DOI: 10.5937/bankarstvo2602190V

Datum prijema: 11.05.2026.
Datum prihvatanja: 14.10.2026.

DEFINISANJE PROTOKOLA VEŠTAČKE INTELIGENCIJE U REŠAVANJU NEDOSTAJUĆIH PODATAKA U OBLASTI BANKARSTVA

Vladimir Vasić, redovni profesor, Beogradska bankarska akademija – Fakultet za bankarstvo, osiguranje i finansije
Imejl: vladimir.vasic@bba.edu.rs
ORCID: 0000-0001-9694-8071

Milica Bugarčić, vanredni profesor, Beogradska bankarska akademija – Fakultet za bankarstvo, osiguranje i finansije
Imejl: milica.bugarcic@bba.edu.rs
ORCID: 0000-0002-9327-9965

Rezime: *U ovome radu se na jedan sistematičan i sveobuhvatan način rešava problematika nedostajućih podataka u bankarstvu. O ovom ozbiljnom i značajnom izazovu, veoma se malo pisalo, i to sporadično nad nekim konkretnim poslovnim problemima, ali nikad na ovaj način – a to je kreiranje tj. definisanje protokola. Uključivanjem modela veštačke inteligencije u proces rešavanja nedostajućih podataka, ovaj i dan danas "nerešiv problem" za analitičare podataka se stavlja pod kontrolu, i počinje da predstavlja rutinu, tj. još jedan od obaveznih koraka u pripremi i analizi podataka.*

Ključne reči: veštačka inteligencija, mašinsko učenje, nedostajući podaci, analitika podataka, bankarstvo.

JEL klasifikacija: C14, G21

Uvod

U bankarskom sektoru, u uslovima sve veće konkurencije i oštrije tržišne utakmice, poslovnim bankama je sve teže da održe dobre poslovne performanse i rezultate. Iz tog razloga, okreću se poslovnoj analitici, koja im može ukazati, koje su im slabe tačke u poslovanju, a koje su im razvojne šanse. Integracija veštačke inteligencije i mašinskog učenja u finansijskim uslugama ubrzano transformiše interne operacije i upravljanje rizicima, stavljajući fokus na neophodnost korišćenja naprednih prediktivnih modela nad internim bazama podataka (Godolja i Domi, 2025). Dakle ključ uspeha, ili uopšte opstanka, su kvalitetni rezultati različitih poslovnih analiza. Međutim, za kvalitetan rezultat važi princip GIGO (garbage in, garbage out), to jest u suštini, kvalitet ulaznih podataka za analizu mora biti visok. Međutim, svi koji se bave prikupljanjem i obradom podataka, vrlo brzo se susretnu sa problemom nedostajućih podataka, i tu dolazimo do suštine ovog rada.

Dosadašnji načini rešavanja problematike nekompletnih podataka

U većini slučajeva, istraživači na ovaj ili onaj način otklanjaju problem nekompletnih podataka, a da pri tome uopšte ne naznačavaju koliko je takvih bilo podataka. Verovatno su neprimetno isključeni nekompletni podaci iz skupa podataka koji se analiziraju. Ovakav postupak je donekle razumljiv, jer istraživači, prilikom pisanja stručnih radova, s pravom strahuju da to može predstavljati ozbiljno ograničenje njihovog istraživanja. S razlogom se plaše da bi recenzenti rada mogli odbiti njihov rad, uz obrazloženje da je istraživanje prilično slabo i nepouzdano, usled značajnog prisustva nekompletnih podataka. Prema tome, šanse za objavljivanje rada znatno se povećavaju ukoliko se problem nekompletnih podataka ne iznese na videlo. Kao što su primetili Orchard i Woodbury (1972), najboji način za suočavanje s nekompletnim podacima jeste jednostavno ne imati ih.

Naravno, u takvim radovima, postojanje nekompletnih podataka nigde se ne naznačava. Tada se podrazumevana metoda rešavanja problema nekompletnosti (to jest uklanjanje nekompletnih podataka) primenjuje bez ikakvog pominjanja. Štaviše, tabele u kojima su prikazani rezultati sprovedenih kvantitativnih analiza tada se zasnivaju na različitim veličinama uzoraka.

Razlog zbog kojeg istraživači pribegavaju zastarelim procedurama za rešavanje nekompletnosti podataka delom leži u tome što su im te procedure teorijski jasne. Te procedure svakako obuhvataju brisanje opservacija s nekompletnim podacima, zatim brisanje opservacija samo u slučaju kada imaju nekompletne podatke za promenljive koje se trenutno analiziraju, kao i zamenu nekompletnih podataka aritmetičkom sredinom posmatrane promenljive. Savremene procedure veštačke inteligencije za tretman nekompletnih podataka izuzetno su kompleksne i u većini slučajeva nerazumljive prosečnom istraživaču, te ih stoga izbegavaju.

Protokol veštačke inteligencije

Ukoliko u podacima koji su planirani za analizu postoji prisustvo nekompletnosti, postoji mogućnost da će ocene nepoznatih parametara u kvantitativnoj analizi koja se planira biti pristrasne. Iz tog razloga, od ključne je važnosti pristupiti problemu nekompletnih podataka pažljivo i sveobuhvatno, što bi se, u slobodnijem govoru, moglo nazvati protokolom. U datom protokolu, nulti korak bi predstavljao polaznu tačku u rešavanju problema nekompletnih podataka. U prvom koraku postavilo bi se pitanje:

- Da li posmatrane promenljive, tj. da li u analiziranom skupu podataka postoji prisustvo nekompletnih podataka?

Ukoliko je odgovor da, prelazi se na drugi korak. Ukoliko je odgovor ne, prelazi se direktno na poslednji korak. U drugom koraku, postavlja se pitanje:

- Da li podaci nedostaju potpuno slučajno?

Ukoliko je odgovor pozitivan, prelazi se na treći korak. Ukoliko je odgovor negativan, prelazi se na četvrti korak.

Treći korak predstavlja rešavanje nekompletnih podataka, bilo primenom tradicionalnih tehnika (brisanje opservacija koje sadrže nedostajuće podatke ili umetanje aritmetičke sredine posmatrane promenljive umesto nedostajućih podataka), bilo primenom savremenih metoda. Nakon toga, prelazi se na poslednji korak.

U četvrtom koraku, postavlja se pitanje:

- Da li kvantitativno modelovanje koristi nadgledani ili nenadgledani model veštačke inteligencije?

Ukoliko je odgovor nadgledani, prelazi se na peti korak. Ukoliko je odgovor nenadgledani, prelazi se na osmi korak.

U petom koraku ispituje se:

- Da li je nedostupnost podataka za objašnjavajuće promenljive uzrokovana vrednostima zavisne promenljive?

Ukoliko je odgovor pozitivan, prelazi se na šesti korak. Ukoliko je odgovor negativan, prelazi se na sedmi korak. U šestom koraku konstatuje se da nedostajući podaci to čine prema ne slučajnom mehanizmu. Iz tog razloga, dalja analiza podataka se obustavlja i prelazi na poslednji korak.

U sedmom koraku konstatuje se da podaci koji su nedostupni nedostaju na slučajan način, te da se problem nedostajućih podataka može rešiti isključivo primenom savremenih tehnika. Potom se prelazi na poslednji korak.

U osmom koraku ispituje se:

- Da li nedostupnost podataka za neku promenljivu zavisi od vrednosti sopstvenih podataka, ili od vrednosti podataka ostalih analiziranih promenljivih?

Ukoliko je odgovor od sopstvenih vrednosti, prelazi se na šesti korak. Ukoliko je odgovor od vrednosti ostalih promenljivih, prelazi se na sedmi korak. U poslednjem koraku zaključuje se da je rešavanje problema nekompletnosti podataka završeno.

Tipovi mehanizama

U većini slučajeva, podaci koji treba statistički da se analiziraju imaju, u većoj ili manjoj meri, problem nekompletnosti. Sam taj nedostatak naizgled nije ozbiljan, međutim, ukoliko je mehanizam nekompletnosti određenog tipa, može doći do ozbiljnog problema. Iz tog razloga, od ključne je važnosti uvek ispitati tip mehanizma koji dovodi do nekompletnih podataka. Za sada je definisano da postoje tri tipa mehanizama koji dovode do nekompletnosti podataka (Allison, 2002), a to su: potpuno slučajan, slučajan i neslučajan.

Da bi se mogli opisati dati tipovi mehanizma, promenljiva s nedostajućim podacima označava se sa Y. Zatim, njen raspoloživi deo označava se sa Yo, dok se nedostupni deo označava sa Ym. Sada se može definisati indikatorska promenljiva nedostupnosti, koja uzima vrednosti nula za one opservacije koje su raspoložive za promenljivu Y, i vrednost jedan za one opservacije koje su nedostupne za promenljivu Y. Data indikatorska promenljiva može se označiti sa M. Ukoliko u analiziranom skupu podataka postoje promenljive koje nemaju nedostajuće vrednosti, označavaju se sa X.

