Prediktivna analiza prodaje

Bokulić, Tena

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO - MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Tena Bokulić

PREDIKTIVNA ANALIZA PRODAJE

Diplomski rad

Voditelj rada: prof. dr. sc. Robert Manger

Zagreb, rujan, 2021.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Ovaj diplomski rad posvećujem svojoj obitelji za beskrajnu podršku i razumijevanje tijekom cijelog studiranja. Posebna i najveća hvala mom Luki na strpljenju i beskompromisnoj vjeri u mene koja je vrijedila i više nego što mogu izraziti riječima.

Zahvaljujem i vlasniku tvrtke Data Sense, gospodinu Mariu Korbaru na ukazanim smjernicama u pisanju rada i kontinuiranoj motivaciji koja je doprinijela mom osobnom i profesionalnom rastu. Hvala i mom mentoru prof. dr. sc. Robertu Mangeru na pomoći u stvaranju ovog diplomskog rada. Na kraju hvala i mojim prijateljima i svima koji su bili uz mene tijekom studenskog putovanja.

~ Without data you're just another person with an opinion. ~

William Edwards Deming (1900–1993)

Sadržaj

Uvod

Klasični pristupi poslovanju, odnosno poslovni modeli su nedovoljno brzi i u konačnici sve više ne zadovoljavaju potrebe potrošača. Kako bi organizacija opstala na današnjem tržištu potrebna je njihova prilagodba i prihvaćanje novih načela koji podrazumijevaju poslovnu inteligenciju, umjetnu inteligenciju ili općenito rečeno, digitalizaciju. Digitalizacija poslovanja predstavlja uvođenje novih tehnologija s ciljem dobivanja kvalitetnijih ali i brzih rezultata, što omogućava poduzimanje pravovremenih aktivnosti u skladu s promjenama na tržištu i unutar organizacije.

Upotreba napredne analitike za predviđanje budućih događaja pomoću širokog spektra tehnika, kao što su rudarenje podacima, strojno učenje, statistički algoritmi i umjetna inteligencija sažimaju se u proces prediktivne analitike. Prediktivna analitika predstavlja metodu koja služi za prognoziranje budućih događaja na način da utvrđuje povezanost među povijesnim podacima i nudi prognozu o vjerojatnosti ustupanja određene veličine ili događaja. S obzirom da je primjena prediktivne analitike uspješno integrirana u mnoštvo poslovnih područja, u radu se promatra konkretan slučaj navedene integriranosti u poslovno područje prodaje.

U nastavku se definira pojam prediktivne analitike kao sastavnog dijela digitalizacije, navode se metode koje prediktivna analitika primjenjuje u svojim izračunima, te je opisan postupak implementacije prediktivne analitike u poslovanje. Rad sadrži i studijski primjer razvoja prediktivnog modela prodaje za zamišljenu organizaciju.

Poglavlje 1

Poslovna analitika

Poslovna analitika (*engl. Business Analytics*) obuhvaća vještine, tehnologije i praksu kontinuiranog istraživanja ostvarenih performansi poslovanja radi uvida i podrške u planiranju. Poslovna analitika je proces transformiranja podataka u akcije kroz analizu i uvid u kontekstu organizacijskog odlučivanja i rješavanja problema. Alati i tehnike poslovne analitike našli su primjenu u brojnim oblastima širokog spektra organizacija u cilju unapređenja upravljanja odnosima s klijentima te unapređenja financijskih i ostalih funkcionalnih cjelina organizacije [14]. Poslovna analitika bavi se razvojem novih saznanja i razumijevanjem proteklog poslovanja na osnovu podataka i statističkih metoda. Koristi statističke analize, deskriptivne i prediktivne modele. Daje odgovore na pitanja o uzrocima pojava, razvojnim trendovima, predviđa i optimizira buduće događaje. Ono što poslovanje u današnje vrijeme čini posebno kompleksnim je prevelika količina nestrukturiranih ili slabo strukturiranih podataka i informacija. Podaci potrebni za donošenje odluka, uključujući one koje organizacija samostalno prikupi iz internih izvora ili one koje prikupi putem Interneta i društvenih mreža, rastu eksponencijalno i stoga ih je sve teže razumjeti i koristiti. Ovo je jedan od razloga zašto je analitika postala veoma značajna u suvremenom poslovnom okruženju.

Termin analitika koristi se za klasičan spoznajni proces koji počinje od prikupljanja i analize podataka u cilju utvrđivanja činjenica, zatim uključuje otkrivanje zakona (obrazaca) koji postoje između promjenljivih veličina predstavljenih podacima, a završava se formiranjem teorije. S obzirom na sve veće količine podataka (*tzv. Big Data*) koje se analiziraju, kao i veliki broj sofisticiranih znanstvenih metoda i moćnih računarskih resursa za njihovo skupljanje i obradu, u literaturi i praksi se kao sinonim za analitiku često koristi termin znanost o podacima (*engl. Data Science*), dok se o podacima iz područja poslovanja govori u okviru analitike poslovanja ili poslovne analitike. Poslovna analitika je primjena podatkovne analitike u poslovanju.

1.1. Podatkovna analitika

Podatkovna analitika (*engl. Data Analytics*) predstavlja skup statističkih i matematičkih metoda koje u kombinaciji sa strojnim učenjem analiziraju i obrađuju podatke iz skladišta podataka i *Big Data* te ih čine raspoloživim poslovnim organizacijama. Strojno učenje predstavlja umjetno generiranje znanja i iskustava budući da sustav uči na primjerima te ih generalizira. Dakle, sustav se ne temelji isključivo na skupu analiziranih primjera, već otkriva uzorke i povezanosti među njima. Područje djelovanja podatkovne analitike je upravljanje podacima, odnosno njihovo prikupljanje, organiziranje i izvještaji o rezultatima. Tako dobiveni podaci služe kao oslonac za donošenje poslovnih odluka [27].

Podatkovna analitika osim same analize podataka podrazumijeva i sve faze upravljanja podacima koje prethode analizi, kao što su prikupljanje, čišćenje, organiziranje, pohrana i izvješćivanje o rezultatima [27].

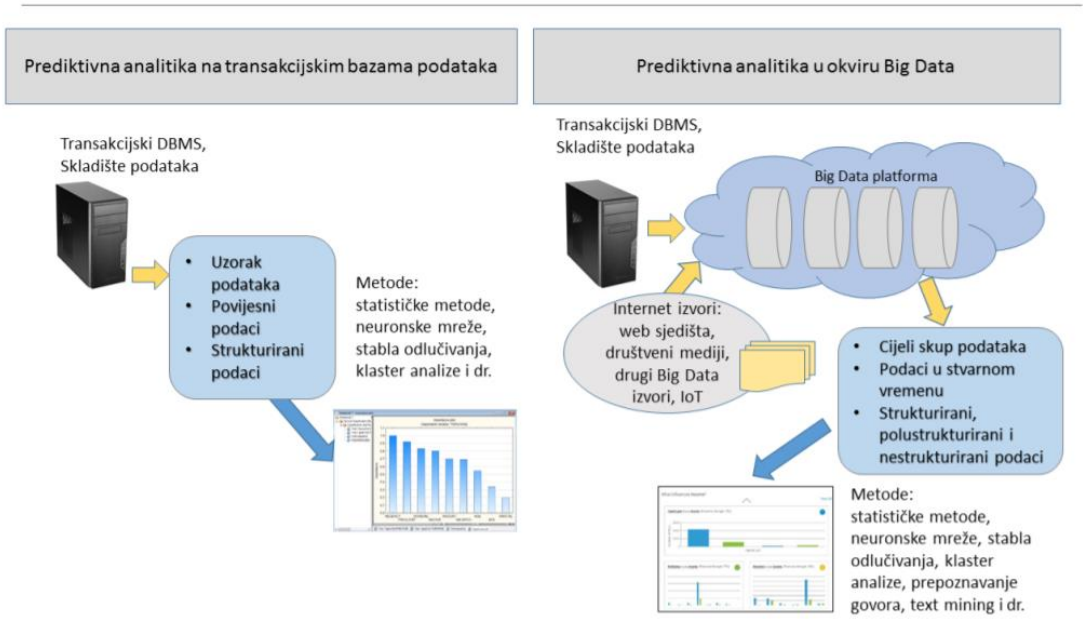
Često primjenjivana podjela podatkovne analitike je podjela prema Gartner Inc. [11] koja razlikuje četiri stepenice:

- *Reaktivna ili deskriptivna analitika*. Bavi se prošlošću i objašnjava utjecaj događaja iz prošlosti na sadašnjost. Primjena deskriptivne analitike zahtjeva znatne vremenske utroške budući da se većinski postotak analize obavlja ručnim unosom podataka. Karakteriziraju je tradicionalni principi poslovne inteligencije i vizualizacija rezultata pomoću tabličnih prikaza i dijagrama. Koristi jednostavne statističke metode koje opisuju jednu varijablu i njezinu raspodjelu, na primjer prosječne vrijednosti, učestalost pojavljivanja, smjer kretanja prihoda, troškova, profita i slično.

- *Proaktivna ili dijagnostička analitika*. Naprednija analitika koja kroz analiziranje podataka iz prošlosti dolazi do razloga, utjecaja i posljedica događaja na trenutnu situaciju. Na primjer, ova analitika dati će odgovor na pitanje zašto jedna od prodavaonica ima učestalo manji promet od ostalih.

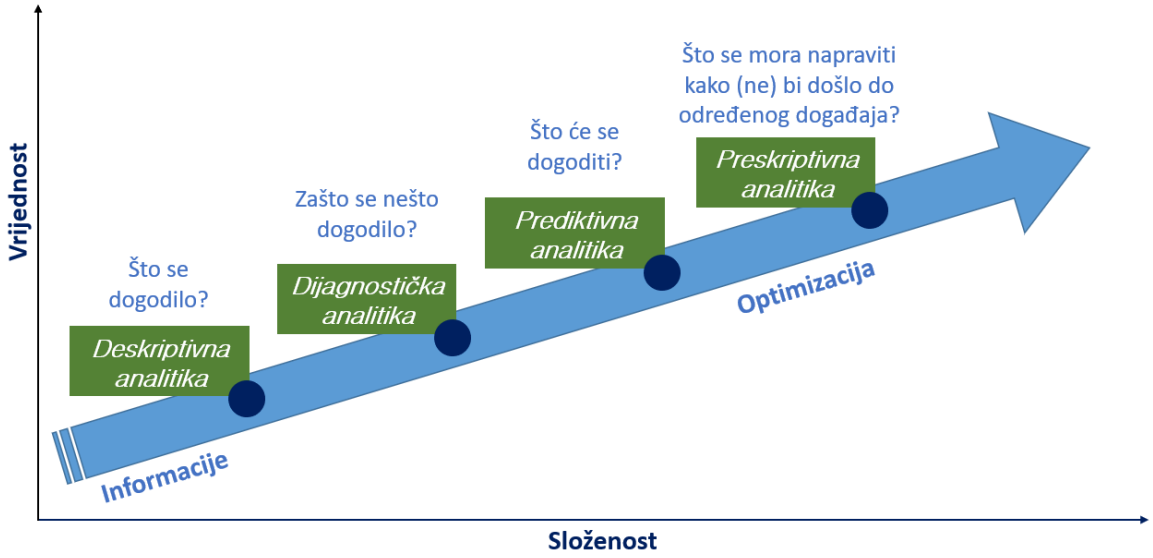
▪ *Prediktivna analitika*. Pruža informaciju o tome što će se, s kojom vjerojatnošću, dogoditi u budućnosti. Provodi se na višedimenzionalnim podacima s pomoću online analitičkog procesiranja i sličnih metoda. Cilj je odrediti vrijednost nekog obilježja koje će se vjerojatno pojaviti u budućnosti. Primjeri problema koji se rješavaju metodama prediktivne analitike su predviđanje prodaje, predviđanje rasta organizacije, predviđanje sutrašnje cijene dionice, ali i segmentiranje kupaca prema vjerojatnosti buduće kupnje nekog proizvoda i slično.

Kako bi producirala modele koji s visokom točnošću mogu predvidjeti buduću vrijednost, prediktivna analitika traži puno podataka iz prošlosti na kojima će ustanoviti određene veze među podacima. Podatkovna analitika fokusira se na upravljanje velikim količinama podataka uz uporabu skalabilnih distribuiranih tehnologija, a može obrađivati i nestrukturirane podatke, poput slika, audio i videozapisa, podatke iz uređaja putem Interneta [23].



Slika 1.1. Prediktivna analitika na transakcijskim bazama i u okviru *Big Data* platforme (slika preuzeta iz [22])

- *Preskriptivna analitika*. Predstavlja kombinaciju deskriptivne i prediktivne analitike. Na temelju podataka iz navedene dvije analitike usmjerava i sugerira koje aktivnosti treba poduzeti kako bi se ostvarili temeljni ciljevi. Karakteristike ove analize su tehnike poput grafičke analize, simulacija, neuronske mreže, strojno učenje i slično [22].



Slika 1.2. Podjela podatkovne analitike prema složenosti

1.2. Razvoj podatkovne analitike

Pojam podatkovna analitika prvi put se pojavljuje sredinom 90-tih godina prošloga stoljeća, pod nazivom analitika 1.0. Početak je predstavljala deskriptivna analitika, koja pruža izvješća na temelju internih, strukturiranih podataka. Podaci su prikazivani pomoću proračunskih tablica. Najveći udio procesiranja programa bio je u pripremi podataka umjesto njihovoj konkretnoj analizi. Dio analize se bavio povijesnim podacima koji su pružali uvid u uzročno-posljedične odnose na tržištu. Razvojem informacijske tehnologije i novih programa za analiziranje razvijala se i podatkovna analitika. Unatoč tome, deskriptivna analitika je i dalje dominantna disciplina za izvještaje [39].

Analitika 2.0 javlja se početkom 2000-tih godina u Silicon Valley-u. Obilježje ove ere je početak analiziranja i rada s nestrukturiranim podacima, pojava *Big Data* te nove računalne mogućnosti i visok stupanj povezanosti putem Interneta. Razdoblje karakterizira početak primjene prediktivne i preskriptivne analitike. Povećanjem količine podataka s kojima se posluje, povećava se i potreba za stručnjacima na tom području tako se pojavljuje pojam i uloga podatkovnog znanstvenika (*engl. Data Scientist*) [21].

Slijedeća etapa analitike razvijala se sukladno analitici 2.0. Najznačajnije obilježje analitike 3.0 je mogućnost svih organizacija da stvaraju proizvode i usluge temeljene na podacima i podatkovnoj analitici. Analitika prestaje biti samo alat za donošenje poslovnih odluka, postaje sastavni dio strategije organizacija, te je izvor prihoda [35].

Poglavlje 2

Prediktivna analitika

Prediktivna analitika (*engl. Predictive Analytics*) postaje nezaobilazna u poslovnom odlučivanju. Danas organizacije više nisu prisiljene donositi odluke na temelju vlastite intuicije već na temelju obrađenog znanja u sustavu koji se generira metodama podatkovne analitike. Zekić-Sušac [23] navodi kako intuicija više nije dovoljna, posebno ako se konkurencija neke organizacije oslanja na sofisticirane alate za poslovnu inteligenciju i rudarenje podataka koji se povlače iz skladišta podataka ili *Big Data* baza.

Lotfi Zadeh [18], otac neizrazite (fuzzy) logike, još od 1965. godine ističe kako tradicijske statističke i matematičke metode ne pružaju zadovoljavajuća rješenja za stvarne probleme u kojima ima puno nesigurnosti, rizika, nedostajućih ili nepreciznih podataka. Mnogo menadžera susretalo se sa situacijama u kojima su trebali donijeti odluku na temelju vlastite intuicije budući da nisu imali nikakav odgovarajući alat za potporu. Danas sofisticirani alati za poslovnu inteligenciju i rudarenje podataka povlače podatke iz skladišta podataka ili *Big Data* baza podataka te na taj način omogućuju da menadžeri nisu prepušteni sami sebi već da donose odluke na temelju obrađenog znanja u sustavu. Znanje se generira metodama podatkovne analitike što u posljednje vrijeme sve više uključuje ne samo statističke metode već i metode strojnog učenja [23].

2.1. Definicija prediktivne analitike

U literaturi pronalazimo različite definicije prediktivne analitike.
Prema izvoru [31] prediktivna analitika je primjena matematike u svrhu analiziranja obrazaca unutar podataka iz prošlosti kako bi se moglo predvidjeti buduće ponašanje. Radi se o tipu prognoze koja traži veze između prošlih i budućih događaja.
Compton [12] prediktivnu analitiku definira kao "tehnike, alate i tehnologije koje koriste podatke za pronalazak modela koji mogu predvidjeti ishode sa značajnom vjerojatnosti točnosti."
Ona se koristi kako bi predvidjela buduće nepoznate događaje. Koristi mnoge tehnike kao što su statistički algoritmi, rudarenje podataka, strojno učenje i umjetnu inteligenciju kako bi analizirala trenutnu bazu podataka i predvidjela budućnost. Ima za cilj identificirati vjerojatnost budućih ishoda na temelju raspoloživih povijesnih podataka. Cilj je ići dalje od saznanja o tome što se dogodilo pružajući najbolju procjenu onoga što će se dogoditi u budućnosti [1].
Nadalje, prema izvoru [5] prediktivna analitika je upotreba podataka, statističkih algoritama i tehnika strojnog učenja kako bi se utvrdila vjerojatnost budućih ishoda na temelju povijesnih podataka.

