Univerzitet u Beogradu

Matematički fakultet

Miloš Ristić

Kursna lista za euro/dinar

Seminarski rad

ime i prezime	Miloš Ristić
broj indeksa	1018/2011
predmet	Istraživanje podataka
školska godina	2011/2012
nastavnik	dr. Nenad Mitić
datum	21.05.2012

1. Zadatak

Na adresi http://www.nbs.rs/kursnaListaModul/inputPeriod.faces?lang=cir nalazi se kurna lista za evro od 2002. godine. Na osnovu podataka odrediti koji je mesec najbolji za kupovinu/prodaju evra? Da li postoje i koji su meseci u kojima opada ili raste kurs? Da li se na početku i kraju meseca kurs menja na određen način.

2. Priprema podataka

Podaci su preuzeti sa date adrese i nalaze se u tabeli KURSNA_LISTA. Kreiranje i popunjavanje tabele dati su u skript fajlu create_and_charge.txt. Potrebno je uraditi data exploration radi pripreme podataka za fazu obrade.

3. Čišćenje podataka

Nakon kreiranja baze podataka i potrebne tabele neophodno je očistiti podatke od dupliranih i nedostajućih vrednosti. Pošto je urađen data exploration uočeno je sledeće:

1. Vrednosti atributa DATUM_FORMIRANJA i DATUM_PRIMENE su uvek jednake.

```
SELECT COUNT(*)
FROM KURSNA_LISTA
WHERE DATUM_FORMIRANJA != DATUM_PRIMENE

REZ=0
```

2. Vrednost atributa VAZI_ZA je uvek jednaka 1.

```
SELECT DISTINCT VAZI_ZA
FROM KURSNA_LISTA


REZ=1
```

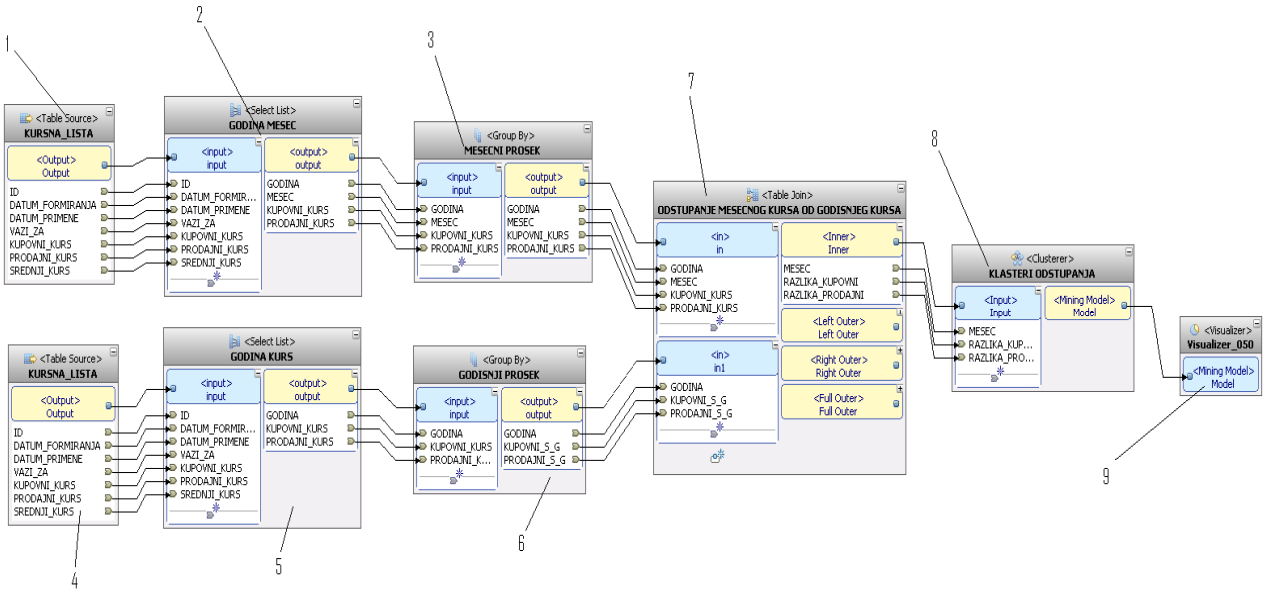
3. Nema nedostajućih vrednosti.