Prema tome, kaže se da je mehanizam nedostajućih podataka za promenljivu Y potpuno slučajan ako podaci nedostaju na potpuno slučajan način, te stoga ne zavise ni od dela promenljive Ym, ni od dela promenljive Yo, ni od ostalih prisutnih promenljivih X. Korišćenjem indikatora verovatnoće nedostupnosti, dati mehanizam nedostajućih podataka (Fitzmaurice i sar., 2015) može se iskazati kao:

$$P(M|Y,X) = P(M) \tag{1}$$

Kaže se da je mehanizam nedostajućih podataka za promenljivu Y slučajan ako nedostajući podaci ne zavise od samih nedostajućih vrednosti (Ym). Korišćenjem indikatora verovatnoće nedostupnosti, dati mehanizam može se iskazati kao:

$$P(M|Y,X) = P(M|Y^o,X) \tag{2}$$

Konačno, kaže se da je mehanizam nedostajućih podataka za promenljivu Y neslučajan, ako nedostajući podaci zavise i od vrednosti nedostajućih podataka (Ym). Korišćenjem indikatora nedostupnosti i njegove verovatnoće, dati mehanizam može se prikazati kao:

$$P(M|Y,X) = P(M|Y^o,Y^m,X) \tag{3}$$

Tradicionalni načini tretiranja nekompletnih podataka, kao što su isključivanje opservacije koja nema kompletne podatke ili procenjivanje nedostajuće vrednosti aritmetičkom sredinom date promenljive, prikladni su jedino u slučaju kada je mehanizam nekompletnosti podataka potpuno slučajan. Ukoliko je mehanizam nekompletnosti podataka slučajan, rešavanje problema nekompletnosti podataka je ispravno isključivo primenom savremenih procedura, kao što su: algoritam ekspektacije i maksimizacije (EM algoritam), metoda kompletnih informacija sa maksimalnom verodostojnošću, algoritam specifikacije potpune uslovljenosti, algoritam augmentovanih podataka, kao i ostale savremene procedure. Važno je napomenuti da se navedene procedure mogu koristiti umesto tradicionalnih procedura i u slučaju kada podaci nedostaju potpuno slučajno.

Nadgledani model veštačke inteligencije

Dakle, kada je reč o nadgledanim modelima veštačke inteligencije, priroda nedostajućih podataka utvrđuje se na sledeći način (Tabachnick i Fidell, 2014): najpre se testira da li nedostajući podaci nedostaju na potpuno slučajan način (Little i Rubin, 2002). Ukoliko se data hipoteza odbaci, vrši se naredno testiranje, kojim se testira da li nedostajući podaci nedostaju na slučajan ili neslučajan način. U zavisnosti od odbacivanja ili neodbacivanja date hipoteze, donosi se zaključak da podaci nedostaju na neslučajan ili slučajan način.

U literaturi je već uobičajeno testirati da li nedostajući podaci nedostaju potpuno slučajno korišćenjem Littleovog testa (IBM, 2017). Zatim, ukoliko se testirana nulta hipoteza odbaci, prelazi se na testiranje da li nedostajući podaci nedostaju na slučajan način. Dato testiranje (uz pomoć statističkog softvera) sprovodi se uz pomoć jednodimenzionalnih statističkih t-testova. Suština je ispitati da li raspoloživi podaci predstavljaju slučajan uzorak iz planiranog uzorka (koji je postavljen na početku istraživanja). Ukoliko je to ispunjeno, tada nedostajući podaci imaju istu raspodelu kao i raspoloživi podaci (Enders, 2010).

Jednakost raspodela podrazumeva da grupa raspoloživih podataka i grupa nedostajućih podataka imaju jednake aritmetičke sredine i varijanse. Kim i Bentler (2002) nazvali su tu osobinu homogenošću aritmetičkih sredina i kovarijansnih matrica. Neki autori proveravali su homogenost aritmetičkih sredina i varijansnih komponenti (Thommes i Enders, 2007), dok su drugi autori smatrali da je teško pretpostaviti da su podaci homogeni u pogledu aritmetičke sredine, a heterogeni u pogledu varijanse. Iz tog razloga, oni su ispitivali isključivo homogenost aritmetičkih sredina (IBM, 2017). Prema tome, testiranje homogenosti aritmetičkih sredina sprovedeno je korišćenjem jednodimenzionalnih t-testova.

Model veštačke inteligencije zasnovan na neparametarskoj statistici

Kao što je naznačeno, pri ispitivanju da li podaci nedostaju slučajno, koristi se jednodimenzionalni t-test, kojim se ispituje jednakost aritmetičkih sredina dva nezavisna uzorka (implementiran u većini kvantitativnih softvera koji poseduju modul za analizu nedostajućih podataka). Dva nezavisna uzorka za svaku promenljivu u analizi formiraju se na osnovu nedostupnosti podataka za neku drugu promenljivu u analizi. Naime, ukoliko neka promenljiva (označena sa Xi, i = 1,...,n) ima nedostajuće podatke, za nju se kreira indikatorska promenljiva nedostupnosti (označena sa Mi), koja uzima vrednost 1 ako je dati podatak nedostupan, i vrednost 0 ako je dati podatak raspoloživ. Dati indikator nedostupnosti može se predstaviti u

$$M_i = \begin{cases} 0, ukoliko\ je\ X_i\ raspoloživo \\ 1, ukoliko\ X_i\ nedostaje \end{cases},\ i = 1, \dots, n \qquad (4)$$

Dakle, da bi se ispitalo da li neka promenljiva (Xi) ima nedostajuće podatke na slučajan način, najpre se za datu promenljivu formira indikator nedostupnosti (Mi), a zatim se on koristi za podelu svake ostale promenljive u analizi na dva nezavisna uzorka, nakon čega se primenjuje jednodimenzionalni t-test radi testiranja jednakosti aritmetičkih sredina.

Dati test ima određene nedostatke; na primer, nezavisni uzorci mogu biti veoma neuravnoteženi (Enders, 2010) (u smislu da je uzorak koji predstavlja raspoložive podatke znatno veći od uzorka koji predstavlja nedostajuće podatke). Takođe, nezavisni uzorak koji predstavlja nedostajuće podatke može imati svega par elemenata. Uz to, ono što se može pripisati ograničenju testiranja aritmetičkih sredina dva nezavisna uzorka jeste i činjenica da čak i ako podaci nedostaju neslučajno, dva nezavisna uzorka formirana na osnovu indikatora nedostupnosti mogu imati jednake aritmetičke sredine.

Iz tog razloga, mnogo je bolje predložiti neki drugi način provere da li nedostajući podaci nedostaju slučajno. Upravo to je predmet ovog dela rada – predložiti alternativni način provere. Drugi način testiranja, da li je mehanizam nedostajućih podataka slučajan, mogao bi biti neparametarski test, Mann-Whitney U test dva nezavisna uzorka. Velika prednost ovog neparametarskog testa u poređenju s parametarskim jednodimenzionalnim t-testom jeste u tome što Mann-Whitney test ne testira jednakost aritmetičkih sredina, već upravo ono što je potrebno – da li su dve raspodele jednake. Dakle, nema potrebe za zaobilaznim pristupom kojim bi se sugerisalo da ako su dve raspodele jednake, onda su im jednake i aritmetičke sredine i varijanse; naprotiv, jednakost dve raspodele testira se direktno neparametarskim Mann-Whitney testom dva nezavisna uzorka.

Dalje unapređenje predstavljalo bi korišćenje egzaktne metode za izračunavanje p-vrednosti neparametarskog Mann-Whitney testa, umesto opšteprihvaćene asimptotske metode. Prednost primene egzaktne metode pri računanju p-vrednosti je u tome što (Mehta i Patel, 2017) ona izračunava pouzdane vrednosti čak i u situacijama kada uslovi za primenu t-testa nisu ispunjeni, bez obzira na veličinu nezavisnih uzoraka, kao i na uravnoteženost podataka. Dakle, primena egzaktne metode računanja p-vrednosti u neparametarskom Mann-Whitney testu, pri testiranju da li je mehanizam nedostajućih podataka slučajan, prevazišla bi sve nedostatke i ograničenja t-testa koji se koristi u praksi.

Naravno, pored očiglednih prednosti koje predložena procedura pruža, postoje i određene teškoće. Naime, primena metode egzaktnog računanja p-vrednosti izuzetno je zahtevna, tj. računarski intenzivna, zasnovana na svim mogućim permutacijama realizovanog skupa dva nezavisna uzorka koji treba testirati.
Zbog toga ćemo početi lagano, sa teorijskim objašnjenjem predložene procedure. Dakle, kada je reč o neparametarskom Mann-Whitney testu dva nezavisna uzorka, pretpostaviće se da su dva nezavisna uzorka veličina n1 i n2. Nulta hipoteza koja se testira predstavlja se kao

$$H_0: F_1(x) = F_2(x) \tag{5}$$

što znači da data dva nezavisna uzorka imaju iste raspodele.

