Primjena podatkovne analitike u poduzetništvu

Diminić, Matija

SVEUČILIŠTE U ZAGREBU

FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

Matija Diminić

Primjena podatkovne analitike u poduzetništvu

ZAVRŠNI RAD

Varaždin, 2020.

SVEUČILIŠTE U ZAGREBU

FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

Matija Diminić

Matični broj: 44033/15-R

Studij: Informacijski sustavi

Primjena podatkovne analitike u poduzetništvu

ZAVRŠNI RAD

Mentorica:

Doc. dr. sc. Oreški Dijana

Varaždin, rujan 2020.

Matija Diminić

Izjava o izvornosti

Izjavljujem da je moj završni/diplomski rad izvorni rezultat mojeg rada te da se u izradi istoga nisam koristio drugim izvorima osim onima koji su u njemu navedeni. Za izradu rada su korištene etički prikladne i prihvatljive metode i tehnike rada.

Autor/Autorica potvrdio/potvrdila prihvaćanjem odredbi u sustavu FOI-radovi

Sažetak

Ovaj rad odnosi se na primjenu podatkovne analitike u poduzetništvu. Označava metode i načine pomoću kojih se može analizirati odabrani skup podataka kako bismo mogli donositi neke značajne poslovne odluke u poduzetništvu. Pokazat ćemo nekoliko sličnih primjera već provedenih istraživanja o primjeni analitike u poduzetništvu. Opisati metodologiju i skup podataka vezan uz prodaju u supermarketima kako bismo uz pomoć podatkovne analitike predvidjeli zadovoljstvo kupaca i jediničnu cijenu proizvoda. Na kraju će se provesti interpretacija dobivenih rezultata analize.

Ključne riječi: Skupovi podataka; Podatkovna analitika; Rudarenje podataka; Poduzetništvo.

Sadržaj

1. Uvod

Svakodnevnim korištenjem interneta i aplikacija proizvodimo veliku količinu podataka smještenih u različitim bazama podataka. 2017. godine *The Economist* objavljuje članak naziva „Najvrijedniji svjetski resurs nije više nafta već podaci“ iz čega je uslijedilo mišljenje da su podaci nova nafta odnosno podaci koje proizvedemo se prikupljaju, čiste, organiziraju i analiziraju te naposljetku koriste za donošenje poslovnih odluka. Svi navedeni procesi su dio podatkovne analitike koja se danas koristi kao uobičajena praksa današnjih poduzeća za donošenje poslovnih odluka.

Korištenjem podatkovne analitike poduzeće može uspješno predvidjeti kretanja tržišnih cijena i odabrati najpovoljnije trenutke za određene poteze, pravodobno otkriti i smanjiti poslovni rizik i donijeti pravovremene i točnije poslovne odluke iz čega proizlazi i važnost podatkovne analitike za poduzeća.

Motivacija za odabir ove teme je to što smatram da je vještina podatkovne analitike veoma tražena u današnjem svijetu te je poželjno poznavati metode i tehnike kako pripomoći pojedinom poduzeću prilikom donošenja odluke.

Rad će obuhvatiti poglavlje u kojem ćemo pregledati već slična istraživanja koja su provedena kako bi mogli vidjeti primjenu podatkovne analitike u različitim područjima. Nadalje će biti navedeno poglavlje u kojem ćemo obraditi metodologiju pomoću koje ćemo odabrati i pročistiti podatke. Podatke ćemo opisati kako bismo znali za što i s kojim ciljem ćemo obrađivati te podatke. Sljedeće poglavlje će biti analiza pripremljenih i odabranih podataka gdje ćemo provesti podatkovnu analitiku o prodaji u 3 velika supermarketa, a za kraj će nam ostati interpretacija nakon analize kako bismo mogli evaluirati dobiveni model i vidjeti na koje sve načine možemo dobivene rezultate primijeniti u našu korist.

Na kraju rada ćemo navesti zaključak koji sam izvukao nakon izrade i provedbe svih potrebnih postupaka za podatkovnu analitiku. Također ćemo dati svoj osvrt na zadanu temu.

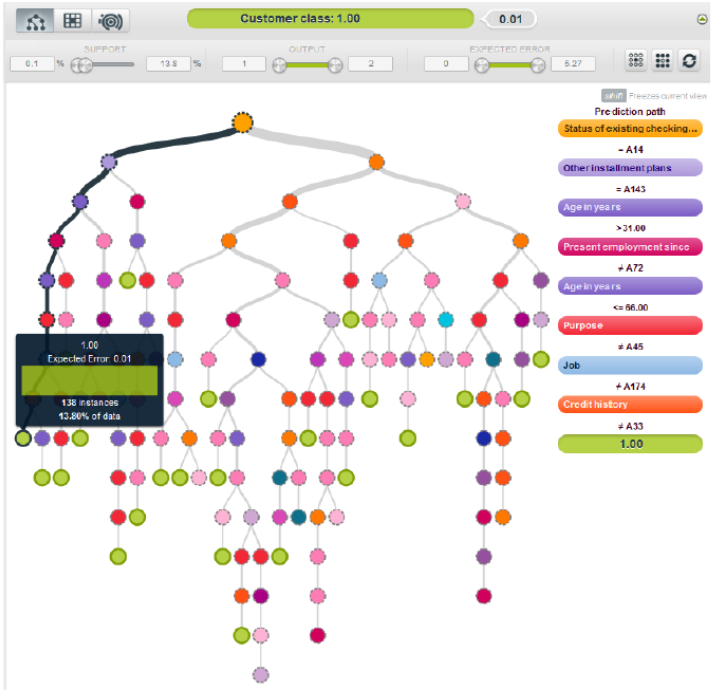
2. Pregled sličnih istraživanja

Podatkovna analitika u poduzetništvu je česta tema istraživanja. U ovom poglavlju opisano je nekoliko istraživanja koja su provedena na sličnu temu analize podataka za primjenu u poduzetništvu. Navedeno je nekoliko istraživanja različitih namjena za različita područja.

2.1. Procjena kreditnog rizika primjenom rudarenja podataka

Prvo istraživanje otkriva analizu kandidata za kreditiranje. Banke i slične organizacije dolaze do raznih rješenja kako bi prepoznale pojedinca koji je dobar kandidat za kredit pri čemu se služe rudarenjem podataka. Rudarenjem podataka se na kraju odlučuje je li pojedinac kreditno sposoban. Istraživanje je provedeno nad bazom podataka, sa skupom numeričkih i kategorijskih podataka koji pripadaju njemačkoj banci. [1]

Za istraživanje je korištena metoda stabla odlučivanja za analiziranje podataka koja je primjenjiva za potrebe klasificiranja, predviđanja, procjenjivanja, klasteriranja, opisivanja i vizualizaciju podataka. Zbog jednostavnosti i drugih prednosti prvo se selektira reprezentativni uzorak, iz podataka koji su već prošli kroz čišćenje netočnih podataka, integraciju, redukciju i transformaciju.



Slika 1. Stablo odlučivanja u alatu BigML [1]

Baza podataka za istraživanje sastoji se od 1000 instanci i 20 vrsta atributa. Za precizniji opis skupa podataka u istraživanju je korišten alat BigML pri čemu su generirane slike koje odgovaraju svakom atributu. [1]

Prema stablu odluke na slici izvedeni su podaci za svakog klijenta određenih atributa. Iz tih podataka napravljen je presjek dobrih kandidata – odnosno koji su njihove karakteristike iste i prema tome je zaključak da sve dobre klijente karakterizira nepostojanje otvorenog tekućeg računa, nepostojanje dužnika ili jamaca, nemaju kašnjenja prilikom isplaćivanja. Naspram toga lošim klijentima su zajedničke sljedeće karakteristike: otvoren tekući račun, prijavljeno prebivalište duže od godinu dana, posjed automobila. [1]

2.2. Rudarenje podataka u poduzetništvu

Tema ovog rada je bilo rudarenje podataka u poduzetništvu. Isto tako opisano je u detalje GEM (**G**lobal **E**ntrepreneurship **M**onitor) te su navedeni primjeri rudarenja podataka.

GEM je najvažnija svjetska studija o poduzetništvu. Omogućuje pristup visoko kvalitetnim informacijama, izvješćima i pričama koje omogućuju bolje razumijevanje podataka. Neke od najvažnijih svjetskih organizacija koriste skupove podataka od GEM-a (Ujedinjeni narod, Svjetski ekonomski foru, Svjetske banke …). U svakoj ekonomiji GEM promatra dva elementa, a do su:

1. Poduzetničko ponašanje i stavovi pojedinaca
2. Nacionalni kontekst i kako to utječe na poduzetništvo

Informacije koje se dobiju se pažljivo analiziraju od strane GEM istraživača koji nakon analize dobivaju saznanja o kvaliteti okruženja za poduzetništvo. Cilj GEM istraživanja je kako bi se otkrilo zašto su neke ekonomije odnosno države više poduzetničke od drugih.

Rudarenje podataka je proces za koji ne postoji recept kako bi uvijek bio uspješan to jest kako bi rezultirao uspješnim pronalaskom važnih informacija. Proces rudarenja podataka moguće je opisati u 4 važna koraka pomoću kojih se može povećati vjerojatnost uspjeha, a to su redom:

1. Definicija poslovnog problema
2. Priprema podataka
3. Modeliranje
4. Implementacija

Od primjera bih izdvojio primjer pod poglavljem 4.2. koji se odnosi isključivo na rudarenje podataka u poduzetničkoj analizi. Navedeno istraživanje koristi podatke CCEEmprende projekta. Metode rudarenja podataka koji su uključeni u istraživanje su [4]:

1. Pravila asocijacije
2. Stabla odlučivanja
3. Logičku regresiju

Istraživanje je imalo za cilj pokazati odnosno predvidjeti uspjeh odnosno identificirati najvažnije čimbenike koji su povezani s uspjehom ili neuspjehom projekata u poduzetništvu. [4]

Prema primjeru ovog projekta možemo vidjeti da u rezultatima istraživanja stoji da se uz pomoć metode asocijacije došlo do zaključka kako su mogućnost financiranja i neovisna radna situacija najvažniji sporedni slučajevi kod pokretanja poduzeća i kako će se ukoliko su te dvije stavke zadovoljene najvjerojatnije i otvoriti poduzeće. Iz toga slijedi da ukoliko nemamo novca za financiranje poduzeća i ukoliko imamo ovisnu radnu situaciju poduzeće vrlo teško otvoriti. Prema modelu logičke regresije najvažnije je kod otvaranja poduzeća imati izvor financiranja te već postojeća zaposlenost kao poduzetnik. [4]

Na kraju za zaključak bi se moglo izvesti iz rezultata da postoje dva čimbenika o kojima ovisi uspješnost poduzetničkog projekta. Oni su poduzetnikovo financiranje i poduzetnikov prethodni status zaposlenja. Uz pomoć tog rada otvorile su se nove mogućnosti za istraživanje kao što su promatranje uspješnosti projekta prema drugim načinima. Mogućnost primjene ostalih metoda rudarenja podataka te proširenja istraživanja na ostale aspekte projekata. [4]

3. Metodologija i opis podataka

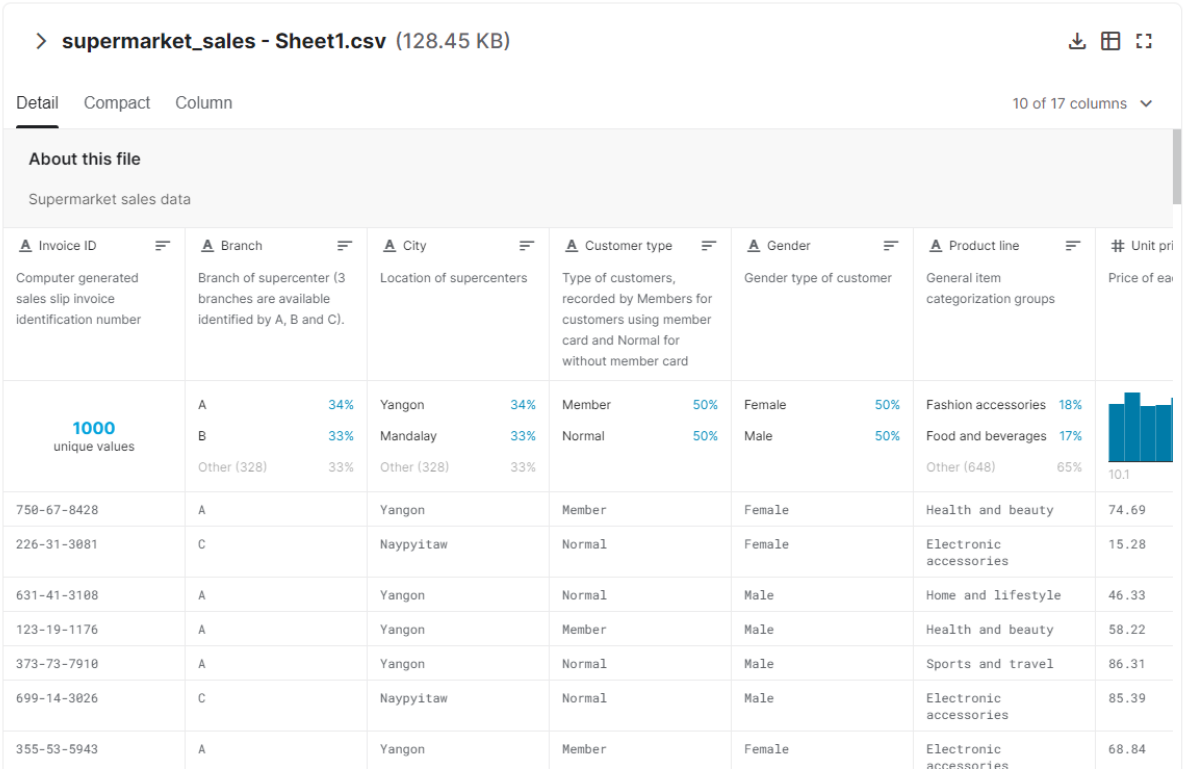
Kao primjer primjene podatkovne analitike u poduzetništu odlučio sam analizirati podatke o prodaji iz supermaketa. Za skup podataka odabrao sam popis prodaje 3 supermarketa različitih branši u najmnogoljudnijim gradovima Yangon, Mandalay te Naypyitaw. Popis podataka sastavljen od atributa kao što su identifikacijski kod računa, branša, grad, tip kupca, spol kupca, cijena proizvoda, količina proizvoda i dr. Kompletni popis atributa navest ćemo u nastavku ovog poglavlja. Popis prodaje ima 1000 jedinstvenih računa sa svim svojim informacijama. [3]

Popis ima 17 atributa. Pripremom skupa podataka uočio sam kako su svi podaci ispravno uneseni te ne bi trebalo biti problema sa kasnijom obradom. Samim time obrada i predviđanja će biti točnija ukoliko nema praznih zapisa u skupu podataka.

Atributi koji su sačinjeni u samom popisu podataka navedeni su u nastavku:

1. ID računa (engl. *Invoice ID*) – računalno generiran ID svakog izdanog računa
2. Branša (engl. *Branch*) – naziv branše kojoj trgovački supermarket pripada (A, B, C)
3. Grad (engl. *City*) – grad u kojemu se nalazi određeni supermarket
4. Tip kupca (engl. *Customer Type*) – tip kupaca koji mogu biti članovi kluba potrošača ili obični kupci
5. Spol kupca (engl. *Gender*) – spol kupca, muški ili ženski
6. Skupina proizvoda (engl. *Product line*) – skupina proizvoda kojima pripada određeni proizvod
7. Cijena jedinice proizvoda (engl. *Unit price*) – cijena po jedinici proizvoda određenog proizvoda
8. Količina proizvoda (engl. *Quantity*) – količina kupljenih proizvoda
9. Takse (engl. *Tax 5%)* – PDV na cijenu proizvoda
10. Ukupna cijena (engl. *Total*) – ukupna cijena računa nakon uračuna PDV-a
11. Datum kupovine (engl. *Date*) – datum izdavanja računa na dan kupovine
12. Vrijeme kupovine (engl. *Time*) – vrijeme kupovine na računu
13. Način plaćanja (engl. *Payment*) – način plaćanja moguć je gotovinski, kartično te elektronski
14. Cijena računa bez poreza (engl. *Cost of goods sold*) – cijena računa umanjena za vrijednost poreza
15. Postotak bruto marže (engl. *Gross margin percentage*) – kolika je marža na vrijednost prodanog proizvoda
16. Bruto dohodak (engl. *Gross income*) – krajnja zarada supermarketu

17. Ocjena (engl. *Rating*) – ocjena kupaca na zadovoljstvo usluge supermarketa



supermarket_sales - Sheet1.csv (128.45 KB)

Detail Compact Column 10 of 17 columns

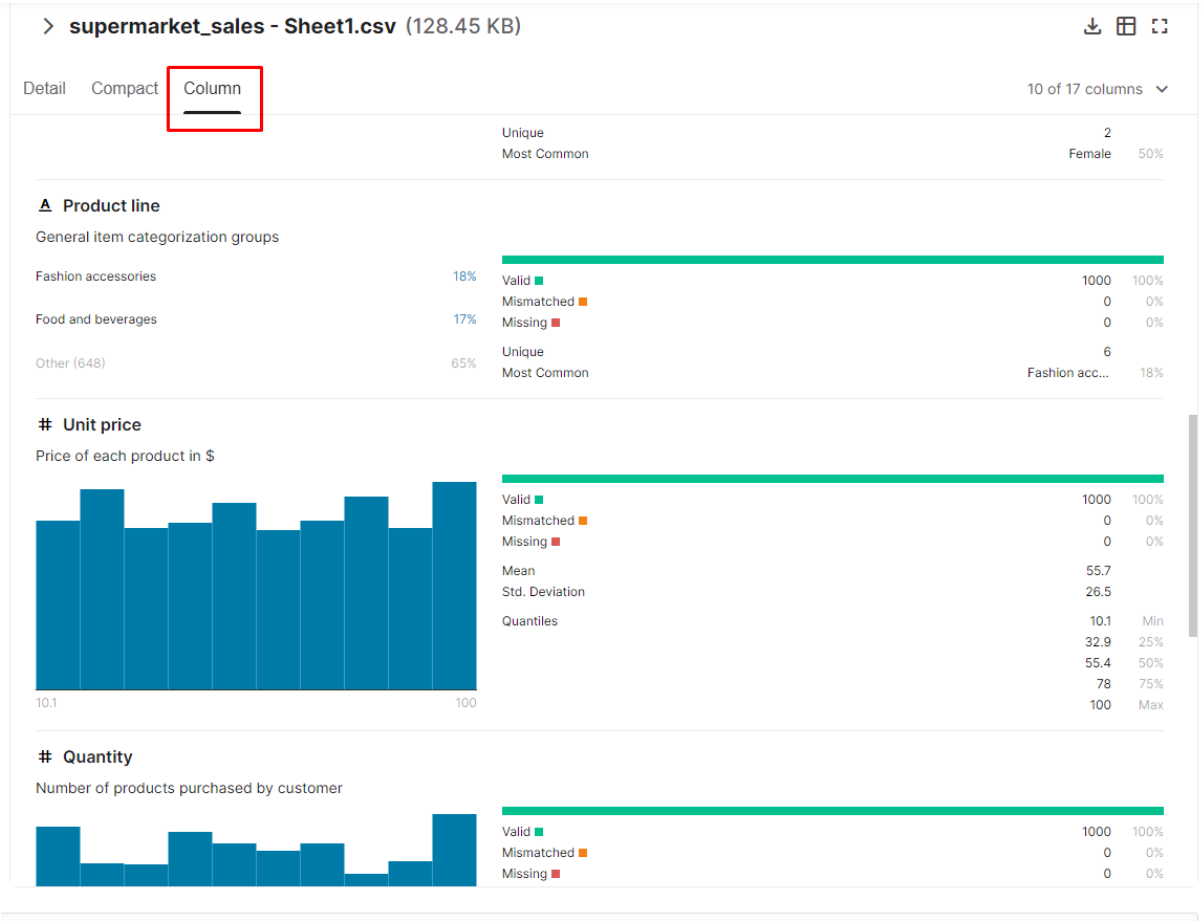
About this file

Supermarket sales data

Invoice ID	Branch	City	Customer type	Gender	Product line	Unit pri
Computer generated sales slip invoice identification number	Branch of supercenter (3 branches are available identified by A, B and C).	Location of supercenters	Type of customers, recorded by Members for customers using member card and Normal for without member card	Gender type of customer	General item categorization groups	Price of ea
1000 unique values	A 34% B 33% Other (328) 33%	Yangon 34% Mandalay 33% Other (328) 33%	Member 50% Normal 50%	Female 50% Male 50%	Fashion accessories 18% Food and beverages 17% Other (648) 65%	10.1
750-67-8428	A	Yangon	Member	Female	Health and beauty	74.69
226-31-3081	C	Naypyitaw	Normal	Female	Electronic accessories	15.28
631-41-3108	A	Yangon	Normal	Male	Home and lifestyle	46.33
123-19-1176	A	Yangon	Member	Male	Health and beauty	58.22
373-73-7910	A	Yangon	Normal	Male	Sports and travel	86.31
699-14-3026	C	Naypyitaw	Normal	Male	Electronic accessories	85.39
355-53-5943	A	Yangon	Member	Female	Electronic accessories	68.84