Možemo zaključiti kako prediktivna analitika predstavlja metode i tehnike, koje na temelju podataka iz prošlosti izračunavaju vjerojatnost ustupanja određenog događaja ili pojave u budućnosti. Pomoću alata prediktivne analitike razrađuju se tehnologije koje uče iz iskustva kako bi predvidjele buduće ponašanje i na taj način pridonijele pravim, pouzdanim odlukama. U osnovi to znači da trebamo informacije iz prošlosti iz kojih ćemo naučiti šta će se dogoditi u budućnosti.

2.2. Prediktivna analitika u poslovanju

Zbog velike količine podataka koje organizacije danas posjeduju mogućnosti primjene prediktivne analitike u današnje vrijeme doista su opsežne u kontekstu raznih ključnih pitanja vezanih uz poslovanje svake pojedine organizacije. Prediktivna analitika nadogradnja je poslovne analitike. Primjenom spomenutih statističkih i matematičkih metoda na strukturirane i nestrukturirane podatke dobivamo uvid u uzorke i odnose među podacima te možemo procijeniti vjerojatnost da se u budućnosti određeni događaj dogodi ili ne dogodi. Korištenjem prediktivne analitike možemo dobiti odgovore na neka od ključnih pitanja vezana uz klijente, poslovanje, kapital i prijevare. Kako bi se uspješnost prodaje povećala nastoji se "dekodirati" prošle kupovne navike potrošača i projicirati njihove buduće kupovne navike kako bi mogli donijeti odluke na temelju tih spoznaja. Stvaranjem baza podataka kupaca otvara se mogućnost predviđanja njihovog budućeg ponašanja. Modeli predviđanja, koji su se u praksi koristili i prije pojave današnje tehnologije, svoju opravdanost temelje na osnovnoj premisi da potencijalni i postojeći kupci reagiraju na predvidiv način. Prikupljeni podaci u prošlosti i sadašnjosti koriste se kako bi se predvidjelo ponašanje u budućnosti. Pri tome prediktivna analiza traži korelaciju između prošlih i budućih događanja i pomaže nam provjeriti da li su ta predviđanja točna. Rezultati tih analiza su prediktivni modeli.

Naglašava se da niti jedan od alata, metoda ili tehnika nije u stanju "predvidjeti budućnost", ali postoji mogućnost donijeti prognozu s 90-95% točnosti. Ovaj postotak nastupa u slučaju da se u odnosu na podatke iz prošlosti nije ništa znatno promijenilo.

Prediktivna analitika bavi se slijedećim područjima [19]:

- *Segmentacija* - grupiranje analiziranih objekata na temelju sličnosti.
- *Asocijacija* - identifikacija učestalosti pojavljivanja određenog događaja i donošenje zaključaka "A i B vode prema C".
- *Klasifikacija* - predviđanje pripadnosti pojedinoj grupi elemenata.
- *Regresijska analiza* - identifikacija veza među pojedinim elementima.
- *Prognoziranje* - predviđanje budućih iznosa.

2.3. Alati prediktivne analitike

S obzirom na veliku brojnost metoda podatkovne analitike i njihovih primjena, fokusirat ćemo se na metode prediktivne analitike i njihovu implementaciju u okviru Big Data platformi. U kontekstu *Big Data*, podatkovna analitika fokusira se na upravljanje velikim količinama podataka uz uporabu skalabilnih distribuiranih tehnologija, a može obrađivati i nestrukturirane podatke, poput slika, audio i videozapisa i podatke iz uređaja putem Interneta.

Prilikom izrade modela prediktivne analitike, jedan od izazova s kojim se organizacije susreću je odabir alata i njegovo povezivanje s postojećom transakcijskom bazom ili integriranje u *Big Data* platformu.

Pri odabiru alata ne postoje rješenja za sve, stoga organizacija treba razmotriti [23]:

- Za što joj je potrebna prediktivna analitika?
- Koje će probleme rješavati?
- S kakvim podacima će se susresti?
- Koji budžet ima na raspolaganju?
- Kakva je postojeća infrastruktura (platforma koja omogućuje povezivanje s postojećim bazama)?
- Koju razinu znanja i vještina ima osoblje odnosno, jesu li upoznati s tumačenjima rezultata metoda prediktivne analitike?
- Koliku razinu vizualizacije, fleksibilnosti i skalabilnosti žele imati menadžeri i sl.

Prema izvoru [23] odabir alata često otežava činjenica da pokrivaju sličan opseg metoda, no ohrabrujuća je njihova skalabilnost te moguća zamjena drugim alatom. Prvi odabir ne mora biti i posljednji pa tako organizacije mogu biti sigurne da će manje pogriješiti odaberu li bilo koji alat za prediktivnu analitiku, nego da odluče ne koristiti nijedan.
Kako bi se odabrao odgovarajući alat nužno je utvrditi za što će se primjenjivati prediktivna analitika, odnosno koje ciljeve se teži ostvariti. Također je potrebno utvrditi s kakvim podacima će se raditi i koji podaci se žele analizirati.

Podloga za odabir odgovarajućeg alata je *Gartnerov kvadrant za analitiku i poslovnu inteligenciju* jer se u njemu mogu naći detaljne analize i korisni savjeti koji pružaju uvid u smjer i tržišnu zrelost pružatelja usluga za analitiku i poslovnu inteligenciju. Sastoji se od dvodimenzionalne matrice kojom se pružatelja usluga ocjenjuje s obzirom na učinkovitost i zastupljenost njihovog rješenja, ali i na cjelovitost njihove vizije i dugoročne strategije.

Pružatelji usluga su u magičnom kvadrantu raspoređeni u četiri kategorije:

1. **Lideri** (*engl. Leaders*) – visoko pozicionirani na tržištu, dobro izvršavaju trenutne vizije i dobro su pozicionirani u budućnosti.

2. **Vizionari** (*engl. Visionares*) – imaju snažnu razvojnu viziju, ali su još nedovoljno prisutni na globalnom tržištu.

3. **Igrači** (*engl. Niche Players*) – uspješno se fokusiraju na mali specijalizirani segment na tržištu, ali nemaju kompetentna rješenja i slabi su u inovacijama.

4. **Izazivači** (*engl. Challengers*) – dobro izvršavaju današnje potrebe i mogu dominirati velikim segmentima, ali ne pokazuju razumijevanje kretanja smjera tržišta.

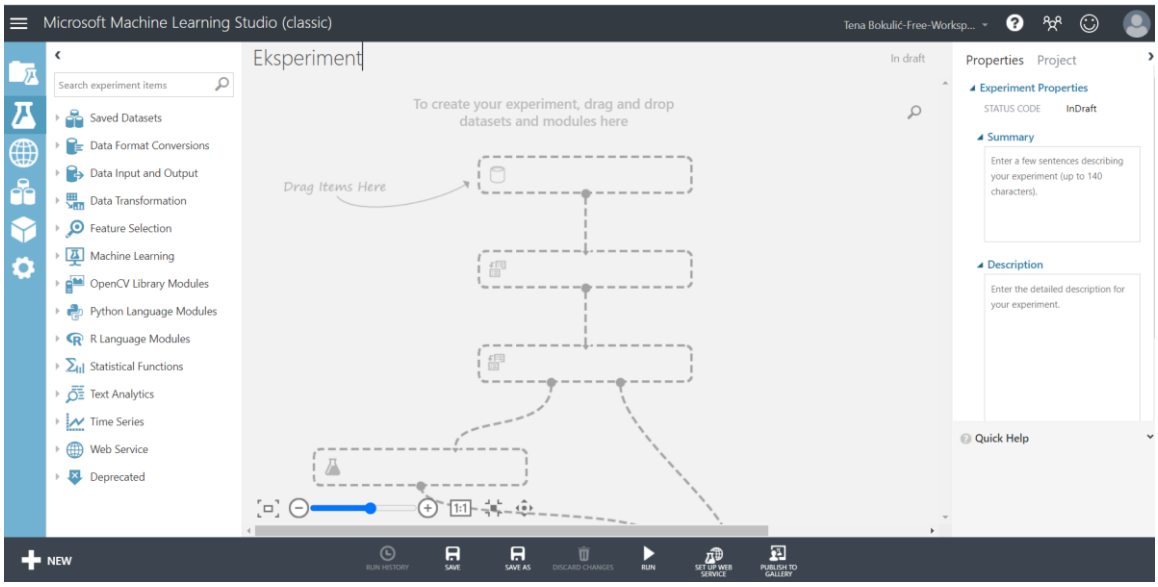
Slika 2.1. Gartnerov kvadrant za analitiku i poslovnu inteligenciju (slika preuzeta iz [3])

Prema Gartnerovom magičnom kvadrantu na tržištu 2021. godine najbolje se plasira *Microsoft*, zatim *Tableau* te *Qlik.*

S obzirom da će se u studijskom primjeru primjene prediktivne analize u prodaji koristiti Cloud verzija Microsoft alata za prediktivnu analitiku, kratko ću se osvrnuti na sami alat i njegove funkcionalnosti.

2.3.1. **Microsoft Azure Machine Learning Studio**

Microsoft Azure Machine Learning Studio je Cloud platforma koja podatkovnim analitičarima i programerima omogućuje jednostavnu izradu prediktivnih modela kroz web sučelje. Microsoft Azure ML Studio je *drag&drop* alat koji koristimo za izradu, testiranje i implementaciju prediktivnih analitičkih rješenja te implementiranje istih unutar poslovnih aplikacija.

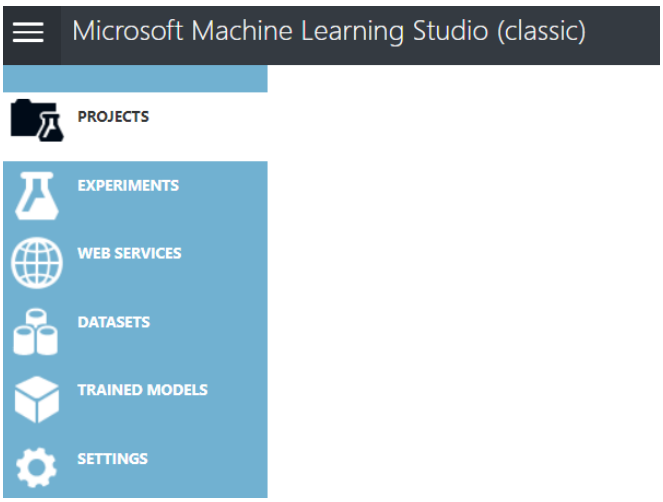


Slika 2.2. Alat MS Azure Machine Learning Studio

Microsoft Azure ML Studio dostupan je besplatno uz određene limite koji ne bi trebali predstavljati problem ni zahtjevnijim korisnicima. Samo uz registraciju moguće je besplatno koristiti osnovne funkcionalnosti alata među kojima je razvoj modela i to na neograničeno vrijeme, što je i bio jedan od razloga odabira ovog alata. Dostupnost ovog alata ne znači da po kvaliteti zaostaje među drugim alatima na tržištu, već naprotiv, u usporedbi s drugima, Microsoft Azure ML Studio donosi zavidan broj naprednih metoda strojnog učenja koje se mogu koristiti za izradu prediktivnih modela. Pretplata donosi dodatne funkcionalnosti u smislu uporabe API sučelja i web-usluga te deploy u informacijski sustav organizacije [26].

Microsoft Azure ML Studio sastoji se od slijedećih komponenti:

- **Projekti** (*engl. Projects*) - kolekcije eksperimenata, baza podataka, bilježnica i drugih resursa koji predstavljaju pojedini projekt.
- **Eksperimenti** (*engl. Experiments*) - eksperimenti, odnosno modeli koje korisnik kreira.
- **Web servisi** (*engl. Web services*) - web servisi koji su razvijeni od kreiranih eksperimenata.
- **Baze podataka** (*engl. Database*) - baze podataka koje korisnik dodaje u studio.
- **Trenirani modeli** (*engl. Trained models*) - modeli koje korisnik uči u eksperimentima i pri završetku sprema.



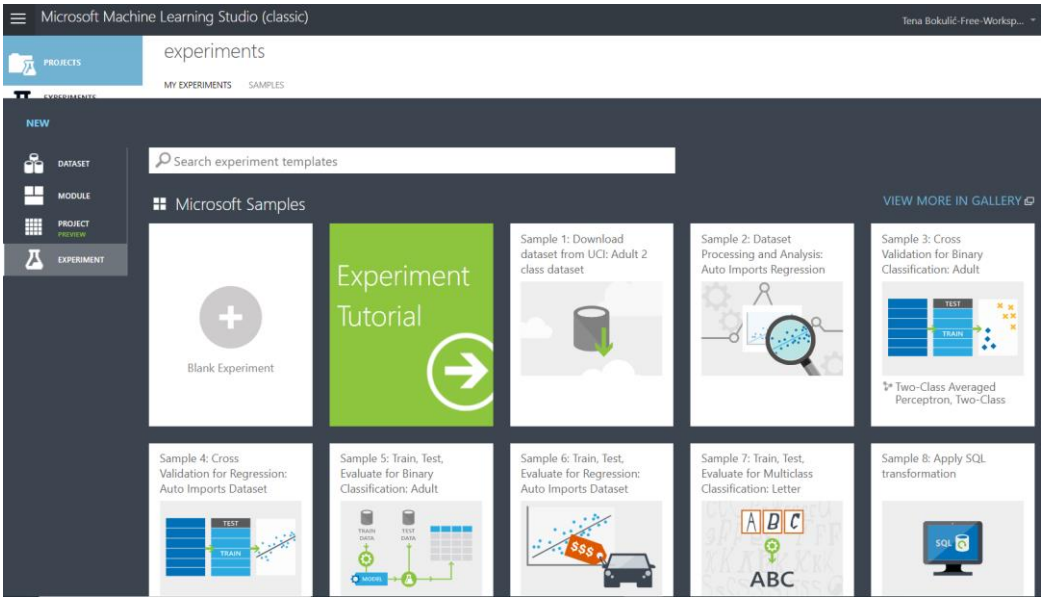
Slika 2.3. Sučelje MS Azure Machine Learning Studia

Proces implementacije prediktivne analize u poslovanje, pa tako i u MS Azure ML Studio, sastoji se od nekoliko koraka koji će biti detaljnije opisani u poglavlju 2.4.

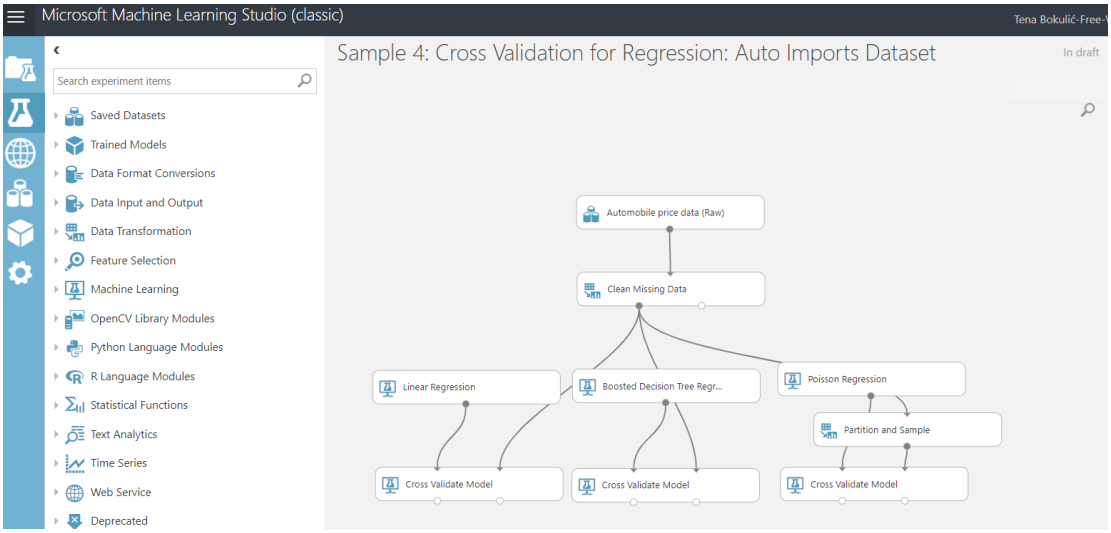
Prvi i najbitniji korak je postavljanje poslovnog cilja, zatim slijedi prikupljanje i transformacija podataka koji je u većini slučajeva i vremenski najzahtjevniji proces. MS Azure ML Studio nudi mogućnost učitavanja podataka u *.csv* ili *.txt* formatu, Excel dokumentu ili iz skladišta podataka. Također to mogu biti i baze podataka iz CRM sustava, odnosno sustava koji služe za upravljanje odnosa s kupcima. Podaci su ključna stavka svakog eksperimenta, a osim količine istih, bitna je i njihova kvaliteta. MS Azure ML Studio pojednostavljuje proces transformacije podataka

primjenom modula za procesiranje podataka kao što su filtracija, normalizacija i grupiranje. Sadrži i mnoštvo unaprijed instaliranih paketa te za potrebe detaljnijih analiza omogućava korištenje Python i R skripti. Nadalje, alat sadrži niz metoda za predikciju, neke od njih su: logistička regresija, Bayesove mreže, metoda potpornih vektora za klasifikaciju, linearna regresija, regresijska stabla, Bayesova linearna regresija i druge metode za predviđanje, zatim metoda klasteriranja, traženja anomalija (npr. za otkrivanje prijevara), sustavi preporuka, prognoze vremenskih nizova i druge [26]. Nakon kreiranja odgovarajućeg modela, alat nudi mogućnost evaluacije istog kako bi se dobila predodžba o uspješnosti. Kreirani modeli mogu se objaviti kao web servisi te ih možemo vizualizirati koristeći alate kao što su Microsoft Power BI ili Excel.

