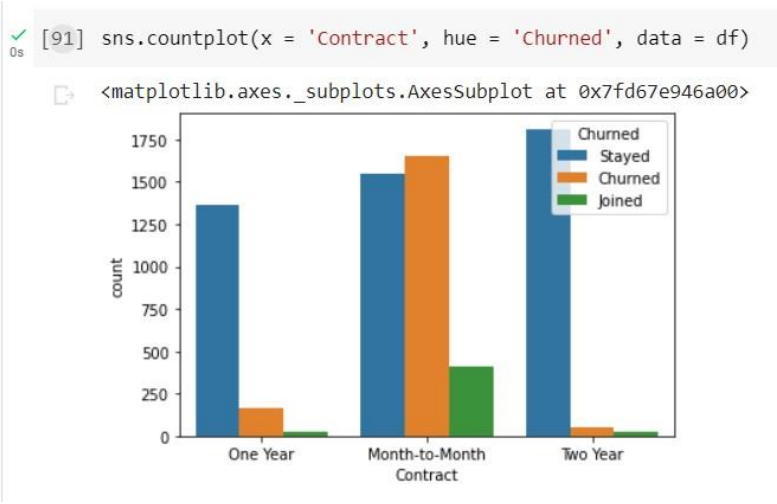
Slika 3.6 Distribucija vrsta ugovora za korisnike sa statusom "*Joined*"

Nadalje, uvidom u podatke saznaje se i da je među korisnicima koji su po statusu kategorizirani kao *Stayed* najmanji *TenureInMonths* četiri mjeseca. To znači da kompanija sustavno novopridošle korisnike čija je vjernost kompaniji manja ili jednaka tri mjeseca, a nisu *churnali,* kategorizira kao *Joined*, što poslovno ima smisla. S druge strane, iz histograma u nastavku vidljivo je da ugovorni odnos češće prekidaju korisnici koji su kraće s kompanijom, tj. imaju manji *tenure.* Drugim riječima, što je korisnik duže s kompanijom, to je manja vjerojatnost da će otkazati ugovor.



Slika 3.7 Distribucija vjernosti u mjesecima za *churnere* (crveno) i *nechurnere* (zeleno)

Istovremeno, iz distribucije statusa korisnika prema vrsti ugovora, koju prikazuje Slika 3.8, saznaje se da najviše *churnaju* korisnici koji uslugu plaćaju iz mjeseca u mjesec.



Slika 3.8 Distribucija statusa korisnika prema vrsti ugovora

Stoga se s korisnicima sa statusom *Joined* čini sljedeće: oni koji kao vrstu ugovora imaju *One Year* ili *Two Year* - njih 46 - bit će prebačeni u status *Stayed,* dok će se 408 korisnika s ugovorom *Month-to-Month* izbrisati iz *dataframea* jer je prema dostupnom domenskom znanju vjerojatnost da će *churnati* i ostati podjednaka. Jednostavno nije jasno kamo ih smjestiti.

Po završetku ove faze čišćenja podataka *dataframe* izgleda ovako:

Slika 3.9 Krajnji *dataframe* ima 6635 redaka i 31 stupac

3.2.3. Nedostajuće vrijednosti

Na prvi pogled u nekim stupcima ima nedostajućih vrijednosti, no nakon detaljnije analize uviđa se da se podaci koji nedostaju zapravo mogu zamijeniti s „Nije primjenjivo“ ili *No Phone Service*/*No Internet Service* jer se tiču daljnjih usluga za koje su preduvjeti telefonska linija ili internet, a korisnik ih nema (vrijednost za *PhoneService* ili *InternetService* mu je *No*).

```
[26] #Provjera ima li nedostajućih vrijednosti
     df.isna().sum()

Gender                             0
Age                                0
Married                            0
NumberOfDependents                 0
NumberOfReferrals                  0
TenureInMonths                     0
Offer                              0
PhoneService                       0
AvgMonthlyLongDistanceCharges    682
MultipleLines                    682
InternetService                    0
InternetType                    1526
AvgMonthlyGBDownload            1526
OnlineSecurity                  1526
OnlineBackup                    1526
DeviceProtectionPlan            1526
PremiumTechSupport              1526
StreamingTV                     1526
StreamingMovies                 1526
StreamingMusic                  1526
UnlimitedData                   1526
Contract                           0
PaperlessBilling                   0
PaymentMethod                      0
MonthlyCharge                      0
TotalCharges                       0
TotalRefunds                       0
TotalExtraDataCharges              0
TotalLongDistanceCharges           0
TotalRevenue                       0
Churned                            0
dtype: int64
```

Slika 3.10 Stupci s nedostajućim podacima

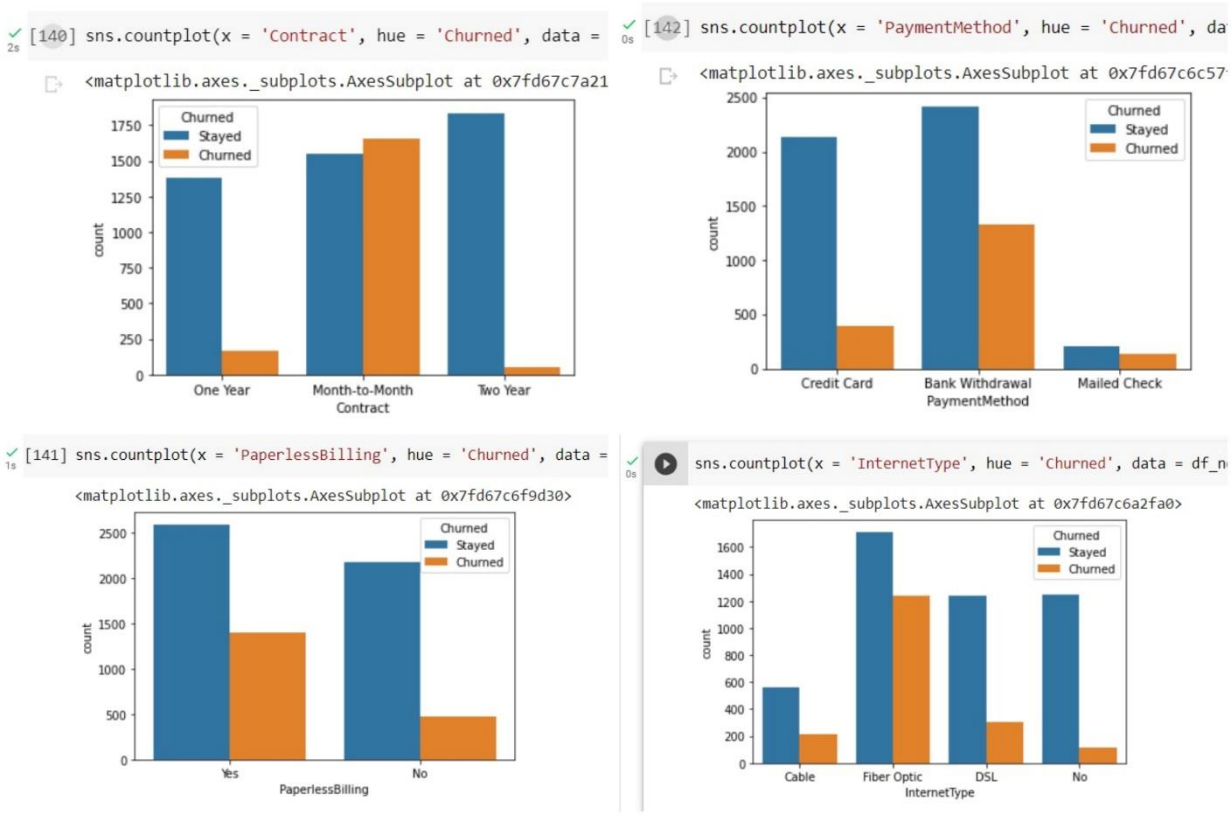
Zapravo se za potrebe daljnje analize te vrijednosti mogu zamijeniti izravno s *No*, što je i učinjeno. Time je riješen problem nedostajućih podataka.

Na kraju će se još sve vrijednosti *Yes* i *No* u skupu podataka pretvoriti u 1 i 0 kao priprema za predstojeću analizu relevantnosti atributa.

3.3. Neke pretpostavke na temelju vizualizacija

Tijekom pripreme i čišćenja skupa podataka otkrivena su neka pravila i trendovi koji su postali posebno uočljivi nakon vizualizacije podataka.

Tako je uočeno da najveći broj korisnika koji je prekinuo ugovorni odnos ima vrstu ugovora iz mjeseca u mjesec, dok korisnici s jednogodišnjim i dvogodišnjim ugovorom vrlo malo *churnaju*. Nadalje, uočena je značajna veza između prekida ugovornog odnosa i upotrebe usluge optičkog interneta, koju prikazuje Slika 3.11 - stupčasti grafikon dolje desno: narančasti stupac iznad oznake *Fiber Optic* ukazuje na probleme s tom uslugom. U stvarnom životu ta informacija potaknula bi niz dodatnih analiza i aktivnosti u pokušajima otkrivanja što je točno problematično s uslugom optičkog interneta tog operatera.



Slika 3.11 Još neke vizualizacije za stvaranje zaključaka i domenskog znanja

Uočeno je i da ima manje *churnera* među korisnicima dodatnih usluga kao što su *Online Security*, *Online Backup, Premium Tech Support* i ostalih. Ako korisnik ima već jednu od tih usluga, manja je vjerojatnost da će prekinuti ugovorni odnos. Dobar daljnji smjer istraživanja bio bi pokušaj konstrukcije izvedenih varijabli – raznih kombinacija više dodatnih usluga i utvrđivanje koliko točno povećanje broja usluga i proizvoda koje korisnik upotrebljava utječe na smanjenje vjerojatnosti da on prekine ugovorni odnos.