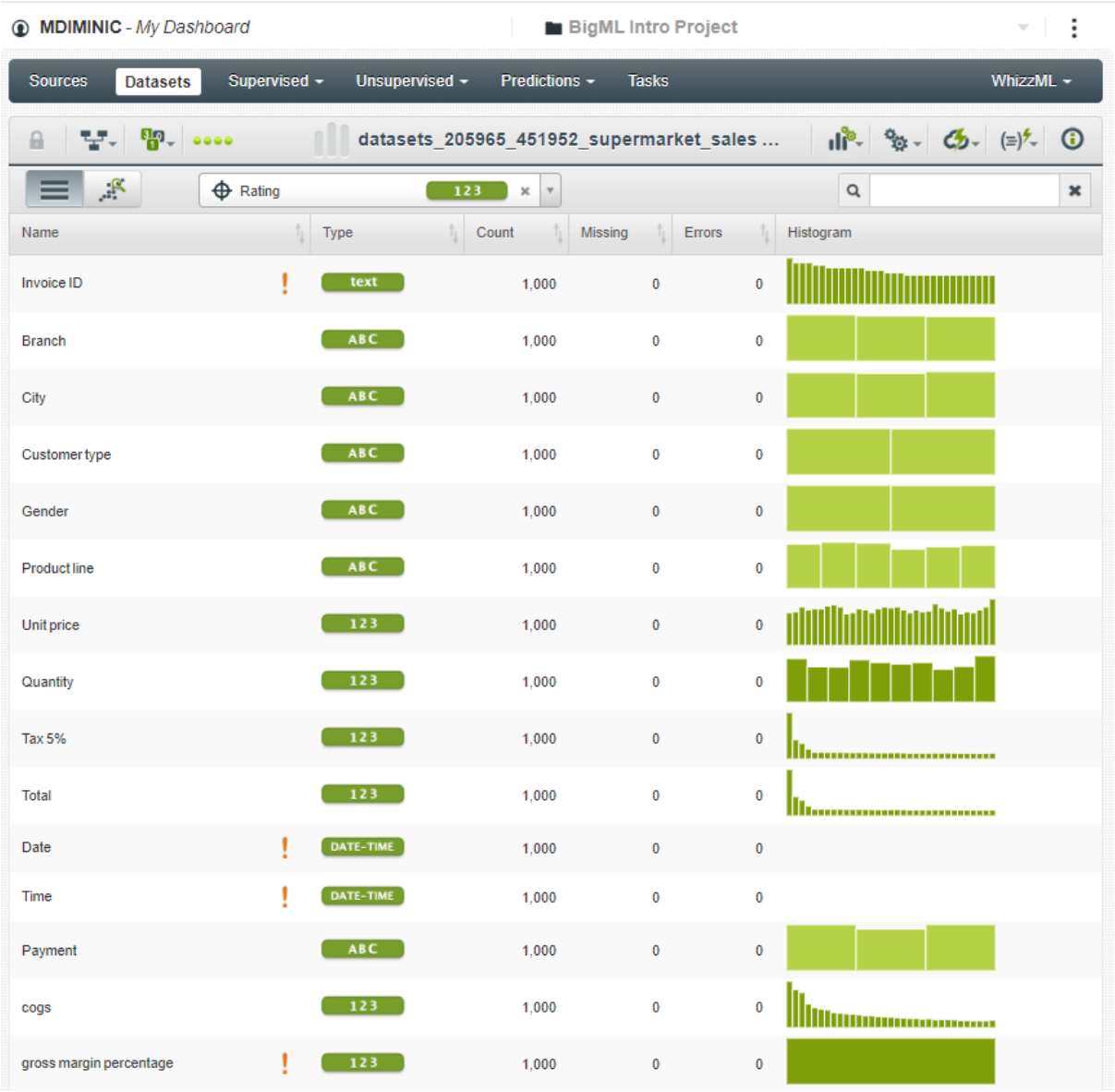
Slika 2. Prikaz popisa podataka *supermarket_sales* s web stranice Kaggle

Kao što možemo vidjeti pregledom danih podataka na web stranici Kaggle (slika 2) svi podaci su ispravno uneseni te je svaki atribut objašnjen posebno za lakše razumijevanje čitatelju odnosno korisniku. Također svaki atribut možemo i posebno pregledati što se može vidjeti na slici 3 ukoliko odaberemo opciju „Column“ pokazati će nam se detalji za svaki atribut. Njegova ispravnost podataka, podudaranje, nedostajanje informacija te isto tako količine, aritmetička sredina, najčešća vrijednost, standardna devijacija, minimalna te maksimalna vrijednost ovisno o kojem tipu atributa se radi.



Slika 3. Prikaz s popisom detalja 3 atributa

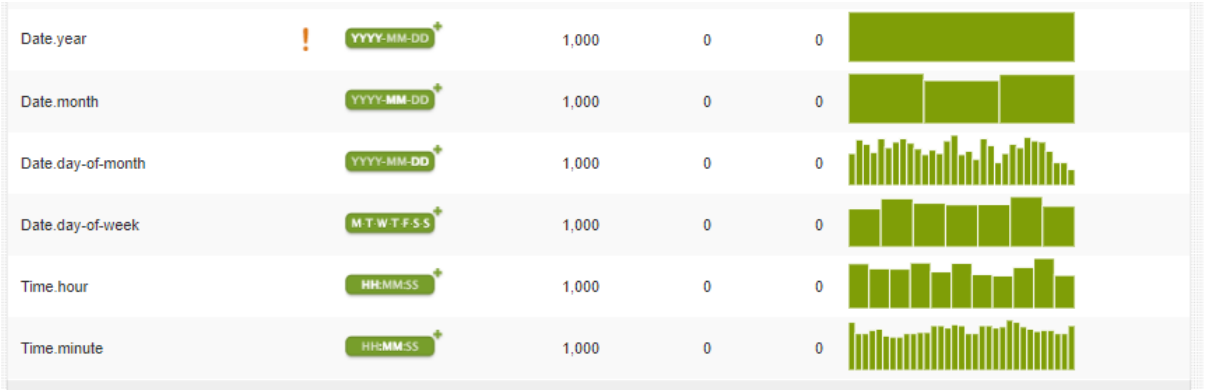
Nakon što smo preuzeli skup podataka u obliku Excel datoteke pod nazivom supermarket_sales.csv sa zapisom i izgledom podataka kao na slici 2 možemo ga učitati u alat BigML kako bismo pogledali pobliže što sve sačinjava popis podataka te kakva je distribucija unutar različitih atributa.



Slika 4. Prikaz skupa podataka *supermarket_sales* učitan u alat BigML

Pregledom učitanih podataka u BigML možemo vidjeti kako imamo numeričke, vremenske, tekstualne te kategorijske vrste atributa (Slika 4). Postoji 1 tekstualni tip podatka te je to *Invoice ID* koji označava broj računa te je kao takav nepotreban za analize, 6 kategorijskih tipova podataka te su to *Branch, City, Customer type, Gender, Product line, Payment*. Sljedeće imamo 8 numeričkih tipova podataka te su oni redom *Unit price, Quantity, Tax 5%, Total, cogs, gross margin percentage, gross income, Rating.* Za kraj nam ostaju vremenski tipovi podataka kojih je 2 i oni su *Date, Time*. Također na slici 4. možemo vidjeti da se svaki atribut sastoji od točno tisuću zapisa te da nijedan atribut nema nedostajućih podataka.

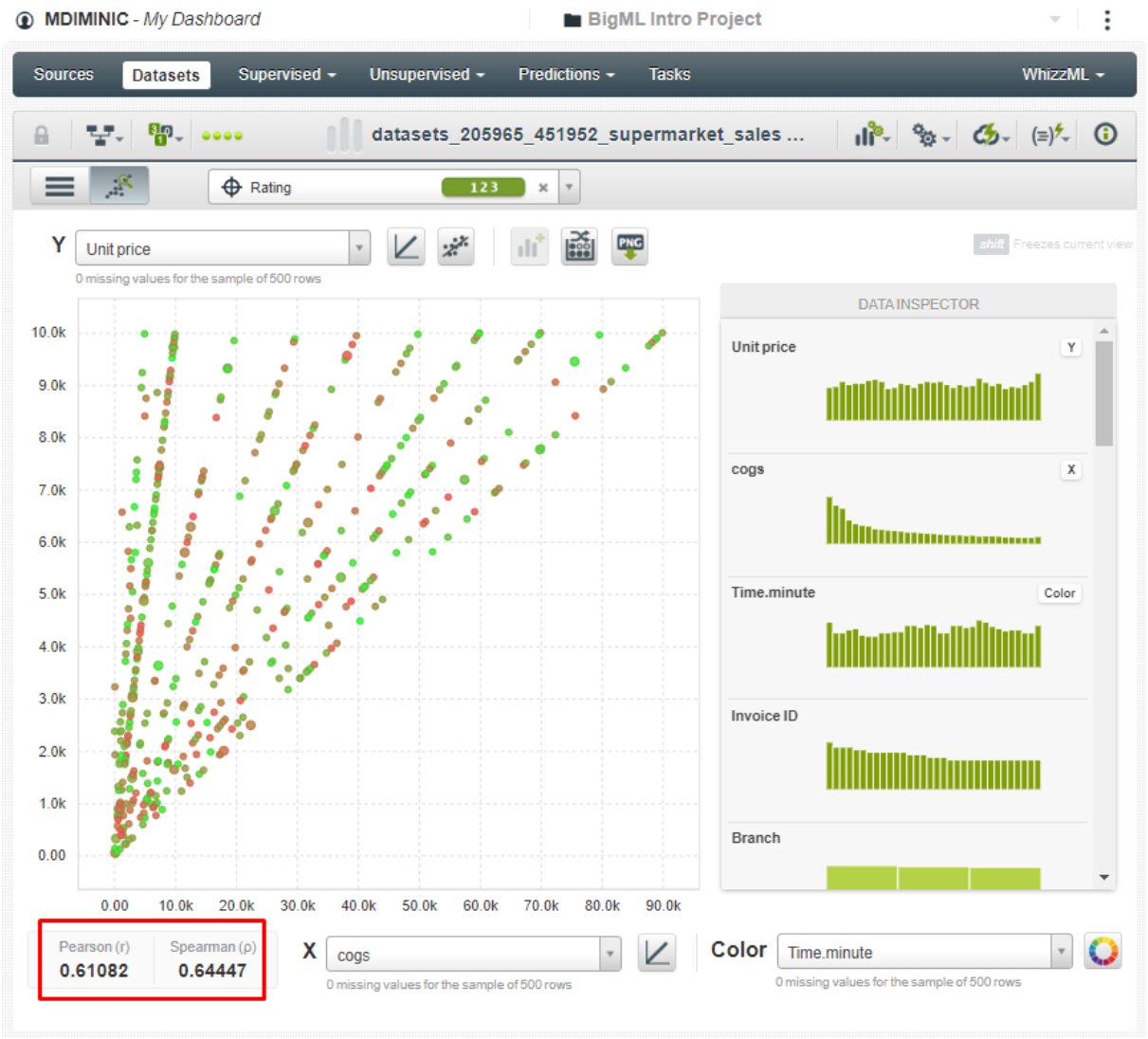
Uz pomoć histograma možemo uočiti da imamo devet atributa s uniformnom distribucijom (*Branch, City, Customer type, Gender, Product line, Unit Price, Quantity, Payment, gross margin percentage*). ID se sastoji od eksponencijalne distribucije isto kao i *Tax 5%, Total* i *cogs*.



Slika 5. Prikaz generiranih atributa

3.1. Dinamički raspršeni graf (eng. *Scatterplot*)

Dinamički raspršeni grafovi (eng. *Scatterplot*) su grafovi koji koriste točke koje predstavljaju vrijednosti različitih atributa. Točka je postavljena između x-osi i y-osi na kojima su vrijednosti dva različita atributa te se pomoću takvih grafova gleda ovisnost nekih atributa o drugome, ovisno o tome koji s kime uspoređujemo. [6]

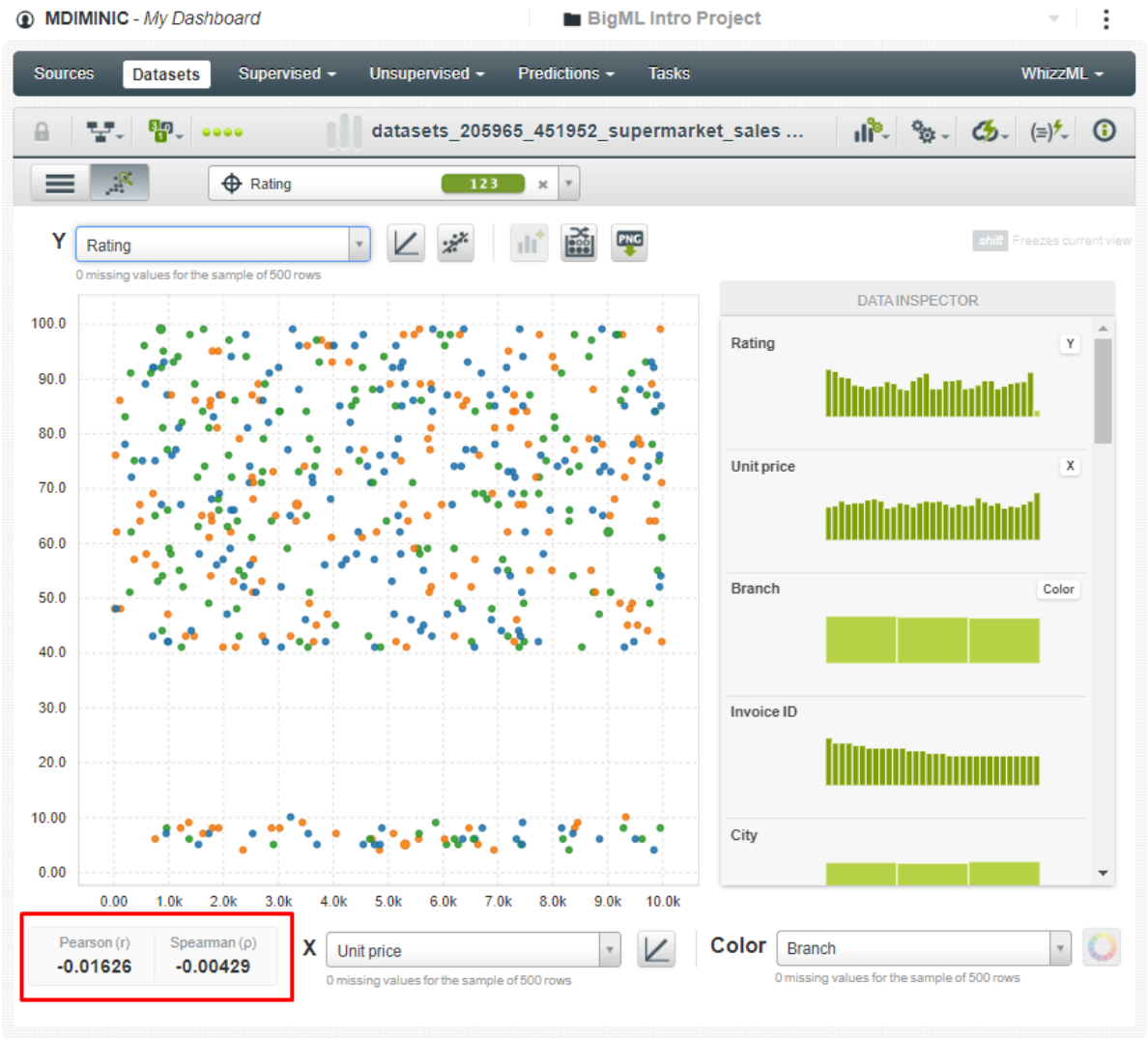


Slika 6. Prikaz ovisnosti varijable cijene proizvoda i nabavne cijene proizvoda

Kako bi prikazali povezanost između dva atributa odabrali smo dva numerička tipa podataka *COGS (**C**ost **o**f **g**oods **s**old)* i cijena po jedinici proizvoda (eng. *Unit price*). Na slici 6 je označeno crvenim pravokutnikom povezanost između ta dva tipa podataka. Pearsonov koeficijent korelacije iznosi 0.61082. Vrijednost koeficijenta prikazuje se u rasponu od -1 do 1 što označava jaku ili slabu povezanost. Možemo donijeti zaključak da su ova dva atributa jako povezani jer je blizu pozitivnoj jedinici. [9]

Spearmanov rank koeficijenta korelacije isto tako pokazuje veoma jaku povezanost zbog toga što je također pozitivan broj i veoma blizu je vrijednosti jedinice. Također se vrijednost Spearmanovog koeficijenta kreće između 1 i -1 te označava jaku ili slabu povezanost između dva atributa ovisno o predznaku. [8]

Logičan zaključak je taj da su jako povezani zbog toga što se oboje odnose na cijenu prodanih proizvoda i trošak prodanih dobara je sadržan u cijeni proizvoda samoga. Odnosno trošak nabave nekog proizvoda i marža su sadržani u finalnoj cijeni nekog proizvoda.



Slika 7. Prikaz ovisnosti varijable cijene jedinice proizvoda i ocjene kupca

Na slici 7 možemo vidjeti slabe ovisnosti varijabli ocjena kupca i cijena jedinice proizvoda. Možemo vidjeti da se ocjena kupca rasporedila na skali od 0 do 100 iako je u skupu podataka bila spremljena od 0 do 10.

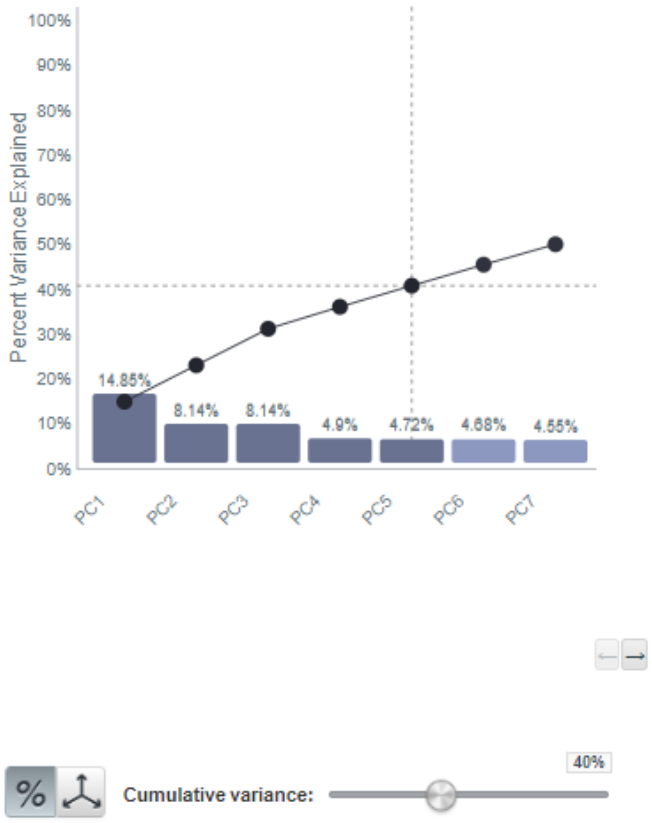
Odabrane dvije varijable nisu povezane, odnosno imaju jako nisku povezanost zbog toga što im koeficijenti Pearson i Spearman odlaze u negativne vrijednosti. Isto to se može reći i za realni svijet jer prodajna cijena proizvoda ne ovisi o ocjeni kupca koji ocjenjuje uslugu u supermarketu.