MS Azure ML Studio nudi i razne predloške koji nam pomažu u izradi s već ponuđenim načinima uvoza podataka, izbora metoda za preprocesiranje, samu izgradnju modela kao i dodatnih izvješća.



Slika 2.4. Predlošci koji mogu olakšati izgradnju modela u MS Azure ML Studiju



Slika 2.5. Primjer predloška iz alata MS Azure Machine Learning Studio

MS Azure ML Studio je odličan alat za analitičare podataka ili developere koji žele u kratkom vremenu razviti prediktivni model, a da im pri tome nije nužna pomoć nekog eksperta za prediktivnu analitiku. Prednost alata definitivno je velika fleksibilnost u uvozu i izvozu podataka te brojnost metoda i obrada podataka koje se mogu napraviti. Jednostavnost izrade prediktivnih modela u Cloud inačici te mnoštvo gotovih primjera također je prednost ovog alata.

Kao jedan od nedostataka ovog alata može se uočiti nedovoljna intuitivnost koraka pri izgradnji modela i izboru metoda, pri čemu korisnik treba dobro poznavati metodologiju kako bi izgradio model. Uz navedeno, postoji i "ograničenje" na količinu podataka koja se može uvesti (10GB) i nedostatak podrške za neke izvore podataka.

2.4. Implementiranje prediktivne analitike u poslovanje

Neke poslovne organizacije koriste prediktivnu analitiku desetljećima. One imaju dobro uspostavljene IT infrastrukture, sustave i procese za izgradnju i implementaciju predvidljivih modela.

Prema Finlay [31] za organizacije koje su nove u prediktivnoj analitici, izgradnja prediktivnog modela prvi put, može biti vrlo težak zadatak. Ponekad je potrebno nekoliko mjeseci da potrebni podaci budu spremni i da prediktivni proces može započeti. Samo prikupljanje pravih podataka i izrada prediktivnog modela predstavljaju lakši dio projekta prediktivne analitike. Razlog tome je, što na kraju dana, prediktivni model nije ništa više nego skup jednadžbi zarobljenih u proračunskoj tablici ili drugom softveru. Model treba biti operacionaliziran kako bi bio od koristi.

Glavni izazov za organizacije koje nikada prije nisu koristile prediktivnu analitiku je da prihvate korištenje automatskog odlučivanja. Potrebno je postaviti infrastrukturu kako bi omogućili prediktivnim modelima da postanu ključni dio sposobnosti organizacijskog donošenja odluka i da budu zadovoljni odlukama koje se donose na temelju predviđanja modela. Ovo uključuje uspostavljanje postupaka upravljanja kako bi se osiguralo da se odluke temeljene na modelu ponašaju prema namjeravanoj svrsi i da se odluke donositelja odluka ne zanemaruju ili nadilaze, osim u određenim unaprijed dogovorenim situacijama [31].

Prediktivna analitika može biti nešto što može dodati vrijednost na ono što organizacija radi, ali s druge strane i ne mora. Primjerice, kako navodi Finlay [31] korisna analogija je izgradnja automobilskog motora. Inženjeri mogu provesti dosta vremena na izgradnji vrlo moćnog i učinkovitog motora. Međutim, taj motor neće pružiti nikakvu korist ukoliko netko nije razmišljao o dizajnu automobila, u koji bi stavili motor, montaži motora i tako dalje. Bez ostatka automobila, motor je beskoristan. Isto vrijedi i za prediktivnu analitiku. Dok ne postoji poslovni proces u kojem će se postaviti model, model neće imati puno koristi.

Za uspješno integriranje prediktivne analitike u poslovni proces, literatura navodi više mogućnosti implementacije prediktivne analitike. Prema Siegelu [7] i Iffertu [19] proces od šest koraka se pokazao kao najefikasniji. Proces se temelji na jednom od najpopularnijih

modela za rudarenje podataka - **CRISP-DM model** (*engl. Cross Industry Standard Process for Data Mining*).

CRISP-DM model sadrži šest koraka i veze među koracima su dvosmjerne što znači da se možemo vraćati iz trenutnog u prethodni korak i obrnuto. Veze također kreiraju životni ciklus koji nam govori, da se proces nakon dobivanja traženih informacija vraća na prvu fazu, te se procjenjuje jesu li ti podaci točni, odnosno zadovoljavaju li kriterije prve faze, pa u slučaju da rezultati nisu u skladu poslovnog cilja, postupak se ponavlja [10].

Slika 2.6. Model procesa CRISP-DM (slika preuzeta iz [34])

Poslovno razumijevanje (*engl. Business Understandig*)

Početna faza CRISP-DM modela čini spoznaja oko potrebe za implementiranjem prediktivne analitike u poslovanje te utvrđivanje prilika i prednosti koje alati prediktivne analitike donose. Postavlja se pitanje za što će se to prediktivna analitika koristiti. Dakle, treba postojati pravi problem kojem prediktivna analitika može pomoći u rješavanju. Navedeno se može odnositi na pitanja vezana uz tržišni rast, određivanje proizvodnog asortimana prema potrebama kupaca, utvrđivanje najprofitabilnijih dobavljača i sl. U ovom koraku treba jasno definirati početne situacije i ciljeve te odrediti očekivanja po pitanju ishoda nakon implementacije. Potrebno je utvrditi odmah raspolaže li organizacija sa osobljem sposobnim za uspostavljanje i održavanje metoda koji će koristiti te u konačnici tumačiti rezultate, i moći donijeti zaključke temeljene na njima. Bitno je i da organizacija prihvati prediktivnu analitiku te joj dopusti da pokreće automatizirano odlučivanje.
Zadnji korak u ovoj fazi je jasno definiran cilj i način razvoja prediktivnog modela.

Razumijevanje podataka (*eng. Data Understandig*)

Nakon što je upoznato problemsko područje i određeni su ciljevi i kriteriji uspješnosti potrebno je odrediti, prikupiti i pripremiti podatke koji će se tijekom procesa analizirati. Faza razumijevanja podataka počinje sa inicijalnim skupljanjem podataka. Ti podaci mogu biti jednostavne baze podataka, tekstovi, Excel dokumenti i drugo. Sve podatke bez obzira na izvor potrebno je ujednačiti i povezati na smislen način. Ključno je iz podataka zaključiti jesu li oni potrebne kvalitete i kvantitete za rudarenje podataka. Specijalisti nakon pregleda podataka iste moraju istražiti detaljnije, a to rade preko traženja istih varijabli i uzorka u podacima te ih testiraju na jednostavnim hipotezama. Nakon testiranja specijalisti se moraju uvjeriti da su podaci realni i da nemaju anomalije npr. da im ne nedostaju neke vrijednosti [33].
Rezultat ove faze je osnovni skup podataka koji će se u slijedećoj fazi pripremiti i pročistiti.

Priprema podataka (*engl. Data Preparation*)

Nakon što smo odlučili koji su naši poslovni cilj/-evi, vrijeme je da zavirimo u izvore podataka koji nam mogu dati odgovore na naša poslovna pitanja. To uglavnom mogu biti podaci u *.csv* ili *.txt* formatu, Excel-ice ili podaci iz skladišta podataka. Također to mogu biti i baze podataka iz CRM sustava, odnosno sustava koji služe za upravljanje odnosa s kupcima. Ova faza zahtijeva najviše vremena te podrazumijeva detaljnu pripremu i odabir podataka za daljnju analizu. Obuhvaća početno prikupljanje podataka, opis podataka, istraživanje podataka i verifikaciju kvalitete podataka iz projektne dokumentacije. Utvrđuje se postoje li izvori strukturiranih i nestrukturiranih podataka koji će omogućiti izgradnju prediktivnih modela. Kako je već spomenuto, preporučuje se korištenje *Big Data* i skladišta podataka budući da se informacijske tehnologije razvijaju u smjeru sve većeg korištenja navedenih izvora podataka. U ovom koraku se traže mogući izvori podataka pomoću kojih će se doći do ostvarivanja ciljeva postavljenih u prethodnom koraku.
Nakon što su utvrđeni izvor podataka i tipovi podataka, određuje se alat koji će se koristiti za pohranu i analizu tih podataka. Ovisno o cilju koji se postavio i na pitanja na koja se želi odgovoriti, utvrđuju se npr. minimalne, srednje i maksimalne vrijednosti varijable, prema kojim kriterijima se grupiraju varijable i slično. Nadalje se utvrđuje koji od dobivenih podataka se smije koristiti, te ima li podataka koji otkrivaju više od dozvoljenog. U ovom koraku se utvrđuju problemi do kojih je moglo doći nejasnim postavljanjem cilja, koje rezultira skupom nepovezanih podataka koji u konačnici neće dati željene i objektivne rezultate. Također se u ovoj fazi utvrđuju prve skrivene povezanosti među podacima. Nakon što se podatke analiziralo prema željenim parametrima, prevodi ih se u odgovarajući format za daljnju analizu i stvaranje modela, popularan format su jednostavne tablice.

Krajnji rezultat ove faze je pripremljen i pročišćen skup podataka spreman za rudarenje podataka.

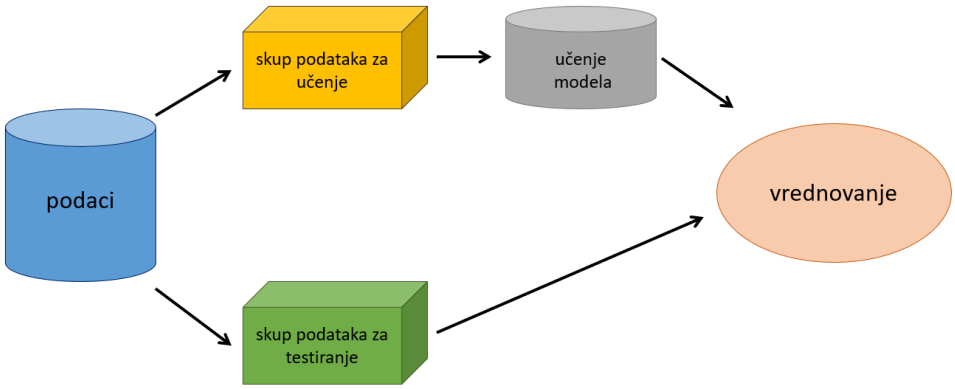
Modeliranje (*engl. Modeling*)

U prošlom koraku podaci su pripremljeni za daljnju analizu koja se odvija u odgovarajućem alatu za koji se organizacija odluči. Organizacije su u mogućnosti stvarati modele prema željenim varijablama, ovisno o cilju analize, te dobivaju uvid u uzorke ponašanja kupaca i dobavljača, predviđanja prodaje, životnog ciklusa proizvoda itd. Rješavanje jednostavnijih problema je moguće primjenom skupa formula unutar Excela, dok se za složenije pothvate primjenjuju posebno programirani alati. Cilj ove faze je pronaći odgovarajući model za predviđanje. Potrebno je odabrati metodu rudarenja podataka koja će u najvećoj mjeri iskoristiti dostupne podatke i omogućiti postizanje zadanih ciljeva. Rijetko se odlučuje samo za jednu metodu već se obično koristi više metoda. Tek nakon njihove usporedbe na konkretnom uzorku podataka primjenjuje se odgovarajuća metoda. Detaljnije o metodama rudarenja podataka u poglavlju 3.1.

Prilikom provedbe odgovarajuće metode prikupljeni podaci razdvajaju se u dvije skupine podataka, *skup podataka za učenje* i *skup podataka za testiranje*.
Iz dostupnog skupa podataka, potrebno je nasumično odrediti koji podaci pripadaju skupu podataka za učenje, a koji skupu podataka za testiranje. U literaturi [21] nisu precizno definirani omjeri podjele podataka u dva skupa; spominju se omjeri 70:30 (70% podataka je u skupu za učenje, a 30% u skupu za testiranje), kao i pridruživanje dvije trećine podataka skupu podataka za učenje, dok preostalu jednu trećinu čini skup podataka za testiranje. Skup podataka za učenje, kako i sam naziv govori, koristi se za učenje modela. Skup podataka za testiranje koristimo za testiranje učinkovitosti modela i vrednovanje njegovih rezultata. Predstavlja nezavisni skup podataka s kojim se model po prvi put susreće.

Podjela podataka u skup za učenje i za testiranje smanjuje rizik da se model bazira isključivo na podacima za učenje. Svrha modela je utvrditi skrivene povezanosti i veze među varijablama, te izbjeći pogreške. Overfitting je termin koji opisuje pojavu kada se model upotrebljava na isključivo jednom skupu podataka, iz čega slijedi da su povezanosti i veze među varijablama isključivo precizne za navedeni jedan skup podataka. Isti model

vjerojatno neće davati jednako precizne rezultate bude li se primjenjivao na drugim podacima. Korištenjem skupa podataka za testiranje se eliminira pojava overfitting-a, te je prediktivni model precizniji. Također, skup daje prvu sliku o tom kako će model funkcionirati kad se stavi u uporabu, ako model daje dobre rezultate na temelju podataka iz skupa testiranja, vjerojatno će pružati jednako kada se stavi u uporabu.
Nakon obrade podataka, alat pruža numeričke prikaze rezultata, grafikone i vizualne prikaze rezultata te tekstualne dijelove gdje se objašnjava rezultat.



Slika 2.7. Modeliranje prediktivnog modela

Evaluacija (*engl. Evaluation*)

Evaluacija modela je ključna faza u procesu otkrivanja znanja. Tijekom ove faze ocjenjuje se u kojoj mjeri kreirani model može riješiti zadani problem. Ako model nije zadovoljavajuće riješio zadan problem, potrebno je promijeniti primijenjenu metodu ili uskladiti skup podataka sa zahtjevima izabrane metode. U tom slučaju slijedi povratak na neku od prethodnih faza s ciljem korektivnih radnji. Naravno, sve ove faze su iterativne i nema konačnog rezultata iz prvog puta te dobra praksa pokazuje da iste podatke moramo rudariti na različite načine kako bismo dobili optimalni rezultat. Puno se puta događa da se revidiraju poslovni ciljevi jer smo rudarenjem podataka dobili neočekivane rezultate. Ako je model zadovoljavajuće riješio zadani problem on se može implementirati na prethodno definiran način.

Implementacija (*engl. Deployment*)

Kreirani model prolazi kroz proces implementacije. Nakon implementacije potrebno je ocijeniti učinkovitost modela. Navedeno se obavlja sa poslovnog i tehničkog stajališta. Menadžer treba procijeniti koliko su saznanja dobivena modelom relevantna za poslovanje i prvobitno postavljenje ciljeve. Implementacija bi trebala olakšati donošenje odluka i ubrzati vrijeme reagiranja organizacije. Sa tehničkog stajališta ocjenjuju se jesu li odabrane odgovarajuće metode za prediktivni model te je li vremenski period u granicama unaprijed određenog. Stvaranje konkurentske prednosti nakon prihvaćanja rezultata dobivenih postupkom je glavni pokazatelj uspješnosti modela. Jednom otkriveno i zabilježeno znanje može se naknadno koristiti na bilo koji način kada to bude potrebno.
Kako bi model dugoročno bio efikasan potrebno ga je kontinuirano održavati. Modeli imaju tendenciju zastarjeti u slučaju da se ne pokreću u određenim vremenskim periodima, te ako se promjene podataka ne prate. U slučaju da dođe da promjena okolnosti unutar kojih organizacija posluje, potrebno je model ponovno trenirati s podacima relevantnim za novonastalu situaciju.
Promjene okolnosti podrazumijevaju nove trendove na tržištu, sustizanje organizacija od strane konkurencije ili promjena poslovnih ciljeva.

Možemo zaključiti kako je cijeli proces iterativan, tj. ponavlja se mnogo puta. Većina analitičara identificira i ispituje mnoge kombinacije varijabli kako bi vidjeli koje imaju najveći utjecaj. Većina ih započinje korištenjem statističkih i OLAP alata za prepoznavanje značajnih trendova u podacima kao i prethodnog analitičkog rada interno ili uz stručne konzultante. Također, može se razgovarati s poslovnim korisnicima koji su bliski predmetu i oslanjati se na vlastito znanje o organizaciji da se pridruže najvažnijim varijablama za uključivanje u model. Kao rezultat toga, većina analitičara uklanja popis varijabli od nekoliko stotina u početnoj verziji do nekoliko desetaka u konačnom modelu. Usput, testiraju se različite metode rudarenja podataka kako bi se vidjelo što najbolje funkcionira na skupu podataka za učenje. Moguće je na različite načine dodati nove vrste podataka ili rekombinirati postojeća polja radi poboljšanja točnosti modela. Ovaj iterativni proces omogućuje stvaranje modela koji zahtijevaju intenzivnu i dugotrajnu upotrebu [40].