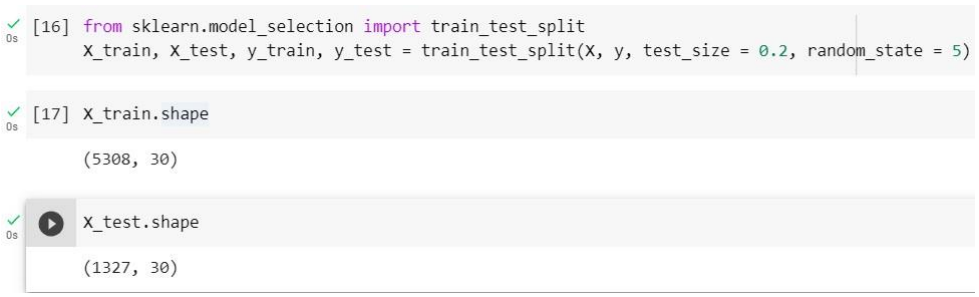
Iz toga je očito da domensko znanje i izvedene varijable idu pod ruku s tehnikama i modelima rudarenja podataka u stvaranju rješenja za sprječavanje prekida ugovornih odnosa.

3.4. Uzorkovanje skupa podataka u omjeru 80 : 20

Vrlo važna faza pripreme i čišćenja podataka završava dijeljenjem skupa podataka na dio koji će se koristiti za treniranje budućih prediktivnih modela i dio za testiranje njihove uspješnosti u predviđanju. U literaturi se navodi da je pravi trenutak za dijeljenje skupa podataka prije izvođenja postupaka odabira značajki (engl. *features*) – u slučaju ovog rada to je analiza relevantnosti atributa putem mjera IV i WoE. Razlog je tome kako ne bi došlo do „curenja" podataka iz testnog skupa u *pipeline* za treniranje (Brownlee, 2020).

To se obično radi u omjeru 80 : 20 iako je u literaturi primijećeno da neki dijele skup podataka na tri dijela. Pored dijelova za treniranje i testiranje, neki pripremaju i dio za validaciju. U ovom radu korištena je podjela skupa podataka na dio za treniranje i testiranje.

Podjela je izvedena s pomoću funkcije *sklearn.model_selection.train_test_split()* iz bliblioteke scikit-learn.



Slika 3.12 Podjela skupa podataka na dio za treniranje i dio za testiranje

4. Priprema za modeliranje

Nakon što su uz izdašnu upotrebu domenskog znanja i poslovne logike pročišćeni podaci i skup podataka podijeljen u omjeru 80 : 20, slijedi iznimno važan pripremni korak za prediktivno modeliranje: analiza relevantnosti atributa.

Radi se o moćnoj tehnici koja služi za prepoznavanje varijabli koje najjače utječu na prekid ugovornih odnosa. Atributi koji pokazuju najveću razlikovnu moć za *churn* (korisnik je prekinuo ugovorni odnos = 1 ili 0) bit će ovom tehnikom odabrani kao najbolji kandidati za izradu modela za predviđanje prekida ugovornih odnosa (Klepac et al., 2015).

Dva su preduvjeta za analizu relevantnosti atributa (Radečić, 2019):

1. da u skupu podataka nema praznih / nedostajućih vrijednosti
2. da skup podataka ne sadrži kontinuirane varijable.

Prvi preduvjet već je ispunjen u fazi pripreme i čišćenja, a drugi se postiže grupiranjem („binanjem“) kontinuiranih varijabli u 5 – 10 skupina s najmanje 5 % zapisa u svakoj.

Drugi je preduvjet zadovoljen izvršavanjem odgovarajućeg koda kojim su kontinuirane varijable – njih 12 – podijeljene na maksimalno pet grupa („binova“) s pomoću pandas funkcije *pd.qcut()*.

Nakon zadovoljavanja obaju preduvjeta krenulo se s izračunom relevantnosti atributa.

4.1. Analiza relevantnosti atributa na 80 % uzorka

U ovoj točki analize iza nas je detaljna podjela na razrede (engl. *fine classing*) i slijede daljnji koraci analize relevantnosti atributa:

1. računanje IV i WoE
2. gruba podjela na razrede (engl. *coarse classing*)
3. kreiranje *dummy* varijabli
4. korelacijska analiza

Izvedena je analiza relevantnosti atributa na binomnoj ciljnoj varijabli *prekid ugovornog odnosa*, što znači s dvama mogućim ishodima: „ugovorni odnos je prekinut“ (*churn* = 1) ili „ugovorni odnos nije prekinut“ (*churn* = 0).

U programskom kodu su izračunate mjere **težine dokaza** (engl. *weight of evidence,* WoE) i **vrijednosti informacije** (engl. *information value*, IV), i to prema sljedećim formulama:

$$\text{WoE} = \ln\left(\frac{\%\ \text{nonchurn u atributu}}{\%\ \text{churn u atributu}}\right) \tag{2}$$

Mjera vrijednosti informacije izračunava se s pomoću težine dokaza (WoE):

$$\text{IV} = \sum_{i=1}^{n} (\%\ nonchurn\ u\ atributu - \%\ churn\ u\ atributu) * WoE \tag{3}$$

Evo kako te formule izgledaju u kodu (istaknuto žutim):

```
[271] def calculate_woe_iv(dataset, feature, target):
          lst = []
          for i in range(dataset[feature].nunique()):
              val = list(dataset[feature].unique())[i]
              lst.append({
                  'Value': val,
                  'All': dataset[dataset[feature] == val].count()[feature],
                  'Good': dataset[(dataset[feature] == val) & (dataset[target] == 0)].count()[feature],
                  'Bad': dataset[(dataset[feature] == val) & (dataset[target] == 1)].count()[feature]
              })

          dset = pd.DataFrame(lst)
          dset['Distr_Good'] = dset['Good'] / dset['Good'].sum()
          dset['Distr_Bad'] = dset['Bad'] / dset['Bad'].sum()
          dset['WoE'] = np.log(dset['Distr_Good'] / dset['Distr_Bad'])
          dset = dset.replace({'WoE': {np.inf: 0, -np.inf: 0}})
          dset['IV'] = (dset['Distr_Good'] - dset['Distr_Bad']) * dset['WoE']
          iv = dset['IV'].sum()

          dset = dset.sort_values(by='WoE')

          return dset, iv
```

Slika 4.1 Izračun mjera WoE i IV tijekom analize relevantnosti atributa

Po dovršetku izračuna, odnosno nakon izvršenja koda, analiza relevantnosti je kao atribute beznačajne ili slabe prediktivnosti (niske vrijednosti informacije IV) eliminirala spol, dob, korištenje usluge telefonije te paketa Online Backup i Device Protection Plan, zatim streaming televizije, glazbe i filmova, dodatnu potrošnju interneta i dvije varijable s mjesečnim prosjecima (potrošeni gigabajti i troškovi za međugradske pozive).

```
WoE and IV for column: Married
   Value   All  Good   Bad  Distr_Good  Distr_Bad       WoE        IV
1      0  3303  2103  1200    0.441251   0.642055 -0.375061  0.075314
0      1  3332  2663   669    0.558749   0.357945  0.445321  0.089422
IV score: 0.16


WoE and IV for column: Offer
     Value   All  Good   Bad  Distr_Good  Distr_Bad       WoE        IV
1  Offer E   649   223   426    0.046790   0.227929 -1.583371  0.286811
0     None  3625  2574  1051    0.540076   0.562333 -0.040385  0.000899
2  Offer D   602   441   161    0.092530   0.086142  0.071537  0.000457
5  Offer C   415   320    95    0.067142   0.050829  0.278340  0.004541
4  Offer B   824   723   101    0.151700   0.054040  1.032185  0.100803
3  Offer A   520   485    35    0.101762   0.018727  1.692697  0.140555
IV score: 0.53
```

Slika 4.2 Izračuni težine dokaza i vrijednosti informacije za varijable *Married* i *Offer*

A za sljedeće je atribute izračunato da su srednje jaki ili jaki prediktori – mnogi od njih čak i iznimno jaki pa zahtijevaju daljnju provjeru koreliranosti.

Atribut	*IV*	*Prediktivna snaga*
Married	0.16	Srednji prediktor
Number Of Dependents	0.31	Jak prediktor
Number Of Referrals	1	Vrlo jak prediktor
Tenure In Months	1.09	Vrlo jak prediktor
Offer	0.53	Vrlo jak prediktor
Internet Service	0.35	Jak prediktor
Internet Type	0.52	Vrlo jak prediktor
Online Security	0.21	Srednji prediktor
Premium Tech Support	0.19	Srednji prediktor
Unlimited Data	0.14	Srednji prediktor
Contract	1.77	Vrlo jak prediktor
Paperless Billing	0.19	Srednji prediktor
Payment Method	0.26	Srednji prediktor
Monthly Charge	0.24	Srednji prediktor
Tota lCharges	0.52	Vrlo jak prediktor
Total LongDistance Charges	0.53	Vrlo jak prediktor
Total Revenue	0.64	Vrlo jak prediktor

Tablica 4.1 Prediktivnost najjačih atributa iz skupa podataka

Od ukupno 30 atributa (31. stupac je bila ciljna varijabla) 13 ih se pokazalo beskorisnima ili slabim prediktorima, a 17 atributa ima srednju i veliku prediktivnu moć. Ti će atributi nakon

grube podjele u razrede (engl. *coarse classing*) i korelacijske analize koje slijede biti uzeti za temelj prediktivnih modela, a s obzirom na visoku prediktivnost atributa, za očekivati je da će modeli raditi kvalitetna predviđanja.