3.2. Redukcija podataka pomoću ekstrakcije atributa

Redukcija podataka se provodi sa svrhom da maknemo atribute ili karakteristike koje su nevažne za naš problem. Stvorena dimenzija mora biti s manjim brojem atributa. Što manje podataka imamo za obradu i gledanje to će se algoritmi efikasnije provoditi nad danim skupom podataka. Isto tako što su podaci bolje reducirani moguće je primijeniti sve veći broj algoritama različitih funkcija. [7]

U radu je prikazana redukcija podataka koja bi se provodila ukoliko bi odabrani skup podataka bio prevelik ili ukoliko bi bio previše kompliciran za provođenje istraživanja, a s obzirom da je naš skup podataka ne sastoji od previše zapisa za daljnju analizu nije bila potrebna redukcija.

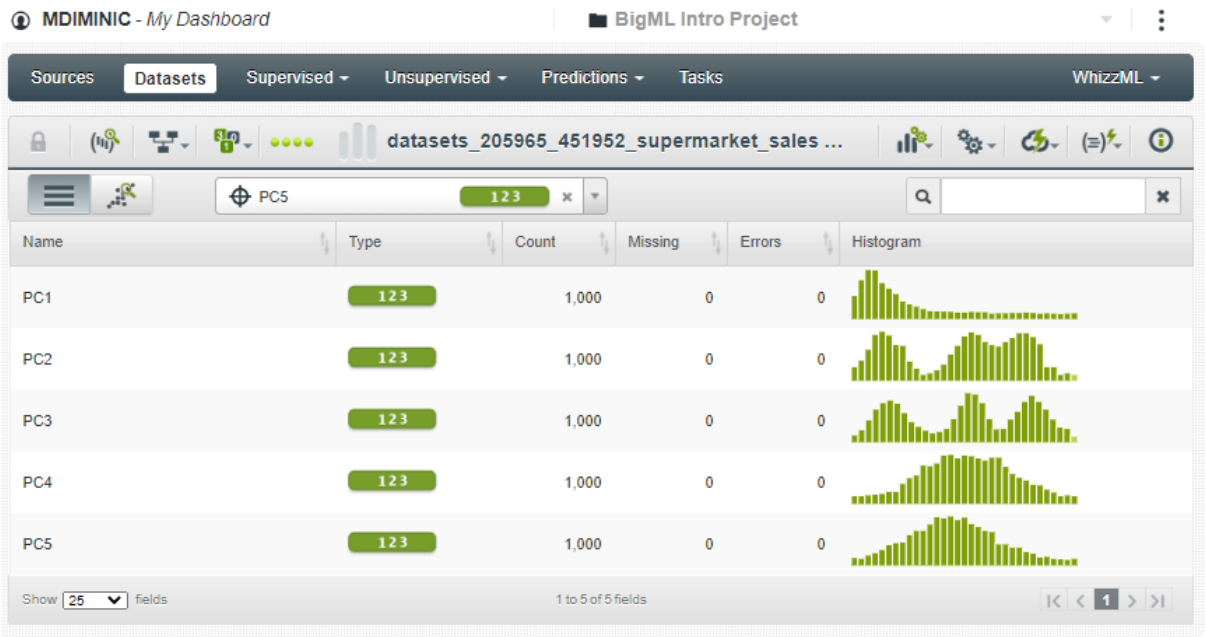
Kako bi reducirali podatke kao mjeru određivanja komponenata odabrali smo aritmetičku sredinu.



Slika 8. Prikaz odabira kumulativne varijance

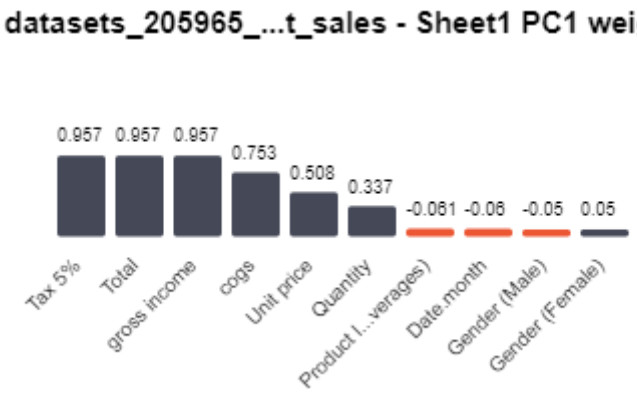
Imamo 17 početnih atributa te s generiranim atributima 23 ukupno no mi ćemo gledati samo početne atribute, a kako bi dobili 5 komponenata postavili smo kumulativnu varijancu na 40% što se može vidjeti na slici 8.
PCA skup podataka (eng. *Dataset*) koji smo dobili nakon generiranja i kreiranja novog izgleda kao što se vidi na slici 9. Sadržan od 5 komponenti koje su sve numeričkog tipa.
PC1 komponenta ima unimodalnu distribuciju s pomakom u lijevo te se kreće od -2.04 do 9.86.
PC2 komponenta ima multimodalnu distribuciju te se kreće od -2.45 do 1.97.
PC3 komponenta ima multimodalnu distribuciju te se kreće od -2.11 do 2.38.
PC4 komponenta ima distribuciju jako sličnu normalnoj, ali nekoliko vrijednosti odskače to jest narušava pravilnu normalnu distribuciju. Zbog tih nekoliko odstupanja može se reći čak i da se radi o unimodalnoj distribuciji. Kreće se od -3.11 do 2.96.
PC5 komponenta ima normalnu distribuciju te se kreće od -2.81 do 3.49.

Što koja komponenta znači te što je određuje opisano je u nastavku ovog završnog rada.



Slika 9. Prikaz dobivenih komponenti nakon provedbe redukcije

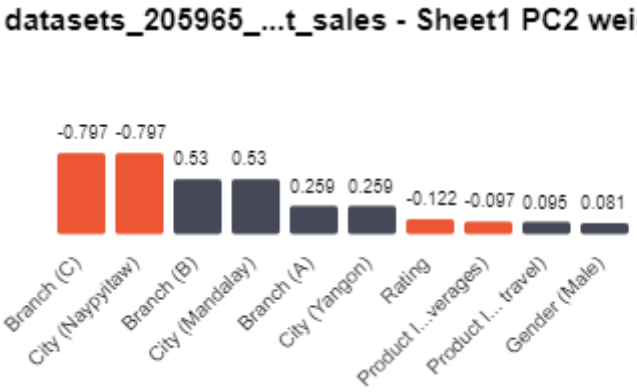
Dobili smo pet komponenata (PC1, PC2, PC3, PC4, PC5) i sastoje se od točno 1 000 zapisa. Isto tako možemo uočiti da su sve komponente koje su nastale numerički tip zapisa. Ne postoje zapisi koji su falični ili imaju grešaka u sebi.



Slika 10. Prikaz komponente PC1

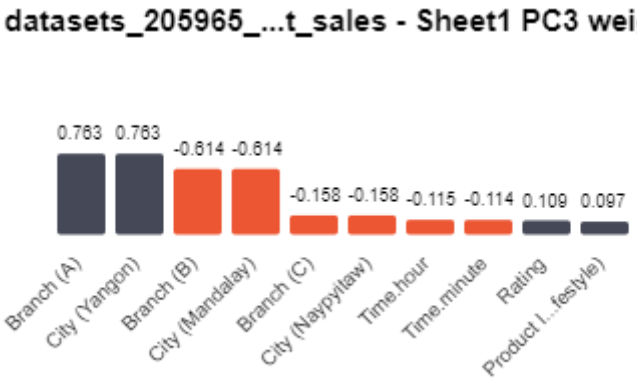
PC1 komponenta sastoji se od Tax (5%), Total, gross income, cogs, Unit price, Quantity, Product line (Food and beverages), Date.month, Gender (Male), Gender (Female). Prikaz komponente PC1 možemo vidjeti na slici 10.

PC1 komponentu određuje Tax (5%), Total, gross income, cogs, Unit price, Quantity.



Slika 11. Prikaz komponente PC2

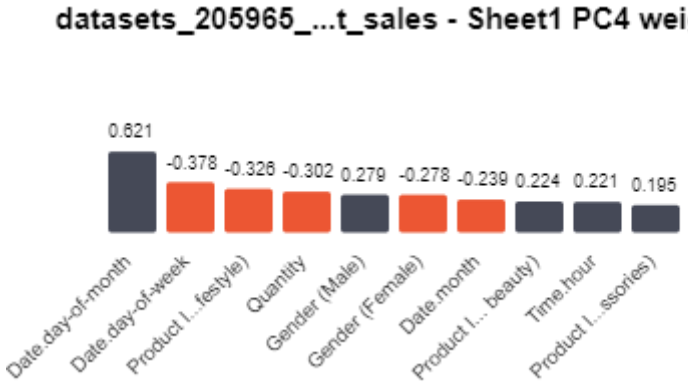
PC2 komponenta sastoji se od Branch (C), City (Naypyitaw), Branch (B), City (Mandalay), Branch (A), City (Yangon), Rating, Product line (Food and beverages), Product line (Sports and travel), Gender (Male). Prikaz komponente PC2 možemo vidjeti na slici 11. Ta komponenta je određena Branch (C), City (Naypyitaw), Branch (B), City (Mandalay).



Slika 12. Prikaz komponente PC3

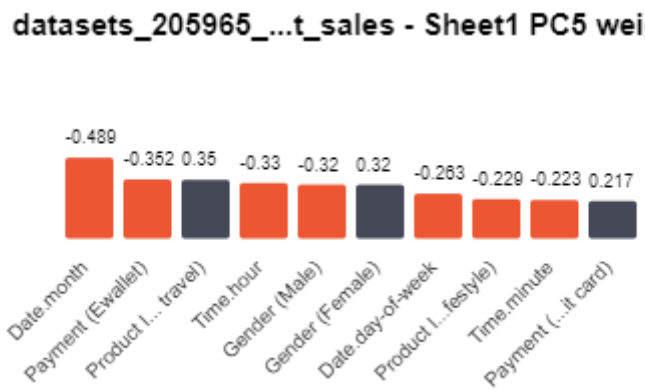
PC3 komponenta sastoji se od Branch (A), City (Yangon), Branch (B), City (Mandalay), Branch (C), City (Naypyitaw), Time.hour, Time.minute, Rating, Product line (Home and lifestyle).

Komponenta je određena uz pomoć Branch (A), City (Yangon), Branch (B), City (Mandalay). Prikaz komponente PC3 možemo vidjeti na slici 12.



Slika 13. Prikaz komponente PC4

Komponenta PC4 sastoji se od Date.day-of-month, Date.day-of-week, Product line (Home and lifestyle), Quantity, Gender (Male), Gender (Female), Date.month, Product line (Health and beauty), Time.hour, Product line (Fashion accessories). Prikaz komponente PC4 možemo vidjeti na slici 13, a ona je određena uz pomoć Date.day-of-month, Date.day-of-week, Product line (Home and lifestyle), Quantity.



Slika 14. Prikaz komponente PC5

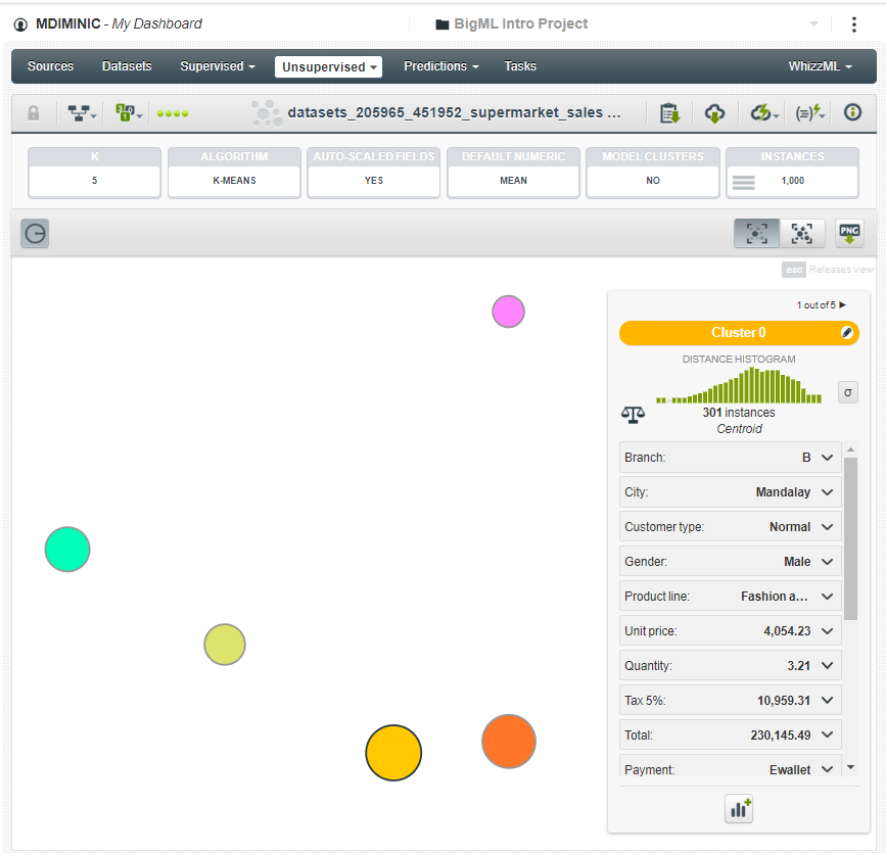
PC5 komponenta sastoji se od Date.month, Payment (Ewallet), Product line (Sports and travel), Time.hour, Gender (Male), Gender (Female), Date.day-of-week, Product line (Home and lifestyle), Time.minute, Payment (credit card). Prikaz komponente možemo vidjeti na slici 14, a ona je određena pomoću Date.month, Payment (Ewallet), Product line (Sports and travel).

4. Analiza podataka

Analizu podataka odabranog skupa o prodaji u 3 različita supermarketa iz 3 različite branše ćemo provesti u ovom poglavlju kako bismo vidjeli s čime radimo i kakve možemo očekivati vrijednosti kod završne interpretacije odnosno koja predviđanja bi se mogla očekivati, a da pomognu u donošenju poslovnih odluka. Sam postupak bit će proveden pomoću klaster analize, stabla odlučivanja te neuronskih mreža kako bismo dobili što bolji uvid s čime radimo. Sam postupak analize svake pojedinačno bit će prikazan u nastavku i objašnjen što smo mogli zaključiti pomoću kreiranih dijagrama.

4.1. Klaster analiza

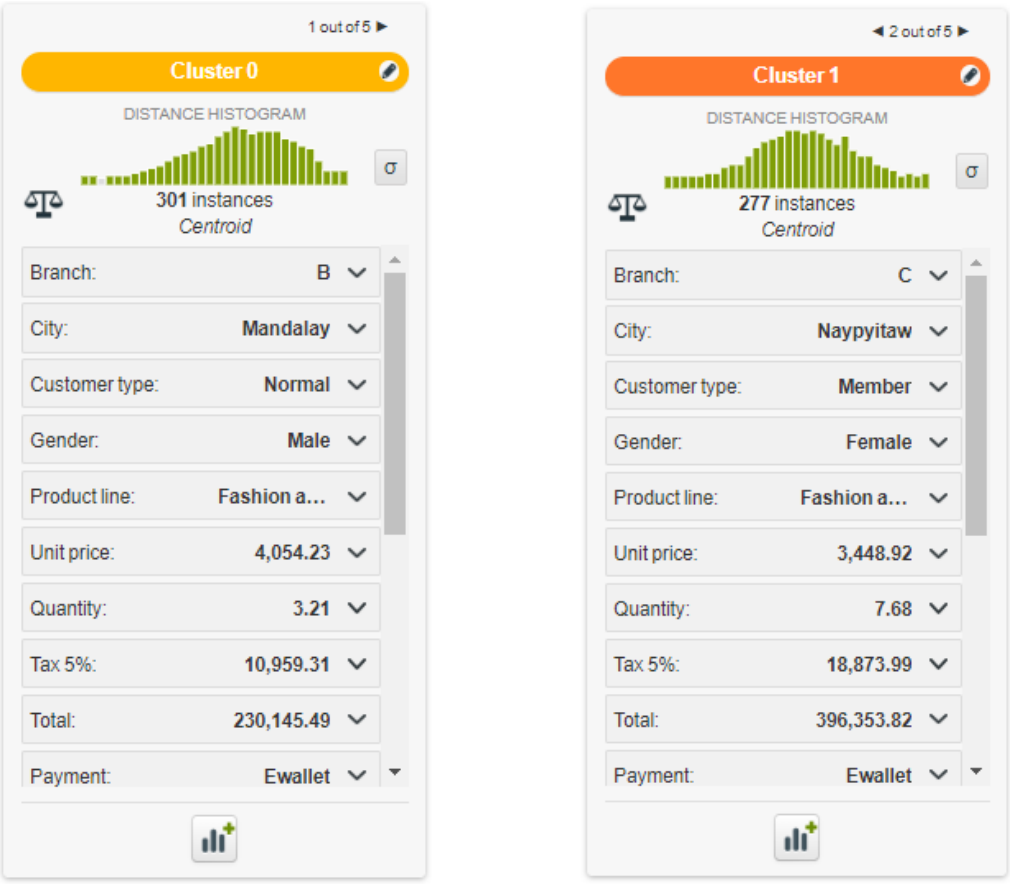
Klaster analiza je skupina tehnika i načina koji su korišteni kako bi klasificirali objekte u povezane skupine zvane klastere. Klaster analiza još se naziva i klasifikacijska analiza ili numerička taksonomija. U klaster analizi ne postoje podaci o grupi ili klasteru kojem bi neki podatak trebao biti smješten prije provođenja same analize. [10]



Slika 15. Prikaz klastera nakon kreiranja

Prije kreiranja klastera morali smo odabrati proizvoljni broj klastera koji želimo da nam generira. Bilo je bitno paziti da se uz pomoć broja klastera mora odrediti optimalni odnosno podjednaki raspored instanci po klasterima. Znači da je distribucija instanci po klasterima podjednaka. Brojku 5 smo smatrali da je najoptimalnija i možemo vidjeti da je broj instanci po klasterima približno jednak odnosno proteže se od 4700 do 6400 instanci po klasteru te jedan klaster oko 450 instanci i još jedan sa samo 34 instance.