2.5. Razlozi korištenja prediktivne analitike

Razloga i prednosti za korištenje prediktivnih modela ima više. Finlay [31] navodi tri osnovne: brzina, bolje i kvalitetnije prognoziranje te dosljednost. Prediktivni modeli sve se više koriste kako bi zamijenili i/ili dopunili stručnu prosudbu i odlučivanje na svim područjima. To je zbog toga što prediktivni modeli nastoje biti [31]:

▪ *Mnogo točniji od ljudskih eksperata*. Prediktivni modeli omogućavaju bolje prognoziranje i predviđanje nego ljudi budući da će uvijek generirati ista predviđanja nad istim podacima, što nije slučaj kod ljudi.

▪ *Objektivni*. Za razliku od ljudi, prediktivni modeli ne prikazuju predrasude prema ljudima zbog njihovog spola, rase, invalidnosti i dr. Modeli mogu prikazati pristranost, ali ako imaju tendenciju dati određenim pojedincima i grupama veće ili niže rezultate od populacije u cjelini, to je iz razloga što se temelji na čvrstim dokazima, a ne na temelju nedokazanih predznaka ili stereotipa.

▪ *Brzi*. Kada se prediktivni modeli koriste kao dio automatiziranog sustava donošenja odluka, milijuni kupaca mogu se rješavati i vrednovati u samo nekoliko sekundi. Da ljudi rade takve prosudbe, u većini slučajeva to bi bilo neizmjerno skupo i dugotrajno.

▪ *Jeftini*. Jednom razvijeni, prediktivni modeli su često povoljniji za implementaciju od njihovih ljudskih kolega. Prediktivna analiza štedi na svim neraspoređenim resursima, dodatno štedeći i na troškovima i na vremenu.

Danas se prediktivna analitika koristi za mnogobrojne probleme. Činjenica je da se koristi u gotovo bilo kojem aspektu života koji uključuje odlučivanje o velikom broju ljudi [31].

Poglavlje 3

Rudarenje podataka

Rudarenje podataka (engl. *Data Mining*), odnosno otkrivanje znanja u bazama podataka je netrivijalni postupak pronalaženja novih, valjanih, razumljivih i potencijalno korisnih informacija. Rudarenje podataka provodi se na velikim količinama podataka iz baza podataka da bi se iz njih otkrilo novo znanje i potom iskoristilo za donošenje boljih poslovnih odluka. To je podatkovno vođeno pronalaženje zanimljivih informacija, koje se znatno razlikuje od tradicionalnog dobivanja informacija ad-hoc pretraživanjima, standardnim izvješćivanjima ili analitičkom obradom podataka, pa se može nazvati i *rudarenjem podataka*.

Rudarenje podataka i njegova primjena u otkrivanju znanja su nove tehnike koje predstavljaju neizostavan dio suvremene analize podataka te su još u fazi intenzivnog razvoja. Ne postoji standardna praksa rudarenja podataka za razliku od na primjer primjene statističkih postupaka. Postoje samo pozitivna i manje pozitivna iskustva sa određenim postupcima i njihovom primjenom na konkretnim domenama.

Iako su metode koje se koriste u rudarenju podataka već poznate iz statistike, matematike i računarstva, za intenzivnije uvođenje rudarenja podataka presudni su bili razvoj informacijske tehnologije i povećana potreba boljeg iskorištenja velike količine podataka koji se u organizacijama prikupljaju godinama. Rudarenjem podataka koriste se poslovni analitičari, kojima na raspolaganju stoji niz programskih alata jednostavnih za upotrebu, za koje nije neophodno detaljno poznavanje metoda rudarenja podataka.

Postoji mnogo sličnih definicija rudarenja podataka tako prema izvoru [29] rudarenje podataka je automatski proces otkrivanja korisnih informacija u velikim repozitorijima podataka. Tehnike rudarenja podataka su razvijene kako bi pronašle nove korisne obrasce koji bi inače ostali nepoznati. Prema izvoru [6] rudarenje podataka je analiza (često velikih) promatranih podataka kako bi pronašli neotkrivene veze i interpretirali podatke na novi način koji je razumljiv i koristan vlasniku podataka.

Friedman [15] navodi kako je rudarenje podataka netrivijalni proces identificiranja validnih, novih, potencijalno korisnih i ultimativno razumljivih obrazaca u podacima.
U rudarenju podataka obično se rješavaju ove vrste problema: sumiranje podataka, segmentacija, klasifikacija, asocijacija (prepoznavanje uzoraka) i predviđanje.

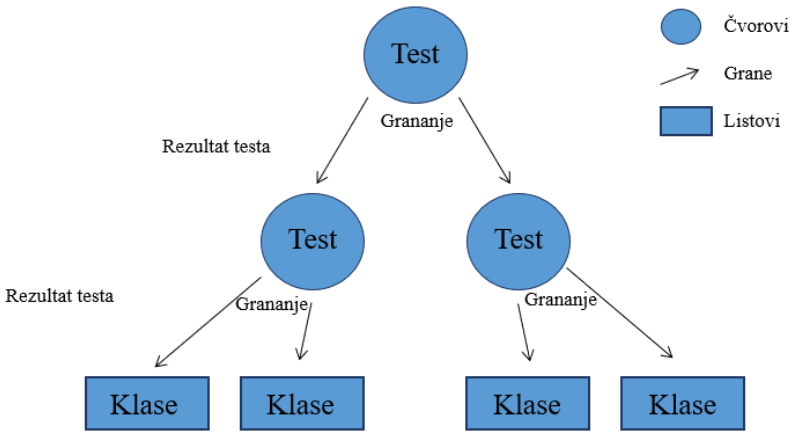
3.1. Metodologija rudarenja podataka

Za potrebe rudarenja podataka (traženja skrivenih zakonitosti) mogu se koristiti različite statističke metode, metode strojnog učenja, asocijacijska pravila i druge metode.
Pod statističke metode podrazumijevamo deskriptivnu i vizualizacijsku tehniku, klaster analizu, korelacijsku analizu, diskriminantnu analizu, faktorsku analizu, regresijsku analizu, logističku regresiju i druge.
Metode strojnog učenja koje se najčešće koriste su stabla odlučivanja, neuronske mreže, metoda potpornih vektora, genetički algoritmi i druge [6].
Posljednjih godina se događa veliki rast i konsolidacija područja rudarenja podataka. Rudarenje podataka kao proces izvlačenja informacija i prepoznavanja obrazaca se sve više koristi u raznim industrijama za rješavanje poslovnih problem dok je početkom tisućljeća njegova primjena u većini bila u akademske svrhe i znanstvena istraživanja. Rast važnosti područja teži utvrđivanju standarda i metoda provođenja rudarenja podataka. Tako se razvijaju dvije metodologije CRISP-DM i SEMMA. Obje se javljaju kao industrijski standardi i definiraju niz uzastopnih koraka kojim se implementira primjena rudarenja podataka. U prethodnom poglavlju 2.4. već je opisana CRISP-DM metodologija za izvlačenje informacija iz podataka. Ovisno o problemu koji se rješava, potrebno je izabrati odgovarajuću metodu za rudarenje podataka. Za potrebe predviđanja prodaje u studijskom primjeru koristiti ćemo dvije prediktivne metode strojnog učenja: stablo odlučivanja i neuronska mreža.

3.2. Stablo odlučivanja

Stablo odlučivanja (*engl. Decision Tree*) jedno je od najčešće korištenih metoda analize podataka. Primjenjuje se za razvrstavanje (*engl. Classification*), predviđanje *(engl. Prediction*), procjenu vrijednosti (*engl. Estimation*), grupiranje (*engl. Clustering*), opisivanje podataka i vizualizaciju. Stabla odlučivanja su prediktivni modeli koji na temelju podataka izvode njihove veze u cilju dobivanja izlaznih vrijednosti. Kao takvi modeli koriste se u rudarenju podataka odnosno traženju skrivenih veza između podataka. Takva stabla temelje se na podacima, a ne na odlukama eksperta [28].
Prema Rokachu i Maimonu [20] stablo odlučivanja jest klasifikator izražen kao rekurzivno particioniranje prostora predviđanja.

Stablo odlučivanja je aciklički usmjereni graf čija se vrhovi zovu čvorovi (*engl. Nodes*), a bridovi grane (*engl. Branches*), čvorovi bez potomaka - listovi (*engl. Leaves*), a korijenski čvor je jedini čvor bez roditelja (*engl. Root node*).



Slika 3.1. Konceptualni model stabla odlučivanja

Svi čvorovi sadrže testove atributa koji se generiraju temeljem kriterija grananja i oni predstavljaju načine prema kojima se određuju grananja podataka prema vrijednostima njihovih atributa. Grane prikazuju rezultate ispitivanja, a listovi predstavljaju distribucije klasa.

Specijalni tip stabla odlučivanja je binarno stablo odlučivanja u kojem svaki čvor osim terminalnih listova ima točno dva potomka. Na taj način se opservacije uvijek dijele u točno dva podskupa. Najvažniji zadatak stabla odlučivanja je prikaz svih mogućnosti i definiranje samog problema odlučivanja. Najčešće se primjenjuje prilikom donošenja odluka u pojedinim rizičnim situacijama koje su povezane s poslovnim svijetom. Sastoji se od niza odluka koje su međusobno povezane i svaka ovisi o prethodnoj.

Nayab i Scheid [16], [17] navode kako su prednosti stabla odlučivanja transparentnost, elastičnost i jednostavno korištenje, dok su nedostaci složenost, nezgrapnost, potrebno obrazovanje, troškovi te previše informacija.

Kako bi se jednostavnije prikazala primjena metode stabla odlučivanja u procesu odlučivanja, koristi se primjer predstavljen od strane autora [30].

Primjer 3.1. (Odluka o novom proizvodu)
Poduzetnik razmatra ideju o lansiranju nove linije proizvoda te do sadašnjeg trenutka nije uložio znatna sredstva u njezin razvoj. Prihvaćanje te ideje o lansiranju nove linije proizvoda podrazumjeva ulaganje u visini 200.000 kn bez sigurnosti u poslovni uspjeh novog proizvoda. Poduzetnik procjenjuje da postoji 50% vjerojatnost da potražnja bude dovoljno velika za dobit od 500.000 kn, 30% da potražnja bude mala uz dobit od 100.000 kn i 20% vjerojatnost da novi proizvod neće ostvariti nikakvu potražnju, odnosno da poduzetnik ostvari poslovni gubitak od 200.000 kn. Situacija odlučivanja predstavljena primjerom može se prikazati jednostavnom tablicom odlučivanja (tablica 3.1.).

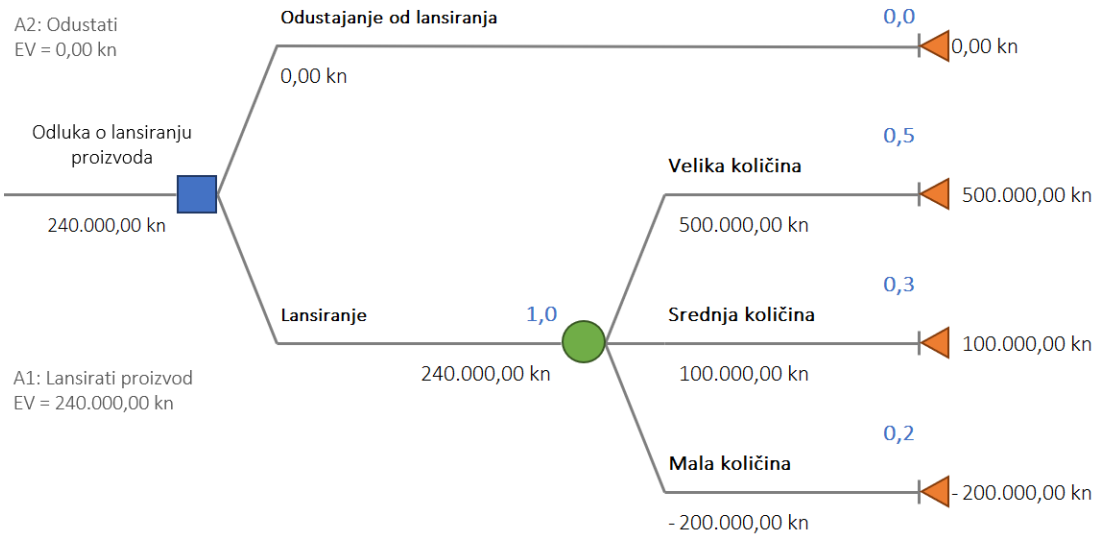
	Velika potražnja (0.5%)	Mala potražnja (0.3%)	Bez potražnje (0.2%)	EV
Lansiranje	500.000,00	100.000,00	-200.000,00	240.000,00
Odustajanje od lansiranja	0	0	0	0

Tablica 3.1. Tablica odlučivanja za problem lansiranja novog proizvoda

Ako primijenimo metodu stabla odlučivanja na zadani problem, prva odluka – lansiranje proizvoda dobiva tri različite grane kao što je prikazano na slici 4.

Simboli koji se koriste u ovom grafičkom modelu su:

- kvadrat - čvor odluke,
- krug - čvor slučaja,
- grane - povezuju čvorove u stablu,
- trokut - krajnji čvor.



Slika 3.2. Stablo odlučivanja za problem lansiranja proizvoda

Čvor odluke je pod kontrolom donositelja odluke, dok čvor slučaja nije pod kontrolom donositelja odluke, odnosno sadrži sve mogućnosti koje se mogu dogoditi te je izvor potencijalnih rizika. Ishodi slučaja navode se na kraju svake grane te su ovisno o alatu prikazani trokutom ili je naznačena numerička vrijednost na kraju grane. Na temelju ishoda i vjerojatnosti slučajeva računa se očekivana vrijednost čvora slučaja. U skladu s kriterijem očekivane vrijednosti, odluka se donosi na način da se u čvoru odluke bira ona grana koja vodi prema čvoru slučaja s najvećom očekivanom vrijednošću [30]. Prema modelu prikazanom na slici 3.2., za problem lansiranja proizvoda ispravna odluka bi bila A1 - lansirati proizvod. Konačna odluka ovisi o tome kako se poduzetnik odnosi prema riziku.

Za nekoga je 20% vjerojatnosti da izgubi 200.000,00 kn prevelik i neprihvatljiv rizik u odnosu na ostale mogućnosti, a za nekoga je taj rizik prihvatljiv.

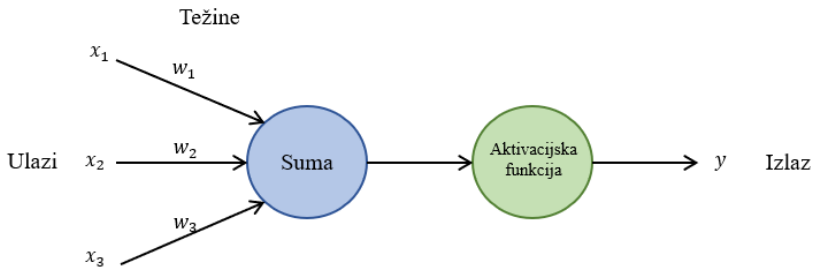
Primjenom metode stabla odlučivanja nije moguće zaobići situaciju povezanih odluka. U nekim djelatnostima nije moguće donositi jednokratne odluke zbog prirode poslovnih procesa. Takve povezane poslovne odluke donose se i modeliraju uz pomoću metode stabla odlučivanja [30].

3.3. Neuronska mreža

Neuronska mreža (*engl. Neural Network*) pripada u inteligentne metode rudarenja podataka čiji je cilj pronaći skrivene veze među podacima. Prema Kevinu Gurney-u neuronska mreža je međusobno povezana nakupina jednostavnih elemenata obrade, jedinica ili čvorova, čiji se načini djelovanja otprilike temelje na neuronima kod živih bića. Sposobnost obrade mreže je posljedica jačine veza među tim jedinicama, a postiže se kroz proces adaptacije ili učenjem iz skupa primjera za učenje [37].

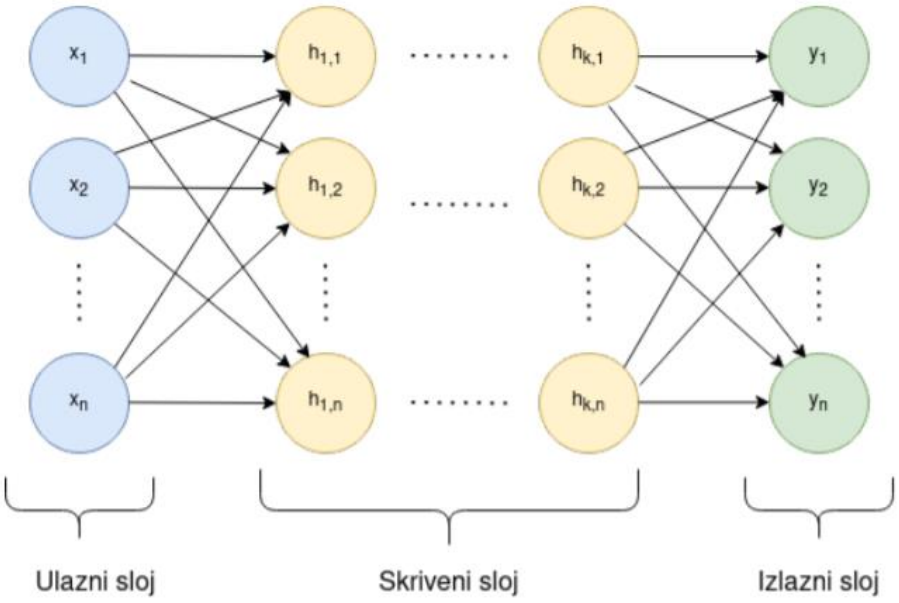
Drugim riječima, neuronske mreže su programi ili hardverski sklopovi koji, najčešće iterativnim postupkom iz prošlih podataka nastoje pronaći vezu između ulaznih i izlaznih varijabli modela, kako bi se za nove ulazne varijable dobila vrijednost izlaza. Obično se umjetne neuronske mreže sastoje od hijerarhije slojeva u kojima su uzduž raspoređeni neuroni. Ulazne i izlazne slojeve čine neuroni koji su povezani s okolinom [36].