Zbog ovih činjenica u daljem radu nije potrebno razmatrati vrednost atributa VAZI_ZA jer on ne nosi nikakvu informaciju. Treba eliminisati jedan od atributa DATUM_FORMIRANJA ili DATUM_PRIMENE, jer oni nose istu informaciju. Nema potrebe za rukovanjem nedostajućim vrednostima.

4. Analiza podataka

4.1. Koji je mesec najbolji za kupovinu/prodaju evra?

Kao ideja za rešavanje ovog zadatka korišćeni su godišnji prosek kupovnog i prodajnog kursa eura i mesečni prosek kupovnog i prodajnog kursa eura za svaku godinu.



1. Table source – tabela KURSNA_LISTA
2. Select list - izdvaja godinu i mesec iz atributa DATUM_PRIMENE kao zasebne atribute i selektuje KUPOVNI_KURS i PRODAJNI_KURS iz tabele KURSNA_LISTA.
3. Group by – izračunava prosečan kupovni i prosečan prodajni kurs za svaki mesec u svakoj godini. Rezultat je grupisan po atributima GODINA I MESEC.
4. Table source – tabela KURSNA_LISTA
5. Select list - izdvaja godinu iz atributa DATUM_PRIMENE kao zaseban atribut i selektuje KUPOVNI_KURS i PRODAJNI_KURS iz tabele KURSNA_LISTA.
6. Group by – izračunava prosečan kupovni i prosečan prodajni kurs za svaku godinu. Rezultat je grupisan po atributu GODINA.
7. Table join – spajanje prethodne dve tabele po jednakosti atributa GODINA. Rezultat je tabela sa atributim MESEC, RAZLIKA_KUPOVNI (razlika godišnjeg i mesečnog proseka kupovnog kursa za datu godinu pomnožena sa 100 radi lakšeg formiranja interval razlike), RAZLIKA_PRODAJNI (razlika godišnjeg i mesečnog proseka prodajnog kursa za datu godinu pomnožena sa 100 radi lakšeg formiranja interval razlike).

8. Clusterer – klasterovanje datih podataka. Klasterovanje se vrši na osnovu atributa RAZLIKA_PRODAJNI i RAZLIKA_KUPOVNI Kohonenovim algoritmom u 10 iteracija. Broj klastera je postavljen na 8 (vrednosti atributa RAZLIKA_PRODAJNI i RAZLIKA_KUPOVNI biće predstavljene kao high/medium/low, pa je to 2^3 kombinacija). Prilikom klasterovanja atribut MESEC nije korišćen jer klastere ne treba formirati prema njegovim vrednostima.
9. Visualizer – vizualizacija podataka.

Klasterovanje i parametri: Prilikom klasterovanja korišćen je Kohonenov algoritam jer je prilagođen neprekidnim vrednostima atributa. Klasterovanje je izvršeno sa 10, 20 ,30 i 40 prolaza sa maksimalnim brojem klastera 8. U sva četiri slučaja dobijeni su dosta slični rezultati, obzirom da je potrebno pronaći samo one klastere kod kojih je odstupanje prodajnog i kupovnog kursa najvece u odnosu na godišnji prosek. Sa promenom broja klastera na 3 i testom sa 10, 20, 30 i 40 iteracija takođe se dobijaju slični rezultati. Meseci sa najvećim odstupanjem uvek se javljaju u jednom klasteru. Kao zvanično rešenje može se uzeti bilo koje od rezultata klasterovanja.