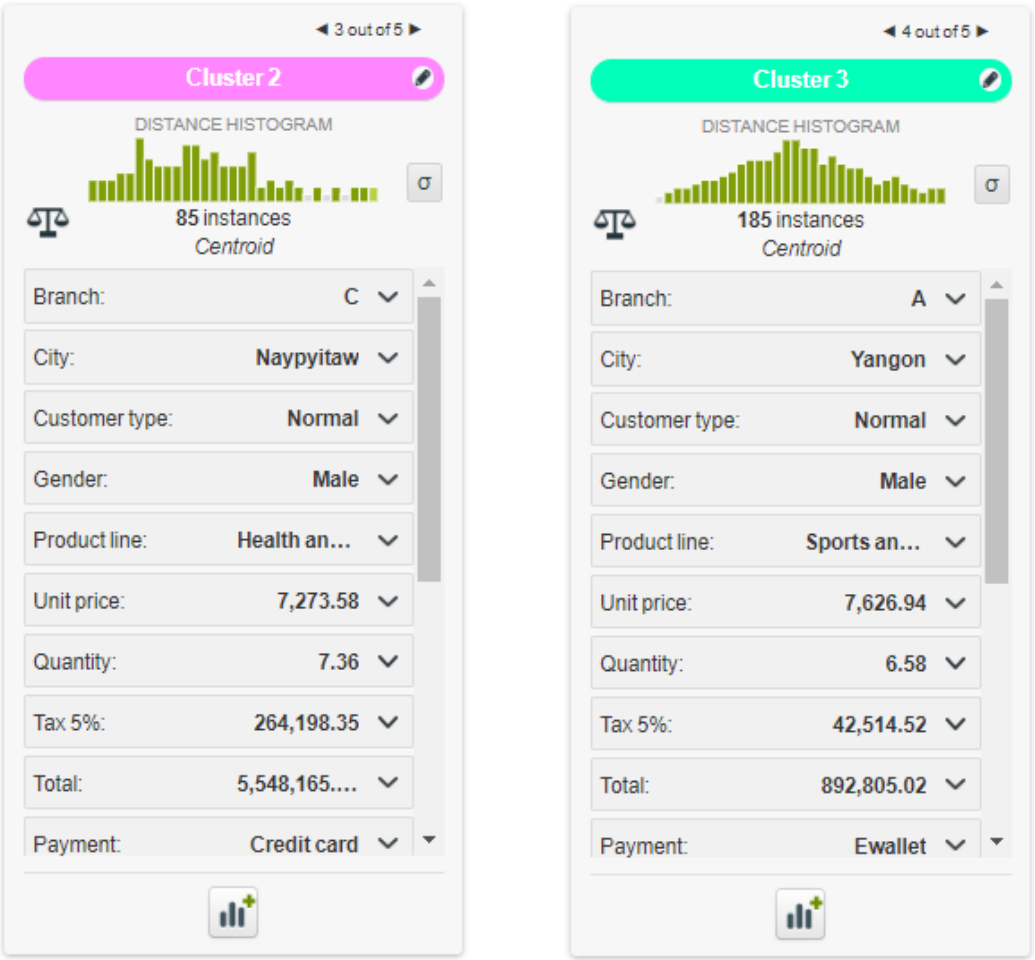
Iako je **klaster broj 2** klaster s najmanje instanci u sebi, ne možemo ga nazvati stršilom jer i dalje sadrži dovoljan broj instanci u usporedbi s ostalim klasterima te smatram da ga nije potrebno izbaciti iz analize.



Slika 16. i 17. Prikaz detalja o klasteru 0 i 1

Na slici 16 možemo vidjeti detalje o klasteru 0 te da ima 301 instancu unutar sebe te njegov histogram pokazuje da se radi o unimodalnoj s pomakom u desno distribuciji. Minimum mu iznosi 0.18, a maksimum 0.48 dok srednja vrijednost iznosi 0.36. Možemo vidjeti da je je spol oko kojeg su grupirani podaci muški (eng. *Male*) te da je vrsta proizvoda odjeća i obuća, isto tako vidimo da je vrsta kupca normalna što označuje da se ne radi o članu kluba supermarketa. Grad oko kojeg su grupirani podaci je Mandalay te način plaćanja je elektroničko plaćanje.

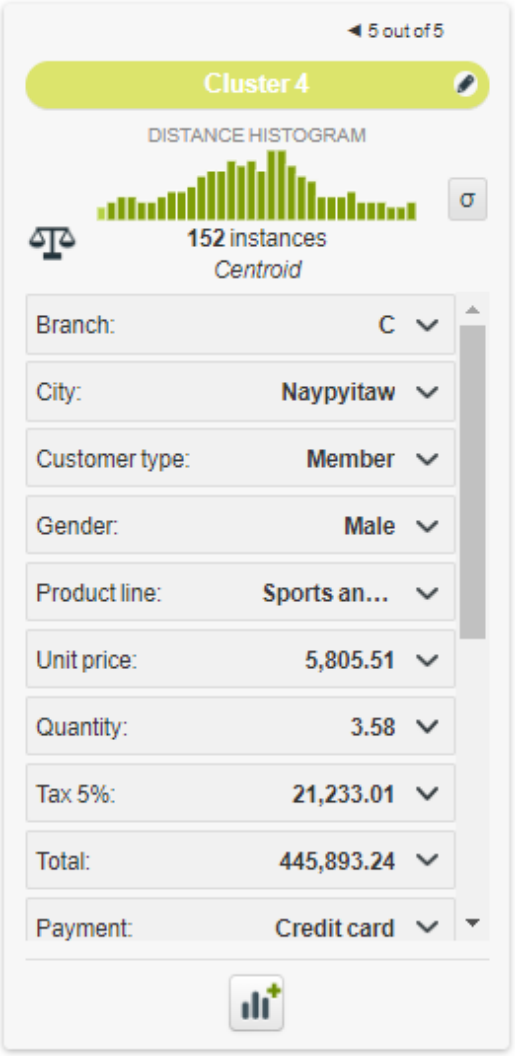
Slika 17 nam prikazuje detalje o klasteru 1 koji ima 277 instanci što je već pošten broj instanci odnosno možemo sa sigurnošću reći da to nije stršilo, histogram nam pokazuje da se radi o normalnoj distribuciji te nekoliko vrijednosti odskače. Grad oko kojeg su grupirani podaci je Naypyitaw te se radi o kupcima ženskog spola. Vidimo da je također odjeća i obuća zajedničko podacima te da je način plaćanja elektroničko plaćanje. Minimum mu iznosi 0.24, a maksimum 0.50 dok je srednja vrijednost 0.38.



Slika 18. i 19. Prikaz detalja o klasteru 2 i klasteru 3

Instanci u klateru 2 imamo 85 te je moguće da je ovaj klaster stršilo jer nije na skupu s ostalim klasterima i ima nešto manji broj instanci. Preko histograma možemo reći da se radi o uniformnoj distribuciji. Prikaz detalja o klasteru 2 možemo vidjeti na slici 18. Instancama je zajedničko da su iz grada Naypyitaw te da im je vrsta kupca normalna odnosno da nisu članovi nikakvih klubova, isto tako možemo vidjet da su pretežno muški kupci i da se najviše kupuje na odjelu za zdravlje i ljepotu (end. *Health and beauty*), plaćanje se najčešće obavlja kreditnom karticom. Minimum iznosi 0.31, maksimum 0.67, a srednja vrijednost iznosi 0.44.

Klaster 3 možemo vidjeti na slici 19 te vidimo da on ima samo 185 instanci u sebi. Broj instanci nam govori da ovaj klaster najvjerojatnije nije stršilo. Isto tako minimum iznosi 0.25, maksimum 0.53, a srednja vrijednost 0.39. Preko histograma možemo vidjeti da se radi o normalnoj distribuciji koja nije potpuno pravilna. U ovom klasteru su podaci grupirani iz grada Yangon te je vrsta kupca nije bila član nekog kluba sa posebnim pogodnostima. Kupci su muškog spola te se kupuje na sporstkom odjelu. Najčešće sredstvo plaćanja je elektronički novac.

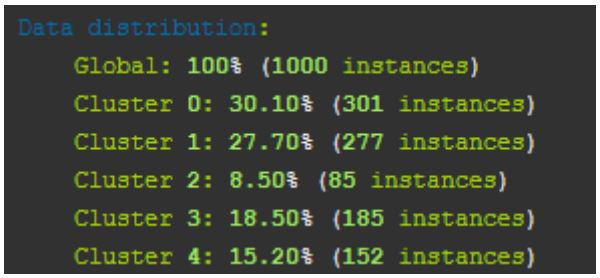


Slika 20. Prikaz detalja o klasteru 4

Klaster 4 ima 152 instance u sebi. Prikaz detalja o klasteru 4 nalazi se na slici 20, a prema histogramu možemo reći da se radi o nepravilnoj normalnoj distribuciji koju narušava nekoliko stupaca. Instancama je zajedničko da su iz grada Naypyitaw te da su kupci bili članovi klubova s posebnim pogodnostima. Isto tako kupci su muškog spola i najviše se kupovalo na

sportskom odjelu. Najčešće sredstvo plaćanja je kreditna kartica. Minimum iznosi 0.21, maksimum 0.52, a srednja vrijednost je 0.39.

Na slici 21 možemo vidjeti raspored podataka u svim klasterima po postotcima i koliki je postotak u kojem klasteru od ukupnog broja instanci prebačen.



Slika 21. Raspored podataka

4.2. Neuronske mreže

Prema Kevinu Gurney-u neuronska mreža je međusobno povezana nakupina jednostavnih elemenata obrade, jedinica ili čvorova, čiji se načini djelovanja otprilike temelji na neuronima kod životinja. Sposobnost obrade mreže je posljedica jačine veza među tim jedinicama, a postiže se kroz proces adaptacije ili učenjem iz skupa primjera za uvježbavanje.[11]

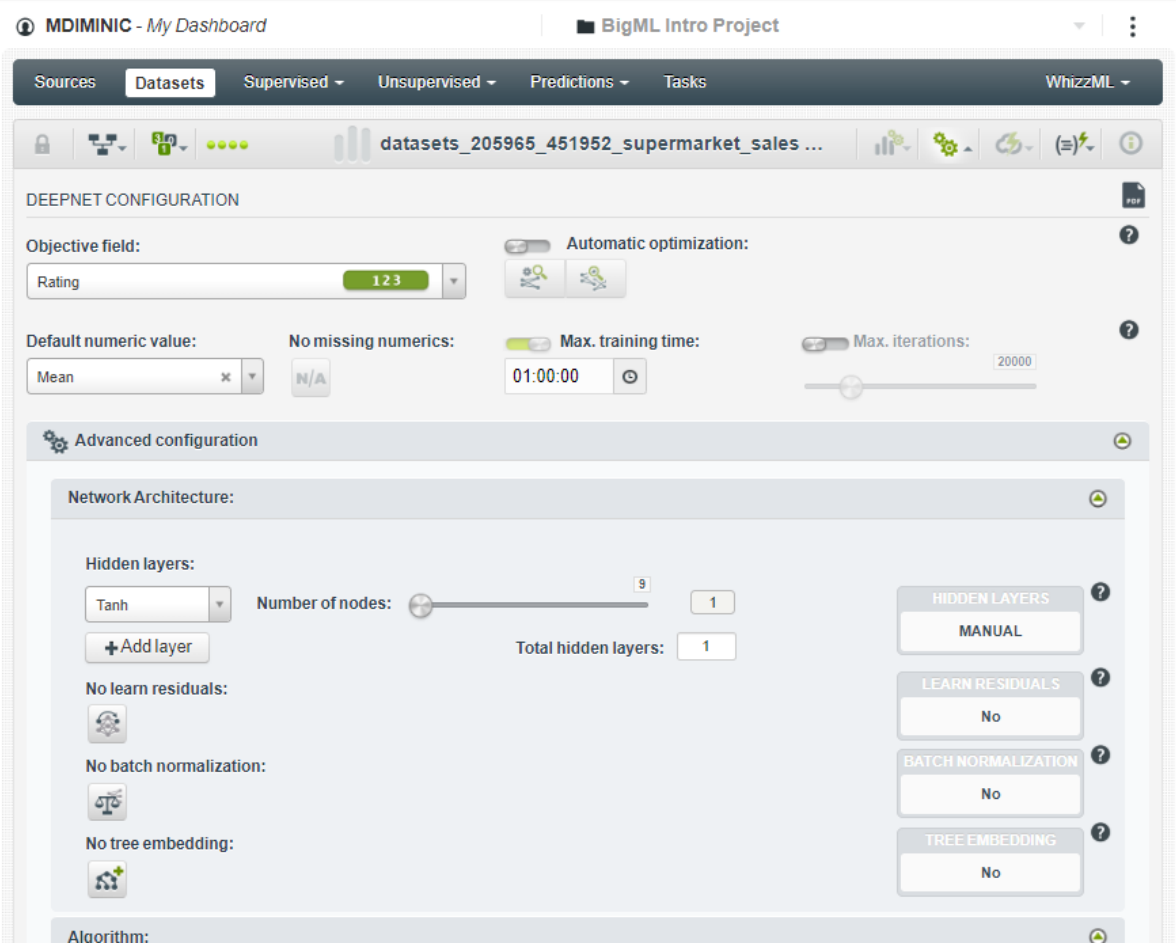
U ovom dijelu rada prikazani su prediktivni modeli primjenom neuronskih mreža. Za dobivanje broja skrivenih neurona moramo izračunati aritmetičku sredinu. S obzirom da imamo 17 ulaznih i 1 izlazni atribut njihov zbroj je 18. Aritmetička sredina od 18 je 9 (dijeljeno s dva). Broj 9 će nam određivati broj skrivenih neurona u konfiguracijama za postavljanje neuronske mreže.

Obradit ćemo dva predviđanja korisna za poslovanje, predviđanje zadovoljstva kupaca danom uslugom te predviđanje kretanja jedinične cijene nekog proizvoda. Predviđanja se rade nad originalnim skupom podataka jer ne sadrži previše zapisa te je s njim moguće napraviti predviđanja bez redukcije podataka.

4.2.1. Predviđanje zadovoljstva kupaca danom uslugom

Kako bismo predvidjeli zadovoljstvo kupaca danom uslugom moramo prvo postaviti potrebnu konfiguraciju neuronske mreže prije kreiranja iste. Prvotno odabiremo ciljani atribut prema kojem želimo odraditi predviđanje, a to će kod nas biti ocjena kupaca odnosno

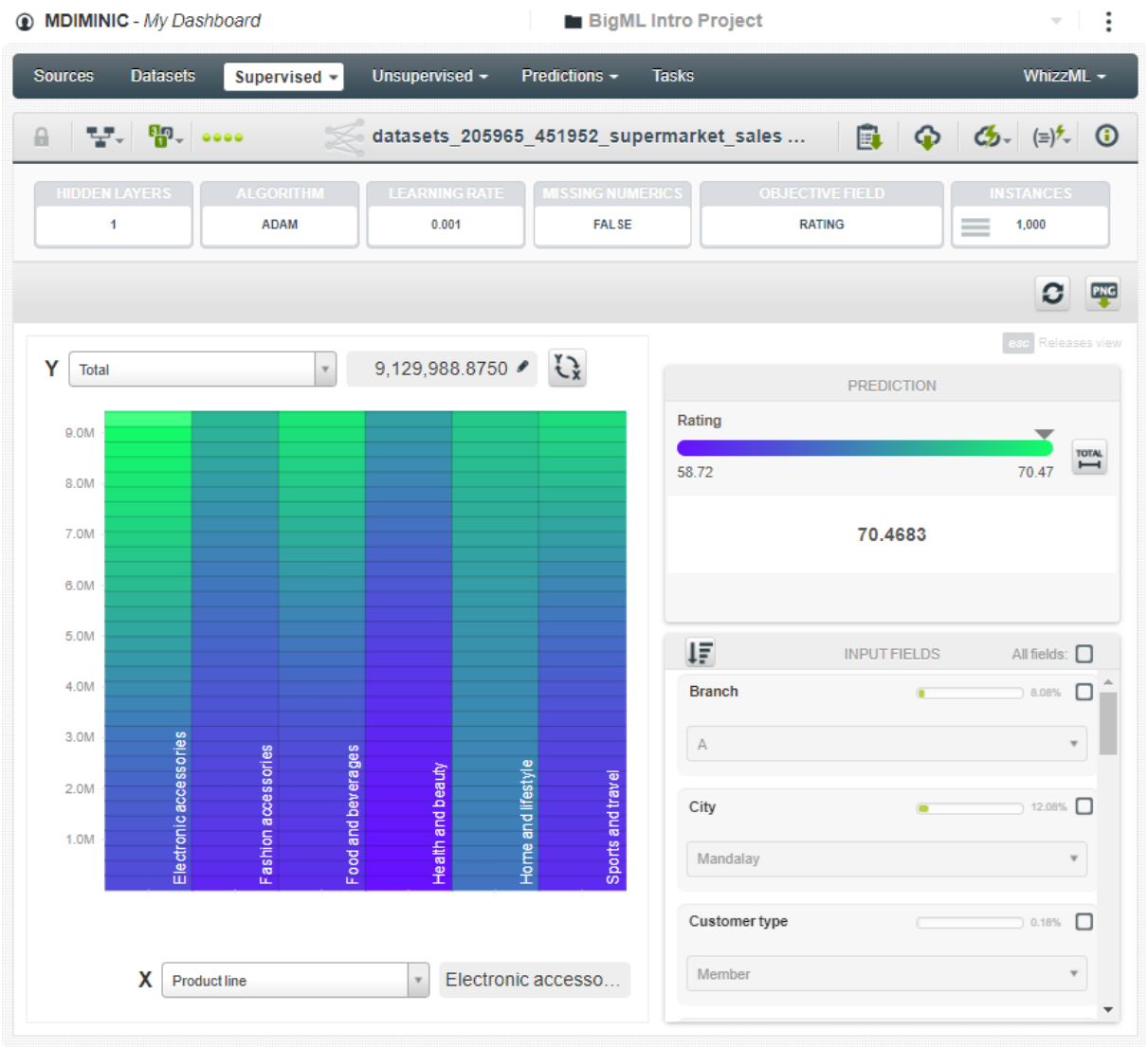
zadovoljstvo (eng. *Rating*) i taj atribut je numerički tip podatka. Za postavke ćemo staviti da se prema aritmetičkoj sredini vrše izračuni, a broj neurona u mreži je 9 što smo izračunali u uvodnom dijelu ovog poglavlja. Konfiguraciju možemo vidjeti na slici 22 te smo nakon toga pokrenuli samu izradu neuronske mreže koju ćemo koristiti za ovo predviđanje



Slika 22. Prikaz postavki neuronske mreže za zadovoljstvo kupaca danom uslugom

Na slici 23 možemo vidjeti kreiranu neuronsku mrežu za prikaz prometa prema vrsti proizvoda koji se prodaju. Y-os nam je ukupna prodaja, a x-os nam označava kategoriju proizvoda. Za primjer smo uzeli kategoriju proizvoda s najvećom prodajom, a to je kategorija elektroničkih naprava (eng. *Electronic accessories*) koja ima najveći porast u zadovoljstvu kupaca kroz vrijeme. Također veliki porast u zadovoljstvu kupaca imaju i kategorije proizvoda za hranu i piće (eng. *Food and beverages*) te kućna i životna pomagala (eng. *Home and lifestyle*) i sportska i putna oprema (eng. *Sport and travel*).