Era neuronskih mreža započinje još 1943. godine kada je prvi put prezentiran matematički model biološkog neurona [40]. Nadalje, 1958. godine se u [9] objavljuje perceptron - najjednostavnija neuronska mreža za klasifikaciju uzoraka koji su linearno separabilni. Perceptron se sastoji od jednog umjetnog neurona. Umjetni neuron je jedinica za obradu podataka (varijabla) koja prima ponderirane ulazne vrijednosti $(x_1, x_2, \dots, x_n)$ od drugih varijabli, korištenjem aktivacijske funkcije transformira primljenu vrijednost, te šalje izlaz drugim varijablama. Učenje se odvija promjenom vrijednosti "težina" među varijablama (težine w_i su ponderi kojima se množe ulazne vrijednosti u neki "neuron") [24].



Slika 3.3. Arhitektura perceptrona

Budući da se učenje se odvija samo u dva sloja, perceptron nije mogao rješavati probleme klasifikacije koji nisu linearno djeljivi (npr. XOR problem). Za rješenje tog problema kreiran je više slojni perceptron (*engl. Multilayer Perceptron*) što zapravo dobivamo slaganjem višestrukih perceptrona. Osim ulaznog i izlaznog sloja sastoji se od mnogo takozvanih skrivenih slojeva koji se nalaze između njih (vidi slika 3.4.), oni skladno rade u svrhu što boljeg rješavanja konkretnog problema. Skrivenih slojeva može biti od jedan do *k* te ako ih ima više od tri takva mreža se smatra algoritmom strojnog učenja.



Slika 3.4. Primjer prikaza neuronske mreže (slika preuzeta iz [4])

Rad neuronske mreže može se opisati tako da svaki čvor promatramo kao perceptron što znači da je za dobivanje određene vrijednosti iz čvora potrebno napraviti sumu svih ulaza u taj čvor pomnoženih s njihovim pripadajućim težinama. Nakon toga pomoću aktivacijske funkcije, kojoj se predaje prethodna suma, određuje se izlaz čvora. Takva procedura ponavlja se za svaki čvor u svim slojevima te se dobiva rezultat u izlaznom čvoru.

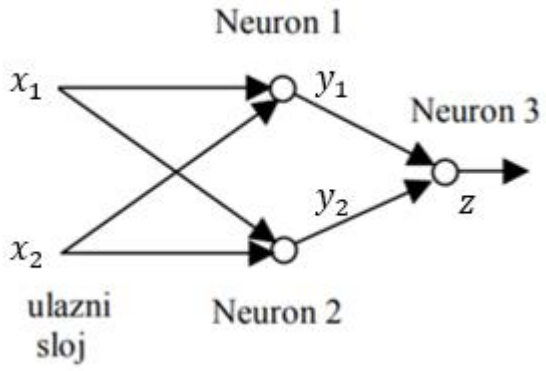
Primjer 3.2. (XOR problem)

Ekskluzivni ILI (*engl. XOR*), poznati je logički problem Booleove algebre. Najjednostavniji je primjer s linearno nedjeljivim uzorcima [32] i zato je često primjenjivan za ispitivanje svojstava različitih modela umjetnih neuronskih mreža. Dokazano je da jednoslojna perceptronska mreža ne može riješiti ovako jednostavan problem te se upravo zato uvode skriveni slojevi. Tablica istine ove logičke operacije ekskluzivno ILI prikazana je tablicom 3.2.

ulaz		*izlaz*
$\boldsymbol{x_1}$	$\boldsymbol{x_2}$	$\boldsymbol{x_1 \oplus x_2}$
0	0	**0**
0	1	**1**
1	0	**1**
1	1	**0**

Tablica 3.2. XOR problem – tablica istine

Ulaz je određen s dvije binarne varijable koje mogu poprimiti vrijednost nula ili jedan. Izlaz se sastoji od jedne binarne varijable koja treba poprimiti vrijednost jedan ukoliko je samo jedna od binarnih ulaznih varijabli jednaka jedan, odnosno nula ukoliko obje ulazne varijable imaju istu vrijednost. Budući da ne postoji jednostavni perceptron koji realizira logičku funkciju XOR, problem možemo riješiti na primjer s dva skrivena neurona i jednim izlaznim neuronom (vidi slika 3.5.).



Slika 3.5. Primjer realizacije XOR problema

Definirajmo ulazne vrijednosti skrivenog sloja sa:

$$y_1 = g\left(-x_1 + x_2 - \frac{1}{2}\right),$$
$$y_2 = g\left(x_1 - x_2 - \frac{1}{2}\right)$$

i vrijednost izlaznog sloja sa:

$$z = g\left(y_1 + y_2 - \frac{1}{2}\right),$$

pri čemu je g funkcija (*tzv. funkcija skoka*) definirana na sljedeći način:

$$g : \mathbb{R} \to Y = \{0,1\}$$
$$g = \begin{Bmatrix} 1\,, ako\ je\ z\ \geq\ 0 \\ 0\,, ako\ je\ z\ <\ 0 \end{Bmatrix},$$

to jest vrijedi:

$$g\big(s(x)\big) = g\big(s(x_1, \dots, x_n)\big) = \left\{ \begin{matrix} 1\,, ako\ je\ \sum_{j=1}^{n} \omega_j x_j\ \geq\ \theta \\ 0\,, ako\ je\ \sum_{j=1}^{n} \omega_j x_j\ <\ \theta \end{matrix} \right\}.$$

Na ovaj način dobivamo mrežu koja realizira funkciju XOR.

Rudarenje podataka temeljeno na neuronskim mrežama počinje "učenjem" mreže pomoću podataka za koje je poznata vrijednost koju želimo prognozirati. Nakon toga se naučeno znanje mreže provjerava na drugoj skupini testnih podataka, za koje je također poznata vrijednost koja se želi prognozirati. Postupak učenja i provjere ponavlja se sve dok rezultati provjere ne budu zadovoljavajući. Nakon toga je neuronska mreža spremna za upotrebu, tj. za prognoziranje nama nepoznatih vrijednosti. Učenje u biološkim sustavima podrazumijeva prilagođavanje na sinaptičkim vezama koje postoje među neuronima, a to pravilo vrijedi i za umjetne neuronske mreže.

Prednosti neuronske mreže su slojevi koji omogućuju da se rezultat pojedinog sloja dodatno obrađuje te na taj način stvara kompleksni sustav. Razlozi što neuronske mreže često daju bolje rezultate nego statističke metode leže u njihovoj mogućnosti da analiziraju nedostatne podatke, podatke sa smetnjama, zatim da uspijevaju rješavati probleme koji nemaju jasno jednoznačno rješenje, te da uče na prošlim podacima. Neuronske mreže su robusne, tj. neosjetljive su na pogreške u podacima i mogu uočavati neke opće značajke predočenih podataka, tj. mogu generalizirati. Zbog takvih prednosti neuronske mreže su pokazale uspjeh u predviđanjima različitih serija podataka koje imaju visok stupanj variranja i fluktuacije. Od nedostataka neuronskih mreža potrebno je spomenuti nedostatak testova statističke značajnosti modela neuronskih mreža i procijenjenih parametara te dugotrajnost algoritama treniranja koji na taj način ne osiguravaju konvergenciju. Nadalje, ne postoje utvrđene paradigme za odlučivanje o tome koja arhitektura neuronskih mreža je najbolja za određene probleme i tipove podataka, te ovaj rad testira više različitih algoritama neuronskih mreža na jednom problemu. U odnosu na stabla odlučivanja, glavni nedostatak neuronskih mreža je njihov rad po principu "crne kutije". Neuronska mreža nam služi za otkrivanje i primjenu neke zakonitosti u podacima, no ne omogućava nam da jasno artikuliramo tu zakonitost i ne objašnjavaju pravila po kojima je određena odluka donesena. Unatoč navedenim nedostacima, rezultati mnogih istraživanja pokazuju da neuronske mreže mogu riješiti gotovo sve probleme učinkovitije nego tradicionalne metode modeliranja i statističke metode.

Poglavlje 4

Studijski primjer – Primjena prediktivne analize u prodaji

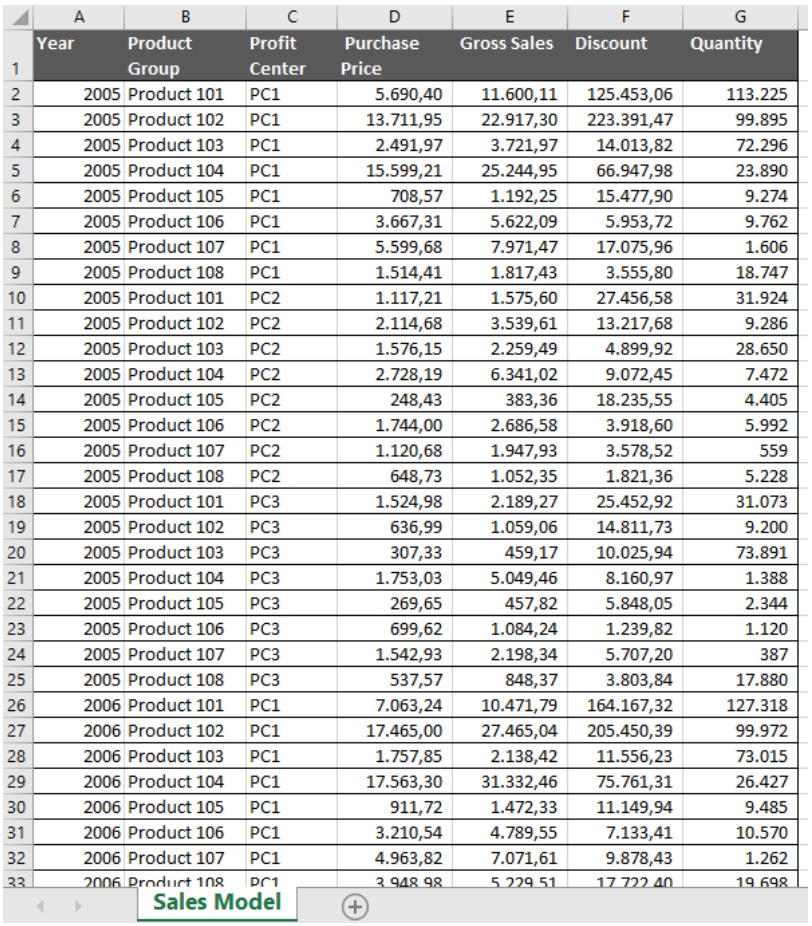
U prošlim smo poglavljima prikazali što je to prediktivna analitika – objasnili smo pojmove koje vežemo uz pojam prediktivne analize i objavili pregled metoda i alata koji se primjenjuju u tom području. Ukratko utvrdili smo da je prediktivna analitika dio podatkovne analitike uz deskriptivnu, dijagnostičku i preskriptivnu analitiku, a cilj joj je predvidjeti buduće vrijednosti nekih pojava, snagu i smjer veza, trendove, uzorke i izuzetke. Rezultat prediktivne analitike su modeli koji pomažu pri donošenju strateških odluka, naprimjer o uvođenju novih karakteristika proizvoda, o promjeni cijene, nabavci novog postrojenja, uvođenju novog distribucijskog kanala, o određivanju trenutka promocije, preporuci proizvoda ili usluge, detektiranju prijevara i slično.
Sad prilazimo korak bliže – idemo zaviriti malo dublje u neke od alata i vidjeti konkretno kako napraviti model pomoću metoda prediktivne analitike.

Za kreiranje prediktivnog modela prodaje korištena je vlastita baza podataka koja sadrži podatke prodaje za zamišljenu organizaciju. Raspolažemo s podacima prodaje unazad 16 godina te za svaku godinu, određenu grupu proizvoda i određeni profitni centar promatramo ukupan iznos nabavne cijene, ukupan iznos bruto prodajne cijene, ukupan iznos prodajnog popusta i količinu prodajnih proizvoda.

Varijable sadržane u bazi su:

	Naziv varijable	**Opis varijable**
1.	*Year*	godina
2.	*Product Group*	naziv grupe proizvoda
3.	*Profit Center Code*	naziv profitnog centra
4.	*Purchase Price*	ukupan iznos nabavne cijene u određenoj godini
5.	*Gross Sales Price*	ukupni iznos bruto prodajne cijene u određenoj godini
6.	*Discount*	ukupan iznos prodajnog popusta u određenoj godini
7.	*Quantity*	ukupna količina prodajnih proizvoda u određenoj godini

Tablica 4.1. Ulazne varijable korištene baze podataka



	A	B	C	D	E	F	G
1	Year	Product Group	Profit Center	Purchase Price	Gross Sales	Discount	Quantity
2	2005	Product 101	PC1	5.690,40	11.600,11	125.453,06	113.225
3	2005	Product 102	PC1	13.711,95	22.917,30	223.391,47	99.895
4	2005	Product 103	PC1	2.491,97	3.721,97	14.013,82	72.296
5	2005	Product 104	PC1	15.599,21	25.244,95	66.947,98	23.890
6	2005	Product 105	PC1	708,57	1.192,25	15.477,90	9.274
7	2005	Product 106	PC1	3.667,31	5.622,09	5.953,72	9.762
8	2005	Product 107	PC1	5.599,68	7.971,47	17.075,96	1.606
9	2005	Product 108	PC1	1.514,41	1.817,43	3.555,80	18.747
10	2005	Product 101	PC2	1.117,21	1.575,60	27.456,58	31.924
11	2005	Product 102	PC2	2.114,68	3.539,61	13.217,68	9.286
12	2005	Product 103	PC2	1.576,15	2.259,49	4.899,92	28.650
13	2005	Product 104	PC2	2.728,19	6.341,02	9.072,45	7.472
14	2005	Product 105	PC2	248,43	383,36	18.235,55	4.405
15	2005	Product 106	PC2	1.744,00	2.686,58	3.918,60	5.992
16	2005	Product 107	PC2	1.120,68	1.947,93	3.578,52	559
17	2005	Product 108	PC2	648,73	1.052,35	1.821,36	5.228
18	2005	Product 101	PC3	1.524,98	2.189,27	25.452,92	31.073
19	2005	Product 102	PC3	636,99	1.059,06	14.811,73	9.200
20	2005	Product 103	PC3	307,33	459,17	10.025,94	73.891
21	2005	Product 104	PC3	1.753,03	5.049,46	8.160,97	1.388
22	2005	Product 105	PC3	269,65	457,82	5.848,05	2.344
23	2005	Product 106	PC3	699,62	1.084,24	1.239,82	1.120
24	2005	Product 107	PC3	1.542,93	2.198,34	5.707,20	387
25	2005	Product 108	PC3	537,57	848,37	3.803,84	17.880
26	2006	Product 101	PC1	7.063,24	10.471,79	164.167,32	127.318
27	2006	Product 102	PC1	17.465,00	27.465,04	205.450,39	99.972
28	2006	Product 103	PC1	1.757,85	2.138,42	11.556,23	73.015
29	2006	Product 104	PC1	17.563,30	31.332,46	75.761,31	26.427
30	2006	Product 105	PC1	911,72	1.472,33	11.149,94	9.485
31	2006	Product 106	PC1	3.210,54	4.789,55	7.133,41	10.570
32	2006	Product 107	PC1	4.963,82	7.071,61	9.878,43	1.262
33	2006	Product 108	PC1	3.948,98	5.229,51	17.722,40	19.698

Sales Model

Slika 4.1. Excel file koji sadrži korištenu bazu podataka (*Sales Model.xlsx*).

Za izradu prediktivnog modela prodaje koristi se već spomenuti alat Microsoft Azure Machine Learning Studio te metode stabla odlučivanja i neuronske mreže. Za izradu modela slijeđeni su već ranije spomenuti koraci implementiranja prediktivne analitike u poslovanje (Poglavlje 2.4.).