U Mann-Whitney testu, originalne realizovane vrednosti zamenjuju se rangiranim vrednostima, tj. vrednost xi zamenjuje se sa rang(xi). Štaviše, radi izračunavanja egzaktne p-vrednosti, neophodno je poznavati egzaktnu verovatnoću realizacije dva nezavisna uzorka s rangiranim vrednostima (Mehta i Patel, 2017). Data verovatnoća zadaje se izrazom

$$h(w) = \frac{\prod_{i=1}^{2} n_i!}{(n_1+n_2)!} \tag{6}$$

gde n1 i n2 predstavljaju veličine prvog i drugog nezavisnog uzorka, a dizajn w predstavlja realizaciju data dva nezavisna uzorka s rangiranim umesto originalnim vrednostima (ili bilo koju permutaciju toga). Nadalje, zakon verovatnoće Mann-Whitney statistike U da uzme egzaktnu vrednost u jeste

$$Pr(U = u) = \sum_{U=u} h(w) \tag{7}$$

gde se sumiranje vrši po svim mogućim permutacijama dizajna w. Egzaktna p-vrednost (i to za desni rep raspodele statistike U) je

$$p_1 = Pr(U \geq u) = \sum_{U \geq u} h(w) \tag{8}$$

Prethodno definisane formule predstavljaju opšte formule za dobijanje egzaktne p-vrednosti za neparametarske testove dva nezavisna uzorka. Naravno, za svaki specifični neparametarski test (dva nezavisna uzorka) postoji poseban numerički algoritam, uz čiju se pomoć računa egzaktna p-vrednost. Time se bavi i nastavak ovog dela.

Stoga će biti prikazana metodologija za egzaktno računanje p-vrednosti za neparametarski Mann-Whitney test dva nezavisna uzorka. Korišćenjem date izračunate p-vrednosti, moguće je testirati nultu hipotezu da nedostajući podaci nedostaju slučajno.

Na početku izračunavanja potrebno je najpre izračunati zbir rangova. To se postiže spajanjem oba uzorka u jedan uzorak, čiji se elementi sortiraju po veličini. Zatim se elementima objedinjenog uzorka dodeljuju rangovi. Ukoliko postoje jednaki elementi, uzima se prosečan rang za njih. Nakon toga (IBM, 2017), izračunavaju se zbirovi rangova za dva nezavisna uzorka (grupa 1 (ili G1) i grupa 2 (ili G2)):

i

$$\begin{aligned} S_1 &= \sum_{i \in G_1 \geq u} rank(x_i; D_1 \cup D_2) \cdot I(x_i \in D_1) \\ S_2 &= \sum_{i \in G_2 \geq u} rank(x_i; D_1 \cup D_2) \cdot I(x_i \in D_2) \end{aligned} \tag{9}$$

gde Di predstavlja opservacije iz i-tog nezavisnog uzorka (i = 1, 2), dok I(·) predstavlja indikator koji uzima vrednost 0 ukoliko izraz u zagradama nije tačan, i vrednost 1 ukoliko izraz u zagradama jeste tačan.

Naredni korak je izračunavanje test statistike kao i odgovarajuće p-vrednosti. Mann-Whitney U test statistika za prvi nezavisni uzorak (ili za grupu 1) zadaje se izrazom

$$\begin{aligned} U &= \sum_{i \in G_1} \sum_{j \in G_2} I(x_i < x_j) + \frac{1}{2} \cdot \sum_{i \in G_1} \sum_{j \in G_2} I(x_i = x_j) \\ &= n_1 \cdot n_2 + \frac{n_1 \cdot (n_1+1)}{2} - S_1 = S_2 - \frac{n_2 \cdot (n_2+1)}{2} \end{aligned} \tag{10}$$

gde n1 i n2 predstavljaju veličine prvog i drugog nezavisnog uzorka.

Nakon definisanja U test statistike, prelazi se na izračunavanje p-vrednosti. Za izračunavanje p-vrednosti, biće potrebna frekvencijska raspodela uzorkovanja test statistike U za vrednost u i za veličine uzoraka i i j za dva nezavisna uzorka (u oznaci fi,j(u)). Nadalje, koristiće se sledeće svojstvo, pomoću kojeg se frekvencijska raspodela test statistike U može izraziti kroz zbir raspodela nižeg reda, kao

$$f_{n_1,n_2}(u) = f_{n_1-1,n_2}(u-n_2) + f_{n_1,n_2-1}(u). \qquad (11)$$

Svaka raspodela nižeg reda je simetrična u odnosu na različite vrednosti test statistike U. Takođe, zbir raspodela nižeg reda daje rezultat koji je takođe simetričan.

Izračunavanje frekvencijske raspodele za test statistiku U (označenu sa fn1,n2(u)) počinje od poznate raspodele test statistike U za i = 1. (Ukoliko se želi, rezultat je isti i kada se počne s raspodelom test statistike U za j = 1.) Zatim se koristi svojstvo zadato izrazom (11), kao i svojstvo simetrije, da bi se izvela frekvencijska raspodela za i = 2. Ova procedura se ponavlja sve dok se ne izvede frekvencijska raspodela za i = n1.

Nakon izračunavanja frekvencijske raspodele za U test statistiku (IBM, 2017), moguće je izračunati (dvostranu) egzaktnu p-vrednost, korišćenjem sledećih obrazaca:

$$p_1 = \begin{cases} \sum_{u=0}^{trunc(U)} f_{n_1,n_2}(u), \text{ukoliko je } U \le \frac{n_1 \cdot n_2 + 1}{2} \\ 1 - \sum_{u=0}^{trunc(U)} f_{n_1,n_2}(u), \text{ukoliko je } U > \frac{n_1 \cdot n_2 + 1}{2} \end{cases}$$

$$p - value = 2 \cdot p_1$$

gde trunc(U) predstavlja celobrojnu vrednost realizovane vrednosti test statistike U.

Podaci banke i poslovna implementacija

Podaci korišćeni u ovom delu predstavljaju odgovore potencijalnih klijenata jedne od vodećih stranih banaka u Republici Srbiji. Istraživanje je sprovedeno u maju 2019. godine u glavnoj poslovnici date banke za grad Novi Sad. Svim potencijalnim klijentima banke koji su se raspitivali o uslovima za dobijanje gotovinskog kredita od strane zaposlenih u poslovnici, ponuđena je anonimna, veoma kratka anketa u vezi s proizvodom gotovinskog kredita. Naime, analitičari odeljenja za upravljanje odnosima s klijentima nameravali su da ispitaju (analiziraju) koji period otplate gotovinskog kredita potencijalne klijente banke najviše zanima. Interesovalo ih je i da li na broj meseci otplate gotovinskog kredita mogu uticati osnovne socioekonomske karakteristike klijenata. Iz tog razloga, sama anketa imala je 6 pitanja, koja glase: „Odgovarajući period otplate gotovinskog kredita" (označeno Y), „Godine starosti (u godinama)" (označeno X1), „Godine stanovanja na trenutnoj adresi" (označeno X2), „Ukupan prihod domaćinstva (u 000 RSD)" koji su logaritamski transformisani (označeno X3), „Godine radnog staža kod trenutnog poslodavca" (u oznaci X4) i „Broj članova domaćinstva" (u oznaci X5).

Za ovo istraživanje, planirana veličina uzorka bila je 200; toliko je i prikupljeno popunjenih upitnika. Međutim, nisu svi klijenti odgovorili na sva pitanja u anketi, te imamo problem nekompletnih podataka. U Tabeli 1 prikazan je broj i procenat nedostajućih podataka, gde su promenljive sortirane u rastućem redosledu u odnosu na procenat nedostajućih podataka.

Tabela 1. Raspoloživost podataka analiziranih promenljivih

		X3	X2	X4	X1	Y	X5
Missing	N	37	29	22	7	7	4
	Percent	18.5%	14.5%	11.0%	3.5%	3.5%	2.0%
Valid N	27.7	163	171	178	193	193	196

Izvor: Rezultat analize autora sproveden uz upotrebu IBM SPSS Statistics 26 softvera.

Može se primetiti da ispitanici u najvećoj meri nisu odgovorili na pitanje označeno X3, i to u iznosu od 18,5%, dok pitanje u anketi, označeno X5, ima najmanji broj nedostajućih odgovora, svega 2%.

Inicijalni korak u statističkoj analizi podataka bio bi ispitivanje prirode nedostajućih podataka. U takvom ispitivanju, prvi korak bi bio testiranje nulte hipoteze, a to je da je mehanizam nekompletnih podataka potpuno slučajan. Kao što je već naznačeno, za testiranje ove hipoteze koristi se Littleov MCAR test. Za sprovođenje ovog testa, kao i svih ostalih statističkih izračunavanja, korišćen je domenski specifičan jezik softvera IBM SPSS Statistics 26 (IBM, 2019). Rezultati sprovedenog testiranja su:

$$\textit{Chi-Square} = 95.465,\ \textit{DF} = 72,\ \textit{Sig.} = 0.034$$

te se nulta hipoteza, da je mehanizam nekompletnosti podataka potpuno slučajan, odbacuje. Iz toga se zaključuje da bi klasično rešavanje problema nedostajućih podataka, kao što je brisanje nekompletnih opservacija ili umetanje aritmetičke sredine, bilo pogrešno.