4.2. Profil tipičnog *„churnera“*

Osim stupca IV u izračunu su praćene i vrijednosti WoE po pojedinim atributima i binovima. Vrijednosti težine dokaza (WoE) govore koliko neka kategorija utječe na prekid ugovornih odnosa: što je vrijednost WoE bliža nuli, to je ta kategorija irelevantnija za predikciju, a što ima veće negativne vrijednosti, to znači da kategorija više utječe na prekid ugovornog odnosa (Klepac, 2019).

Gledajući upravo vrijednosti težine dokaza može se zaključiti o karakteristikama tipičnog *churnera* telekoma čiji su podaci analizirani. Ovo je profil tipičnog korisnika „našeg“ telekoma koji je prekinuo ugovorni odnos:

- nije u braku i sam je u kućanstvu
- koristi usluge tog telekoma manje od 9 mjeseci
- ima optički internet
- ima ugovor koji se produžuje iz mjeseca u mjesec
- račun prima u digitalnom obliku, a mjesečni račun mu je između 79 i 94 dolara
- za plaćanje tog računa NE koristi kreditnu karticu
- ukupan prihod telekoma od tog korisnika manji je od 592 dolara

Sve smo to saznali izračunom vrijednosti IV i WoE na temelju danog skupa podataka korištenjem tehnike analize relevantnosti atributa.

4.3. Kreiranje *dummy* varijabli

Nakon što je napravljena detaljna podjela po razredima, provjereno je u kojim su razredima vrijednosti WoE bliske, pa je nad tim varijablama i razredima napravljena gruba podjela po razredima – dodatno su grupirani razredi sa sličnim vrijednostima težine dokaza.

```
[54] totalcharges_df, total_charges_iv = calculate_woe_iv(ara_df, 'TotalCharges_Bins', 'Churned')
```

```
totalcharges_df
```

	Value	All	Good	Bad	Distr_Good	Distr_Bad	WoE	IV
1	TCharges_lt_387	1327	570	757	0.119597	0.405029	-1.219831	0.348179
0	TCharges_387_to_1095	1327	977	350	0.204994	0.187266	0.090450	0.001603
4	TCharges_2274_to_4635	1327	1033	294	0.216744	0.157303	0.320539	0.019053
2	TCharges_1095_to_2274	1327	1043	284	0.218842	0.151953	0.364778	0.024400
3	TCharges_4635_to_8684	1327	1143	184	0.239824	0.098448	0.890372	0.125877

```
[56] totalcharges_df = coarse_classer(totalcharges_df, 2, 4)
     totalcharges_df
```

```
/usr/local/lib/python3.8/dist-packages/numpy/core/fromnumeric.py:3438: FutureWarning: Dropping of nu
  return mean(axis=axis, dtype=dtype, out=out, **kwargs)
```

	Value	All	Good	Bad	Distr_Good	Distr_Bad	WoE	IV
0	TCharges_4635_to_8684	1327.0	1143.0	184.0	0.239824	0.098448	0.890372	0.125877
1	NaN	1327.0	1088.0	239.0	0.228284	0.127876	0.605455	0.072465
2	TCharges_387_to_1095	1327.0	977.0	350.0	0.204994	0.187266	0.090450	0.001603
3	TCharges_lt_387	1327.0	570.0	757.0	0.119597	0.405029	-1.219831	0.348179

Slika 4.3 Grupiranje dvaju razreda sa sličnom vrijednosti WoE u jedan razred *TCharges_1095_to_4635*

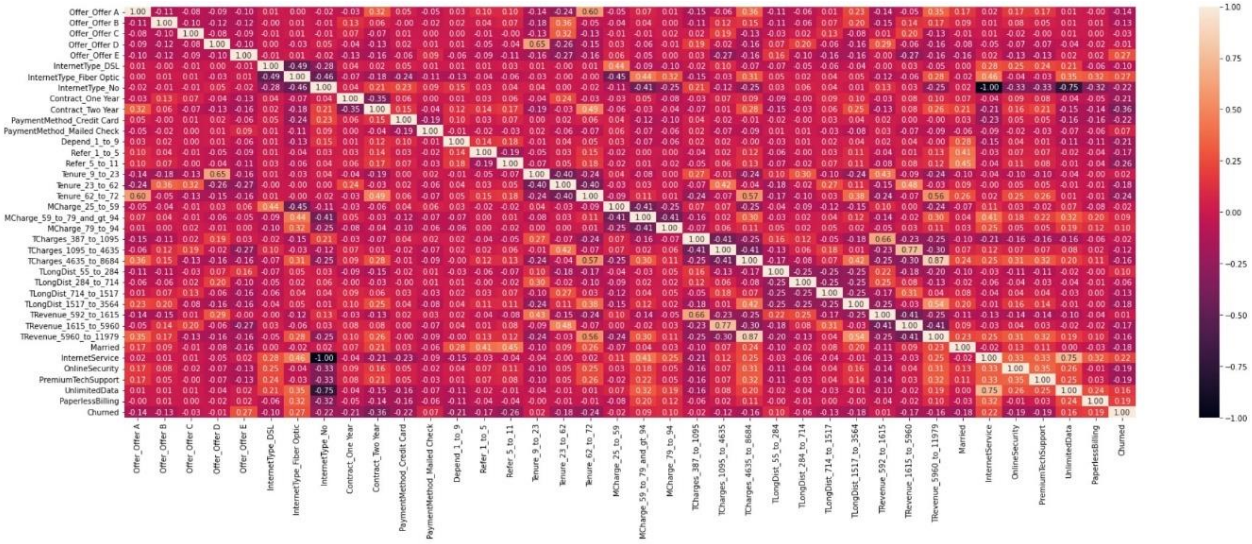
Slijedi kreiranje *dummy* varijabli kao preduvjeta za korelacijsku analizu. Rezultat izrade *dummy* varijabli skup je pročišćenih visokoprediktivnih atributa i razreda (binova), a vrijednosti su 0 ili 1.

TLongDist_1517_to_3564	TRevenue_592_to_1615	TRevenue_1615_to_5960	TRevenue_5960_to_11979	Married	InternetService	OnlineSecurity	PremiumTechSupport	UnlimitedData	PaperlessBilling	Churned
0	1	0	0	1	1	0	1	1	1	0
0	1	0	0	0	1	0	0	0	0	0
0	0	0	0	0	1	0	0	1	1	1
0	1	0	0	1	1	0	0	1	1	1
0	0	0	0	1	1	0	1	1	1	1

Slika 4.4 Finalna tablica u svim stupcima kao vrijednosti sadrži samo nule i jedinice

4.4. Korelacijska analiza

Na kraju smo s pomoću biblioteke seaborn iscrtali tablicu korelacija, koja pokazuje da su neki atributi visoko korelirani. Njih se izbacuje iz modeliranja jer bi prejako utjecale na model.



Slika 4.5 Korelacijska matrica za izdvojene *dummy* varijable

4.5. Osvrt na važnost mjera IV i WoE

U ovom je poglavlju pokazano kako tehnika analize relevantnosti atributa kroz mjere vrijednosti informacije i težine dokaza pomaže suziti skup atributa prema kriteriju snage prediktivnosti, čime se osigurava dobar model za predviđanje. No, jednako su važni i benefiti za poslovanje u vidu odgovora na pitanja zašto je neki korisnik prekinuo ugovorni odnos, razumijevanja motivacije korisnika, profiliranje korisnika i slično. Poslovni korisnici imaju važnu ulogu u tumačenju analize relevantnosti atributa jer bi trebali poslovnom logikom objasniti i podržati odnose između ciljne varijable i atributa koji su prepoznati kao važni prediktori.

U sklopu sveobuhvatnog rješenja za problem prekida ugovornih odnosa, nakon faze analize relevantnosti atributa u borbi protiv *churna* slijede dva jednako važna puta:

- razvoj prediktivnog modela i
- istraga o uzorcima *churna*.

Da model za predviđanje prekida ugovornih odnosa ne bi bio puki kalkulator vjerojatnosti *churna* (Klepac at al., 2015), potrebno je tretirati uvide stečene putem modela, ali i tehnike analize relevantnosti atributa i svih ostalih metoda rudarenja podataka kao smjerokaze za generiranje hipoteza o uzrocima prekida ugovornih odnosa jer je bez dubljeg razumijevanja uzroka nemoguće osmisliti dobru strategiju za sprječavanje *churna.*

Analiza relevantnosti atributa važna je iz sljedećih razloga:

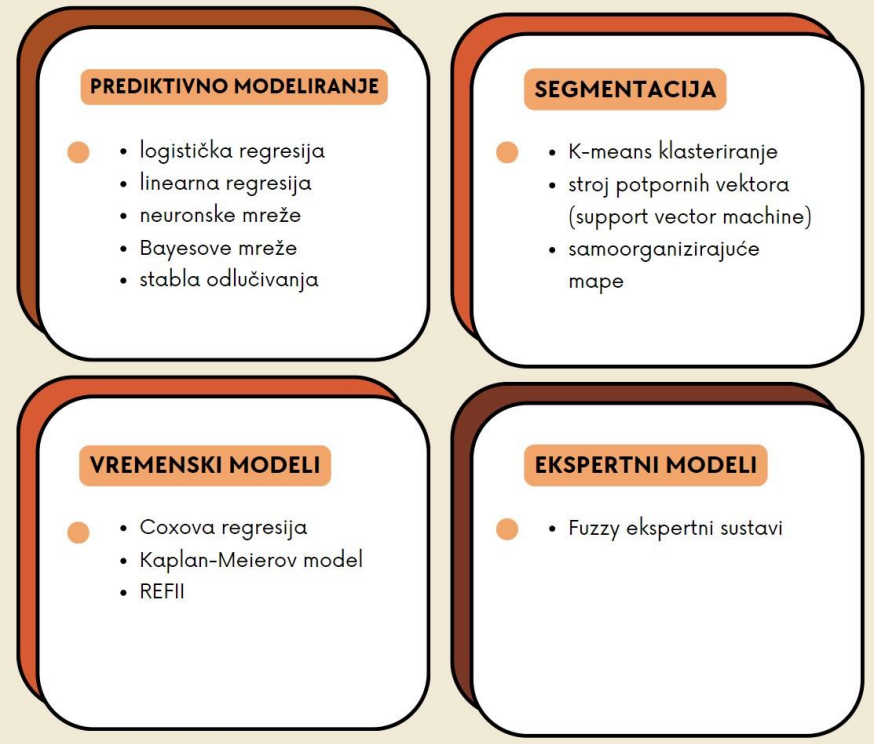
1. Osigurava odabir odgovarajućih varijabli i sprječava izradu modela na atributima koji nisu relevantni.
2. S poslovne strane predstavlja temelj za stvaranje hipoteza o razlozima prekida ugovornih odnosa ako to nije izravno vidljivo (odgovor na pitanje zašto je korisnik *churnao*).
3. Omogućuje pronalaženje važnih odnosa i uzoraka.

Drugim riječima, „Prepoznavanje najvažnijih varijabli, onih koje najjače utječu na ciljnu varijablu, smanjuje redundantnost i neizvjesnost u fazi razvoja modela.“ (Klepac et al., 2015) A upravo je faza razvoja modela ona koja sada slijedi.

5. Kreiranje prediktivnih modela

Prediktivna analitika može se definirati kao proces identifikacije trendova u podacima s pomoću matematičkih algoritama. Izrada modela za predviđanje temelji se na prethodnom ponašanju pretplatnika (Klepac et al., 2015) i samo je dio rješenja za otkrivanje i ublažavanje prekida ugovornih odnosa.

Prediktori *churna* kombinacija su svojstava korisnika poznatih u trenutku sklapanja ugovora (dob, spol, kanal akvizicije itd.) i stvari koje su se dogodile tijekom korisnikova korištenja usluga, npr. tehničke poteškoće, neočekivano veliki računi, loša usluga, loš dojam pri kontaktu s korisničkom službom itd. Već je pokazano kako se skup podataka koji sadrži takva svojstva, statička i bihevioralna, transformira radi čim bolje pripreme za modeliranje, pri čemu je vjerojatno najvažniji korak bila primjena tehnike analize relevantnosti atributa. Zahvaljujući njoj iz skupa atributa izuzete su varijable s najvećom prediktivnom snagom kako bi se na njima temeljilo predviđanje.



Slika 5.1 Tehnike rudarenja podataka koje se koriste za prediktivno modeliranje churna

Iz perspektive rudarenja podataka, dvije su vrste modeliranja i one odgovaraju na različita pitanja (Berry et al., 2014):

1. **Predviđanje s binarnim ishodom** - odgovara na pitanje tko će od korisnika otići. Preduvjet je za takva predviđanja vremenski horizont, koji najčešće iznosi 60 ili 90 dana, dok je rezultat određena ocjena ili **„skor“** (engl. *score*) – vjerojatnost da će korisnik prekinuti ugovorni odnos unutar vremenskog horizonta modela. Definira se prag te ocjene i korisnici s rezultatom iznad praga uključuju se u programe za zadržavanje korisnika.

Predviđanja s binarnim ishodom prikladna su za kratkoročna bavljenja prekidima ugovornih odnosa.

2. **Analiza preživljenja** (engl. *survival* analiza) – njome se procjenjuje preostalo vrijeme korisnika kao pretplatnika predmetnog telekoma (engl. *tenure*). Ta vrsta modeliranja odgovara na pitanje koliko će još korisnik ostati s telekomom.

Analiza preživljenja prikladnija je za dugoročno predviđanje prekida ugovornih odnosa, a kao tehnika temelj je i za izradu modela vrijednosti korisnika za cijelog pretplatničkog vijeka (engl. *customer lifetime value model*) kao i za izračun ocjene vjernosti korisnika (engl. *customer loyalty score*) (Berry et al., 2014).

U sklopu ovog rada rađena su predviđanja s binarnim ishodom, a metode prediktivnog rudarenja podataka primijenjene u ovom radu jesu logistička regresija, Bayesova vjerojatnost i neuronska mreža.

Postoji još cijeli niz alata prikladnih za modeliranje *churna* i konstrukciju prediktivnih modela. U tablici u nastavku navedene su najčešće korištene metode s prednostima i manama.

Metoda rudarenja podataka	Prednosti	Mane
Stabla odlučivanja	• vrlo jednostavna tehnika • daje pouzdane rezultate • daje konkretna pravila	• teško je izdvojiti pravila za klasifikaciju • stabilnost ne jamči optimalno rješenje
Neuronske mreže	• sposobnost preciznog predviđanja	• teško ih je konstruirati • netransparentnost u tumačenju krajnjih rezultata

Regresija	• jednostavnost implementacije modela • bogata literatura o upotrebi modela	• nemogućnost prepoznavanja i izražavanja uzoraka ponašanja skrivenih u podacima
Asocijacijska pravila	• mogućnost otkrivanja skrivenih odnosa među bihevioralnim podacima • sposobnost ulančavanja događaja i korisničkih ponašanja	• ukupan broj stavki koje se rijetko pojavljuju
Stroj potpornih vektora (SVM)	• točnost podataka uz puno bolje rezultate • mogućnost kontrole učestalosti pogrešaka	• odnosi se samo na trenutno stanje
Klasteriranje	• najčešće korištena metoda • početna procjena korisničkih podataka	• performanse same te metode nisu dovoljne za predviđanje ponašanja korisnika
Slučajna šuma (engl. *random forest*)	• stabilnost i ravnomjernost • dobre performanse	• teško konstruirati
Novi Bayes	• veći broj nominalnih varijabli radi boljih performansi	• puno manja preciznost u slučaju binarnih varijabli

Tablica 5.1 Prednosti i mane metoda korištenih za modeliranje *churna* (Reza et al., 2012)

Ne prilagođavaju se samo alati i metode problemu: i rješenje za *churn* ovisi o raznim čimbenicima.

Neki od tih čimbenika su:

- stabilnost tržišta - je li stabilno ili turbulentno
- status konkurencije – primjerice, priprema li se za nadolazeću konkurenciju ili postojeći konkurent postaje agresivniji na tržištu i treba adekvatno odgovoriti
- kojeg se korisnika želi zadržati - potrebno je s pomoću *fuzzy* ekspertnih sustava izračunati i pratiti prospektivnu vrijednost korisnika (engl. *prospective customer value*) (Klepac et al., 2015).

Da bi bilo kvalitetno, prediktivno modeliranje treba se planirati i odvijati u kontekstu poslovne situacije, u skladu s domenskim znanjem i uz praćenje smjernica poslovne logike.

U nastavku će biti pokazani primjeri konstrukcije jednostavnih modela za predviđanje prekida ugovornih odnosa.

5.1. Rješavanje problema nebalansiranosti skupa podataka i ostale pripremne radnje

Kad je u nekom skupu podataka svaka klasa ciljne varijable predstavljena istim brojem uzoraka, kaže se da je taj skup balansiran (Duca, 2021), a u slučaju skupa podataka na kojem se temelji ovaj rad, ciljnu varijablu činilo je puno više *nechurnera* od *churnera.*

Slika 5.2 Nebalansirana ciljna varijabla

Konstrukcija prediktivnog modela nad takvim podacima dala bi iskrivljen rezultat.

5.1.1. Naduzorkovanje

Problem nebalansiranosti klasa (engl. *class imbalance problem*, CIP) velika je prepreka u prediktivnom modeliranju i obično se rješava naduzorkovanjem - dodavanjem zapisa radi usklađivanja s brojnijom klasom, ili poduzorkovanjem – smanjenjem broja zapisa radi usklađivanja s klasom s manje zapisa. Skup podataka korišten u ovom radu bio je vrlo nebalansiran: sadržavao je zapise o 4766 *nechurnera* i samo 1869 *churnera,* pa smo prije modeliranja izbalansirali broj zapisa u klasama ciljne varijable metodom naduzorkovanja (engl. *oversampling*) s pomoću tehnike SMOTE (*Synthetic Minority Oversampling Technique*, tehnika umjetnog naduzorkovanja manjinske klase) koja je dostupna u biblioteci imbalanced_learn. Tehnika SMOTE se u usporednom testu šest tehnika za naduzorkovanje pokazala drugom najboljom u uklanjanju problema nebalansiranosti klasa, odmah iza genetskog algoritma utemeljenog na tehnici MTDF (Amin et al., 2016).