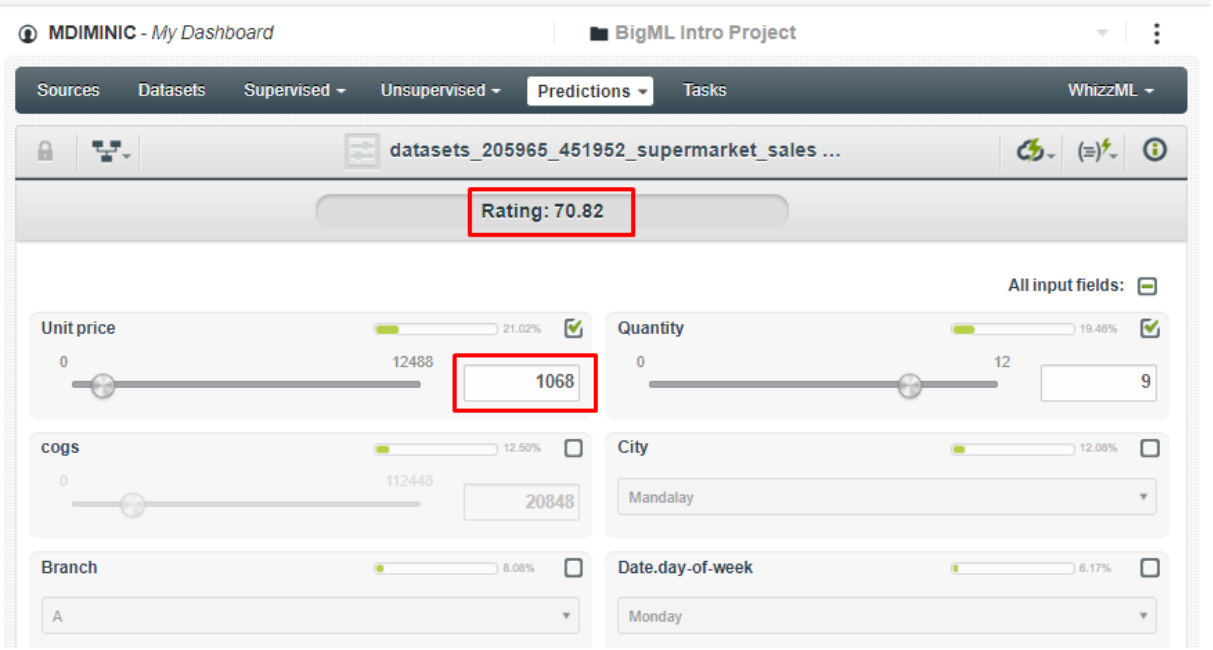
Najmanji porast će imati kategorije proizvoda vezanih uz modu i dodatke odjevne (eng. *Fashion and accessories*) te kategorija proizvoda zdravlje i ljepota (eng. *Health and beauty*).



Slika 23. Prikaz neuronske mreže prema vrsti proizvoda i ukupnoj zaradi

Najvažnija varijabla za odlučivanje je cijena po jedinici proizvoda (*eng. Unit price*) i ona iznosi 21.02%, sljedeća varijabla po važnosti je varijabla količina proizvoda (eng. *Quantity*) te ona iznosi 19.46% prilikom donošenja odluke i još od bitnijih varijabli za donošenje odluke je cijena računa bez poreza odnosno u tablici *cogs* s postotkom od 12.50%. Ostale varijable imaju nešto manje postotke u važnosti prilikom donošenja odluke.

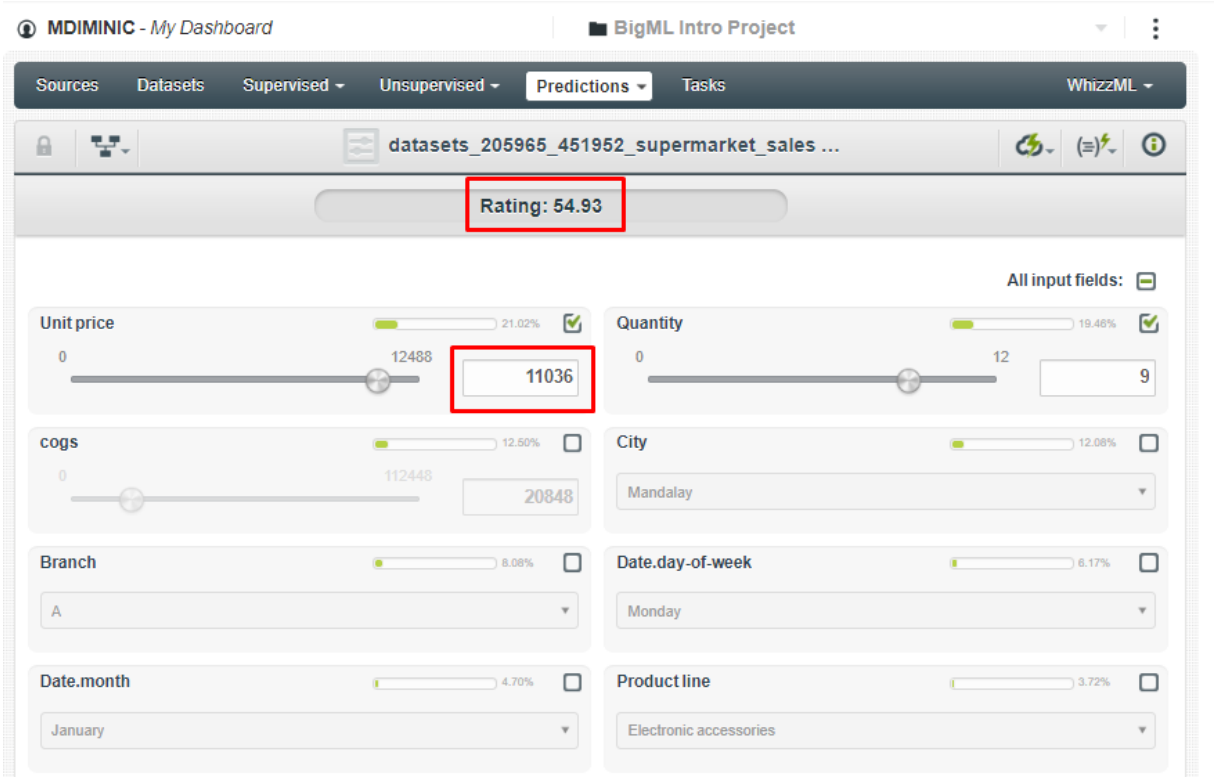
Možemo vidjeti na slici 24 predviđanje za zadovoljstvo kupaca ovisno o cijenama. Za jediničnu cijenu proizvoda postavljenu na 1068 te količinu proizvoda postavljenu na 9 dobili smo zadovoljstvo koje je iznosilo 70.82 (na skali od 0 do 100) što znači da bi kupci bili prilično zadovoljni s takvim jediničnim cijenama proizvoda. Pomoću samo ova dva parametra možemo primijetiti ovisnost jedinične cijene proizvoda o količini kupljenih proizvoda.



Slika 24. Prikaz predviđanja zadovoljstva kupaca danom uslugom ukoliko je cijena jedinice proizvoda manja

Pomoću slike 25 možemo vidjeti što se događa s zadovoljstvom kupaca ukoliko se jedinična cijena proizvoda poveća u neke krajnje granice, a količina kupljenih proizvoda ostane ista. Možemo primijetiti kako je zadovoljstvo kupaca veoma opalo odnosno iznosi 54.93 od 100 što je skoro upola manje. Ovime smo dokazali da kao i u realnom svijetu možemo vidjeti da će kupci biti zadovoljniji obavljenom kupnjom ako su cijene proizvoda manje te ukoliko se cijene proizvoda povećaju ljudi postaju sve više nezadovoljniji obavljenom kupnjom.

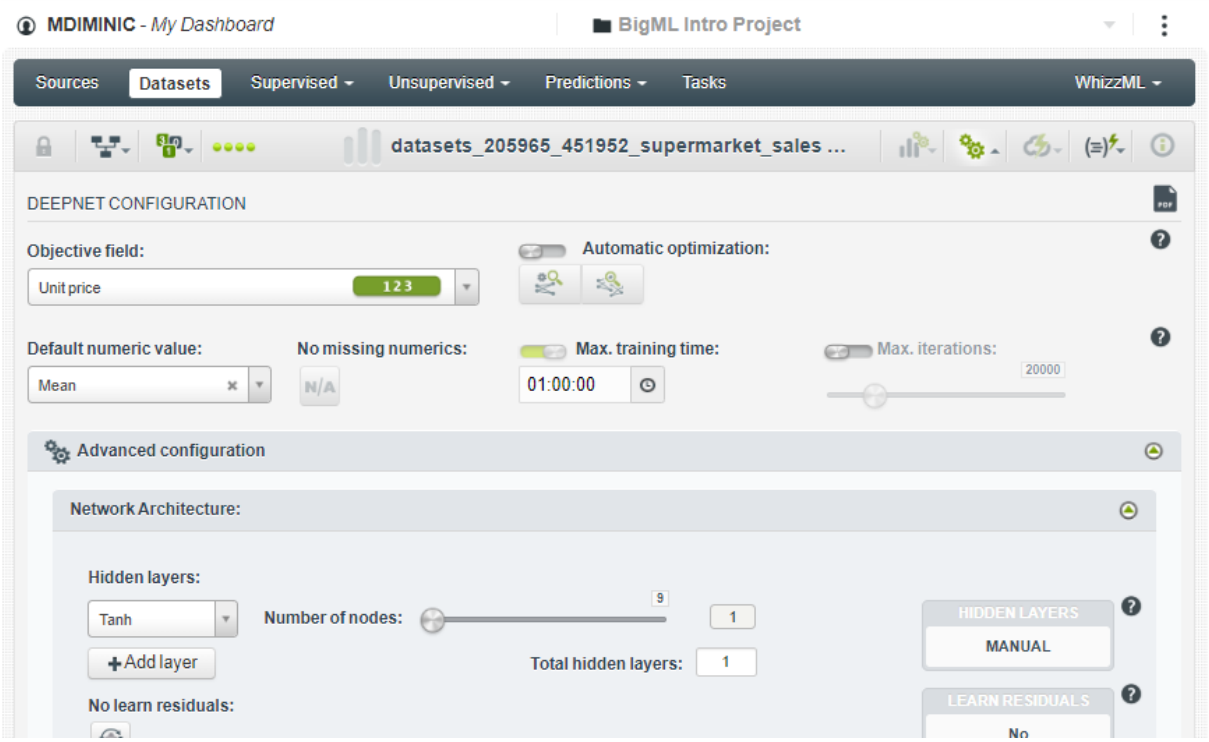
Možemo zaključiti da bi trgovačkim supermarketima bilo bolje držati cijene što je moguće niže od konkurencije kako bi zadovoljstvo kupaca bilo što više te samim time povećalo dolazak novih klijenata u trgovinu.



Slika 25. Prikaz predviđanja zadovoljstva kupaca danom uslugom ukoliko je cijena jedinice proizvoda veća

4.2.2. Predviđanje kretanja jedinične cijene proizvoda

Kao i u prethodnom primjeru moramo za izračun predviđanja postaviti vrijednosti za konfiguraciju neuronske mreže. Za atribut koji želimo promatrati u predviđanju odabrali smo jediničnu cijenu proizvoda (eng. *Unit price*). Varijabla je numeričkog tipa te smo odabrali da će se po aritmetičkoj sredini provoditi predviđanje. Isto tako odabrali smo 9 za broj neurona, tu brojku smo dobili na isti način kao i u prethodnom primjeru. Nakon što smo sve unijeli možemo pokrenuti izračunavanje predviđanja i izrade neuronske mreže. Prikaz konfiguracije neuronske mreže možemo vidjeti na slici 26.

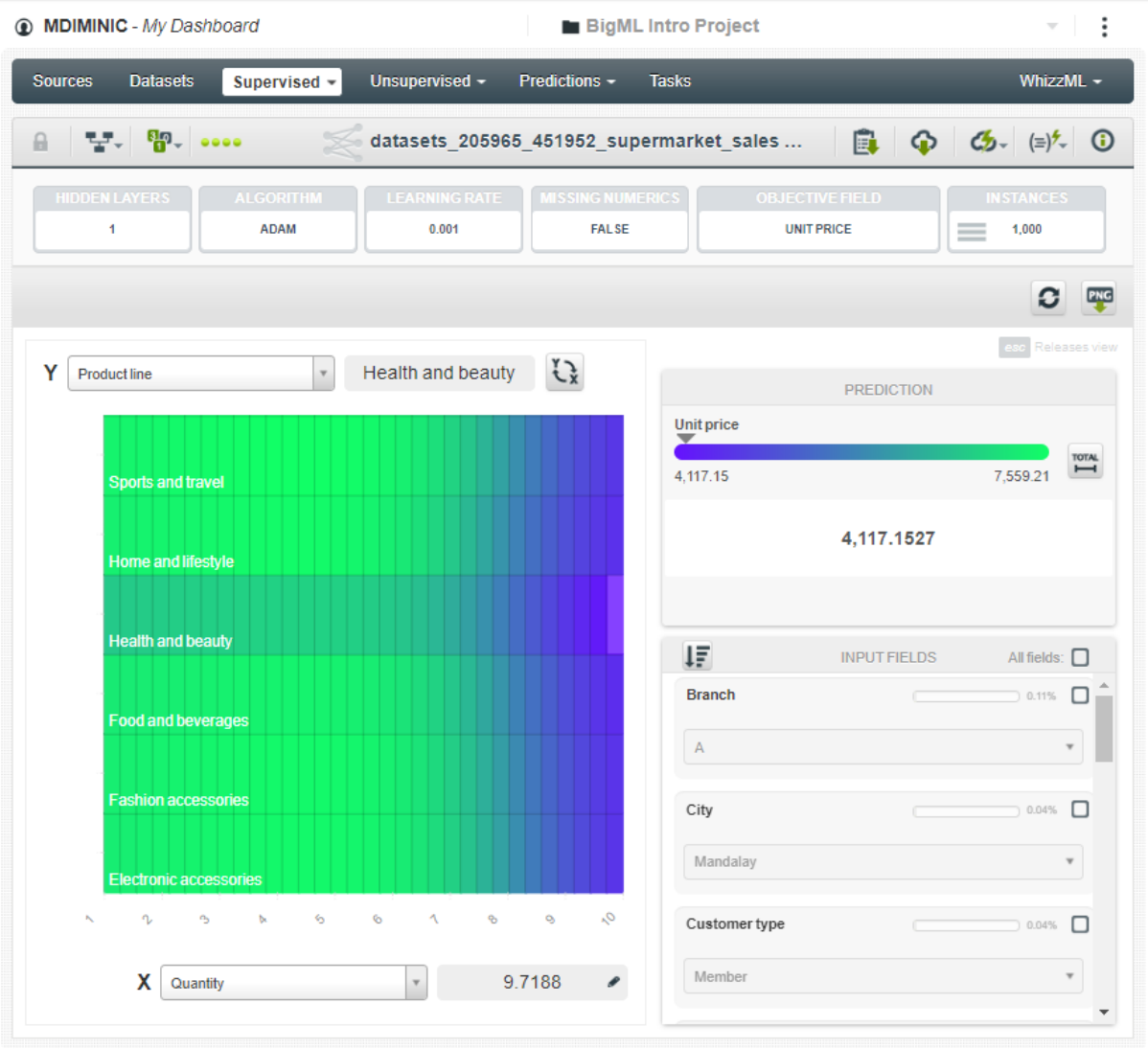


Slika 26. Prikaz postavki neuronske mreže za kretanje jedinične cijene proizvoda

Slika 27. prikazuje također kreiranu neuronsku mrežu, ali za kretanje jedinične cijene proizvoda. Na y-osi, odabran je atribut kategorija proizvoda (eng. *Product line*) koji se odnosi na kategorije proizvoda kojima određeni proizvod pripada, a na x-osi odabran je atribut količine (eng. *Quantity*) koji se odnosi na količinu kupljenih proizvoda.

Ako promatramo kategoriju proizvoda na y-osi te količinu proizvoda na x-osi možemo uočiti jednu zanimljivost, a to je da će proizvodi u kategoriji za zdravlje i ljepotu (eng. *Health and beauty*) doživljava pad vrijednosti jedinične cijene proizvoda mnogo brže nego ostale jedinične cijene proizvoda iz drugih kategorija proizvoda. Sve ostale kategorije proizvoda imaju postepeni pad u vrijednosti jedinične cijene proizvoda.

Može se zaključiti da bi se u budućnosti supermarketi trebali fokusirati na sve preostale kategorije proizvoda osim kategorije vezane za zdravlje i ljepotu.

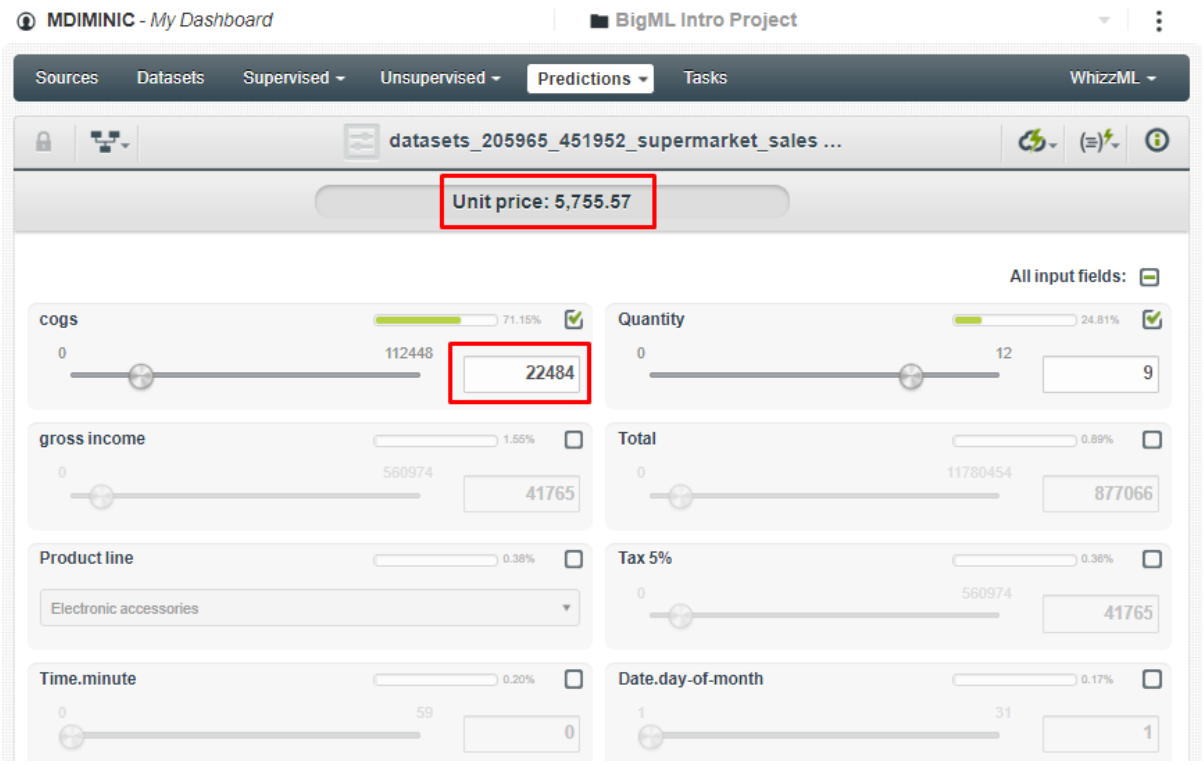


Slika 27. Prikaz neuronske mreže prema kategoriji proizvoda i količini proizvoda te jedinična vrijednost proizvoda

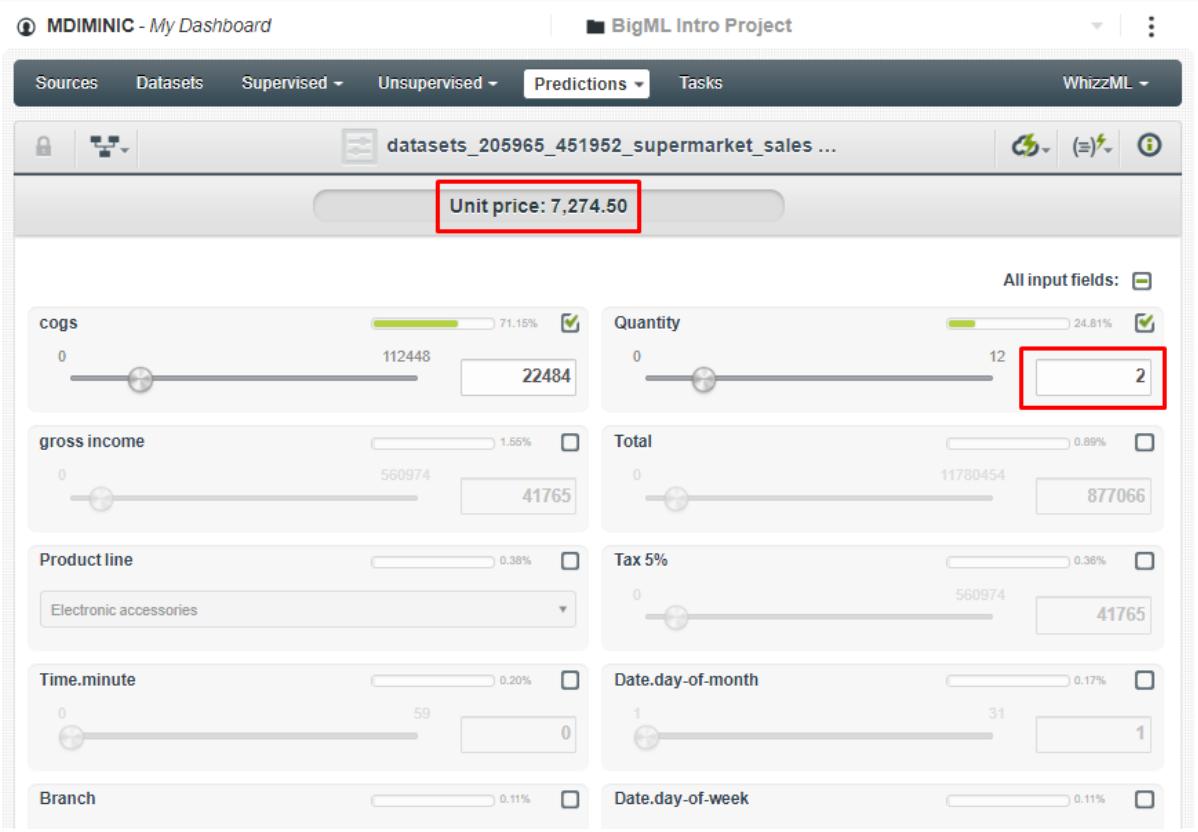
Kod predviđanja jedinične vrijednosti proizvoda najvažnija varijabla za odlučivanje je vrijednost prodanih proizvoda bez poreza dodanog odnosno u tablici *cogs* te on iznosi 71.15% važnosti prilikom donošenja predviđanja. Sljedeći atribut po važnosti za donošenje odluke je količina proizvoda koji ima važnost od 24.81% što je isto veoma velika vrijednost. Ostali atributi nemaju toliku ulogu o jediničnoj vrijednosti proizvoda te smo ih prilikom predviđanja isključili kako bismo dobili što točnija predviđanja.