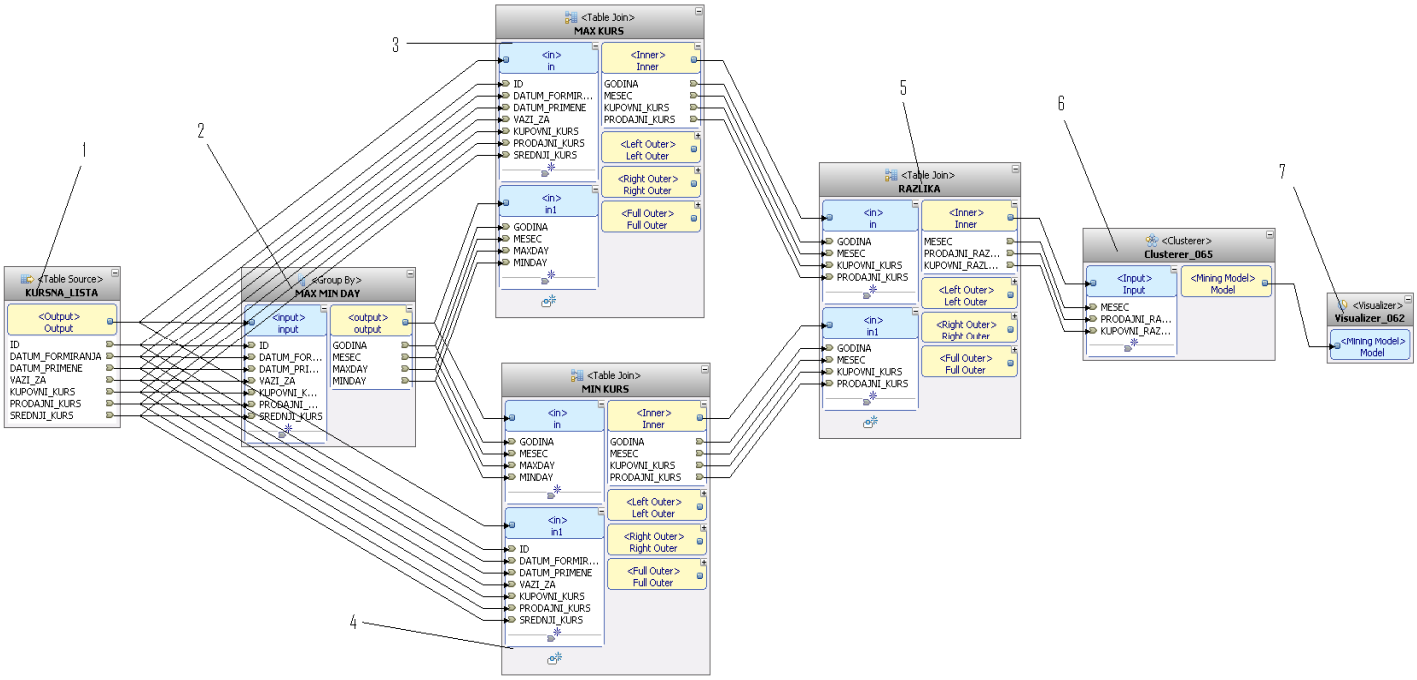
Zaključak: Potrebno je pronaći klaster koji je napravljen na osnovu vrednosti RAZLIKA_PRODAJNI = high i RAZLIKA_KUPOVNI = high, jer je ovde prosečan mesečni kupovni/prodajni kurs bio znatno niži od prosečnog godišnjeg kupovnog/prodajnog kursa. Postoje dva takva klastera. Meseci koji dominiraju su januar, februar i mart (videti mesec_kupovina_prodaja.vis). Kako je razlika u odstupanju mala, ova dva klastera mogu se posmatrati kao jedan, pa bi najdominantniji mesec bio januar.

Komentar:

Kupovni i prodajni kurs se slično ponašaju. Kada jedan raste, raste i drugi i obratno. Bilo je dovoljno posmatrati proseke samo jednog od njih.

4.2. Da li postoje i koji su meseci u kojima opada ili raste kurs?

Za rešavanje ovog zadatka korišćena je razlika kupovnog i prodajnog kursa eura sa početka i sa kraja svakog meseca svake godine.



1. Table source – tabela KURSNA_LISTA
2. Group by - izdvaja godinu, mesec, maksimalan i minimalan dan iz atributa DATUM_PRIMENE. Grupisanje se vrši po atributima GODINA i MESEC.
3. Table join – spaja tabelu KURSNA_LISTA i tabelu dobijenu prethodnim group by operatorom po atributu DATUM_PRIMENE. Cilj je dobijanje vrednosti kupovnog i prodajnog kursa sa kraja svakog meseca svake godine.
4. Table join – spaja tabelu KURSNA_LISTA i tabelu dobijenu prethodnim group by operatorom po atributu DATUM_PRIMENE.Cilj je dobijanje vrednosti kupovnog i prodajnog kursa sa početka svakog meseca svake godine.
5. Table join – spaja tabele iz prethodna dva join operatora po atributima MESEC i GODINA. U rezultatu se nalaze novi atributi RAZLIKA_KUPOVNI i RAZLIKA_PRODAJNI koji predstavljaju razliku kupovnog/prodajnog kursa sa kraja i početka svakog meseca svake godine.