Nakon primjene tog algoritma u skupu podataka bilo je 4766 *nechurnera* i 4766 *churnera*, čime je riješen problem nebalansiranosti.

```
[ ] X = os_df.drop('Churned', axis = 'columns')
    y = os_df['Churned']

[ ] y.value_counts()

    0    4766
    1    1869
    Name: Churned, dtype: int64

[ ] from imblearn.over_sampling import SMOTE

    smote = SMOTE(sampling_strategy = 'minority')
    X_sm, y_sm = smote.fit_resample(X, y)

    y_sm.value_counts()

    0    4766
    1    4766
    Name: Churned, dtype: int64
```

Slika 5.3 Balansiranje skupa podataka s pomoću tehnike SMOTE i krajnji rezultat

5.1.2. Stvaranje *dummy* varijabli

Prije modeliranja trebalo je i sve kategoričke atribute pretvoriti u *dummy* varijable. To je već rađeno u sklopu analize relevantnosti atributa, a sada je napravljeno na podskupovima podataka za modeliranje radi uspješnije konstrukcije klasifikacijskih modela.

5.1.3. Skaliranje

Velik dio atributa u skupu podataka dosad je poprimio vrijednosti 1 ili 0, a ostali su bili numeričke varijable cjelobrojnog ili decimalnog tipa. No, primjerice, ukupan prihod od korisnika (*Total Revenue*) izražen je u stotinama i tisućama dolara, dok su u drugim stupcima vrijednosti 1 i 0. Takav nesrazmjer veličina negativno bi utjecao na konstrukciju modela, pa je zato provedeno skaliranje. Rezultat je skup podataka sa svim vrijednostima skaliranim između 0 i 1.

	Married	NumberOfDependents	NumberOfReferrals	TenureInMonths	InternetService	OnlineSecurity	PremiumTechSupport	UnlimitedData	PaperlessBilling
0	1	0.0	0.181818	0.112676	1	0	1	1	1
1	0	0.0	0.000000	0.112676	1	0	0	0	0
2	0	0.0	0.000000	0.042254	1	0	0	1	1
3	1	0.0	0.090909	0.169014	1	0	0	1	1
4	1	0.0	0.272727	0.028169	1	0	1	1	1

Slika 5.4 Skup podataka nakon skaliranja

Time su uklonjene sve karakteristike skupa podataka koje su mogle prouzročiti probleme pri modeliranju, podaci su podijeljeni na skup za treniranje i skup za testiranje, pa se može pristupiti izradi modelâ za predviđanje prekida ugovornih odnosa.

5.2. Razvoj modela utemeljenog na logističkoj regresiji

Model utemeljen na logističkoj regresiji predviđa zavisnu varijablu u podacima analizirajući odnose između nezavisnih varijabli. Logistička regresija prikladna je za binarne atribute, a upravo je takav podatak o prekidu ugovornog odnosa jer može imati dvije vrijednosti. Stoga se logistička regresija smatra dobrim algoritmom za predviđanje *churna* (Lazarov et al., 2010).

Tradicionalni model za predviđanje prekida ugovornih odnosa utemeljen na logističkoj regresiji izračunava vjerojatnost da će pretplatnik *churnati* u sljedećih nekoliko mjeseci. Da bi se konstruirao takav model, upotrijebljena je scikitlearn biblioteka LogisticRegression.

Praktično izvješće o klasifikaciji (engl. *classification report*) pokazalo je odmah da je **točnost** (engl. *accuracy*) tako dobivenog modela 84 %, što znači da je toliki udio točno klasificiranih korisnika u skupu svih korisnika (Dalbelo Bašić et al, 2011). Točnost se smatra najintuitivnijom mjerom za vrednovanje klasifikatora i računa se po sljedećoj formuli:

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{4}$$

pri čemu su TP, TN, FP i FN poznati pojmovi iz matrice zabune (engl. *confusion matrix*). Točnost je testirana još jednom metodom za izračun točnosti iz scikitlearnova inventara te je potvrđena točnost logističkoregresijskog modela od 84 %.

```
[66] import sklearn.metrics as sm

     accuracy = sm.accuracy_score(y_test, y_pred)
     accuracy

     0.8484530676455165
```

Slika 5.5 Rezultat točnosti i K-S testa za logističkoregresijski model

Spomenimo i druge mjere za vrednovanje klasifikatora, također vidljive na izvješću o klasifikaciji dobivenom u kodu. **Preciznost** (engl. *precision*) je udio točno klasificiranih primjera u skupu pozitivno klasificiranih primjera.

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{5}$$

Odziv (engl. *recall*) je udio točno klasificiranih primjera u skupu svih pozitivnih primjera. Odziv se još naziva i osjetljivost (engl. *sensitivity*).

$$\mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{6}$$

Specifičnost (engl. *specificity*) je udio točno klasificiranih primjera u skupu svih negativnih primjera.

$$\text{Specifičnost} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{7}$$

I na kraju, **f-mjera** (engl. *f-1 score*) je harmonijska sredina preciznosti i odziva, što znači da u jednoj metrici simetrično prikazuje preciznost i odziv.

$$\mathrm{F} = \frac{2\mathrm{PR}}{P + R} \tag{8}$$

Harmonijska sredina se koristi jer je stroža od aritmetičke i geometrijske (Dalbelo Bašić et al, 2011).

```
[27] from sklearn.linear_model import LogisticRegression

[29] lg_model = LogisticRegression(max_iter=200)
     lg_model.fit(X_train, y_train)

     LogisticRegression(max_iter=200)

[30] y_pred = lg_model.predict(X_test)
     print(classification_report(y_test, y_pred))

                  precision    recall  f1-score   support

               0       0.87      0.79      0.83       954
               1       0.81      0.88      0.84       953

        accuracy                           0.84      1907
       macro avg       0.84      0.84      0.84      1907
    weighted avg       0.84      0.84      0.84      1907
```

Slika 5.6 Izrada modela utemeljenog na logističkoj regresiji i izvješće o klasifikaciji

Sada će se protumačiti ostali dobiveni rezultati izvješća o klasifikaciji za model utemeljen na logističkoj regresiji:

- **Preciznost:** od svih korisnika za koje je model predvidio da će biti *churneri*, 81 % je doista *churnalo*.

- **Odziv:** od svih korisnika koji su zbilja *churnali*, model je predvidio taj ishod za 88 % njih.
- **F-mjera** iznosi 0,84 što je relativno blizu 1, a to znači da je model utemeljen na logističkoj regresiji dobar u predviđanju tko će *churnati*.

5.3. Razvoj modela utemeljenog na naivnom Bayesovom teoremu

Za razvoj tog modela korišten je naivni Bayesov algoritam gausovskog tipa koji pretpostavlja da svaki razred koristi Gaussovu distribuciju. U skladu s Bayesovim teoremom, koji pripada u područje vjerojatnosti, naivni Bayesov algoritam pretpostavlja nezavisnost atributa i time je puno jednostavniji od Bayesovih mreža. One imaju vrlo složenu strukturu i čine ih izravni neciklički grafovi i tablice uvjetnih vjerojatnosti.

Izradili smo model utemeljen na naivnom Bayesu s pomoću scikitlearn biblioteke GaussianNB i on je u izvješću o klasifikaciji pokazao **točnost** od 80 %. Time je po udjelu točno klasificiranih korisnika lošiji od modela utemeljenog na logističkoj regresiji.

```
[158] from sklearn.naive_bayes import GaussianNB

[159] nb_model = GaussianNB()
      nb_model = nb_model.fit(X_train, y_train)

[160] y_pred = nb_model.predict(X_test)
      print(classification_report(y_test, y_pred))
```

	precision	recall	f1-score	support
0	0.81	0.79	0.80	954
1	0.80	0.82	0.81	953
accuracy			0.80	1907
macro avg	0.81	0.80	0.80	1907
weighted avg	0.81	0.80	0.80	1907

Slika 5.7 Model utemeljen na naivnom Bayesu

Evo i tumačenja preostalih rezultata izvješća o klasifikaciji za model utemeljen na naivnom Bayesu:

- **Preciznost:** od svih korisnika za koje je model predvidio da će biti *churneri*, 80 % je doista *churnalo*.
- **Odziv:** od svih korisnika koji su zbilja *churnali*, model je predvidio taj ishod za 82 % njih.

- **F-mjera** iznosi 0,81 što je relativno blizu 1, a to znači da je model utemeljen na naivnom Bayesovom teoremu dobar u predviđanju tko će *churnati*, ali slabiji od logističke regresije.

5.4. Razvoj modela utemeljenog na neuronskoj mreži

U matematičkom modelu neuronske mreže osnovna jedinica dizajnirana je po uzoru na biološki neuron (Klepac, 2019). Slično kao povezanost bioloških neurona, i neuronska mreža sastoji se od neurona koji formiraju slojeve tako da postoji ulazni sloj, skriveni slojevi i izlazni sloj. Osnovna je ideja iza neuronskih mreža da je svaki atribut povezan s određenim ponderom (engl. *weight*), a kombinacije ponderiranih atributa sudjeluju u zadatku predviđanja. Tijekom ciklusa učenja ti se ponderi stalno ažuriraju i tako korigiraju utjecaj atributa.

U razvoju modela utemeljenog na neuronskoj mreži korišten je početni sloj s 32 čvora i aktivacijskom funkcijom *relu* te izlaznom funkcijom *sigmoid*. Učenje se odvijalo kroz 100 ciklusa ili epoha. U prvoj iteraciji krajnja točnost iznosila je iznimno visokih 90 %. Gubitak (engl. *loss*) je, očekivano, kroz cikluse učenja stalno padao, a točnost rasla: s 84 % u prvoj epohi, preko 88 % u pedesetoj, do 90 % u zadnjoj, stotoj epohi.