Tabela 2. Statistike testa i egzaktne p-vrednosti

	Y	X1	X2	X3	X4	X5
Grouping Variable: M_X1						
Mann-Whitney U	549.500	-	281.000	454.000	507.000	661.500
Exact Sig. (2-tailed)	0.935	-	0.227	0.886	0.945	1.000
Grouping Variable: M_X2						
Mann-Whitney U	2074.500	1861.000	-	1384.000	1959.000	2244.500
Exact Sig. (2-tailed)	0.391	0.159	-	0.185	0.749	0.690
Grouping Variable: M_X3						
Mann-Whitney U	2435.000	1860.000	2142.000	-	2169.000	2672.000
Exact Sig. (2-tailed)	0.197	0.001	0.747	-	0.223	0.371
Grouping Variable: M_X4						
Mann-Whitney U	1858.000	1635.000	1366.000	1390.000	-	1344.500
Exact Sig. (2-tailed)	0.927	0.484	0.492	0.842	-	0.073
Grouping Variable: M_X5						
Mann-Whitney U	193.500	296.500	124.000	199.500	99.000	-
Exact Sig. (2-tailed)	0.097	0.476	0.139	0.213	0.319	-

Izvor: Rezultat analize autora sproveden uz upotrebu IBM SPSS Statistics 26 softvera.

Nakon prethodno odbačene hipoteze, prelazi se na testiranje naredne nulte hipoteze, a to je da li je mehanizam nekompletnosti podataka slučajan ili neslučajan. Kao što je predloženo (gde smo ukazali na prednosti predložene procedure u poređenju s postojećom), koristiće se neparametarski Mann-Whitney test s egzaktnim računanjem p-vrednosti, umesto t-testa. Rezultati serije testova prikazani su u Tabeli 2. Za svaku promenljivu koja ima nedostajuće podatke, kreiran je indikator nedostajućih podataka, u oznaci M_X#. Takođe, za svaku promenljivu, testiranje je primenjeno u odnosu na preostalih 5 promenljivih u nadgledanom statističkom modelu. Ovo se naznačava zbog potrebe korekcije alfa-rizika za broj testova po promenljivoj. Dakle, nivo značajnosti umesto 0,05 biće 0,05/5 = 0,01.

Analizom realizovanih egzaktnih p-vrednosti u odnosu na korigovani nivo značajnosti, zaključuje se da mehanizam nekompletnosti podataka nije slučajan, jer nedostupnost podataka za nezavisne promenljive nije povezana sa zavisnom promenljivom (analizirati isključivo prvu kolonu Y u tabeli 2).

Ova analiza pokazala je da je mehanizam nekompletnosti podataka slučajan, te da pri ocenjivanju nepoznatih parametara posmatranog statističkog modela mora biti korišćen savremeni način rešavanja problema nekompletnih podataka. Ukoliko bi se, slučajno, problem nekompletnih podataka rešavao tradicionalnim procedurama, ocene nepoznatih parametara posmatranog statističkog modela bile bi pristrasne.

Drugi model veštačke inteligencije zasnovan na veličini efekta

Dakle, ukoliko mehanizam nekompletnosti podataka nije potpuno slučajan, i ukoliko je statistički model nadgledani, tada višestruka primena t-testa dva nezavisna uzorka (Tabachnick i Fidell, 2014) može utvrditi na koji način podaci nedostaju. Međutim, ukazano je na nedostatke t-testa u ovoj situaciji. Jedan od njih bio bi da će dva nezavisna uzorka biti veoma neuravnotežena (u smislu da, na primer, raspoloživi podaci mogu biti 95%, a nedostajući 5%). Predlog rešenja za nedostatke t-testa dat je u prethodnom delu. Ova procedura pripada oblasti neparametarske statistike. Međutim, ovi nedostaci mogu se prevazići korišćenjem statistike veličine efekta.

Stoga predlažemo da se umesto t-testa koristi statistika veličine efekta ω^2. Donekle sličan pristup može se naći kod Endersa (2010), s tim što on, kada je reč o t-testu dva nezavisna uzorka, koristi statistiku veličine efekta (označenu d), koja nije baš najprikladnija (Sheskin, 2011). U svakom slučaju, mnogo je bolje koristiti praktičnu umesto statističke značajnosti u ovom pitanju (zbog naznačenih nedostataka).

Stoga je, pre primene predložene statistike veličine efekta, neophodno kreirati indikatorske promenljive nedostupnosti. Naime, ukoliko neka promenljiva (označena Xi, i = 1,...,K) ima nedostajuće podatke, za nju se kreira indikatorska promenljiva nedostupnosti (označena Mi), koja uzima vrednost 1 ukoliko je dati podatak nedostupan, i vrednost 0 ukoliko je dati podatak raspoloživ.

Nakon formiranja indikatora nedostupnosti, moguće je izračunati predloženu statistiku veličine efekta, korišćenjem sledećeg obrasca (Sheskin, 2011):

$$\tilde{\omega}^2 = \frac{t^2-1}{t^2+n_1+n_2-1} \qquad (12)$$

gde n1 predstavlja veličinu poduzorka koji se odnosi na raspoložive podatke, a n2 veličinu poduzorka koji se odnosi na nedostupne podatke, dok t2 predstavlja realizovanu vrednost t-testa. Formula za datu statistiku veličine efekta može se prikazati na drugi način, kao što je pokazao Sakai (2018). Intenzitet statistike veličine efekta dat je sledećim obrascem (Cohen, 1988):

$$intezitet\ statistike\ veličina\ efekta = \begin{cases} mali, za\ 0.0099 < \tilde{\omega}^2 \leq 0.0588 \\ srednji, za\ 0.0588 < \tilde{\omega}^2 \leq 0.1379 \\ veliki, za\ \tilde{\omega}^2 > 0.1379 \end{cases} \quad (13)$$

Poslovna implementacija

Podaci koji se analiziraju prikazani su u prethodnom odeljku. Više detalja o ovom istraživanju, kao i o samoj svrsi ankete, dato je u Vasiću (2019). Veličina uzorka bila je oko 200 ispitanika, a naravno, deo ispitanika nije odgovorio na neka od pitanja u anketi.

Tabela 3. Statistike testa

	Y	X_1	X_2	X_3	X_4	X_5
promenljiva indikator neraspoloživosti u odnosu na koju se kreiraju dva nezavisna uzorka M_X1						
t	0	.	1.1	0.2	0	-0.1
n_1	187	193	166	157	172	189
n_2	6	0	5	6	6	7
$\tilde{\omega}^2$	-0.0052		0.0012	-0.0059	-0.0056	-0.0051
promenljiva indikator neraspoloživosti u odnosu na koju se kreiraju dva nezavisna uzorka M_X2						
t	-0.9	-1.4	.	1.4	-0.2	0.9
n_1	165	166	171	139	151	168
n_2	28	27	0	24	27	28
$\tilde{\omega}^2$	-0.0010	0.0049		0.0059	-0.0054	-0.0010
promenljiva indikator neraspoloživosti u odnosu na koju se kreiraju dva nezavisna uzorka M_X3						
t	-1.3	-2.9	-0.6	.	-1.6	-0.5
n_1	157	157	139	163	143	159
n_2	36	36	32	0	35	37
$\tilde{\omega}^2$	0.0036	0.0370	-0.0038		0.0087	-0.0038
promenljiva indikator neraspoloživosti u odnosu na koju se kreiraju dva nezavisna uzorka M_X4						
t	0.1	1	-0.4	0.4	.	-1.6
n_1	171	172	151	143	178	176
n_2	22	21	20	20	0	20
$\tilde{\omega}^2$	-0.0052	0.0000	-0.0049	-0.0052		0.0079

promenljiva indikator neraspoloživosti u odnosu na koju se kreiraju dva nezavisna uzorka M_X_5						
t	-1.6	-1	-4.1	-2.6	-0.8	.
$\mathbf{n_1}$	189	189	168	159	176	196
$\mathbf{n_2}$	4	4	3	4	2	0
$\tilde{\omega}^2$	0.0080	0.0000	0.0846	0.0341	-0.0020	

Izvor: Rezultat analize autora sproveden uz upotrebu IBM SPSS Statistics 26 softvera.

Kao što ukazuje procedura za rešavanje nedostajućih podataka, najpre se proverava da li nedostajući podaci nedostaju potpuno slučajno. Testiranje, sprovedeno u prethodnom odeljku, ukazalo je da mehanizam nekompletnosti podataka nije potpuno slučajan. Stoga dolazimo do zaključka da ukoliko bi se nedostajući podaci rešili na tradicionalan način, ocene parametara posmatranog modela veštačke inteligencije (koji se namerava koristiti) bile bi pristrasne.