4.1. Definiranje cilja

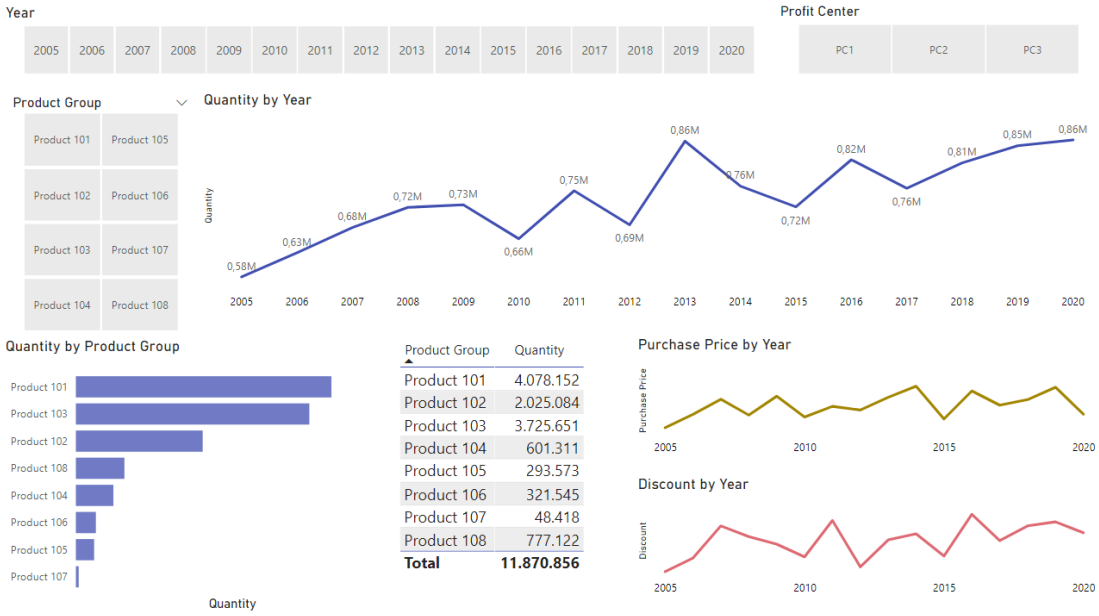
Obzirom da se kod kreiranja planova prodaje kao baza uglavnom uzimaju podaci o prodaji koji su se već dogodili u prošlosti, može se postaviti pitanje kako ćemo iz njih iščitati što će se događati u budućnosti i koliko ti povijesni podatci mogu biti pouzdani i korisni za kreiranje prodajnih planova.

Na temelju zadanih podataka ukupnog iznosa nabavne cijene, ukupnog iznosa bruto prodajne cijene i ukupnog prodajnog popusta unazad 16 godina, cilj je za određenu grupu proizvoda i profitni centar predvidjeti količinu prodaje za narednu godinu. Budući da je izlazna varijabla količina prodaje, prediktivni model se pretvara u problem regresije.

4.2. Razumijevanje podataka

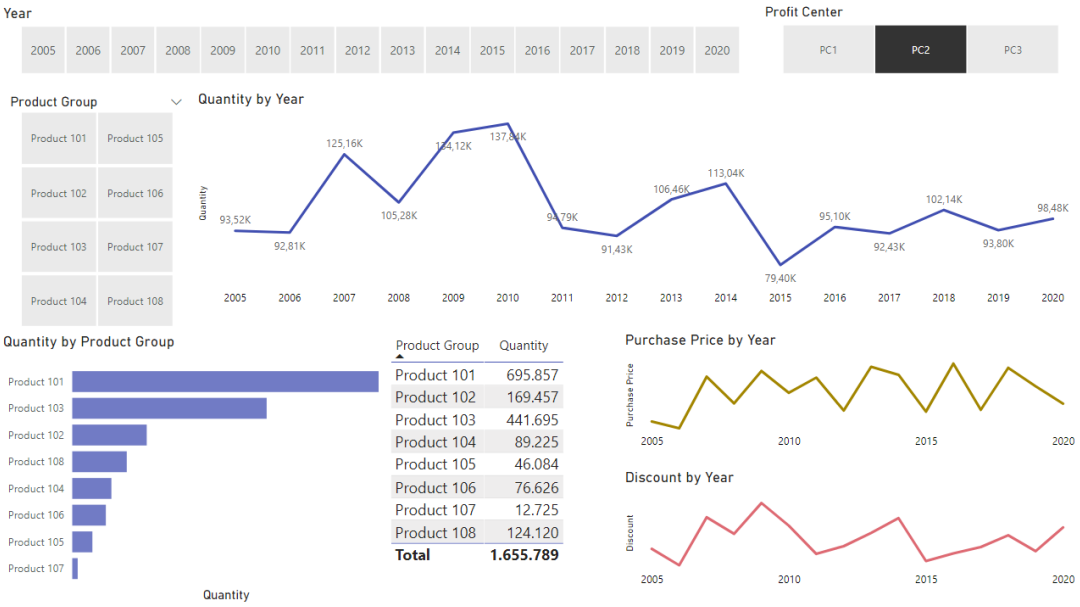
U ovom koraku potrebno je posvetiti se skupu podataka. Raspolažemo s podacima prodaje unazad 16 godina te za svaku godinu promatramo ukupan iznos nabavne cijene, ukupan iznos bruto prodajne cijene, ukupan iznos prodajnog popusta i količinu prodajnih proizvoda prema grupama proizvoda i profitnim centrima.

Tijek kretanja količine prodaje unazad 16 godina vizualno prikazujemo kroz alat Microsoft Power BI (slika 4.2.).

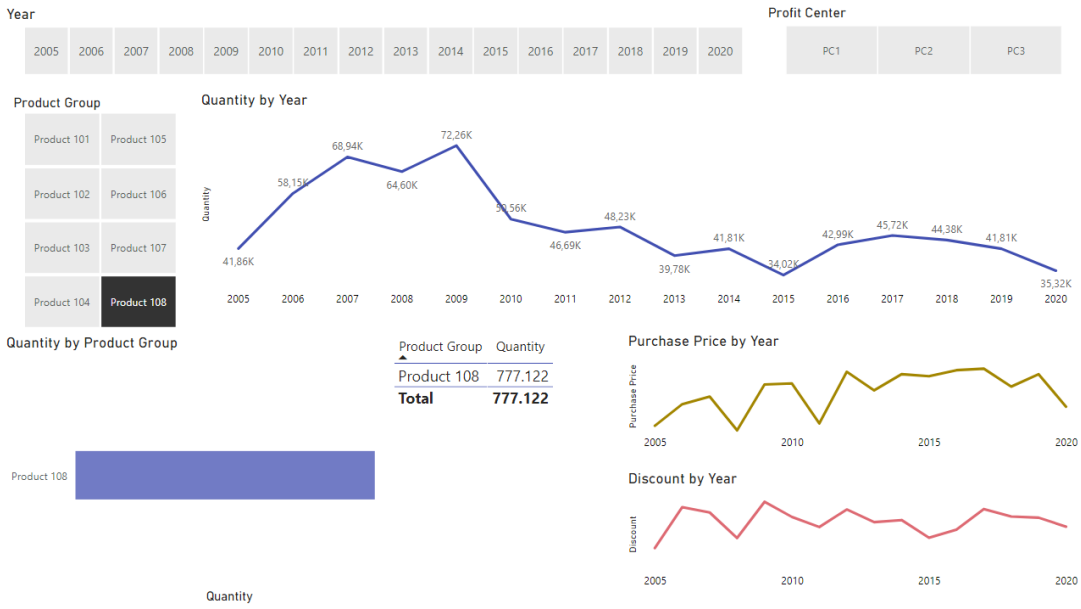


Slika 4.2. Vizualni prikaz baze podataka *Sales Model* kroz Microsoft Power BI

Power BI alat nudi dinamički vizualni prikaz podataka stoga možemo kroz filter odabrati određeni profitni centar ili grupu proizvoda te na taj način promatrati tijek kretanja količine prodaje unazad 16 godina. Slika 4.3. prikazuje tijek kretanja količine prodaje za profitni centar *PC2*, slika 4.4. prikazuje tijek kretanja količine prodaje za grupu proizvoda *Product 108*.



Slika 4.3. Tijek kretanja količine prodaje za profitni centar *PC2*

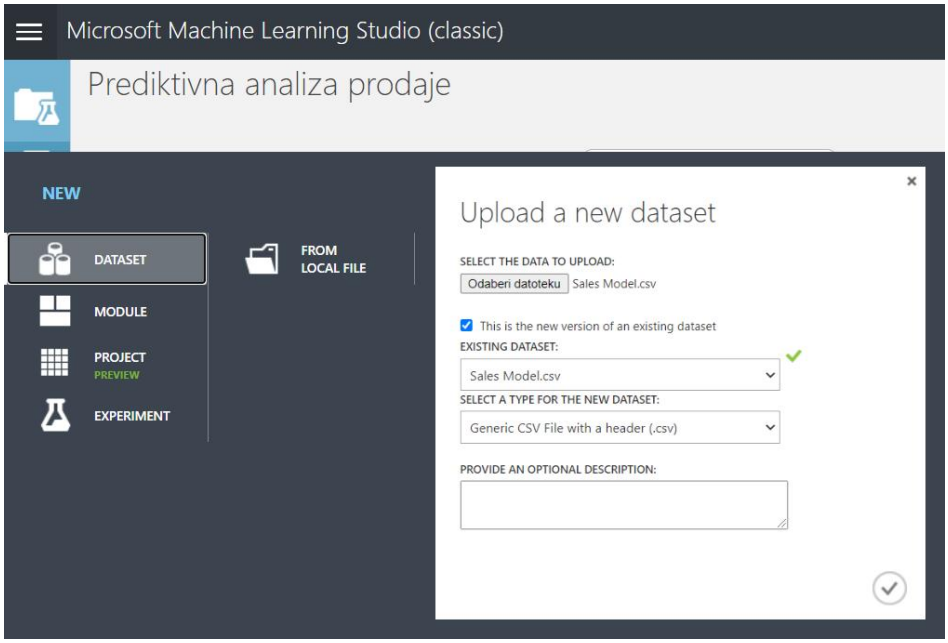


Slika 4.4. Tijek kretanja količine prodaje za grupu proizvoda *Product 108*

Nakon što smo grafički analizirali podatke, potrebno ih je prevesti u odgovarajući format za daljnju analizu i stvaranje modela. Kako bi našu bazu podataka prenijeti u MS Azure ML Studio potrebno ju je prebaciti u *.csv* format. Za te potrebe kreirana je datoteka *Sales Model.csv*.

4.3. Priprema podataka

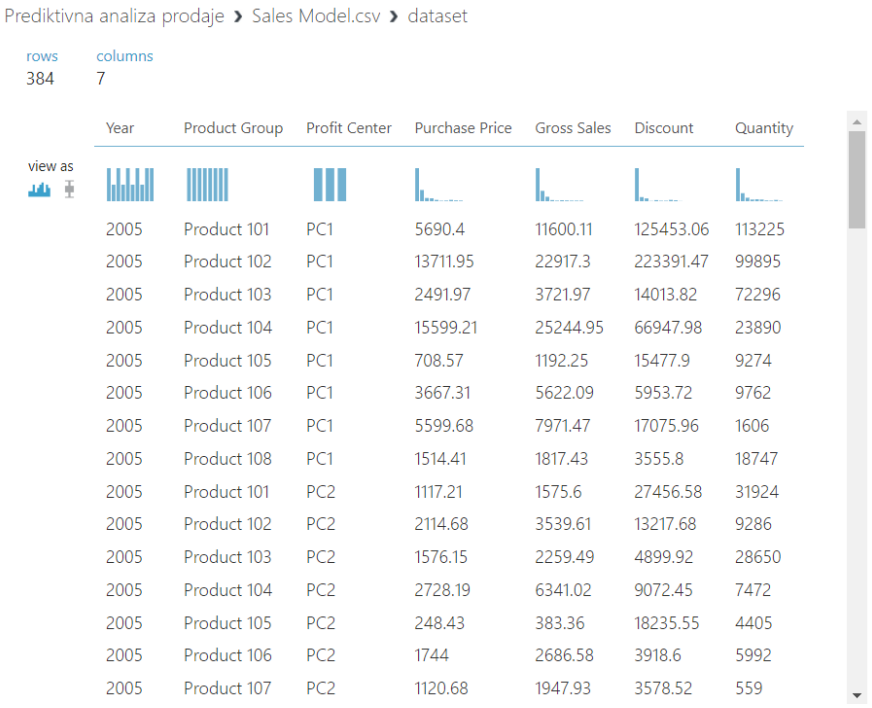
Nakon analize podataka pomoću Excela i grafičkog prikaza pomoću MS Power BI-a podaci su spremni za analizu pomoću MS Azure ML Studija. Bazu podataka *Sales Model.csv* potrebno je učitati u MS Azure ML Studio putem njihovog jednostavnog mehanizma za prijenos podataka.



Slika 4.5. Unos baze podataka *Sales Model.csv* u MS Azure ML Studio

Nakon što su podaci učitani, naš sljedeći korak je stvaranje novog eksperimenta pod nazivom *Prediktivna analiza prodaje*. Aktiviranjem izbornika *Experiments* započinjemo izgradnju modela (*Blank experiment*).

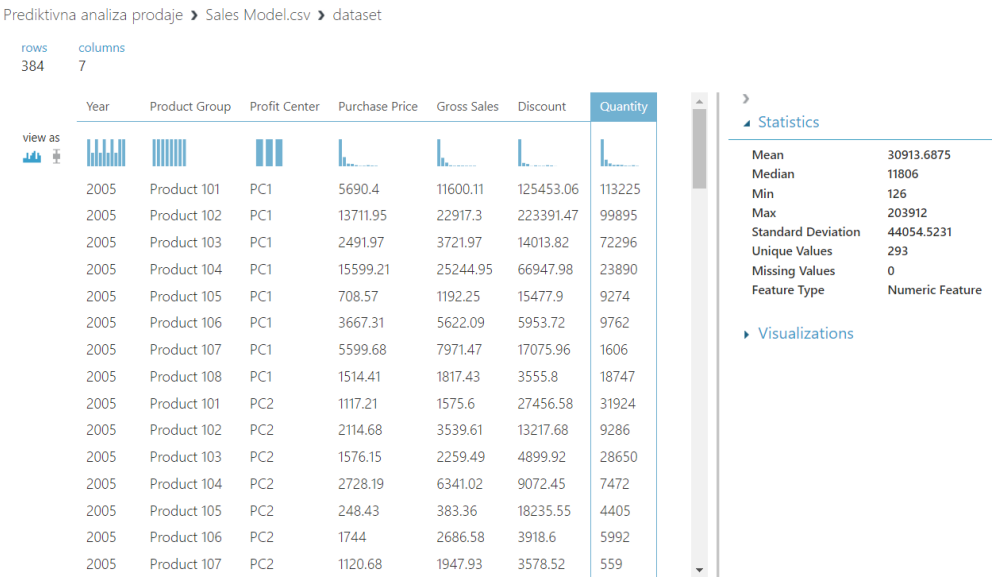
Učitanu bazu podataka uvest ćemo u novokreirani eksperiment i promatriti podatke. Uvid u podatke (*Visualize*) pokazuje nam vrijednosti unutar baze podataka (slika 4.6.).



Year	Product Group	Profit Center	Purchase Price	Gross Sales	Discount	Quantity
2005	Product 101	PC1	5690.4	11600.11	125453.06	113225
2005	Product 102	PC1	13711.95	22917.3	223391.47	99895
2005	Product 103	PC1	2491.97	3721.97	14013.82	72296
2005	Product 104	PC1	15599.21	25244.95	66947.98	23890
2005	Product 105	PC1	708.57	1192.25	15477.9	9274
2005	Product 106	PC1	3667.31	5622.09	5953.72	9762
2005	Product 107	PC1	5599.68	7971.47	17075.96	1606
2005	Product 108	PC1	1514.41	1817.43	3555.8	18747
2005	Product 101	PC2	1117.21	1575.6	27456.58	31924
2005	Product 102	PC2	2114.68	3539.61	13217.68	9286
2005	Product 103	PC2	1576.15	2259.49	4899.92	28650
2005	Product 104	PC2	2728.19	6341.02	9072.45	7472
2005	Product 105	PC2	248.43	383.36	18235.55	4405
2005	Product 106	PC2	1744	2686.58	3918.6	5992
2005	Product 107	PC2	1120.68	1947.93	3578.52	559

Slika 4.6. Prikaz baze podataka *Sales Model* kroz MS Azure ML Studio

Prilikom odabira varijable koja sadrži numeričke vrijednosti, alat prikazuje kratki statistički pregled (slika 4.7.).

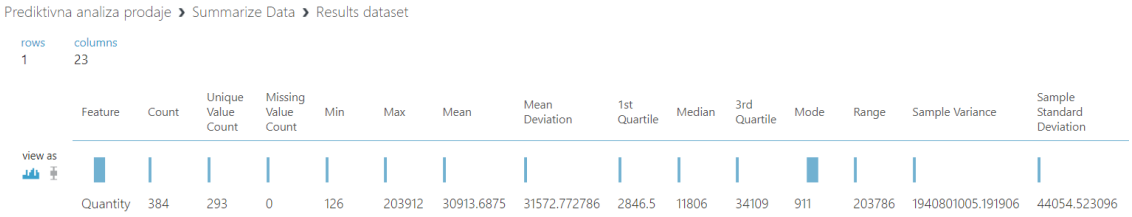


Year	Product Group	Profit Center	Purchase Price	Gross Sales	Discount	Quantity
2005	Product 101	PC1	5690.4	11600.11	125453.06	113225
2005	Product 102	PC1	13711.95	22917.3	223391.47	99895
2005	Product 103	PC1	2491.97	3721.97	14013.82	72296
2005	Product 104	PC1	15599.21	25244.95	66947.98	23890
2005	Product 105	PC1	708.57	1192.25	15477.9	9274
2005	Product 106	PC1	3667.31	5622.09	5953.72	9762
2005	Product 107	PC1	5599.68	7971.47	17075.96	1606
2005	Product 108	PC1	1514.41	1817.43	3555.8	18747
2005	Product 101	PC2	1117.21	1575.6	27456.58	31924
2005	Product 102	PC2	2114.68	3539.61	13217.68	9286
2005	Product 103	PC2	1576.15	2259.49	4899.92	28650
2005	Product 104	PC2	2728.19	6341.02	9072.45	7472
2005	Product 105	PC2	248.43	383.36	18235.55	4405
2005	Product 106	PC2	1744	2686.58	3918.6	5992
2005	Product 107	PC2	1120.68	1947.93	3578.52	559

Slika 4.7. Statistički pregled varijable *Quantity*

Model ćemo izraditi korak-po-korak pri čemu će nam od velike pomoći biti izbornik s lijeve strane u modulu *Experiments* podijeljen po glavnim koracima unutar kojih je moguće izabrati različite opcije ovisno o cilju, metodama koje se žele koristiti i analizama koje se žele napraviti.
Na početku koristimo karticu *Select Columns in Datasets* kako bi odabrali skup podataka koji ćemo podvrgnuti statističkoj obradi. Odabrana je varijabla *Quantity*.