Zbog tako visokog rezultata odlučeno je da se u sljedećoj iteraciji predviđanja s pomoću neuronske mreže upotrijebi tzv. *dropout*-sloj koji nasumce isključuje neke čvorove pri ciklusima učenja i time simulira različite arhitekture, unoseći parametar vjerojatnosti predviđanja koji je u ovom slučaju bio postavljen na 0,2 – sve to radi sprječavanja prenaučenosti (engl. *overfitting*). Rezultat koji je dobiven s *dropout*-slojem ponešto se razlikovao: u prvoj epohi točnost je bila 76 %, u pedesetoj 87 % i u zadnjoj, stotoj 88 %.

```
[ ] import tensorflow as tf
```

```
[ ] nn_model = tf.keras.Sequential([
        tf.keras.layers.Dense(32, activation='relu', input_shape=(29,)),
        tf.keras.layers.Dropout(0.2),
        tf.keras.layers.Dense(32, activation='relu'),
        tf.keras.layers.Dense(1, activation='sigmoid')
    ])

    nn_model.compile(optimizer=tf.keras.optimizers.Adam(0.001), loss='binary_crossentropy', metrics=['accuracy'])
```

```
[ ] nn_model.fit(
        X_train, y_train, epochs=100, batch_size=32
    )
```

Slika 5.8 Model utemeljen na neuronskoj mreži

Nakon treniranja neuronske mreže napravljeno je predviđanje s pomoću nje. Dobivena točnost je iznosila 87 % uz vrlo sličnu preciznost i odziv. Time se model utemeljen na neuronskoj mreži pokazao najboljim u predviđanju.

```
[107] print(classification_report(y_test, y_pred))
```

	precision	recall	f1-score	support
0	0.88	0.86	0.87	954
1	0.87	0.88	0.87	953
accuracy			0.87	1907
macro avg	0.87	0.87	0.87	1907
weighted avg	0.87	0.87	0.87	1907

Slika 5.9 Izvješće o klasifikaciji za neuronsku mrežu

Evo i tumačenja preostalih rezultata izvješća o klasifikaciji za model utemeljen na neuronskoj mreži:

- **Preciznost:** od svih korisnika za koje je model predvidio da će biti *churneri*, 87 % je doista *churnalo*.
- **Odziv:** od svih korisnika koji su zbilja *churnali*, model je predvidio taj ishod za 88 % njih.
- **F-mjera** iznosi 0,87 što je blizu 1, a to znači da je model utemeljen na neuronskoj mreži u predviđanju tko će *churnati* bolji naivnog Bayesa, ali i logističke regresije.

5.5. Određivanje prediktivne moći modelâ

Nakon izrade i implementacije svih triju modela može se zaključiti da se najboljim u predviđanju pokazao model utemeljen na neuronskoj mreži s 87-postotnom točnosti uz upotrebu *dropout*-sloja koji sprječava prenaučenost. Bez upotrebe tog sloja točnost modela utemeljenog na neuronskoj mreži na podskupu podataka za treniranje iznosila je čak 90 %.

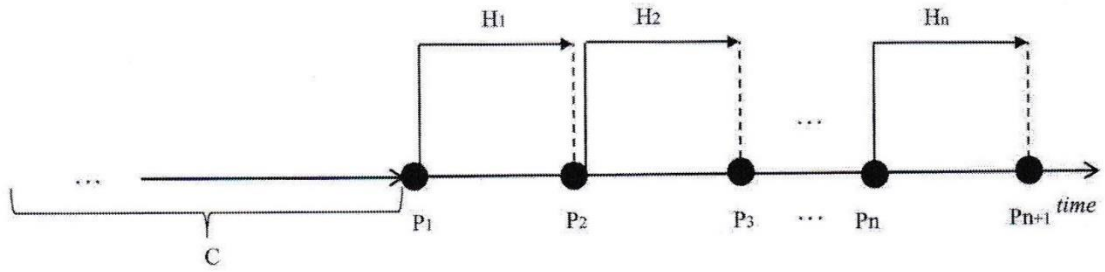
	Logistička regresija	*Naivni Bayes*	*Neuronska mreža*
Točnost	84 %	80 %	87 %
Preciznost	81 %	80 %	87 %
Odziv (osjetljivost)	88 %	82 %	88 %
F-mjera	84 %	81 %	87 %

Tablica 5.2 Usporedba glavnih metrika triju modela

Model utemeljen na Bayesu predviđao je najslabije s točnosti od 80 %, dok je model utemeljen na logističkoj regresiji imao točnost od 84 %. Može se reći da su se sva tri modela pokazala vrlo dobrima u predviđanju prekida ugovornih odnosa, što se vjerojatno dijelom može pripisati i implementaciji analize relevantnosti atributa prije modeliranja.

5.6. Primjena razvijenog prediktivnog modela

Već je u jednom od prethodnih poglavlja objašnjeno kako se konstruira uzorak za modeliranje predviđanja prekida ugovornih odnosa, a sada će se objasniti kako se konstruirani prediktivni model primjenjuje. Upravo to prikazuje shema u nastavku.



Slika 5.10 Shema predviđanja *churna* s pomoću razvijenog modela (Klepac et al., 2015)

Tijekom vremenskog razdoblja C praćeno je ponašanje korisnika i razvijene bihevioralne varijable, dok je u razdoblju H_1 praćen ishod, odnosno reakcije korisnika u vezi s ciljnom varijablom prekida ugovornog odnosa: je li korisnik *churnao* ili ne. Na temelju svega toga razvijen je model za predviđanje prekida ugovornih odnosa.

U narednim razdobljima H_2, H_3... H_n model se iterativno primjenjuje i predviđaju *churneri* za naredne mjesece. Moguće je da prediktivnost modela s vremenom oslabi uslijed promjene uvjeta, primjerice ako telekom odluči povećati cijene pa to počne znatno utjecati na sentiment pretplatnika. Stoga je model potrebno redovito evaluirati i prilagođavati.

6. Dokazivanje prediktivnosti

Osim točnošću, preciznošću, odzivom i f-mjerom, performanse modela dokazuju se i dvama testovima: Kolmogorov-Smirnovljevim testom ili K-S testom te ROC krivuljom i površinom ispod nje (ROC AUC, engl. *area under the curve*). U nastavku je pokazano kakvi su bili rezultati K-S testa za model utemeljen na logističkoj regresiji te ROC i AUC testova za sva tri prethodno razvijena modela.

6.1. K-S statistika i K-S test

Kolmogorov-Smirnovljev (K-S) test dvaju uzoraka mjeri diskriminatornu snagu modela: koliko dobro model može razlikovati događaje od nedogađaja. To radi na temelju usporedbe kumulativnih distribucija opažanja iz dvaju skupova podataka. Postavljaju se dvije hipoteze: nulta hipoteza (H_o) da su vrijednosti obje grupe uzete iz populacija s identičnom distribucijom i alternativna hipoteza (H_a) da su dva skupa podataka iz različitih kontinuiranih distribucija.

Rezultati K-S testa su D i p-vrijednost: D je maksimalna razlika između dviju kumulativnih distribucija, dok se p-vrijednost računa na temelju vrijednosti D i veličina uzoraka. Upravo je p-vrijednost ta koja daje odgovor na pitanje koja hipoteza vrijedi: ako je p-vrijednost mala (5 % ili manje – obično se ta brojka stavlja kao prag statističke značajnosti), nultu hipotezu treba odbaciti i vrijedi alternativna hipoteza – da vrijednosti iz dvaju skupova podataka dolaze iz različitih distribucija. Ili pak ako je p-vrijednost veća od praga, nultu hipotezu treba prihvatiti jer vrijednosti dolaze iz iste distribucije (Teegavarapu, 2019).

$$H_0: \mathrm{P} = P_0, H_1: \mathrm{P} \neq P_0 \tag{9}$$

pri čemu je P distribucija događaja, a P_0 distribucija nedogađaja.

Na temelju podataka dobivenih logističkoregresijskim modelom konstruirana je K-S tablica u kojoj su uspoređene kumulativne distribucije loših i dobrih ishoda (*churnera* i *non-churnera*) kroz populaciju podijeljenu na deset skupina ili binova prema vjerojatnosti *churna* prediktivnog modela, koju smo izračunali s pomoću metode *.predict_proba()*. Zašto podjela na binove? Da se izdvoje korisnici s najvećom vjerojatnosti *churna* kako bi ih se moglo *targetirati* kampanjama za zadržavanje.

probbands	min_prob	max_prob	FREQ	churn	nchurn	% Churner	% NonChurner	% Total	G/B odds	Bad rate	Cum goods	Cum bads	% cum_bads	% cum_goods	% KS	Base line	Cumulative bad rate
1	0.9269	0.9911	190	185	5	19.41%	0.52%	9.96%	0.027027	97.37%	5	185	19.41	0.52	18.89	49.973781	97.368421
2	0.8931	0.9267	191	182	9	19.10%	0.94%	10.02%	0.049451	95.29%	14	367	38.51	1.47	37.04	49.973781	96.325459
3	0.8362	0.8924	191	167	24	17.52%	2.52%	10.02%	0.143713	87.43%	38	534	56.03	3.98	52.05	49.973781	93.356643
4	0.7462	0.8355	191	144	47	15.11%	4.93%	10.02%	0.326389	75.39%	85	678	71.14	8.91	62.23	49.973781	88.859764
5	0.5858	0.7458	190	122	68	12.80%	7.13%	9.96%	0.557377	64.21%	153	800	83.95	16.04	67.91	49.973781	83.945435
6	0.3516	0.5855	190	81	109	8.50%	11.43%	9.96%	1.345679	42.63%	262	881	92.44	27.46	64.98	49.973781	77.077865
7	0.1498	0.3506	192	49	143	5.14%	14.99%	10.07%	2.918367	25.52%	405	930	97.59	42.45	55.14	49.973781	69.662921
8	0.0384	0.1487	190	18	172	1.89%	18.03%	9.96%	9.555556	9.47%	577	948	99.48	60.48	39.00	49.973781	62.163934
9	0.0073	0.0382	191	5	186	0.52%	19.50%	10.02%	37.200000	2.62%	763	953	100.00	79.98	20.02	49.973781	55.536131
10	0.0000	0.0072	191	0	191	0.00%	20.02%	10.02%	inf	0.0%	954	953	100.00	100.00	0.00	49.973781	49.973781

Slika 6.1 K-S tablica - U stupcu „% KS“ izračunate su vrijednosti D za vjerojatnosne binove

Za početak su predviđene vjerojatnosti podijeljene na 10 dijelova, odnosno decila. Nakon toga uslijedilo je računanje kumulativnog postotka događaja i nedogađaja u svakom decilu i provjera gdje je njihova razlika maksimalna – i to je upravo statistička mjera D.