Nakon prethodnog utvrđivanja, prelazi se na određivanje da li nedostajući podaci nedostaju na slučajan ili neslučajan način. U te svrhe, koristiće se formula (12), kao i pravilo odlučivanja zadato izrazom (13). Kompletno testiranje prikazano je u Tabeli 3, koja ukazuje da podaci nedostaju slučajno, budući da ne postoji asocijacija između zavisne promenljive (označene Y) i nedostajućih podataka u nezavisnim promenljivima. Da bi se videla data konstatacija, neophodno je posmatrati prvu kolonu Y u tabeli 3 i analizirati vrednosti statistike $\tilde{\omega}^2$. Kao što se može videti, sve vrednosti statistike $\tilde{\omega}^2$ su po apsolutnoj vrednosti manje od ili jednake 0,0099.

Dakle, nedostajući podaci u ovom primeru nedostaju slučajno, što dalje znači da, ukoliko se želi da u nadgledanom modelu veštačke inteligencije ocene parametara budu nepristrasne, problem nedostajućih podataka treba rešiti savremenim pristupom, koji je, na primer, kada je reč o upotrebi softvera IBM SPSS Statistics, primena algoritma ekspektacije i maksimizacije.

Zaključak

Ovim radom konačno je rešen problem „šta raditi" kada se susretnemo sa izazovom nedostajućih podataka. Definisanjem protokola, kao i upotrebom modela veštačke inteligencije, analitičari podataka su u mogućnosti da reše problem nedostajućih podataka, i to na jedan kvalitetan način, i time izbegnu, već pominjani u uvodu, princip GIGO. Na ovaj način omogućeno je da se dobiju pouzdani rezultati analitike podataka, koja je neophodna da bi poslovna banka mogla da postigne dobre poslovne rezultate i ciljeve.

Literatura

1. Adachi, K. (2016). Matrix-Based Introduction to Multivariate Data Analysis. Singapore: Springer.

2. Allison, D. (2002). Missing Data. Thousand Oaks: Sage Publications.

3. Arbuckle, J. L. (2014). IBM SPSS Amos 23 - User's Guide. Armonk. NY: IBM Corporation.

4. Enders, C. (2010). Applied Missing Data Analysis. New York: The Guilford Press.

5. Fitzmaurice, G. at al. (2015). "Missing Data: Introduction and Statistical Preliminaries", in Molenbergs, G. et al. (ed.). Handbook of Missing Data Methodology. CRC Press, Boca Raton, pp. 3-22.

6. Godolja, M., Domi, L. (2025). "Evolucija veštačke inteligencije u bankarskom sektoru: sistematski pregled literature". Bankarstvo, 54(2). Udruženje banaka Srbije.

7. IBM (2017). IBM SPSS Statistics Algorithms. Armonk, NY: IBM Corporation.

8. IBM (2019). IBM SPSS Statistics 26 Command Syntax Reference. Armonk, NY: IBM Corporation.

9. Kim, J. K. and Shao, J. (2014). Statistical Methods for Handling Incomplete Data. Boca Raton, FL: CRC Press - Taylor & Francis Group.

10. Kim, K. H., & Bentler, P. M. (2002). Tests of homogeneity of means and covariance matrices for multivariate incomplete data. Psychometrika, 67, 609–624.

11. Little, R. & Rubin, D. (2002). Statistical Analysis with Missing Data (2 ed.). Hoboken, New Jersey: John Wiley & Sons, Inc.

12. Mehta and Patel (2017). IBM SPSS Exact Tests 25. Armonk: IBM Corporation.

13. Orchard, T., & Woodbury, M. A. (1972). A missing information principle: Theory and applications. Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability, 1, 697–715.

14. Sakai, T. (2018). Laboratory Experiments in Information Retrieval - Sample Sizes, Effect Sizes, and Statistical Power. Singapore: Springer.

15. Sheskin, D. (2011). Handbook of Parametric and Nonparametric Statistical Procedures (5 ed.). Boca Raton: CRC Press.

16. Tabachnick, B. G. & Fidell, L. S. (2014). Using Multivariate Statistics (6ed). Harlow, Essex: Pearson Education Limited.

17. Thoemmes, F., & Enders, C. K. (2007). A structural equation model for testing whether data are missing completely at random. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

18. van Buuren, S. (2018a). Flexible Imputation of Missing Data (Second Edition). Boca Raton, FL: CRC Press - Taylor & Francis Group.

19. van Buuren, S. (2018b). mice - Multivariate Imputation by Chained Equations. R package version 3.3.0.

20. Vasić, V. (2018). Rešavanje problema multivarijacionih nedostajućih anketnih podataka primenom EM algoritma. Ekonomske ideje i praksa, broj 30, str.35-49.

21. Vasić, V. (2019). Determining the types of missing data under supervised statistical models. 5th International Researchers, Statisticians and Young Statisticians Congress – Congress Book. pp. 285-292. Kusadasi, Turkey.

DOI: 10.5937/bankarstvo2602190V

Date of receival: 11.05.2026.
Date of acceptance: 14.10.2026.

DEFINING THE PROTOCOL OF ARTIFICIAL INTELLIGENCE IN SOLVING MISSING DATA IN THE FIELD OF BANKING

Vladimir Vasić, tenured professor, Belgrade Banking Academy – Faculty of Banking, Insurance and Finance
Email: vladimir.vasic@bba.edu.rs
ORCID: 0000-0001-9694-8071

Milica Bugarčić, associate professor, Belgrade Banking Academy – Faculty of Banking, Insurance and Finance
Email: milica.bugarcic@bba.edu.rs
ORCID: 0000-0002-9327-9965

Summary: *In this paper, the problem of missing data in banking is addressed in a systematic and comprehensive manner. Very little has been written about this serious and significant challenge, sporadically about some specific business problems, but never in this way - namely, the creation, or rather the definition of protocols. By including artificial intelligence models in the process of solving missing data, this "unsolvable problem" for data analysts is brought under control, and begins to represent a routine, i.e. another mandatory step in data preparation and analysis.*

Keywords: artificial intelligence, machine learning, missing data, data analytics, banking.

JEL classification: C14, G21

Introduction

In the banking sector, in conditions of increasing competition and increasingly fierce market competition, commercial banks are finding it increasingly difficult to maintain good business perfor-mance and results. For this reason, they are turning to business analytics, which can show them what their weak points in business are, and what their development opportunities are. The integration of artificial intelligence and machine learning in financial services is rapidly transforming internal operations and risk management, putting a strong focus on the necessity of utilizing advanced predictive models over internal databases (Godolja and Domi, 2025). Therefore, the key to success, or survival in general, is in the quality results of various business analyses. However, for a quality result, the GIGO principle (garbage in, garbage out) applies, that is, in essence, the quality of the input data for analysis must be high. However, anyone who is involved in collecting and pro-cessing data very quickly encounters the problem of missing data, and this is where we come to the essence of this business.

Current Ways to Address the Problem of Incomplete Data

In most cases, researchers remedy the problem of incomplete data in some way, without even indicating how much such data there was. They have probably silently excluded incomplete data from the data set they are analyzing. This procedure is understandable to some extent, because researchers, when writing professional papers, are justifiably afraid that this may represent a major limitation of their research. They justifiably fear that the reviewers of their work could reject their work, with the explanation that the research is quite weak and unreliable, due to the serious presence of incomplete data. Therefore, the chances of publishing a professional work are greatly increased, if the problem of incomplete data is not presented. As noted by Orchard and Woodbury (1972), the best way to deal with incomplete data is simply not to have it.

Of course, in such works, the existence of incomplete data is not indicated anywhere. Then, the default method of solving the incompleteness problem (that is, removing incomplete data) was applied without even being mentioned anywhere. Furthermore, the tables in which the results of the conducted quantitative analyzes are presented, are then based on different sample sizes.

The reason why researchers resort to the use of outdated procedures for solving data incompleteness is to some extent because these procedures are theoretically clear to them. Those procedures certainly include deleting observations with incomplete data, then deleting observations only if they have incomplete data for the variables currently being analyzed, as well as replacing the incomplete data with the arithmetic mean of the observed variable. Modern artificial intelligence procedures for treating data incompleteness are quite complex, and in most cases, incomprehensible to the average researcher, and therefore they avoid using them.

Artificial Intelligence Protocol

If there is a presence of incompleteness in the data that is planned to be analyzed, there is a possibility that the estimates of unknown parameters in the quantitative analysis that is planned to be used will be biased. For this reason, it is important to approach the problem of incomplete data carefully, in a comprehensive way, which, colloquially, could be called a protocol.

In a given protocol, step zero would be the start of solving the problem of incomplete data. In the first step, they would ask the question:

• Whether the observed variables, i.e. in the analyzed data set, is there a presence of incomplete data?

If the answer is yes, they would go to the second step. If the answer is no, they would not go to the last step. In the second step, they would ask the question:

• Is the data missing completely at random?

If the answer is positive, they would go to the third step. If the answer is negative, they would go to the fourth step.

The third step represents solving incomplete data, either by applying traditional techniques (deleting observations that contain missing data; or by inserting the arithmetic mean of the observed variable in place of missing data), or by applying modern methods. Then they would go to the last step.