Slika 28 prikazuje predviđanje za vrijednost jedinice proizvoda ukoliko je varijabla *cogs* (vrijednost računa bez poreza) postavljena veoma nisko odnosno u našem slučaju na 22 484 te varijabla količina (eng. *Quantity*) na 9. Možemo vidjeti da smo dobili jediničnu cijenu proizvoda 5 755 što nije toliko velika cijena. Za samu usporedbu napravili smo dodanu analizu

vezanu uz smanjenje količine na 2. Možemo vidjeti na slici 29 kako je jedinična cijena proizvoda veoma porasla odnosno tada iznosi 7 274 što je dosta više nego u prvom slučaju. Možemo zaključiti da količina proizvoda i vrijednost računa bez poreza utječu na iznos jedinične cijene proizvoda.



Slika 28. Prikaz predviđanja jedinične cijene proizvoda ukoliko je količina veća

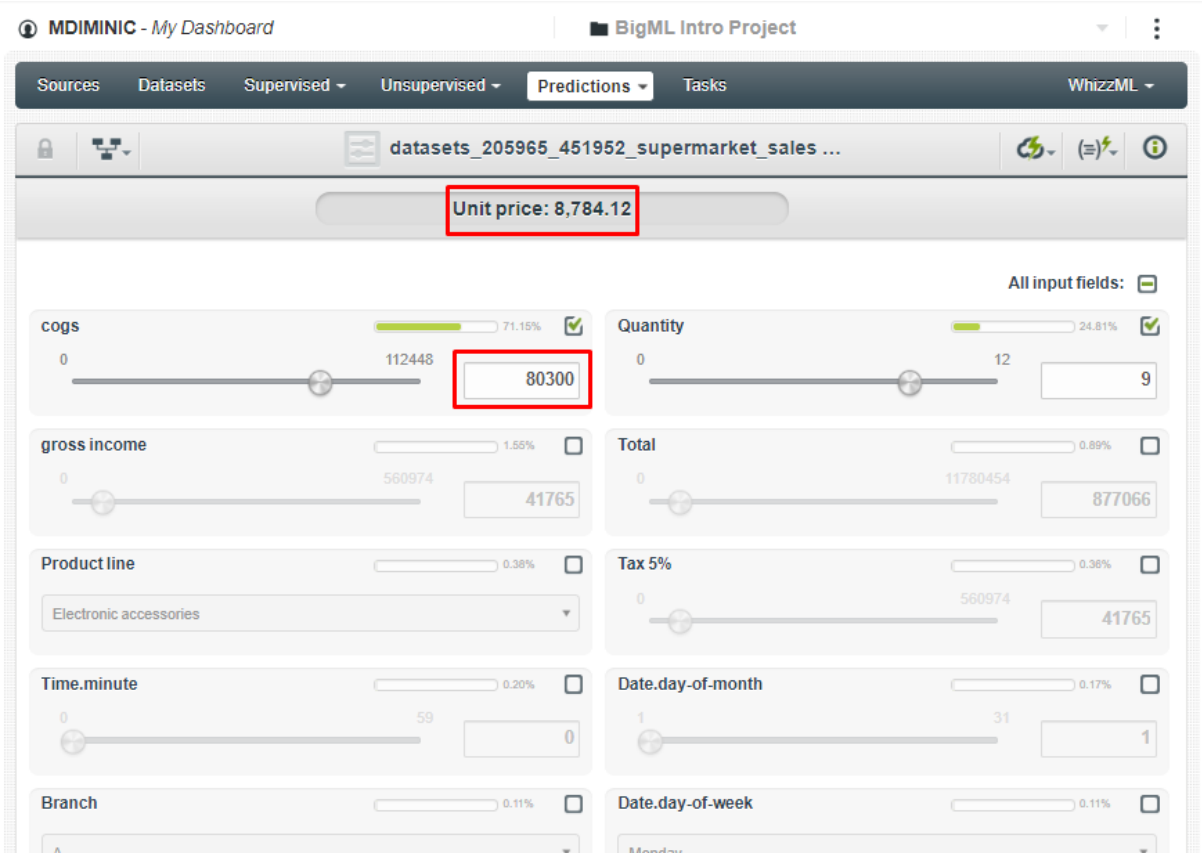


Slika 29. Prikaz predviđanja jedinične cijene proizvoda ukoliko je količina manja

Na slici 29 možemo vidjeti postavke i izračun jedinične cijene proizvoda ukoliko je količina postavljena na manju vrijednost, a *cogs* (vrijednost računa bez poreza) ostavljen na istoj vrijednosti kao i prethodni izračun.

Povećavanjem vrijednosti varijable *cogs* (vrijednost računa bez poreza) povećava se i jedinična vrijednost proizvoda. To znači ukoliko se količina poveća tada će se jedinična cijena smanjiti i obrnuto, a ukoliko se *cogs* (vrijednost računa bez poreza) poveća tada će se povećati i jedinična cijena proizvoda i obrnuto. Prikaz predviđanja vidi se na slici 30.

Možemo zaključiti da će jedinična vrijednost proizvoda rasti ili padati ovisno o količini ili varijabli *cogs* (vrijednost računa bez poreza).



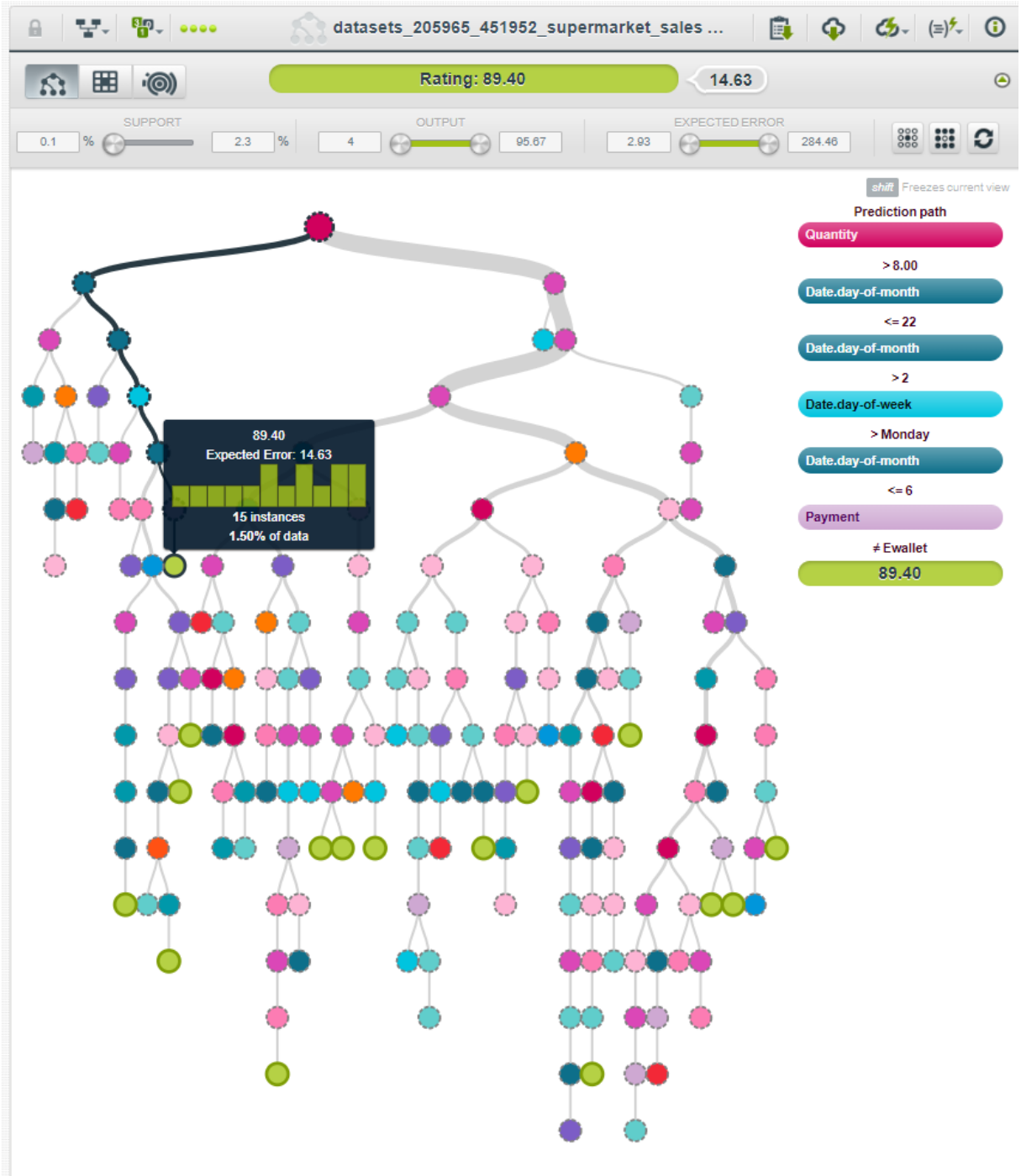
Slika 30. Prikaz predviđanja jedinične cijene proizvoda ukoliko je *cogs* veći

4.3. Stablo odlučivanja

Stabla odlučivanja koriste se u dva područja znanosti, a to su prilikom analize odlučivanja te u strojnom učenju. Nas će zanimati više strojno učenje kako bismo mogli predvidjeti ponašanje naših tržišta. 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. [12]

4.3.1. Predviđanje zadovoljstva kupaca danom uslugom

Za prikaz predviđanja putem stabla odlučivanja, promatrat ćemo varijablu koja opisuje zadovoljstvo kupaca putem ocjena, tj. Rating. Navedena varijabla postavljena je kao Objective field. Kako se radi o numeričkoj varijabli, umjesto pouzdanosti prikazana je očekivana greška koja iznosi 14,63 što možemo vidjeti na sljedećoj slici.

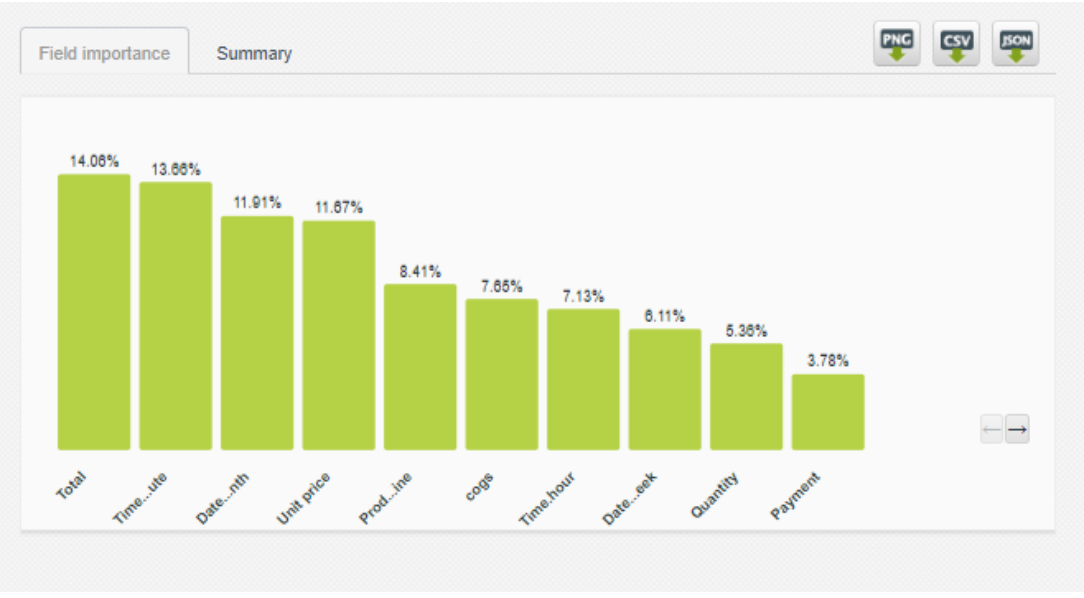


Slika 31. Prikaz puta predviđanja putem stabla odlučivanja – zadovoljstvo korisnika

Slika 31. prikazuje put predviđanja podebljan crnom bojom. Taj put prikazuje zadovoljstvo korisnika od čak 89,40 od maksimalnih 100. Put označen na slici opisuje korake koji u konačnici dovode do visokog zadovoljstva korisnika. Ako promatramo put od vrha prema dolje i pravilo prikazano na desnoj strani slike, možemo vidjeti da je korijen uvjetovan kupljenom količinom nekog proizvoda. Zatim se u sljedeća dva koraka provjerava koji dan u mjesecu je obavljena kupnja proizvoda, odnosno proizvod treba biti kupljen nakon 8-og. i prije ili tijekom 22-og. u bilo kojem mjesecu. Sljedeće pravilo provjerava kojeg dana u tjednu je obavljena kupnja, odnosno to može biti bilo koji dan nakon ponedjeljka, tj. kupnja ne smije biti obavljena ponedjeljkom. Zatim se opet provjerava kojeg dana u mjesecu je obavljena kupnja, odnosno kupnja treba biti obavlja prije ili tijekom 6.-og u mjesecu. Posljednji uvjet određuje da se plaćanje vrši gotovinom ili kartično, dakle ne smije se obaviti kupnja s elektroničnim novcem.

Kao rezultat odabranog puta predviđanja, dobili smo predviđanje zadovoljstva kupaca od 89,40 s očekivanom greškom od 14,63 koja nije od velikog značaja.

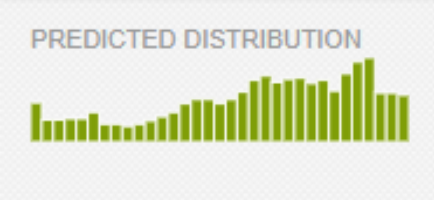
Slika 32. prikazuje koje od varijabli imaju najveći utjecaj prilikom predviđanja putem stabla odlučivanja. Varijable s najvećim utjecajem su: ukupna cijena računa s PDV-om (Total), minuta kupnje proizvoda u satu (Time.minute), mjesec kupnje proizvoda (Date.month), jedinična cijena proizvoda (Unit Price), kategorija proizvoda (Product Line). Navedene varijable najviše utječu na tok grananja. Možemo uočiti da sve varijable iznose iznad 8%. Ostale varijable ispod 8% također imaju utjecaj, ali ne toliko značajan. U prikazanom primjeru, tok grananja određivali su manje značajnije varijable, odnosno varijable s utjecajem manjim od 8% (količina, način plaćanja, dan u mjesecu i dan u tjednu).



Slika 32. Prikaz utjecaja varijabli kod stabla odlučivanja – zadovoljstvo korisnika

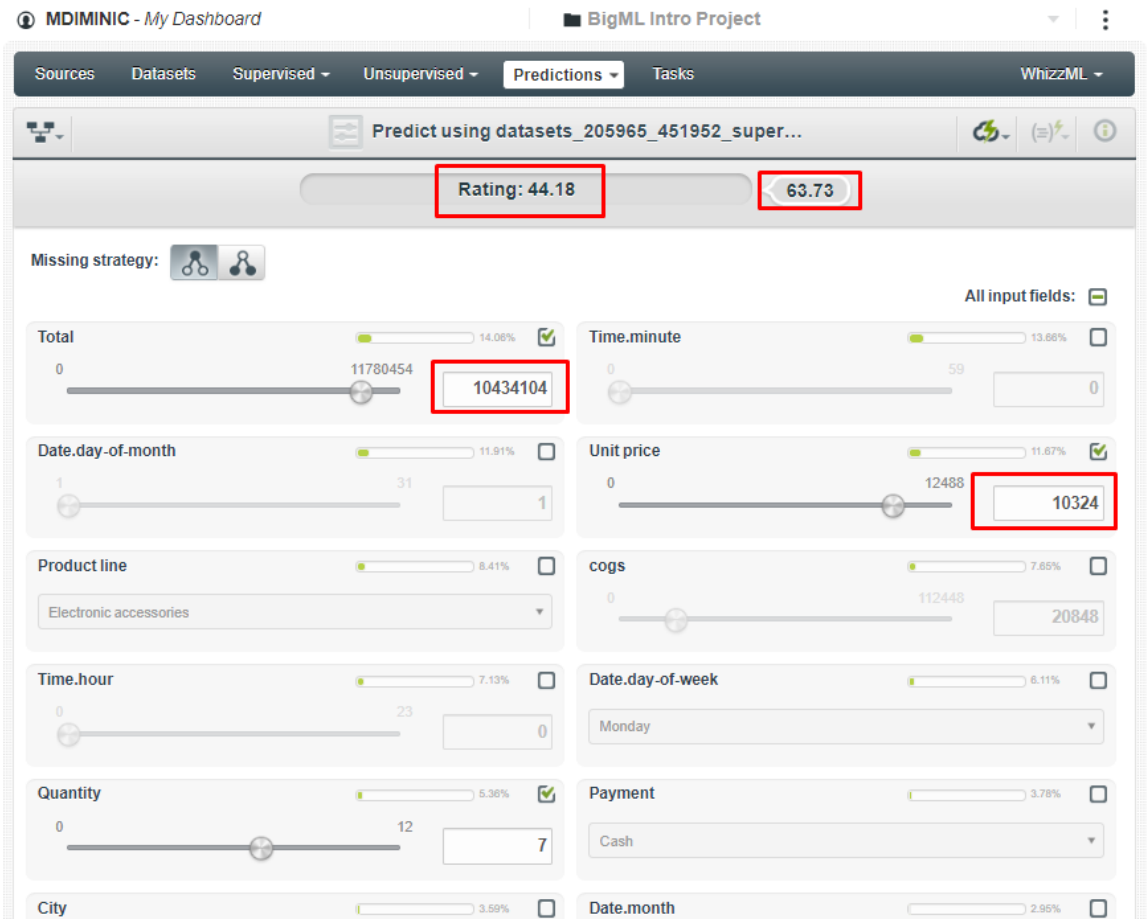
Ako usporedimo slijedeće dvije slike, slika 33 i slika 34, možemo uočiti da su različite. Postojeća distribucija i predviđena distribucija nisu jednake. Postojeća distribucija zadovoljstva korisnika uniformne je distribucije, dok je predviđena nepravilne unimodalne distribucije s pomakom udesno.