U stupcu „% KS“ nalazi se ta razlika između postotka kumulativnih događaja (*churneri*, *% cum_bads*) i postotka kumulativnih nedogađaja (*necuhrneri, % cum_goods*). Za potvrdu kvalitete modela (engl. *goodness of fit*) očekuje se da najveća razlika ne prelazi 70 % te da ona bude u prva tri decila. Vidljivo je da distribucije za model utemeljen na logističkoj regresiji zadovoljavaju prvi uvjet – najveća razlika, tj. D iznosi 68 %, no drugi uvjet nije ispunjen. Najveće su razlike u 4., 5. i 6. decilu.

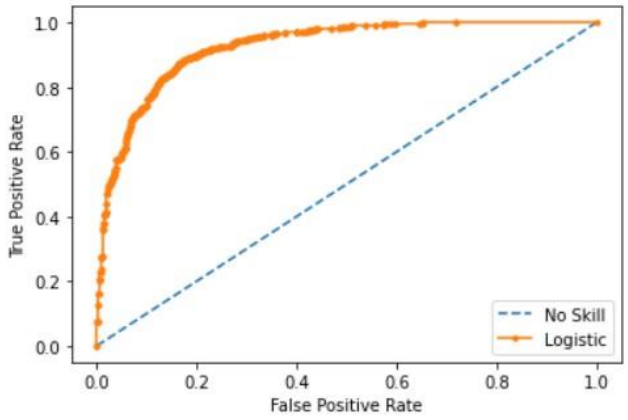
Postoji još jedan uvjet validacije modela koji se može iščitati iz K-S tablice, a to je **redoslijed rangiranja** (engl. *rank ordering*) i on je vidljiv u stupcima s postotkom događaja u pojedinim decilima: te vrijednosti bi se trebale monotono smanjivati. Logističkoregresijski model zadovoljava taj uvjet, kako je vidljivo u stupcu „% Churner“. Suprotno vrijedi za „NonChurner“ – tu se vrijednosti monotono povećavaju i to je dobro i očekivano.

Praktična vrijednost K-S statistike je u podjeli vjerojatnosti obuhvata *churnera* . Telekom bi kampanjama za zadržavanje korisnika trebao ciljati upravo na prvu skupinu jer tu ima najviše korisnika s velikom vjerojatnošću prekida ugovornih odnosa. Kako binovi imaju podjednak broj korisnika, to bi značilo da targetiraju 10 % korisnika, dok u stvarnosti telekomi, zbog resursa i troškova, kampanjama zadržavanja obično targetiraju i manje - primjerice 3 % najvjerojatnijih *churnera*.

6.2. Mjerenje prediktivne snage s pomoću ROC krivulje

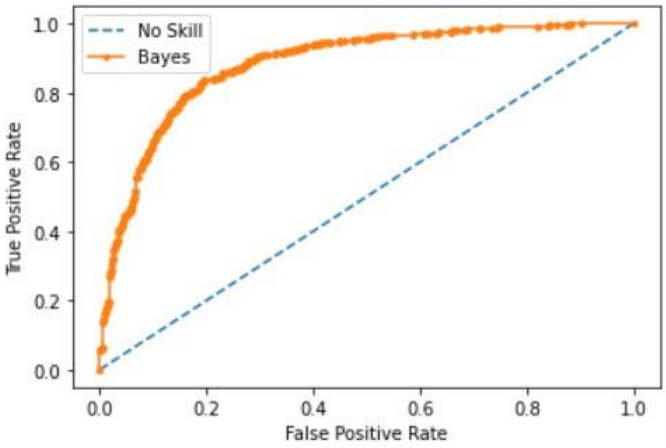
Nakon provjere prediktivne snage s pomoću K-S tablice, provjerena je i prediktivnost triju konstruiranih modela s pomoću ROC krivulje i površine ispod nje (AUC, engl. *area under curve*). ROC krivulja (engl. *Receiver Operating Characteristic*, radna karakteristika primatelja) prikazuje odnos specifičnosti i osjetljivosti klasifikatora poput modela razvijenih tijekom izrade ovog rada. ROC AUC površina je ispod ROC krivulje – što je ta površina veća, to je prediktivna snaga klasifikatora bolja.

Za model utemeljen na logističkoj regresiji izračunata je **ROC AUC vrijednost 0,926**, dok se iscrtana krivulja znatno razlikovala od one koja prikazuje performanse nasumičnog klasifikatora, što svjedoči o njegovoj snazi.



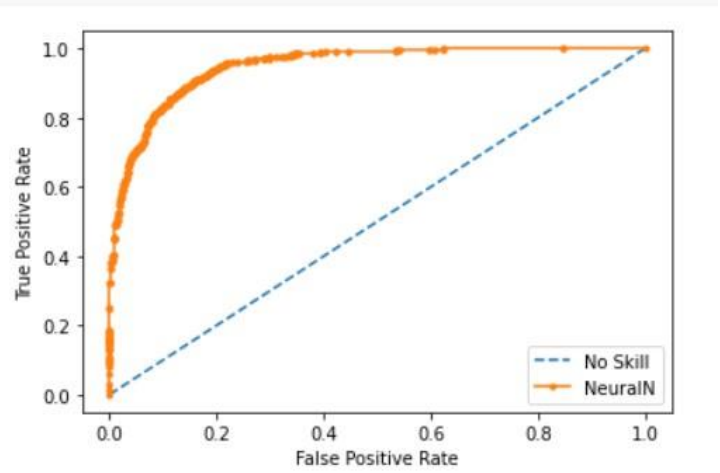
Slika 6.2 ROC krivulja za model utemeljen na logističkoj regresiji

Model utemeljen na naivnom Bayesovom teoremu imao je **ROC AUC vrijednost 0,885**, što je manje od modela s logističkom regresijom. Iz toga se može zaključiti da potonji ima bolje performanse prilikom klasifikacije, odnosno ima veću predikivnu snagu.



Slika 6.3 ROC krivulja za model utemeljen na Bayesu

I na posljetku, za model utemeljen na neuronskoj mreži izračunata je **ROC AUC vrijednost 0,949**, što je vrlo dobar rezultat i ukazuje na bolje performanse od obaju prethodno spomenutih modela. I sam oblik krivulje – njeno asimptotsko približavanje osima – pokazuje da se radi o najboljem klasifikatoru među izrađenim modelima.



Slika 6.4 ROC krivulja za model utemeljen na neuronskoj mreži

6.3. Odabir najboljeg modela i prijedlog strategije zadržavanja korisnika

Prema rezultatima testiranja performansi prvo s pomoću jednostavnog izvješća o klasifikaciji kroz mjere točnosti, preciznosti i odziva, a zatim i s pomoću ROC krivulje i ROC AUC-a utvrđeno je da najveću prediktivnu snagu ima model utemeljen na neuronskoj mreži. U literaturi se navodi superiornost neuronskih mreža u predviđanju *churna*, no također i njihove mane: neuronske mreže ne pokazuju uzorke u lako razumljivom obliku, pa ih se često karakterizira kao svojevrsne „crne kutije“. Uz to, potrebni su im veliki skupovi podataka i puno vremena za izračun kad se konfiguriraju u iole složenijim strukturama. No, njihova točnost predviđanja to opravdava (Lazarov, 2010).

S druge strane, model s logističkom regresijom u ovdje korištenom primjeru nije puno zaostajao u točnosti od modela utemeljenog na neuronskoj mreži, a puno je manje zahtjevan što se tiče resursâ, pa bi vjerojatno bilo najbolje koristiti njega za predviđanje prekida ugovornih odnosa – ovisno o okolnostima.

No, sâm prediktivni model – ma koliko dobar bio – nije dovoljan za sustavno sprječavanje prekida ugovornih odnosa. Pri formuliranju strategija treba voditi računa o ostalim elementima sveobuhvatnog rješenja, kao što su praćenje prospektivnih korisnika, implementacija inputa domenskih stručnjaka i slično.

Stoga bi jedna od mogućih strategija zadržavanja korisnika bila izrada modela za predviđanje prekida ugovornih odnosa na temelju logističke regresije za kratkoročno

sprječavanje *churna*, ali bi svakako bilo dobro planirati i izraditi i model utemeljen na analizi preživljenja za potrebe dugoročnog praćenja i smanjivanja broja prekida ugovornih odnosa. Tijekom razvoja tih modela blisko bi se surađivalo s domenskim stručnjacima, pogotovo na konstrukciji izvedenih varijabli, koje bi se stalno evaluirale tehnikama selekcije atributa i poboljšavale.