In the fourth step, they would ask the question:

• Does quantitative modeling use a supervised or unsupervised artificial intelligence model?

If the answer is supervised, go to the fifth step. If the answer is unsupervised, go to the eighth step.

In the fifth step, it examines:

• Is the unavailability of data on the explanatory variables caused by the values of the dependent variable?

If the answer is positive, go to the sixth step. If the answer is negative, go to the seventh step. In the sixth step, it is stated that the missing data do so by a non-random mechanism. For this reason, further data analysis is stopped, and the last step is taken.

In the seventh step, it is stated that data that are unavailable are missing in a random way, so the problem of missing data can only be solved by applying modern techniques. Then we go to the last step. In the eighth step, it examines:

• Does the unavailability of data for any variable depend on the value of its own data, or on the data value of the other variables being analyzed?

If the answer is from own values, go to the sixth step. If the answer is from the value of the other variables, go to the seventh step. In the last step, we conclude that solving the data incompleteness problem is complete.

Of course, in such works, the existence of incomplete data is not indicated anywhere. Then, the default method of solving the incompleteness problem (that is, removing incomplete data) was applied without even being mentioned anywhere. Furthermore, the tables in which the results of the conducted quantitative analyzes are presented, are then based on different sample sizes.

The reason why researchers resort to the use of outdated procedures for solving data incompleteness is to some extent because these procedures are theoretically clear to them. Those procedures certainly include deleting observations with incomplete data, then deleting observations only if they have incomplete data for the variables currently being analyzed, as well as replacing the incomplete data with the arithmetic mean of the observed variable. Modern artificial intelligence procedures for treating data incompleteness are quite complex, and in most cases, incomprehensible to the average researcher, and therefore they avoid using them.

Types of Mechanisms

Mostly, the data that needs to be statistically analyzed have the problem of incompleteness, to a greater or lesser extent. A given defect does not seem serious in itself, however, if the incompleteness mechanism is of a certain type, there may be a serious problem. For this reason, it is important to always examine the type of mechanism that leads to incomplete data. For now, it is defined that there are three types of mechanisms that lead to data incompleteness (Allison, 2002), and they are: completely random, random, and non-random.

In order to be able to describe the given types of mechanism, the variable with missing data is marked with Y. Then, its available part is marked with Yo, while the unavailable part is marked with Ym. Now an unavailability indicator variable could be defined, which takes the values of zero, for those observations, which are available at variable Y; and the value one, for those observations that are unavailable for variable Y. The given indicator variable can be marked with M. If in the analyzed data set, there are variables that do not have missing values, they are marked with X.

Therefore, the mechanism of missing data for the variable Y is said to be completely random, if the missing data do so in a completely random way, and therefore, neither depends on the part of the variable Ym, nor on the part of the variable Yo, nor on the other present variables X. Using the probability indicator of unavailability, a given mechanism of missing data (Fitzmaurice et al., 2015) can be expressed as:

$$P(M|Y,X) = P(M) \tag{1}$$

The mechanism of missing data for variable Y is said to be random if the missing data do not depend on the missing values (Ym) themselves. Using the probability of unavailability indicator, the given mechanism can be expressed as:

$$P(M|Y,X) = P(M|Y^{o},X) \tag{2}$$

Finally, the mechanism of missing data for the variable Y is said to be non-random, if the missing data also depend on the values of the missing data (Ym). Using the unavailability indicator and its probability, the given mechanism can be represented as:

$$P(M|Y,X) = P(M|Y^{o},Y^{m},X) \tag{3}$$

Traditional ways of dealing with incomplete data, such as excluding an observation that does not have complete data, or estimating a missing value by the arithmetic mean of a given variable, are appropriate, only if the mechanism of data incompleteness is completely random. If the mechanism of data incompleteness is random, then solving the problem of data incompleteness is correct,

only by using modern procedures, such as: the expectation-maximization algorithm, the method of complete information with maximum likelihood, the algorithm of specification of complete conditionality, the algorithm of augmented data, as well as other modern procedures. Note that the given procedures can be used instead of traditional procedures, when the data is missing completely by chance.

Supervised Model of Artificial Intelligence

Thus, when it comes to supervised AI models, the nature of missing data was determined as follows (Tabachnick and Fidell, 2014): first, it is tested whether the missing data are missing in a completely random way (Little and Rubin, 2002). If the given hypothesis is rejected, the following testing is performed, which tests whether the missing data are missing in a random or non-random way. Depending on the rejection or non-rejection of the given hypothesis, a conclusion is made, that the data are missing in a non-random way, or that the data are missing in a random way.

In the literature, it is already common to test whether missing data is missing completely at random with Little's test (IBM, 2017). Then, if the tested null hypothesis is rejected, we move on to testing whether the missing data is missing in a random way. Given testing (with statistical software) is carried out with the help of univariate statistical t-tests. The essence is to examine whether the available data represents a random sample from the planned sample (which was set at the beginning of the research). If this is fulfilled, then the missing data have the same distribution as the available data (Enders, 2010).

Equality of distributions implies that the group of available data and the group of missing data have equal arithmetic means and variances. Kim and Bentler (2002) called this feature the homogeneity of arithmetic means and covariances. Some authors checked the homogeneity of arithmetic means and variances (Thommes and Enders, 2007), while other authors considered that it is difficult to assume that the data are homogeneous in terms of arithmetic mean and heterogeneous in terms of variance. For this reason, they only examined the homogeneity of arithmetic means (IBM, 2017). Therefore, testing the homogeneity of arithmetic means was carried out using univariate t-tests.

Artificial Intelligence Model Based on Non-Parametric Statistics

As mentioned, when conducting, whether data are missing by chance, the univariate t-test is used, which examines the equality of the arithmetic means of two independent samples (implemented in most quantitative software, which have a missing data analysis module). Two independent samples for each variable in the analysis are formed based on the unavailability of data for some other variable in the analysis. Namely, if a variable (labeled Xi, i = 1,...,n) has missing data, an unavailability variable indicator (labeled Mi) is created for it, which takes the value 1 if the given data is unavailable, and the value 0 if the given data is available. A given indicator of unavailability can be presented in the form

$$M_i = \begin{cases} 0, if\ X_i\ is\ available \\ 1, if\ X_i\ is\ missing \end{cases},\ i = 1, \dots, n \qquad (4)$$

So, in order to examine whether some variable (Xi) has missing data in a random way, first an indicator of unavailability (Mi) is formed for a given variable, and then it is used to divide each other variable in the analysis into two independent samples, and then apply a univariate t-test to test the equality of arithmetic means.

The given test has certain disadvantages, for example, independent samples can be very unbalanced (Enders, 2010) (in the sense that the sample representing the available data is much larger than the sample representing the missing data). Also an independent sample, representing missing data, can have only a pair of elements. Furthermore, what can be attributed to the limitation of testing the arithmetic means of two independent samples is the fact that even if the data are missing non-randomly, two independent samples formed on the basis of the unavailability indicator can have equal arithmetic means.

For this reason, it is much better to propose some other way of checking whether the missing data is missing by chance. This is precisely the subject of this section, to suggest another way of checking. Another way of testing, whether the mechanism of missing data is random, could be a non-parametric test, two independent samples Mann-Whitney U test. The great advantage of this non-parametric test, compared to the parametric univariate t-test, is that the Mann-Whitney test does not test the equality of arithmetic means; but just what is needed; that is, whether the two distributions are equal. So there is no need to take a detour, to suggest that if two distributions are equal, then their arithmetic means and variances are also equal; rather, we test the equality of two distributions directly with the non-parametric, two-independent-samples Mann-Whitney test.

A further improvement would be to use the exact method for calculating the p-value of the non-parametric Mann-Whitney test, instead of the generally accepted asymptotic method. The advantage of using the exact method when calculating the p-value is that (Mehta and Patel, 2017) it calculates reliable values, even in situations where the conditions for applying the t-test are not met, regardless of the size of independent samples, as well as the balance of the data. So, using the exact method of calculating the p-value in the non-parametric Mann-Whitney test - when testing whether the mechanism of missing data is random - would overcome all the shortcomings and limitations of the t-test, which is used in practice.

Of course, apart from the obvious advantages provided by the proposed procedure, there are also certain difficulties. Namely, the application of the exact p-value calculation method is very demanding, i.e. computationally intensive, based on all possible permutations of the realized set of two independent samples that need to be tested.

That is why we will start slowly, with a theoretical explanation of the proposed procedure. So, when it comes to the non-parametric, two independent samples Mann-Whitney test, it will be assumed that the two independent samples are of size n1 and n2. The null hypothesis to be tested is represented as

$$H_0: F_1(x) = F_2(x) \tag{5}$$

which means that given two independent samples have the same distributions.