Slika 33. Prikaz distribucije podataka – zadovoljstvo korisnika



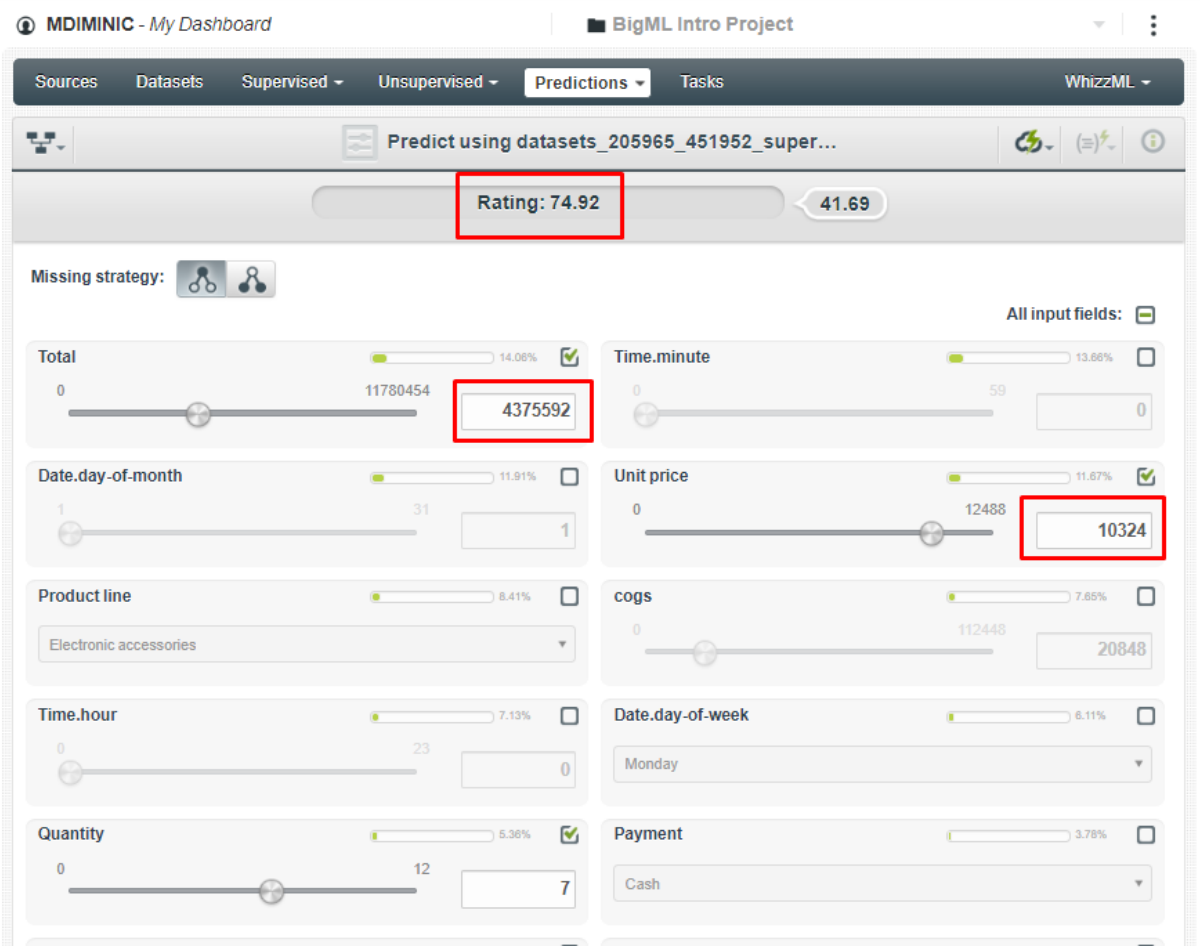
Slika 34. Prikaz predviđene distribucije podataka – zadovoljstvo korisnika

Slika 35 prikazuje predviđanje zadovoljstva korisnika obavljenom kupnjom na temelju stabla odlučivanja. U ovom slučaju, ukupni iznos računa s uračunatim PDV-om postavljen je na 10434104, a jedinična cijena proizvoda postavljena na 10324. S obzirom da su ukupna cijena i količina varijable od najvećeg utjecaja na ishod predviđanja, u sljedećih nekoliko primjera one će se mijenjati. Također od najvećeg utjecaja tu su i varijable vezane uz dan i minutu kupnje, ali nas to u ovom primjeru ne zanima. Dobivena predviđena vrijednost iznosi 44.18, odnosno ta brojka predstavlja ne bas zadovoljnog korisnika.



Slika 35. Prikaz zadovoljstva korisnika putem stabla odlučivanja 1

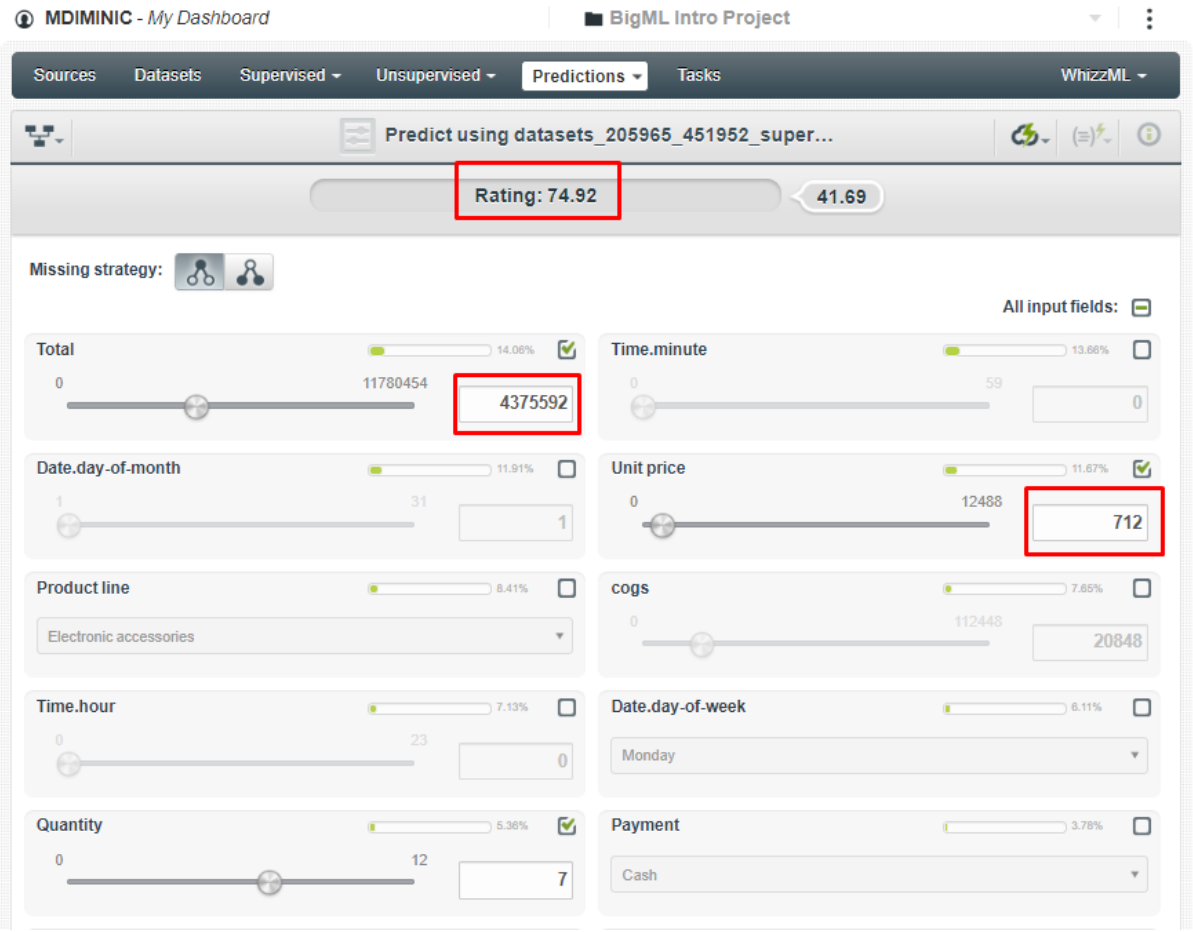
Na sljedećoj slici svi parametri su postavljeni jednako kao i u prethodnom slučaju, osim ukupne cijene računa bez PDV-a. U ovom slučaju ukupna cijena je umjesto na 10434104 postavljena na 4375592, tj. umanjena je za dva. Dobivena predviđena vrijednost zadovoljstva korisnika u ovom slučaju iznosi 74.92 (Slika 36).. Ocjena od 74.92 predstavlja vrlo zadovoljnog korisnika. Dobiveni rezultati u oba primjera imaju smisla jer prikazuju ponašanje korisnika, tj. zadovoljstvo korisnika ovisno o veličini računa, tj. visini ukupne cijene. Ako je cijena računa veća, znači da je kupac kupio više proizvoda iste jedinične cijene. Ako kupac kupuje više jednakih proizvodom vrlo je vjerojatno da je prema takvom proizvodu razvio preferenciju, tj. da je zadovoljan tim proizvodom.



Slika 36. Prikaz predviđanja zadovoljstva korisnika putem stabla odlučivanja 2

U zadnjem primjeru vezanom uz predviđanje zadovoljstva korisnika, prikazanom na slici 37, promatraju se jednaki parametri kao i na prošlom primjeru prikazanom na slici 36. Dakle, ukupna cijena računa s PDV-om iznosi 4375592, ali je jedinična cijena proizvoda postavljena na 712 umjesto 10324. Dakle, u ovom slučaju radi se o jeftinijem proizvodu. S navedenim parametrima, dobivena predviđena vrijednost zadovoljstva korisnika iznosi 74.92. identično kao i u prethodnom primjeru.

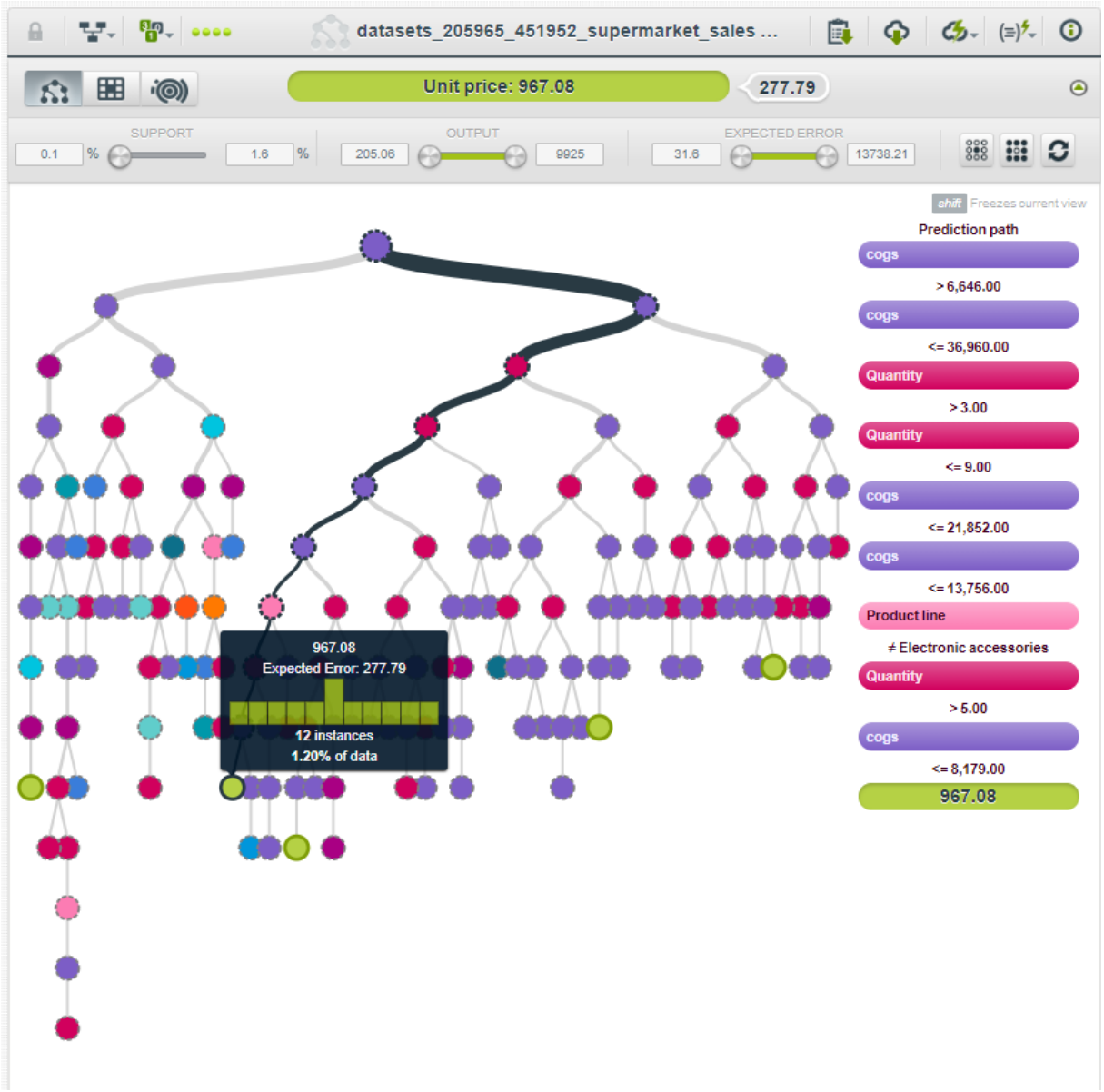
Možemo primijetiti da se zadovoljstvo korisnika nije povećalo s promjenom cijene što zapravo nema smisla. Isto tako možemo primijetiti da je i očekivana greška kod predviđanja velika te iznosi 41.69. Očekivana greška je poprilično velika zbog toga što se predviđa da će zadovoljstvo kupaca biti 74.92, a ukoliko je greška u prosijeku od te vrijednosti 41.69 mjernih jedinica +/- može se dogoditi da zadovoljstvo kupaca bude preko 100 ili ispod 40 što nam ne daje pouzdane rezultate za donošenje poslovne odluke. U pravilu, zadovoljstvo korisnika trebalo bi se povećati ukoliko je proizvod jeftiniji. Međutim, to ovisi o kakvom se proizvodu radi. Ukoliko se radi o proizvodu koji si mogu priuštiti samo osobe visokog statusa u društvu, zadovoljstvo korisnika može se povećati u tom slučaju.



Slika 37. Prikaz predviđanja zadovoljstva korisnika putem stabla odlučivanja 3

4.3.2. Predviđanje kretanja jedinične cijene proizvoda

Nakon predviđanja zadovoljstva korisnika, slijedi predviđanje kretanja jedinične cijene proizvoda. Kao Objective field postavljena je varijabla Unit price, odnosno jedinična cijena proizvoda. Također se radi o numeričkoj varijabli, pa očekivana greška u primjeru prikazanome na slici 38 iznosi 277,79.



Slika 38. Prikaz puta predviđanja putem stabla odlučivanja – jedinična cijena proizvoda

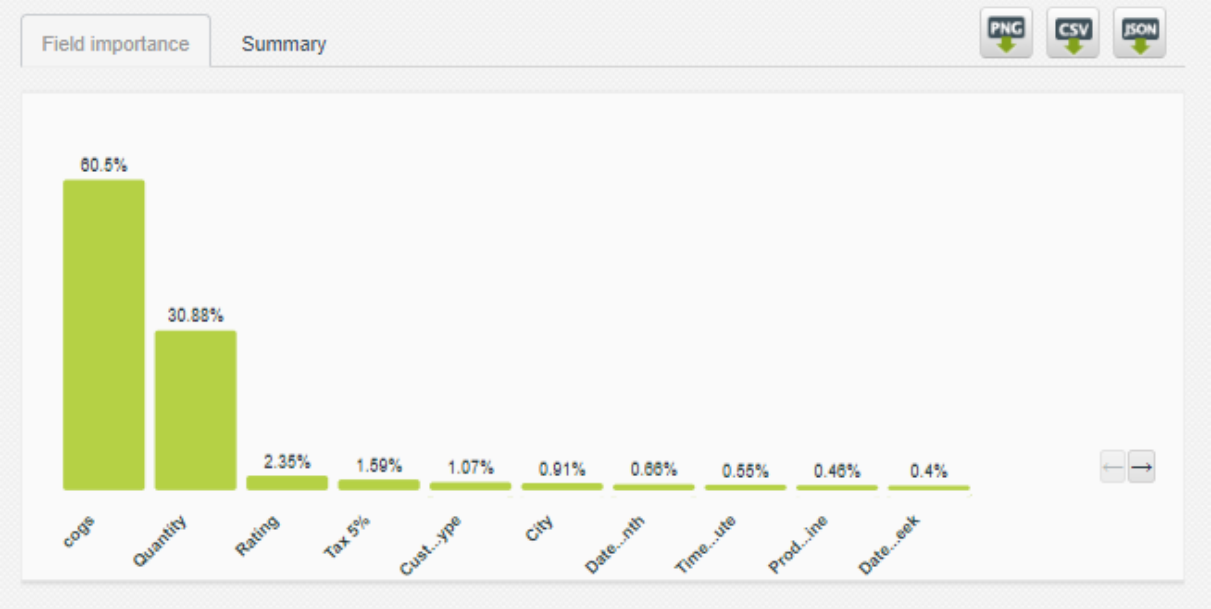
U navedenom primjeru, crnom bojom naglašen je put predviđanja za jediničnu cijenu proizvoda koji nije vezan uz elektroniku te se kupuje u količini od 6 do 9 komada.

Put predviđanja sastoji od sljedećih pravila. U korijenu se provjerava „cogs“, odnosno cijena računa bez uračunatog PDV-a i ona bi trebala iznositi iznad 6,646. Zatim se u sljedećem čvoru također provjerava je li cijena računa bez uračunatog PDV-a manja ili jednaka 36,960. U sljedeća dva čvora provjerava se količina, odnosno ona mora biti iznad 3 i manja ili jednaka 9. Nakon toga opet slijedi provjera cijene računa. Mora iznositi manje od ili točno 21,852 u prvom koraku, a u drugom manje od ili točno 13,756. Sljedeći korak provjerava radi li se o elektroničkom proizvodu, konkretnije nekom dodatku. Ako se ne radi o navedenom proizvodu

put predviđanja se nastavlja. Zatim se provjerava količina koja mora iznositi više od 5. U posljednjem koraku ponovno se provjerava cijena računa bez PDV-a te ona mora iznositi manje od ili točno 8,179.

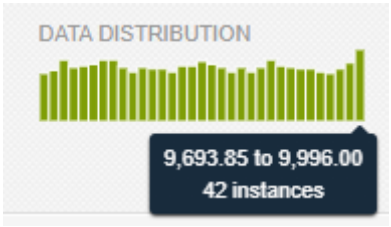
Kao rezultat odabranog puta predviđanja, dobili smo predviđanje jedinične cijene proizvoda u iznosu od 967,08.

Slika 39. prikazuje koje su varijable od najvećeg utjecaja za predviđanje jedinične cijene proizvoda. Od najveće važnosti su cijena računa bez PDV-a i to od čak 60,5% te količina od 31%. Ostale varijable imaju utjecaj manji od 2.5% te nisu od velikog značaja za predviđanje. U navedenom primjeru najčešće su korištene navedene dvije varijable.