Uz to, bilo bi korisno na razini tvrtke osmisliti i koristiti *fuzzy* ekspertni sustav za određivanje vrijednosti kupaca na čijoj bi se izradi angažirali domenski stručnjaci i svojim inputima povećali klasifikacijsku snagu modela. Svi ti modeli u sinergiji omogućili bi adekvatno kratkoročno i dugoročno praćenje *churna* radi njegova pravovremena sprječavanja, uz redovitu evaluaciju i prilagodbu prema potrebama poslovne situacije.

Zaključak

Prekid ugovornih odnosa kao metrika posebno je važan kod zrelih djelatnosti koje su trenutno iza razdoblja eksponencijalnog rasta, a telekomunikacije su izvrstan primjer. To je djelatnost u kojoj je u većini scenarija lako prepoznati početak i kraj odnosa između tvrtke i korisnika, a uz to je i česta njihova interakcija putem različitih kanala, no također i interakcija korisnika s *telco* uslugama i proizvodima. Stoga su telekomunikacije vrelo izvedenih varijabli, a time i „zlatni rudnik" prediktivnog modeliranja prekida ugovornih odnosa.

U ovom je radu razmatrana izrada sustava/rješenja za sprječavanje prekida ugovornih odnosa koje osim ranog prepoznavanja i predviđanja *churna* treba uključivati i njegovo sprječavanje. Nekome neupućenome tko sa strane gleda problematiku prekida ugovornih odnosa moglo bi se činiti da u razvoju takvog rješenja glavnu riječ ima analitičar za rudarenje podataka, odnosno prediktivni model koji on izrađuje, ali kako je višestruko pokazano u korištenoj literaturi (Klepac et al., 2015), to je mit.

Za razvoj kvalitetnog rješenja za sprječavanje prekida ugovornih odnosa nužan je input domenskih stručnjaka, i to u raznim fazama razvoja rješenja. Taj je uvid potvrđen u ovom radu na primjeru razvoja prediktivnog modela za *churn* utemeljenog na javno dostupnom skupu podataka.

Kao prvo, pokazano je već u fazi konstrukcije uzorka, kad je potrebno procijeniti koji je trenutak najbolji za početak promatranja, koji vremenski raspon koristiti itd., da je taj uvid točan. Zatim je u fazi pripreme i čišćenja uzorka podataka na kojemu će se temeljiti prediktivni model bila primijenjena poslovna logika – primjerice, iz skupa je izbačen dio korisnika sa statusom *Joined* koji su počeli koristiti usluge tog telekoma prije tri mjeseca ili manje, a imaju ugovor koji se obnavlja iz mjeseca u mjesec jer je domensko znanje (preuzeto iz prvotnih vizualizacija) pokazalo da što je korisnik kraće s kompanijom, to je veća vjerojatnost da će *churnati,* a isto vrijedi i za korisnike s mjesečnim obnavljajućim ugovorom, za razliku od ugovora sklopljenog na godinu ili dvije. To je bila svojevrsna simulacija inputa domenskog stručnjaka čija je svrha pravilno usmjeriti izradu prediktivnog modela.

Nadalje, skup podataka je uz sociodemografske podatke o korisnicima sadržavao i cijeli niz bihevioralnih varijabli, a to su one koje govore o korisnikovu ponašanju u kontekstu

upotrebe telekomovih proizvoda i usluga. Upravo na takvim, bihevioralnim ili izvedenim varijablama počivaju prediktivni modeli i one nerijetko imaju najveću prediktivnu snagu.

Nad svim je podacima zatim primijenjena jedna od tehnika odabira značajki za modeliranje, a to je analiza relevantnosti atributa preko mjera vrijednosti informacije (IV) i težine dokaza (WoE). Kroz korake te tehnike izračunato je koji atributi imaju dobru prediktivnu snagu te su upravo oni korišteni za konstrukciju prediktivnog modela. Analiza relevantnosti atributa omogućila je i uvid u karakteristike tipičnog *churnera*, odnosno profiliranje. Sve su to postupci koji približavaju telekomunikacijsku tvrtku odgovorima na pitanja tko će najvjerojatnije prekinuti ugovorni odnos i zašto – sve u svrhu preventivnog djelovanja i pokušaja zadržavanja korisnika.

Za izradu modela korišteni su algoritmi logističke regresije, naivnog Bayesova teorema i neuronska mreža. Sva tri modela pokazala su vrlo dobre rezultate i visoku točnost u predviđanju prekida ugovornih odnosa, no najboljim se pokazao model utemeljen na neuronskoj mreži s čak 87 % točnosti, što je potvrđeno K-S testom, mjerom ROC AUC i ROC krivuljom.

I time je zatvoren zadani opseg ovog rada. U stvarnom životu, da bi konstruirano rješenje bilo optimalno kad se radi o troškovima sprječavanja *churna* koji će uslijediti, podatkovno modeliranje trebalo bi komplementirati određivanjem vrijednosti korisnika tijekom njegova vijeka upotrebe usluga telekoma, za što se najčešće koriste tehnike kao što su *fuzzy* ekspertni sustavi i samoorganizirajuće mape.

Izrada ovog rada bilo je zanimljivo putovanje kroz svijet predviđanja prekida ugovornih odnosa, koje mi je potvrdilo da sam odabrala ne samo pravu temu završnog rada, nego i šire područje profesionalnog djelovanja.

Popis kratica

CR	*Churn Rate*	količina prekinutih ugovornih odnosa
ARA	*Attribute Relevance Analysis*	analiza relevantnosti atributa
IV	*Information Value*	vrijednost informacije
WoE	*Weight of Evidence*	težina dokaza
CIP	*Class Imbalance Problem*	problem neuravnoteženosti klase
AUC	*Area Under the Curve*	površina ispod ROC krivulje
ROC	*Receiver Operating Characteristic*	radna karakteristika primatelja

Popis slika

Popis tablica

Literatura

[1] KLEPAC, G., KOPAL, R., MRŠIĆ, L. *Developing Churn Models Using Data Mining Techniques and Social Network Analysis*. Hershey: Information Science Reference, 2015.

[2] BERRY, M., LINOFF, G. *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management* (2nd ed). Indianapolis: John Wiley, 2014.

[3] LAZAROV, V., CAPOTA, M. *Churn Prediction*. http://www.vladislav.lazarov.pro/files/research/papers/churn-prediction.pdf (2010), preuzeto 10. veljače 2023.

[4] CFI TEAM. *Market Saturation.* https://corporatefinanceinstitute.com/resources/economics/market-saturation/ (siječanj 2023), pristupljeno 16. veljače 2023.

[5] KLEPAC, G. *Sustavi potpore odlučivanju: priručnik*. Zagreb: Algebra, 2019.

[6] RUIZ, E. Introducing the Maven Churn Challenge. https://www.mavenanalytics.io/blog/maven-churn-challenge (2022), pristupljeno 10. veljače 2023.

[7] ZUANG, S. L. *Telecom Customer Churn Predicition*. https://www.kaggle.com/datasets/shilongzhuang/telecom-customer-churn-by-maven-analytics?select=telecom_customer_churn.csv (2022), preuzeto 6. srpnja 2022.

[8] RADEČIĆ, D. *Attribute Relevance Analysis in Python — IV and WoE*. https://towardsdatascience.com/attribute-relevance-analysis-in-python-iv-and-woe-b5651443fc04 (2019), pristupljeno 10. veljače 2023.

[9] MAHAJAN, V., MISRA R., MAHAJAN, R. *Review of Data Mining Techniques for Churn Prediction in Telecom*. JIOS, vol. 37, no. 2, 2015.

[10] REZA A. S., KEYVAN V. R. *Applying Data Mining to Insurance Customer Churn Management.* IACSIT Hong Kong Conferences, 2012.

[11] DUCA, A. L. *How to balance a dataset in Python.* https://medium.com/towards-data-science/how-to-balance-a-dataset-in-python-36dff9d12704 (2021), pristupljeno 19. veljače 2023.

[12] BROWNLEE, J. *How to Avoid Data Leakage When Performing Data Preparation*. https://machinelearningmastery.com/data-preparation-without-data-leakage/ (2020), pristupljeno 19. veljače 2023.

[13] AMIN, A. *Comparing Oversampling Techniques to Handle the Class Imbalance Problem: A Customer Churn Prediction Case Study.* IEEE Access, volume 4, 2016.

[14] SCHREIBER, J. *Naive Bayes and Bayes Classifiers.* https://github.com/jmschrei/pomegranate/blob/master/tutorials/B_Model_Tutorial_5_Bayes_Classifiers.ipynb (2020), pristupljeno 19. veljače 2023.

[15] DALBELO BAŠIĆ, B., ŠNAJDER, J. *Vrednovanje klasifikatora.* https://www.fer.unizg.hr/_download/repository/SU-12-VrednovanjeKlasifikatora.pdf (2011), preuzeto 20. veljače 2023.

[16] TEEGAVARAPU, R.S.V. *Methods for Analysis of Trends and Changes in Hydroclimatological Time-Series*. Trends and Changes in Hydroclimatic Variables, 2019.

Prilog

USB memorija koja sadrži:

- završni rad u .docx i .pdf obliku
- korišteni skup podataka u .csv obliku
- kôd svih praktičnih dijelova rada u .ipynb obliku

ALGEBRA VISOKO UČILIŠTE

RAZVOJ I USPOREDBA KOMPETITIVNIH MODELA ZA PREDIKCIJU PREKIDA UGOVORNIH ODNOSA S PRIMJENAMA TEHNIKA PROFILIRANJA

Pristupnica: Kristina Lovrić-Matijević, 0130070847

Mentor: prof. dr. sc. Goran Klepac

Datum: 23. 2. 2023.