Nakon izabranog uzorka odabrane podatke ćemo podvrgnuti osnovnoj statističkoj obradi. Korištenjem kartice *Summarize Data* stvara se skup standardnih statističkih mjera koje opisuju zadani stupac iz baze podataka (slika 4.8.).



Slika 4.8. Statistička obrada varijable *Quantity* korištenjem kartice *Summarize Data*

Slika 4.9. prikazuje dijagram modela nakon provedene statističke obrade podataka.

Slika 4.9. Dijagram modela nakon statističke obrade podataka

4.4. Modeliranje

Primjenom spomenute statističke obrade, dobivamo uvid u podatke i nastavljamo s izgradnjom prediktivnog modela. Ulazne varijable za izradu prediktivnog modela su *Year*, *Product Group*, *Profit Center Code*, *Purchase Price*, *Gross Sales Price*, *Discount* i *Quantity*. Izlazna varijabla je predviđena količina prodaje.

Prije postavljanja metode stabla odlučivanja i neuronske mreže, normalizirati ćemo podatke koristeći karticu *Normalize Data*. Normalizacija parametara uobičajena je i poželjna praksa u strojnom učenju. Postoji više tipova normalizacije parametara, koristit ćemo min-max normalizaciju. Neka je $X = (x_1, x_2, \dots, x_n)$ vektor duljine *n* svih vrijednosti nekog parametra. Min-max normalizacija od svake vrijednosti oduzima minimalnu, a zatim dobiveno broj dijeli s rasponom parametra. Ovaj postupak sužava raspon parametra na segment [0, 1], a nove vrijednosti $\overline{x_i}$ računaju se pomoću formule:

$$\overline{x_i} = \frac{x_i - \max(\{x_i,\ i=1,\dots,n\})}{\max(\{x_i,\ i=1,\dots,n\}) - \min(\{x_i,\ i=1,\dots,n\})}.$$

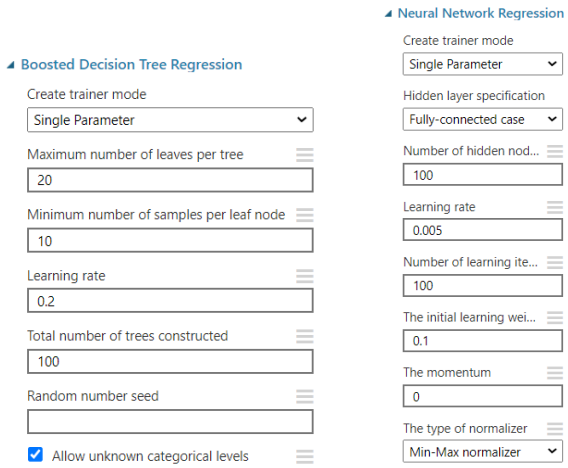
Sada možemo krenuti u izradu prediktivnog modela. Za izradu modela važno je podijeliti podatke na dva dijela - skup podataka za učenje i skup podataka za testiranje. Skup podataka za učenje služi za treniranje (učenje ili procjenu vrijednosti) (*tzv. Train Model*) na kojemu određena metoda pokušava pronaći veze između varijabli dok drugi dio služi za testiranje već naučenog i pohranjenog modela. Skup podataka za testiranje (*tzv. Score Model*) služi kako bi se ocijenila uspješnost modela na podacima koji nisu korišteni za učenje i izračunala greška modela koja se može uzeti kao očekivana na novim podacima u budućnosti pri primjeni modela u praksi. U tu svrhu podjele uzorka koristi se kartica *Split Data*. Pri njezinoj ugradnji u dijagram modela treba proizvoljno odrediti koji dio podataka se koristi za učenje (ne postoji pravilo, obično se veći dio podataka koristi za učenje, a manji za testiranje) pa ćemo za naš model izabrati da se 70 % podataka koristi za učenje, a preostalih 30 % za testiranje.

Iduće je potrebno odabrati modele koje ćemo koristiti za izradu prediktivnog modela prodaje. Koristiti ćemo metodu stabla odlučivanja (*Boosted Decision Tree Regression*) i umjetne neuronske mreže (*Neural Network Regression*). Svaku od navedenih metoda potrebno je dodati u dijagram izgradnje modela.

Prilikom odabira metode stabla odlučivanja, potrebno je u desnom dijelu prozora izabrati potrebne parametre. Odabrano je maksimalno 20 listova po stablu i minimalno 10 slučaja potrebnih za stvaranje bilo kojeg lista u stablu. Stopa učenja postavljena je na 0.2 i ona određuje koliko brzo ili sporo učenje konvergira prema optimalnom rješenju. Unutar modela kreirati će se sveukupno 100 stabala.

Za metodu neuronske mreže potrebno je također postaviti parametre. Postavljeno je sveukupno 100 skrivenih neurona, stopa učenja 0.005, maksimalno 1000 broja ponavljanja učenja, težina čvora na početku procesa učenja 0.1 i vrijednost 0 kao ponder za čvorove iz prethodnih iteracija.

Navedeni se parametri mogu eksperimentalno mijenjati u postupku izgradnje kako bi se pronašao model koji daje najveću točnost.

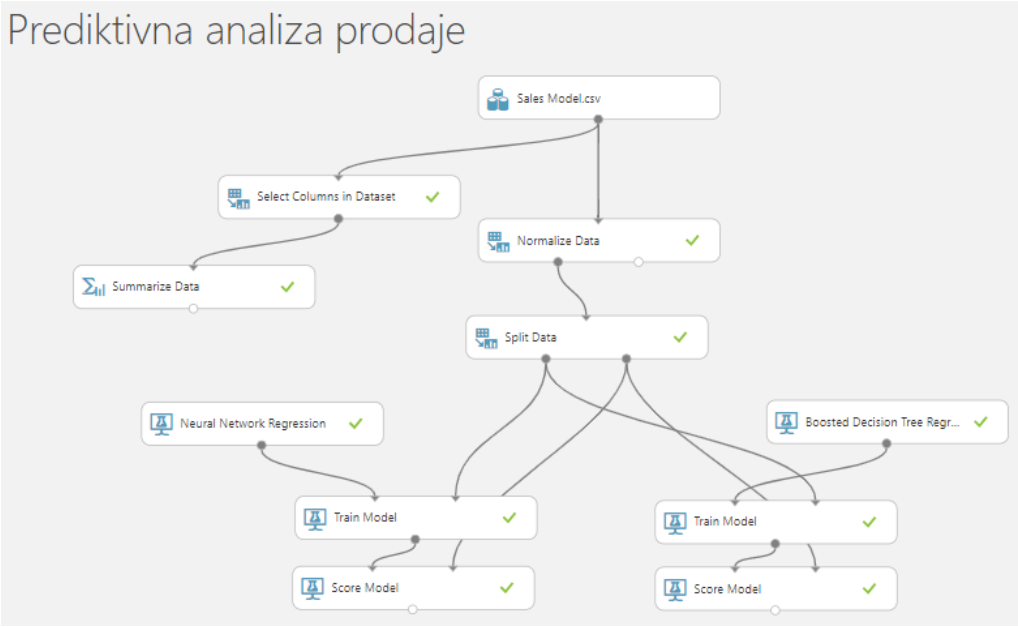


Slika 4.10. Parametri za model stabla odlučivanja i model neuronskih mreža

Nakon što smo odabrali modele strojnog učenja koje ćemo koristiti za predikciju, potrebno ih je povezati s karticom *Train Model* koja će služiti za učenje modela. Na kartici *Train Model* potrebno je u desnom prozoru odabrati varijablu koja će se koristiti kao izlazna u modelu (u ovom slučaju je to varijabla *Quantity* koja označava ukupnu količinu prodajnih

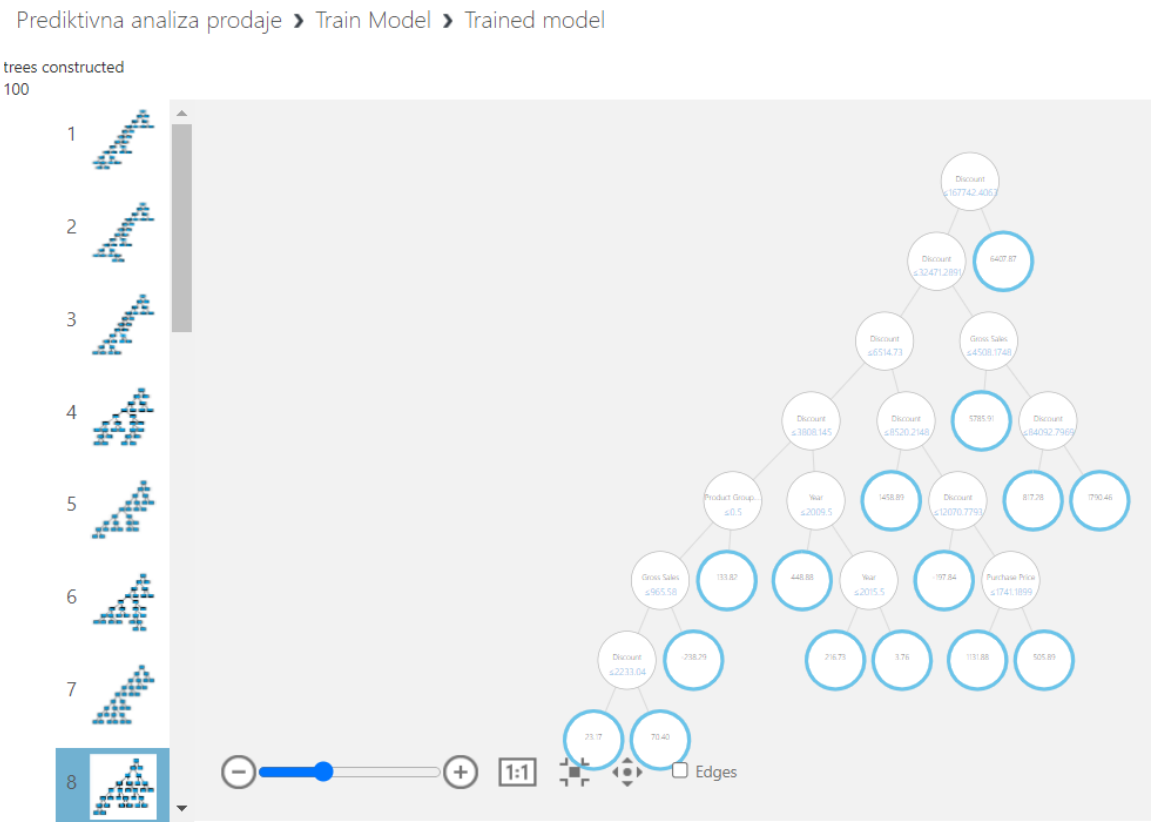
proizvoda u određenoj godini). Kartica *Train Model* povezana je i s karticom *Split Data*, kako bi se za učenje primijenio odgovarajući skup podataka. Prilikom učenja, za obje će se metode provesti iterativni postupak učitavanja svakog retka iz skupa za učenje te pokušati pronaći veze između ulaznih varijabli i izlazne varijable. Nakon toga potrebno je dodati karticu za testiranje *Score Model* koju će model pohranjen u prethodnoj fazi učenja sada primijeniti na skupu za testiranje i to za svaku metodu koja će proizvesti svoje predikcije. *Score Model* potrebno je povezati i sa karticom *Split Data*.

Pritiskom na aktivno polje *Run* započinjemo izvršavanje modela. Izvršava se svaka njegova prikazana faza pa ako nije bilo grešaka označava se zelenom kvačicom. Slika 4.11. prikazuje model nakon izvršavanja.



Slika 4.11. Dijagram modela prediktivne analize nakon izvršavanja

Nakon izvršavanja, u kartici *Train Model* metode stabla odlučivanja možemo pogledati grafički prikaz dijela stabla odlučivanja za predikciju prodaje (slika 4.12.). Rezultati na *Train Modelu* za stablo odlučivanja grafički su prikazani kroz čvorove u kojima su upisane ulazne varijable.



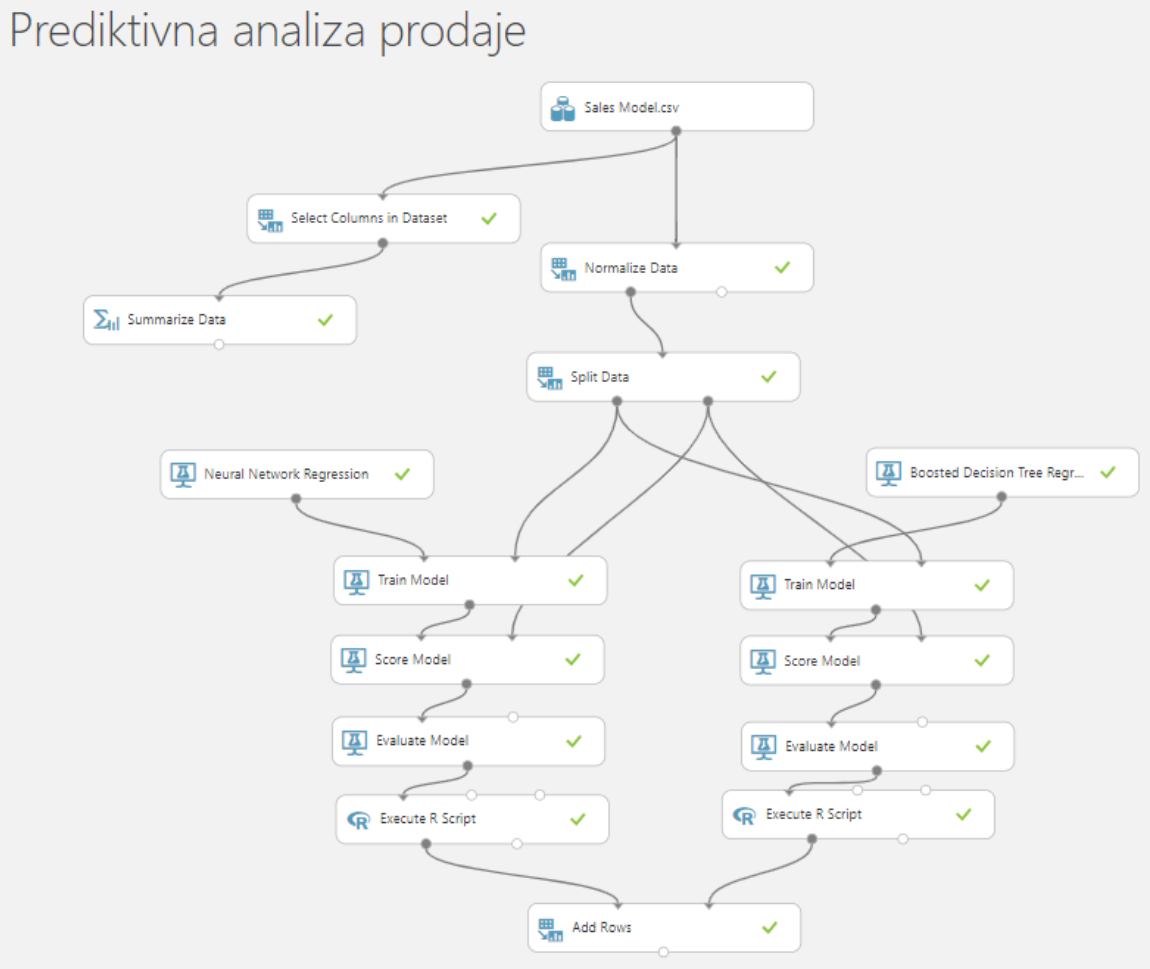
Slika 4.12. Grafički prikaz dijela stabla odlučivanja za prediktivnu analizu

S druge strane neuronske mreže pripadaju metodi crne kutije (tzv. black box) gdje nemamo mogućnost vidjeti slijed donošenja odluke budući da je rezultat izračunat temeljem više funkcija koje se primjenjuju u svakom sloju neuronske mreže od ulaznog sloga, preko jednog ili više skrivenih, do izlaznog sloja.

4.5. Evaluacija

Na kraju zbog usporedbe točnosti predikcija koje proizvodi svaka metoda (i stabla odlučivanja i neuronske mreže) dodaje se u dijagram kartica *Evaluate Model* koja će izračunati stope točnosti modela. U model su ugrađene i *R* skripte radi preglednijeg prikaza rezultata.