In the Mann-Whitney test, the original realized values are replaced with ranked values, i.e. value xi are replaced by rank(xi). Furthermore, to calculate the exact p-value, it is necessary to know the exact probability of the realization of two independent samples with ranked values (Mehta and Patel, 2017). The given probability is given by the expression

$$h(w) = \frac{\prod_{i=1}^{2} n_i!}{(n_1+n_2)!} \tag{6}$$

where n1 and n2 represent the sizes of the first and second independent samples; and the design w represents the realization of the given two independent samples with ranked values instead of the original values (or any permutation thereof). Further, the probability law of the Mann-Whitney statistic U to take the exact value u, is

$$Pr(U = u) = \sum_{U=u} h(w) \tag{7}$$

where the summation is performed over all possible design permutations w. The Mann-Whitney statistic U is defined by formula (10). Similarly, the exact p-value (and that for the right tail of the distribution of the U statistic) is

$$p_1 = Pr(U \geq u) = \sum_{U \geq u} h(w) \tag{8}$$

The previously defined formulas are general formulas for obtaining the exact p-value for two independent samples of non-parametric tests. Of course, for each specific (two independent samples) non-parametric test there is a special numerical algorithm, with the help of which the exact p-value is calculated. This is also dealt with in the continuation of this section.

Therefore, the methodology for the exact calculation of the p-value for two independent samples of the non-parametric Mann-Whitney test will be presented. Using the given calculated p-value, it will be possible to test the null hypothesis that the missing data are missing at random.

At the beginning of the calculation, it is necessary to first calculate the sum of the ranks. This is done by combining both samples into one sample, the elements of which are sorted by size. Then the elements in the unified sample are ranked. If there are equal elements, we take the average rank for them. After that (IBM, 2017), the sum of ranks for two independent samples (group 1 (or G1) and group 2 (or G2)) is calculated.

i

$$\begin{aligned} S_1 &= \sum_{t \in G_1 \geq u} rank(x_t; D_1 \cup D_2) \cdot I(x_t \in D_1) \\ S_2 &= \sum_{t \in G_2 \geq u} rank(x_t; D_1 \cup D_2) \cdot I(x_t \in D_2) \end{aligned} \tag{9}$$

where Di represents observations from the i-th independent sample (i = 1, 2), while I(·) represents an indicator that takes the value 0 if the expression in parentheses is not true, and the value 1 if the expression in parentheses is true.

The next step is to calculate the test statistic as well as the corresponding p-value. The Mann-Whitney U test statistic for the first independent sample (or for group 1) is given by the expression

$$\begin{aligned} U &= \sum_{i \in G_1} \sum_{j \in G_2} I(x_i < x_j) + \frac{1}{2} \cdot \sum_{i \in G_1} \sum_{j \in G_2} I(x_i = x_j) \\ &= n_1 \cdot n_2 + \frac{n_1 \cdot (n_1+1)}{2} - S_1 = S_2 - \frac{n_2 \cdot (n_2+1)}{2} \end{aligned} \tag{10}$$

where n1 and n2 represent the sizes of the first and second independent samples.

After defining the U test statistic, one moves to calculating the p-value. To calculate the p-value, one will need the sampling frequency distribution of the test statistic U for the value u, and for the

sample sizes i and j for two independent samples (in fi,j(u) notation). Furthermore, the following property will be used, by means of which the frequency distribution of the test statistic U can be expressed via the sum of the lower-order frequency distributions, as

$$f_{n_1,n_2}(u) = f_{n_1-1,n_2}(u-n_2) + f_{n_1,n_2-1}(u). \qquad (11)$$

Each lower-order distribution is symmetric with respect to various values of the test statistic U. Also, the sum of the lower-order distributions gives a result that is also symmetric.

The calculation of the frequency distribution for the test statistic U (denoted fn1,n2(u)) starts with the known distribution of the test statistic U for i = 1. (If desired, the result is the same, one can also start with the distribution of the test statistic U for j = 1). Then the property given by expression (11) is used, as well as the symmetry property, to derive the frequency distribution for i = 2. This procedure is repeated until the frequency distribution for i = n1 is derived.

After calculating the frequency distribution for the U test statistic (IBM, 2017), it is possible to calculate the (two-sided) exact p-value, using the following patterns

$$p_1 = \begin{cases} \sum_{u=0}^{trunc(U)} f_{n_1,n_2}(u), ukoliko\ je\ U \le \frac{n_1 \cdot n_2 + 1}{2} \\ 1 - \sum_{u=0}^{trunc(U)} f_{n_1,n_2}(u), ukoliko\ je\ U > \frac{n_1 \cdot n_2 + 1}{2} \end{cases}$$

$$p - value = 2 \cdot p_1$$

where trunc(U) represents the integer value of the realized value of the test statistic U.

Bank's Data and Business Implementation

The data used in this part represent the responses of potential clients in one of the leading foreign banks in the Republic of Serbia. The survey was conducted in May 2019 in the main branch of the given bank for the city of Novi Sad. All potential clients of the bank who inquired about the conditions for obtaining a cash loan from the staff employed in the branch were offered an anonymous, very short survey regarding the cash loan product. Namely, the analysts of the customer relations management department intended to examine (analyze) which cash loan repayment period the bank's potential clients are most interested in. They were also interested in whether the number of months of repayment of the cash loan might be influenced by the basic socio-economic characteristics of the clients. For this reason, the survey itself had 6 questions, which read: "Appropriate cash loan repayment period" (marked Y), "Age (in years)" (marked X1), "Years of residence at the current address" (marked X2), "Total household income (in 000 RSD)" which were logarithmically transformed (marked X3). Then "Years of employment with the current employer" (labelled X4) and "Number of household members" (labelled X5).

For this research, the sample size was planned to be 200; how many completed questionnaires were collected. However, not all clients responded to all survey questions, so we have the problem

of incomplete data. Table 1 shows the number and percentage of missing data, where the variables are sorted in ascending order, in relation to the percentage of missing data.

Table 1. Data Availability of Analyzed Variables

		X3	X2	X4	X1	Y	X5
Missing	N	37	29	22	7	7	4
	Percent	18.5%	14.5%	11.0%	3.5%	3.5%	2.0%
Valid N	27.7	163	171	178	193	193	196

Source: Result of the author's analysis conducted using IBM SPSS Statistics 26 software.

We can notice that the respondents mostly did not answer the question marked X3, in the amount of 18.5%, while the question in the survey, marked X5, has the fewest missing answers, namely only 2%.

The initial step in statistical data analysis would be to examine the nature of missing data. In such an examination, the first step would be to test the null hypothesis, which is that the mechanism of incomplete data is completely random. As already mentioned, to test this hypothesis, Little's MCAR test is used. To perform this test, as well as all other statistical calculations, the domain-specific language of the IBM SPSS Statistics 26 software was used (IBM, 2019). The results of the conducted testing are:

$$\textit{Chi-Square} = 95.465,\ \textit{DF} = 72,\ \textit{Sig.} = 0.034$$

so the null hypothesis, that the mechanism of data incompleteness is completely random, is rejected. This concludes that the classic solution to the problem of missing data, such as deleting incomplete observations or inserting the arithmetic mean, would be wrong.

Table 2. Test Statistics and Exact P-Values

	Y	X1	X2	X3	X4	X5
Grouping Variable: M_X1						
Mann-Whitney U	549.500	-	281.000	454.000	507.000	661.500
Exact Sig. (2-tailed)	0.935	-	0.227	0.886	0.945	1.000
Grouping Variable: M_X2						
Mann-Whitney U	2074.500	1861.000	-	1384.000	1959.000	2244.500
Exact Sig. (2-tailed)	0.391	0.159	-	0.185	0.749	0.690
Grouping Variable: M_X3						
Mann-Whitney U	2435.000	1860.000	2142.000	-	2169.000	2672.000
Exact Sig. (2-tailed)	0.197	0.001	0.747	-	0.223	0.371
Grouping Variable: M_X4						
Mann-Whitney U	1858.000	1635.000	1366.000	1390.000	-	1344.500
Exact Sig. (2-tailed)	0.927	0.484	0.492	0.842	-	0.073
Grouping Variable: M_X5						
Mann-Whitney U	193.500	296.500	124.000	199.500	99.000	-
Exact Sig. (2-tailed)	0.097	0.476	0.139	0.213	0.319	-

Source: Result of the author's analysis conducted using IBM SPSS Statistics 26 software

After the previously rejected hypothesis, we move on to testing the next null hypothesis, that is, that the mechanism of data incompleteness is random or non-random. As suggested (where we pointed out the advantages of the proposed procedure compared to the existing one), the non-parametric Mann-Whitney test with exact p-value calculation will be used, instead of the t-test. The results of the series of tests are shown in Table 2. For each variable that has missing data, a missing data indicator was created, in the label M_X#. Also, for each variable, testing was applied in relation to the remaining 5 variables in the supervised statistical model. This is noted because of the need to correct the alpha-risk for the number of tests, per variable. So, the significance level instead of 0.05 will be 0.05/5 = 0.01.

By analyzing the realized exact p-value in relation to the corrected level of significance, it is concluded that the mechanism of data incompleteness is not accidental, because the unavailability of data for independent variables is not related to the dependent variable (analyze only the first column Y in table 2).