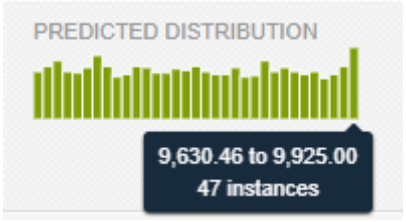


Slika 39. Prikaz utjecaja varijabli kod stabla odlučivanja – jedinična cijena proizvoda

Na sljedeće dvije slike možemo vidjeti distribucije postojeće i predviđene jedinične cijene proizvoda. Što se tiče usporedbe postojeće i predviđene distribucije vezane uz određivanje jedinične cijene, možemo vidjeti da su, za razliku od predviđanja zadovoljstva korisnika, identične. Međutim, ako pogledamo bolje, možemo uočiti da je u postojećoj distribuciji broj instanci najvećeg stupca manji za oko 70 te se kreće u malo drugačije rasponu. Kako se radi o neznatnim razlikama, možemo zaključiti da je predviđanje ispravno. Obje distribucije su uniformne.

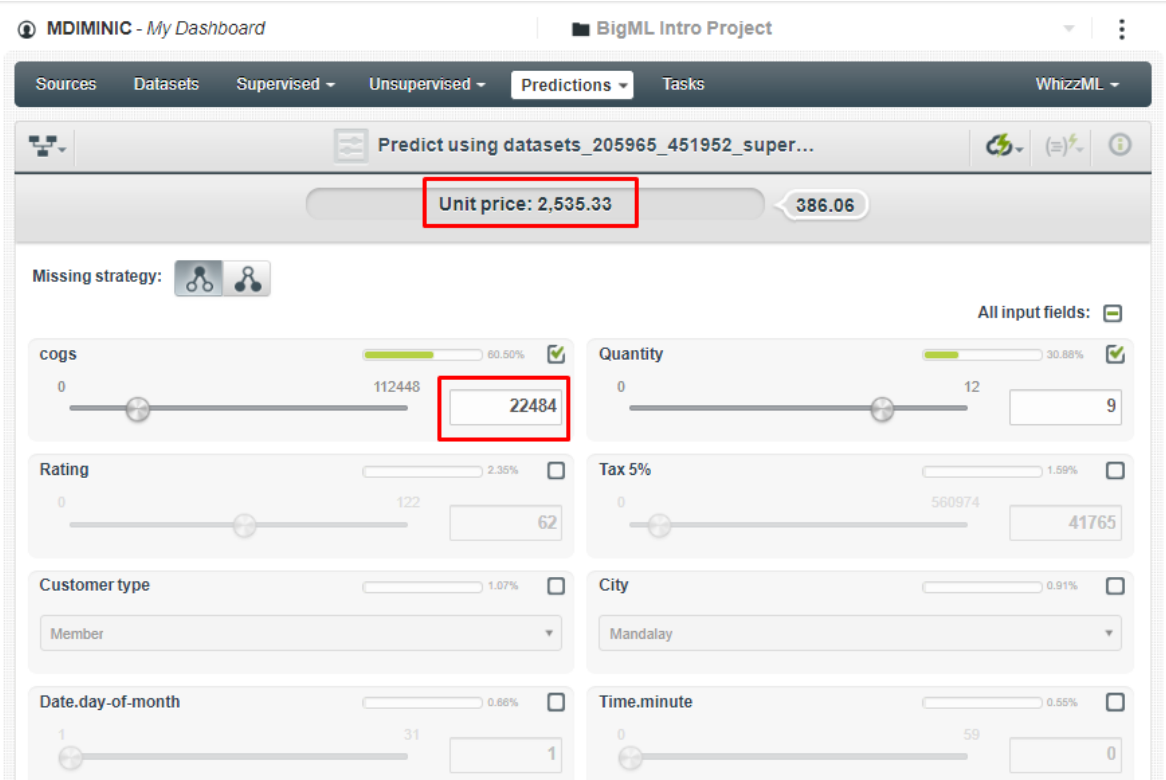


Slika 40. Prikaz distribucije podataka – jedinična cijena proizvoda



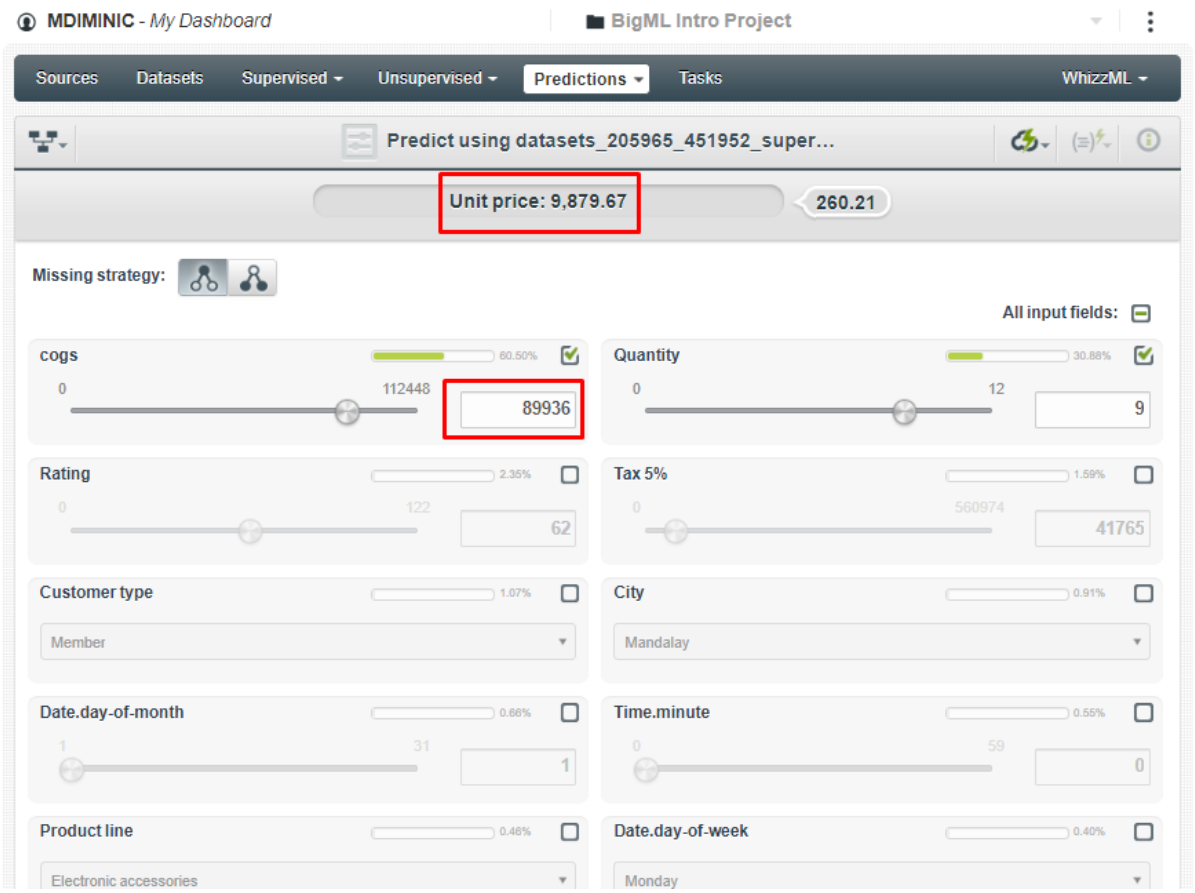
Slika 41. Prikaz predviđene distribucije podataka - jedinična cijena proizvoda

Na slici 42. možemo vidjeti predviđanje jedinične cijene proizvoda. Kako najveći utjecaj ima cijena računa bez PDV-a, u primjeru je postavljena na 22484. Također, velik utjecaj ima i količina koja je u primjeru postavljena na 9. Dobiveno predviđanje pokazalo je da je cijena proizvoda 2,535.



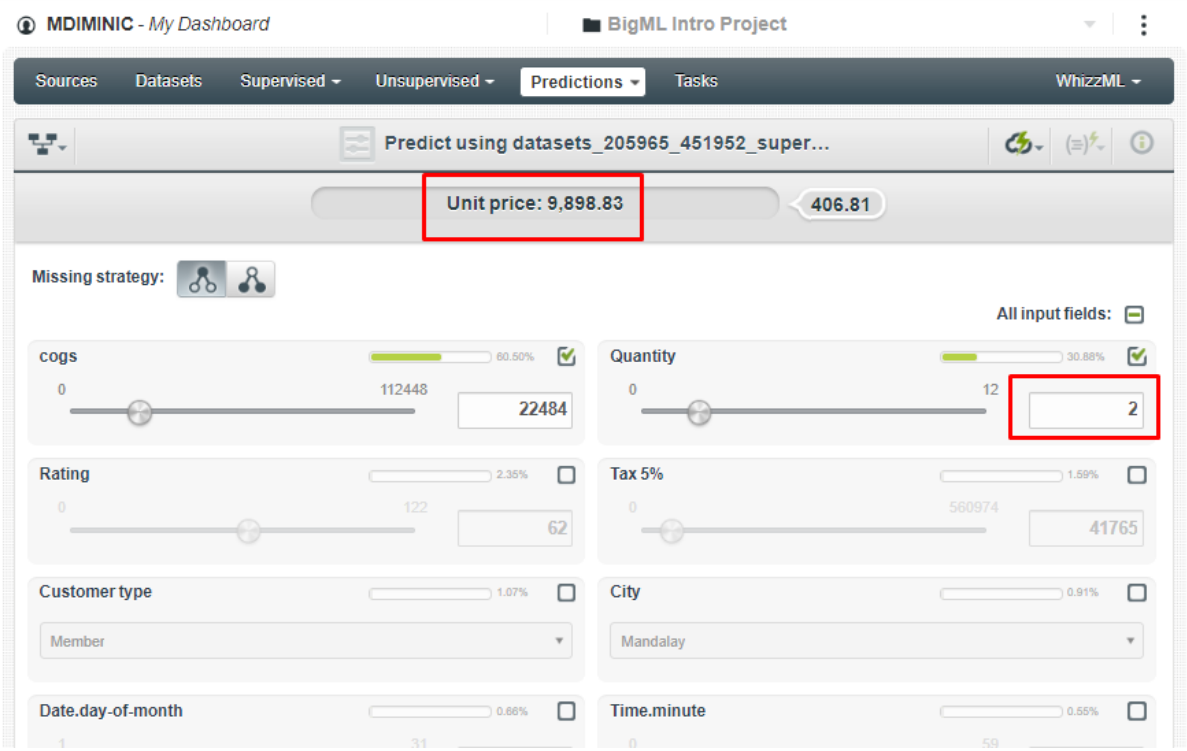
Slika 42. Prikaz predviđanja kretanja jedinične cijene proizvoda putem stabla odlučivanja 1

Idući primjer (Slika 43), također je postavljen kao i prethodni, osim parametra cijene računa bez PDV-a. Cijena računa je u ovom slučaju postavljena na 89936, dakle cijena računa je povećana četverostruko. U ovom primjeru jedinična cijena proizvoda se očekivano povećala i to na 9,879. Očekivano je povećanje jedinične cijene ako se radi o istoj količini proizvoda, a cijena računa se povećala. Isto tako u suprotnom smjeru, ukoliko je količina ostala ista a cijena računa se smanjila, znači da je i jedinična cijena proizvoda pala. Prema tome, predviđanja su ispravna.



Slika 43. Prikaz predviđanja kretanja jedinične cijene proizvoda putem stabla odlučivanja 2

U posljednjem primjeru prikazanom na slici 44 svi parametri postavljeni su jednako kao i na primjeru prikazanom na slici 42, osim količine. Količina je u ovom primjeru smanjena sa 9 na 2. Kao rezultat predviđanja dobivena je jedinična cijena od 9,898. U primjeru gdje je količina postavljena na 9, predviđena jedinična cijena bila je 2,535. Dakle, cijena se povećala poprilično. Prema tome, možemo zaključiti da s porastom kupljene količine, smanjuje se jedinična cijena proizvoda. Isto tako, kod kupnje manjih količina proizvoda, jedinična cijena proizvoda se povećava. Također se takvo ponašanje može usporediti sa popustom na količinu.



Slika 44. Prikaz predviđanja kretanja jedinične cijene proizvoda putem stabla odlučivanja 3

5. Interpretacija i evaluacija modela

Nakon provedbe analiza i procesa predviđanja pomoću stabla odlučivanja te neuronskih mreža možemo izvući potreban zaključak. Stablo odlučivanja te neuronske mreže smo koristili nad istim atributima kako bismo usporedili rezultate oba predviđanja.

Tablica 1. Interpretacija dobivenih predviđanja

	Neuronske mreže	***Stablo odlučivanja***
Zadovoljstvo korisnika	• Predviđanjem se pokazalo da je smanjenjem jedinične cijene proizvoda zadovoljstvo kupaca poraslo, a povećanjem jedinične cijene zadovoljstvo kupaca se smanjilo. • Atribut s najvećim prioritetom prilikom predviđanja je jedinična cijena proizvoda (21,02%) te količina (19,46%)	• Predviđanje zadovoljstva korisnika rađeno je pomicanjem vrijednosti ukupne cijene računa s PDV-om te jedinične cijene nekog proizvoda. Pokazano je da se povećanjem ukupne cijene računa, smanjuje i zadovoljstvo korisnika, a smanjenjem ukupne cijene računa, povećava zadovoljstvo korisnika. Pomicanjem jedinične cijene proizvoda nije došlo do promjene zadovoljstva korisnika što ne mora biti slučaj. • Atributi s najvećim prioritetom su ukupna cijena računa s PDV-om (Total), minuta kupnje proizvoda u satu (Time.minute), mjesec kupnje proizvoda (Date.month), jedinična cijena proizvoda (Unit Price), kategorija proizvoda (Product Line). Sve varijable imaju utjecaj od oko 10%.

Kretanje jedinične cijene proizvoda	• Predviđanjem se pokazalo da će jedinična cijena proizvoda rasti ukoliko se količina smanji. U suprotnom, jedinična cijena proizvoda će se smanjiti ako se poveća količina. Isto to vrijedi i za pomicanje ukupne cijene računa. Ako je ukupna cijena računa veća, jedinična cijena proizvoda je veća, a ako je ukupna cijena računa manja tada je i jedinična cijena proizvoda manja. • Atribut s najvećim prioritetom je ukupna cijena računa bez PDV-a (71.15%), a slijedi je količina (24.81%)	• Predviđanje jedinične cijene proizvoda rađeno je pomicanjem vrijednosti cijene računa bez PDV-a te količine. Povećanjem ukupne cijene računa, povećava se i jedinična cijene proizvoda. Suprotno tome, smanjenjem ukupne cijene računa smanjuje se jedinična cijena proizvoda. Ukoliko se količina poveća, smanjuje se jedinična cijena proizvoda, a ako se količina smanji, jedinična cijena proizvoda se povećava. • Atribut s najvećim prioritetom je cijena računa bez uračunatog PDV-a (60,5%) te ga slijedi količina (30,88%).

Nakon provedene analize i pregledanog skupa podataka mogli smo doći do zaključka da je odabrani model podataka bio dobar izbor. Obavljene analize i predviđanja nad odabranim skupom podataka, pokazale su rezultate koji se mogu poistovjetiti sa stvarnim svijetom. Primijenjene analize, stablo odlučivanja i neuronska mreža, također su pokazale ista predviđanja koja su vidljiva u prikazanoj tablici.

Prilikom provođenja analiza, očekivane greške bile su prisutne, ali ne u velikim postotcima. Na temelju veličine očekivanih grešaka možemo zaključiti da je odabrani skup podataka dobar te se na njemu mogu izraditi razne korisne i zanimljive analize. U ovom završnom radu fokus je bio na proučavanju kretanja jedinične cijene proizvoda te zadovoljstva korisnika.

6. Zaključak

Na kraju ovog završnog rada došao sam do zaključka da bi se podatkovna analitika trebala što više primjenjivati, ne samo u poduzetništvu nego i u drugim aspektima života zbog toga što se dosta stvari može predvidjeti. U današnjem svijetu možemo vidjeti da se dosta poduzeća koristi podatkovnom analitikom kako bi donijeli što bolju odluku vezanu uz korekciju cijena i plasiranje proizvoda na tržište.

Praktičnim dijelom ovog rada smo došli do zaključka da se mogu važne informacije saznati upotrebom podatkovne analitike. Prikazali smo kako je moguće vidjeti da se cijena jedinice proizvoda mijenja ukoliko se mijenja i količina proizvoda. Česta praksa supermarketa je da se cijena proizvoda smanjuje što se veća količina kupuje pa smo to isto tako dokazali i u našoj analizi. Druga stavka predviđanja nam se odnosila primarno na zadovoljstvo kupaca. Primarno smo se bazirali na promatranje zadovoljstva kupaca danom uslugom te ukoliko smo mijenjali cijenu proizvoda koji se prodaje došli smo do zaključka ako je cijena proizvoda prevelika zadovoljstvo kupaca pada, a ukoliko je manja tada njihovo zadovoljstvo raste. To nam govori kako bi supermarketi trebali držati cijene što niže kako bi im kupci bili zadovoljniji, a samim time bi privukli i više novih kupaca. U tu svrhu supermarketi u današnje vrijeme prakticiraju funkciju sniženja i rasprodaja.

Smatram da sam odabrao veoma dobru i zanimljivu temu za završni rad te svakako nadogradio svoje znanje koje sam već imao stečeno kroz fakultetsko obrazovanje. Nakon kreiranja ovog završnog rada probudio sam u sebi znatiželju da svakako kroz svoje slobodno vrijeme pregledam još poneko predviđanje za druge industrije te svakako smatram da ću podatkovnu analitiku koristiti ukoliko otvorim svoje poduzeće kako bih donosio ispravne odluke.

Popis literature

[1.] Procjena kreditnog rizika primjenom rudarenja podataka, Petra Poljak; posjećeno 01.09.2020. na linku: https://repozitorij.unizg.hr/islandora/object/foi:4453/preview

[2.] Analiza otvorenih podataka, Karlo Krznarić , posjećeno 02.09.2020. na linku: https://zir.nsk.hr/en/islandora/object/foi%3A5674

[3.] Supermarket sales, Aung Pyae, posjećeno 04.09.2020. na linku: https://www.kaggle.com/aungpyaeap/supermarket-sales

[4.] Rudarenje podataka u poduzetništvu, Mislav Knez, posjećeno 03.09.2020. na linku: https://repozitorij.foi.unizg.hr/islandora/object/foi%3A5581/datastream/PDF/view

[5.] Značaj industrije video igara s osvrtom na stanje u Republici Hrvatskoj, Mislav Karamatić, posjećeno 03.09.2020. na linku: https://repozitorij.efzg.unizg.hr/islandora/object/efzg:2843

[6.] A Complete Guide to Scatter Plots, Mike Yi; posjećeno 05.09.2020. na linku: https://chartio.com/learn/charts/what-is-a-scatter-plot/

[7.] Redukcija podataka, Nenad Mitić; posjećeno 05.09.2020. na linku: http://poincare.matf.bg.ac.rs/~nenad/ip2/redukcija_podataka.pdf

[8.] Spearman's Rank-Order Correlation, Laerd statistics; posjećeno 06.09.2020. na linku: https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php

[9.] Pearson’s Correlation Coefficient, Statistics Solutions; posjećeno 06.09.2020. na linku: https://www.statisticssolutions.com/pearsons-correlation-coefficient/

[10.] Cluster Analysis, Statistics Solutions; posjećeno 07.09.2020. na linku: https://www.statisticssolutions.com/directory-of-statistical-analyses-cluster-analysis/

[11.] Neuronske mreže, Fakultet organizacije i informatike ERIS; posjećeno 07.09.2020. na linku: https://eris.foi.hr/11neuronske/nn-predavanje2.html

[12.] Stabla odlučivanja, Marijana Zekić-Sušac EFOS; posjećeno 08.09.2020. na linku: http://www.efos.unios.hr/sustavi-poslovne-inteligencije/wp-content/uploads/sites/192/2017/10/P4_Stabla-odlucivanja-2017.pdf

Popis slika

Popis tablica