Slika 4.13. prikazuje cijeli prediktivni model prodaje izgrađen pomoću alata MS Azure ML Studio.



Slika 4.13. Dijagram prediktivnog modela prodaje izgrađenog pomoću alata MS Azure ML Studio

Nakon što je model završen, možemo tumačiti rezultate. Dobivamo evaluaciju obučenog modela izraženu u statističkim vrijednostima. Koeficijent određivanja *Coefficient of Determination*, predstavlja moć predviđanja modela kao vrijednost između 0 i 1. Nula znači da je model slučajan (ne objašnjava ništa); 1 znači da savršeno pristaje. Na slici 4.14. možemo vidjeti kako je za izgrađeni model stabla odlučivanja dobivena stopa točnosti od 90,23 %, a za model neuronskih mreža 66,71 %.

Algorithm	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Relative Squared Error	Coefficient of Determination
Neural Network Regression	0.084437	0.128823	0.518014	0.332949	0.667051
Boosted Decision Tree Regression	0.04158	0.069787	0.255091	0.09771	0.90229

Slika 4.14. Stopa točnosti modela

Nakon provedbe procesa predviđanja pomoću stabla odlučivanja i neuronskih mreža možemo izvući potreban zaključak. Stablo odlučivanja te neuronske mreže koristili smo nad istim podacima kako bismo usporedili rezultate oba predviđanja. Podaci su bili raspoređeni u skup podataka za učenje (70 %) i skup podataka za testiranje (30 %). Podaci su na slučajan način svrstavani u podskup za učenje i testiranje. Analizom izrađenih modela stabla odlučivanja i modela neuronske mreže vidljivo je da je model stabla odlučivanja dao bolje rezultate u odnosu na model neuronske mreže. Stopa točnosti modela stabla odlučivanja iznosi 90,23%, dok je stopa odlučivanja modela neuronskih mreža 66,70%.

Stabla odlučivanja dala su znatno veću točnost od neuronskih mreža, te ih se predlaže koristiti kao točniju metodu na promatranom skupu podataka.

S obzirom na to da je model stabla odlučivanja dao veću stopu točnosti, na slici 4.15. pogledat ćemo predikcije koje je proizvelo stablo odlučivanja na podacima iz testnog uzorka. Kartica *Score Model* daje kratak prikaz podataka i dodaje novi stupac u naš skup podataka, *Scored Labels*. Vrijednosti u stupcu *Scored Labels* približne su vrijednostima

odgovarajućih vrijednosti *Quantity* kada primijenjeni algoritam učenja dobro funkcionira s dostupnim podacima.

Prediktivna analiza prodaje › Score Model › Scored dataset

rows 192 columns 8

view as

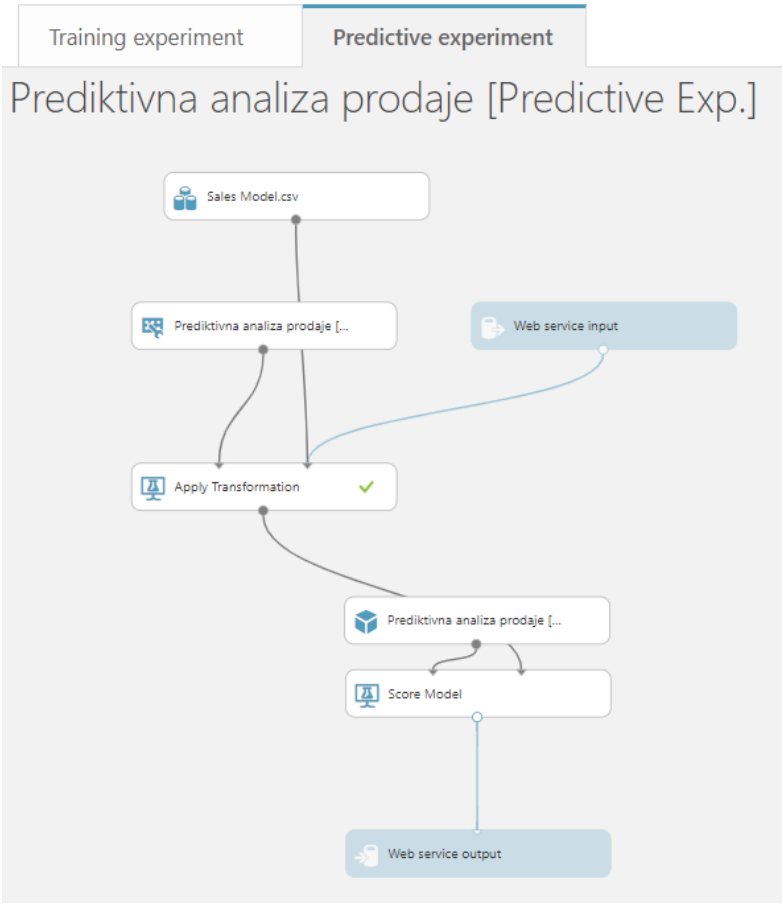
Year	Product Group	Profit Center	Purchase Price	Gross Sales	Discount	Quantity	Scored Labels
2020	Product 108	PC3	0.000383	0.000185	0.005235	0.050784	0.062234
2012	Product 101	PC2	0.093364	0.07447	0.136864	0.191789	0.248309
2017	Product 107	PC2	0.054575	0.043567	0.011949	0.001678	-0.000977
2010	Product 103	PC1	0.06466	0.044117	0.039974	0.413502	0.437458
2010	Product 106	PC3	0.011986	0.00948	0	0.007321	0.006077
2012	Product 105	PC1	0.034658	0.030011	0.062781	0.06786	0.13896
2018	Product 108	PC3	0.00496	0.003996	0.008185	0.058625	0.067421
2018	Product 105	PC3	0.008253	0.00761	0.006676	0.006845	0.005486
2014	Product 101	PC1	0.436249	0.279085	0.724544	0.804226	0.802286
2018	Product 104	PC2	0.284207	0.257934	0.047157	0.031052	0.004963
2018	Product 106	PC2	0.067854	0.057678	0.009137	0.008318	0.01767
2008	Product 105	PC1	0.025094	0.019806	0.080614	0.057202	0.009935
2017	Product 108	PC1	0.371721	0.287716	0.071568	0.111715	0.121961
2020	Product 101	PC3	0.054355	0.040754	0.084766	0.125028	0.160746
2011	Product 106	PC2	0.145274	0.121372	0.01263	0.03381	0.024322

Slika 4.15. Scored Labels prikazuje rezultate predikcije korištenjem metode stabla odlučivanja

Ukoliko nismo zadovoljni sa dobivenim rezultatima, moguće je napraviti novi eksperiment s promijenjenim parametrima za svaku metodu.

4.6. Implementacija modela

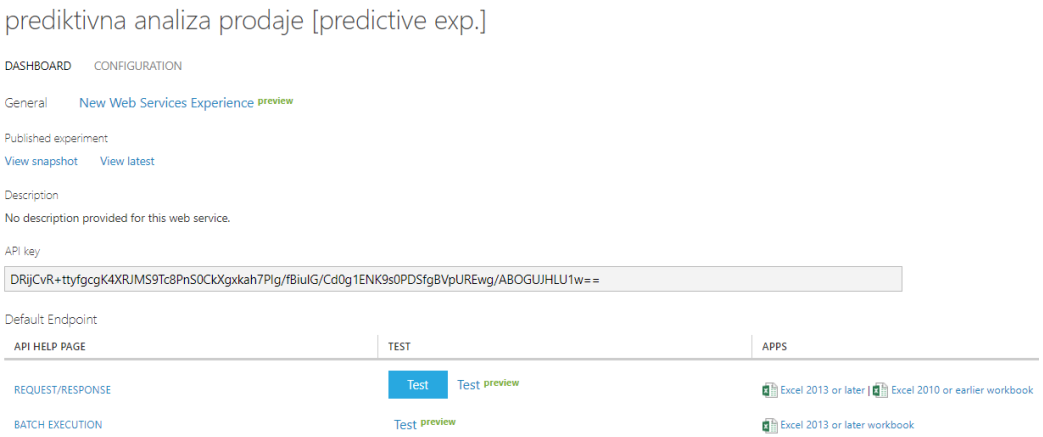
Nakon što smo utvrdili da metoda stabla odlučivanja vraća bolje rezultate, odlučujemo se koristiti tu metodu za buduće predikcije ove baze podataka. Pokrećemo kreiranje prediktivnog modela za odabranu metodu stabla odlučivanja (slika 4.16.).



Slika 4.16. Izgrađen prediktivni model korištenjem stabla odlučivanja

U ovom trenutku možemo spremiti kreirani model u MS Azure ML Studio kako bi ga koristili za buduću upotrebu.

Koristeći opciju *Publish Web Service* možemo stvoriti jednostavnu web uslugu koja se nalazi na Azureovoj infrastrukturi u oblaku.



Slika 4.17. Publish Web Service za kreirani prediktivni model

Nakon toga, možemo testirati naš model predikcije koristeći jednostavan testni obrazac kroz Excel alat.

4.7. Zaključak

Na temelju podataka prodaje unazad 16 godina, cilj je bio za određenu grupu proizvoda i profitni centar predvidjeti količinu prodaje za narednu godinu. Testiranjem modela za predviđanje prodaje na skupu baze podataka za izmišljenu organizaciju, utvrđeno je da su rezultati predviđanja modela u jakoj korelaciji sa stvarnim podacima. Iz toga proizlazi kako navedeni model može relativno kvalitetno procijeniti predviđanje prodaje.

Zaključujemo da korištenje prediktivne analitike može unaprijediti proces donošenja poslovnih odluka u organizaciji.

Bibliografija

[1] A. Zhang, Data Analytics, Amazon Fulfillment, Poljska, 2017.

[2] A. Zoltners, P. Sinha, The Power of Sales Analytics, ZS Associates, 2014

[3] Climber, Qlik a leader in the 2021 Gartner Magic Quadrant, dostupno na https://www.climber.eu/ (kolovoz 2021.)

[4] C. C. H. Michael, Classical Neural Network: What really are Nodes and Layers?, dostupno na https://towardsdatascience.com/ (srpanj 2021.)

[5] D. Fagella, Techemergence: Predictive Analitics for Marketing-What's possible and how it works, dostupno na https://www.techemergence.com/ (lipanj 2021.)

[6] D. J. Hand, H. Mannila, P. Smyth, Principles of Data Mining, Drug safety, London, 2007.

[7] E. Siegel, The power to predict who will click, buy, lie or die., NY John Wiley & Sons, New York, 2013.

[8] FOI, Neuronske mreže, ERIS, dostupno na https://eris.foi.hr/11neuronske/nn-predavanje2.html (kolovoz 2021.)

[9] F. Rosenblatt, The perceptron: A probabilistic model for information storage and organization in the brain, Psychological Review, 65(6):386–408, 1958.

[10] G. P.-S. P. S. U. Fayyad, »Six steps in CRISP-DM the standard data mining process,« AI Magasine, London, 1996.

[11] I. Gartner, Technical Professional Advice, dostupno na https://www.gartner.com/, (kolovoz 2021.)

[12] J. Compton, Data. Strategy. Tehnology. – Six Reasons Why Use Predictive Analytics, dostupno na https://www.dmnews.com/customer-experience/ (srpanj 2021.)

[13] J. D. Kelleher, B. Mac Namee, A. D'arcy, Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies, MIT press, 2020.

[14] M. Kuzmanović, Poslovna analitika i optimizacija, dostupno na http://pa.fon.bg.ac.rs/wp-content/uploads/2019/02/Deskriptivna-analitika.pdf (lipanj 2021.)

[15] J. H. Friedman, Data mining and statistics: What's the connection?, dostupno na https://statweb.stanford.edu/ (srpanj 2021.)

[16] J. S. N. Nayab, A Review of Decision Tree Analysis Advantages, dostupno na https://www.brighthubpm.com/project-planning/106000 (kolovoz 2021.)

[17] J. S. N. Nayab, A Review of Decision Tree Disadvantages, dostupno na https://www.brighthubpm.com/project-planning/ (kolovoz 2021.)

[18] L. A. Zadeh, Electrical engineering at the crossroads, IEEE Transactions on Education, 1965.

[19] L. Iffert, Predictive Analytics richtig einsetzen, Controlling & Management Review, 2016.

[20] L. Rokach, O.Z. Maimon, Data mining with decision trees: theory and applications, World scientific, 2007.

[21] M. Varga, V. Čerić, Informacijska tehnologija u poslovanju, Element, Zagreb, 2004.

[22] M. Zekić-Sušac, A. Has, Predictive analytics in Big Data plathorms- comparison and strategies, MIPRO Proceedings 2016, Rijeka, 2016.

[23] M. Zekić-Sušac, Intuicija više nije dovoljna, dostupno na http://www.infotrend.hr/clanak/2017/3/intuicija-vise-nije-dovoljna (lipanj 2021.)

[24] M. Zekić-Sušac, Neuronske mreže, dostupno na http://www.efos.unios.hr/upravljanje-marketingom/Neuronske-mreze (kolovoz 2021.)

[25] M. Zekić-Sušac, Poslovna inteligencija, dostupno na http://www.efos.unios.hr/upravljanje-marketingom/ (srpanj 2021.)

[26] M. Zekić-Sušac, Prediktivna analitika 2 - korak bliže, dostupno na http://www.infotrend.hr/clanak/prediktivna-analitika-2-korak-blize (kolovoz 2021.)

[27] M. Zekić-Sušak, Podatkovna analitika i BigData, dostupno na http://www.efos.unios.hr/upravljanje-marketingom/ (kolovoz 2021.)

[28] M. Zekić-Sušak, Stabla odlučivanja, dostupno na http://www.efos.unios.hr/sustavi-poslovne-inteligencije/Stabla-odlucivanja (kolovoz 2021.)

[29] P.-N. Tan, M. Steinbach, V. Kumar, Introduction to Data Mining, Pearson, 2016.

[30] P. Sikavica, T. Hunjak, N. Begičević Ređep i T. Hernaus, Poslovno odlučivanje. Školska knjiga d.d., Zagreb, 2014.

[31] S. Finlay, Predictive Analytics in 56 Minutes, 2015.

[32] S. Lončarić, Neuronske mreže: Višeslojni perceptron, dostupno na https://www.aes.hr/06-ViseslojniPerceptron (kolovoz 2021.)

[33] S. V. Europe, Introduction to the CRISP DM data mining methodology, dostupno na: https://www.youtube.com/Introduction to the CRISP DM (srpanj 2021.)

[34] Stteph, Crisp-dm and why you should know about it, Locke Data, dostupno na https://itsalocke.com/blog/crisp-dm-and-why-you-should-know-about-it/ (srpanj 2021.)

[35] T. Davenport, Analytics 3.0. Harvard Business Review, dostupno na https://hbr.org/2013/12/analytics-30 (kolovoz 2021.)

[36] T. Masters, Practical neural network recipes in c++, Morgan Kaufmann, San Diego, 1993.

[37] U. I. Akpan, A. Starkey, Review of classification algorithms with changing inter-class distances, Machine Learning with Applications, 2021.

[38] V. Kotu, B. Deshpande, Predictive analytics and data mining: concepts and practice with rapidminer, Morgan Kaufmann, 2014.

[39] Ž. Panian, Poslovna inteligencija, MASMEDIA, Zagreb, 2003.

[40] W. S. Mcculloch, W. Pitts, A logical calculus of the ideas immanent in nervous activity, The Bulletin of Mathematical Biophysics, 5(4):115–133, 1943.

Sažetak

U diplomskom radu prikazana je i objašnjena izrada prediktivnih modela u prodaji kako bi organizacije mogle predvidjeti ponašanje kupaca u budućnosti. Za stvaranje modela vrlo je bitna prediktivna analitika koja postaje nezaobilazna u analizi postojećih podataka i predviđanju budućnih nepoznatih događaja. Definira se pojam prediktivne analitike kao sastavnog dijela digitalizacije, opisan je postupak implementacije prediktivne analitike u poslovanje i navedene metode koje prediktivna analitika primjenjuje u svojim izračunima. Kroz studijski primjer izrade prediktivnog modela prodaje ustanovljeno je kako velika količina podataka zajedno s analitikom može ponuditi organizacijama brojne mogućnosti za poboljšanje poslovne učinkovitosti.

Summary

The thesis presents and explains the development of predictive models in sales so that organizations can predict customer behavior in the future. For the creation of the model, predictive analytics is very important, which becomes indispensable in the analysis of existing data and the prediction of future unknown events. The concept of predictive analytics as an integral part of digitalization is defined, the process of implementing predictive analytics in business is described and the methods that predictive analytics apply in its calculations are listed.

Životopis

Rođena sam 13. svibnja 1994. godine u Zagrebu, Hrvatska. Nakon završetka XV. Gimnazije (MIOC), upisujem preddiplomski sveučilišni studij Matematika u Zagrebu. Tijekom studija paralelno u visokom učilištu Algebra u Zagrebu završavam program obrazovanja za Web dizajnera. Nakon završenog preddiplomskog studija, upisujem diplomski sveučilišni studij Matematika i informatika na istom fakultetu. U trenutku nastanka rada radim kao Junior Data Management Consultant u tvrtki Data Sense d.o.o. u Zagrebu.