This analysis showed that the mechanism of data incompleteness is random, and that when evaluating the unknown parameters of the monitored statistical model, a modern way of solving the problem of data incompleteness must be used. If, by any chance, the problem of data incompleteness would be solved by traditional procedures, the evaluations of the unknown parameters of the monitored statistical model would be biased.

The Second Artificial Intelligence Model Based on Effect Size

Therefore, if the mechanism of data incompleteness is not completely random, and if the statistical model is supervised, then multiple application of the t-test of two independent samples (Tabachnick and Fidell, 2014) can determine how the data are missing. However, they pointed out the shortcomings of the t-test in this situation. One of them would be that the two independent samples will be very unbalanced (in the sense that, for example, the available data may be 95% and the missing data 5%). A proposal for a solution to the shortcomings of the t-test was given in the previous section. This procedure belongs to the field of non-parametric statistics. However, these shortcomings can be overcome by using effect size statistics.

Therefore, we suggest that instead of the t-test, the effect size statistic – –^2 should be used. A somewhat similar approach can be found in Enders (2010), with the fact that, when it comes to the t-test of two independent samples, he uses the statistic of the effect size (labeled d), which is not exactly the most appropriate (Sheskin, 2011). In any case, it is much better to use practical instead of statistical significance in this issue (due to the indicated shortcomings).

Therefore, before applying the proposed effect size statistics, it is necessary to create variable indicators of unavailability. Namely, if a variable (labeled Xi, i = 1,...,K) has missing data, we create an unavailability indicator variable for it (labeled Mi), which takes the value 1 if the given data is unavailable, and the value 0 if the given data is available.

After forming the unavailability indicator, it is possible to calculate the proposed effect size statistics, using the following formula (Sheskin, 2011):

$$\tilde{\omega}^2 = \frac{t^2-1}{t^2+n_1+n_2-1} \quad (12)$$

where n1 represents the size of the subsample related to the available data, and n2 represents the size of the subsample related to the unavailable data. t2 represents the resulting t-test value. The formula for a given effect size statistic can be presented in another way, as shown by Sakai (2018). The intensity of effect size statistics is given by the following formula (Cohen, 1988):

$$intensity\ of\ effect\ size\ statistics = \begin{cases} small, za\ 0.0099 < \tilde{\omega}^2 \leq 0.0588 \\ medium, za\ 0.0588 < \tilde{\omega}^2 \leq 0.1379 \\ large, za\ \tilde{\omega}^2 > 0.1379 \end{cases} \quad (13)$$

Business Implementation

The data to be analyzed are presented in the previous section. More details about this research, as well as the purpose of the survey itself, are given in Vasić (2019). The sample size was about 200 respondents, and of course, some respondents did not answer some of the survey questions.

Table 3. Test Statistics

	Y	X_1	X_2	X_3	X_4	X_5
variable unavailability indicator against which two independent samples are created						
M_X1						
t	0	.	1.1	0.2	0	-0.1
n_1	187	193	166	157	172	189
n_2	6	0	5	6	6	7
$\tilde{\omega}^2$	-0.0052		0.0012	-0.0059	-0.0056	-0.0051
variable unavailability indicator against which two independent samples are created						
M_X2						
t	-0.9	-1.4	.	1.4	-0.2	0.9
n_1	165	166	171	139	151	168
n_2	28	27	0	24	27	28
$\tilde{\omega}^2$	-0.0010	0.0049		0.0059	-0.0054	-0.0010
variable unavailability indicator against which two independent samples are created						
M_X3						
t	-1.3	-2.9	-0.6	.	-1.6	-0.5
n_1	157	157	139	163	143	159
n_2	36	36	32	0	35	37
$\tilde{\omega}^2$	0.0036	0.0370	-0.0038		0.0087	-0.0038
variable unavailability indicator against which two independent samples are created						
M_X4						
t	0.1	1	-0.4	0.4	.	-1.6
n_1	171	172	151	143	178	176
n_2	22	21	20	20	0	20
$\tilde{\omega}^2$	-0.0052	0.0000	-0.0049	-0.0052		0.0079

variable unavailability indicator against which two independent samples are created

M_X_5						
t	-1.6	-1	-4.1	-2.6	-0.8	.
$\mathbf{n_1}$	189	189	168	159	176	196
$\mathbf{n_2}$	4	4	3	4	2	0
$\tilde{\omega}^2$	0.0080	0.0000	0.0846	0.0341	-0.0020	

Source: Result of the author's analysis conducted using IBM SPSS Statistics 26 software.

As indicated by the procedure for solving missing data, it is first checked whether the missing data is missing completely by chance. The testing, which was carried out in the previous section, indicated that the mechanism of data incompleteness is not completely random. Therefore, we come to the conclusion that if the missing data are solved in a traditional way, the parameter estimates of the monitored artificial intelligence model (which is intended to be used) will be biased.

After the preliminary determination, we move on to determining whether the missing data is missing in a random or non-random way. For these purposes, formula (12) will be used as well as the decision rule given by expression (13). Complete testing is shown in Table 3, which indicates that data are missing at random, as there is no association between the dependent (labeled Y) variable and missing data in the independent variables. In order to see this statement, it is necessary to observe the first column Y in table 3, and analyze the values of the statistic $\tilde{\omega}^2$. As can be seen, all values of the statistic $\tilde{\omega}^2$ are less than or equal to 0.0099 in absolute value.

Therefore, the missing data in this example are missing by chance, which further means that if it is desired that in the supervised model of artificial intelligence, parameter estimates are unbiased, then the problem of missing data should be solved with a modern approach, which is, for example, when it comes to the use of the IBM SPSS Statistics software, the use of the expectation-maximization algorithm.

Conclusion

This work has finally solved the problem of "what to do" when faced with the challenge of missing data. By defining a protocol, as well as using artificial intelligence models, data analysts are able to solve the problem of missing data, in a qualitative way, and thus avoid the GIGO principle already mentioned in the introduction. In this way, it is possible to obtain reliable data analysis results, which are necessary for a commercial bank to achieve good business results and goals.

Literature

1. Adachi, K. (2016). Matrix-Based Introduction to Multivariate Data Analysis. Singapore: Springer.

2. Allison, D. (2002). Missing Data. Thousand Oaks: Sage Publications.

3. Arbuckle, J. L. (2014). IBM SPSS Amos 23 - User's Guide. Armonk. NY: IBM Corporation.

4. Enders, C. (2010). Applied Missing Data Analysis. New York: The Guilford Press.

5. Fitzmaurice, G. at al. (2015). "Missing Data: Introduction and Statistical Preliminaries", in Molenbergs, G. et al. (ed.). Handbook of Missing Data Methodology. CRC Press, Boca Raton, pp. 3-22.

6. Godolja, M., & Domi, L. (2025). Artificial Intelligence Evolution in the Banking Sector: A Systematic Literature Review. Bankarstvo, 54(2). Association of Serbian Banks

7. IBM (2017). IBM SPSS Statistics Algorithms. Armonk, NY: IBM Corporation.

8. IBM (2019). IBM SPSS Statistics 26 Command Syntax Reference. Armonk, NY: IBM Corporation.

9. Kim, J. K. and Shao, J. (2014). Statistical Methods for Handling Incomplete Data. Boca Raton, FL: CRC Press - Taylor & Francis Group.

10. Kim, K. H., & Bentler, P. M. (2002). Tests of homogeneity of means and covariance matrices for multivariate incomplete data. Psychometrika, 67, 609–624.

11. Little, R. & Rubin, D. (2002). Statistical Analysis with Missing Data (2 ed.). Hoboken, New Jersey: John Wiley & Sons, Inc.

12. Mehta and Patel (2017). IBM SPSS Exact Tests 25. Armonk: IBM Corporation.

13. Orchard, T., & Woodbury, M. A. (1972). A missing information principle: Theory and applications. Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability, 1, 697–715.

14. Sakai, T. (2018). Laboratory Experiments in Information Retrieval - Sample Sizes, Effect Sizes, and Statistical Power. Singapore: Springer.

15. Sheskin, D. (2011). Handbook of Parametric and Nonparametric Statistical Procedures (5 ed.). Boca Raton: CRC Press.

16. Tabachnick, B. G. & Fidell, L. S. (2014). Using Multivariate Statistics (6ed). Harlow, Essex: Pearson Education Limited.

17. Thoemmes, F., & Enders, C. K. (2007). A structural equation model for testing whether data are missing completely at random. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

18. van Buuren, S. (2018a). Flexible Imputation of Missing Data (Second Edition). Boca Raton, FL: CRC Press - Taylor & Francis Group.

19. van Buuren, S. (2018b). mice - Multivariate Imputation by Chained Equations. R package version 3.3.0.

20. Vasić, V. (2018). Rešavanje problema multivarijacionih nedostajućih anketnih podataka primenom EM algoritma. Ekonomske ideje i praksa, broj 30, str.35-49.

21. Vasić, V. (2019). Determining the types of missing data under supervised statistical models. 5th International Researchers, Statisticians and Young Statisticians Congress – Congress Book. pp. 285-292. Kusadasi, Turkey.