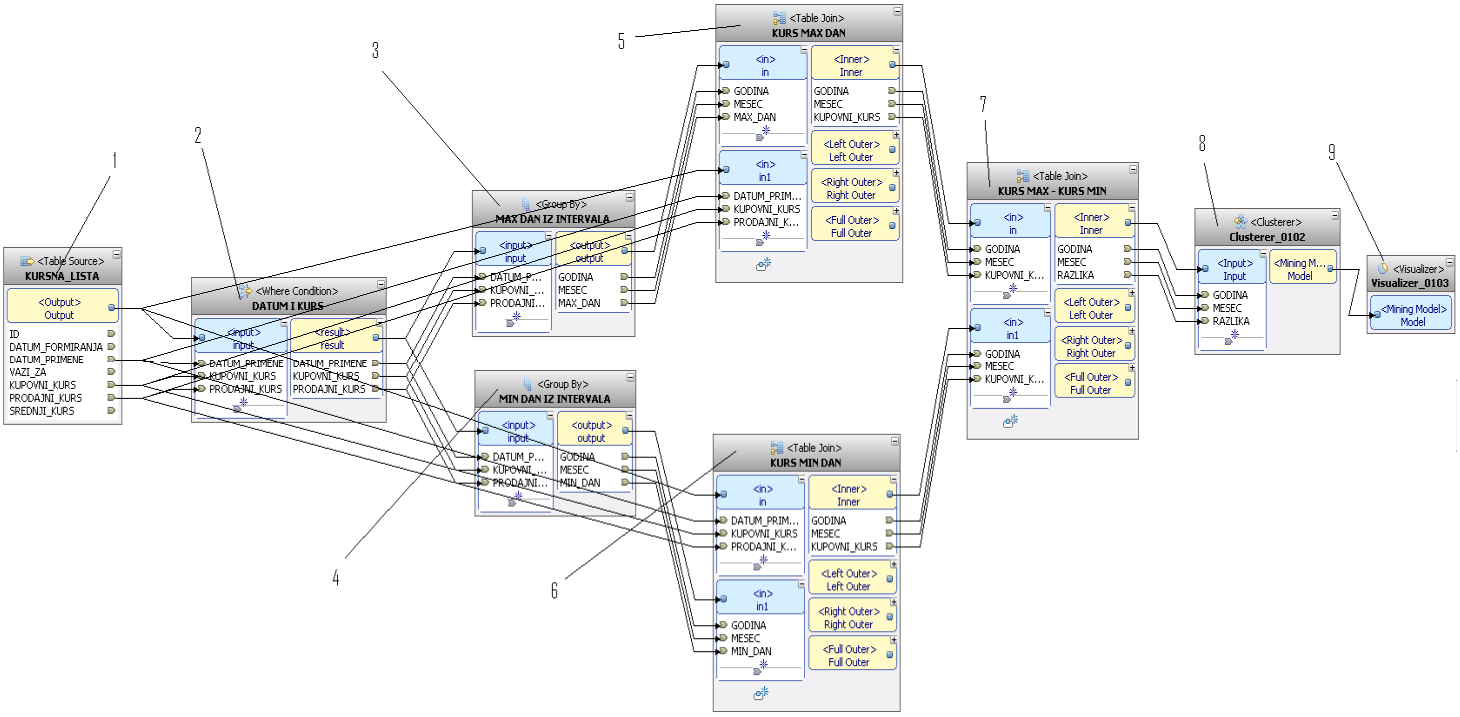
6. Clusterer – klasterovanje datih podataka. Klasterovanje se vrši na osnovu atributa RAZLIKA_PRODAJNI Kohonenovim algoritmom u 10 iteracija. Broj klastera je postavljen na 5 (vrednosti atributa RAZLIKA_PRODAJNI biće posmatrane kao very-high/high/medium/low/very-low). Prilikom klasterovanja atribut MESEC nije korišćen jer klastere ne treba formirati prema njegovim vrednostima, a atribut RAZLIKA_KUPOVNI nije korišćen iz razloga visokog stepena korelacije sa atributom RAZLIKA_PRODAJN (verovatnoća da se isto ponašaju iznosi 0.999999999016728).
7. Visualizer – vizualizacija podataka.

Klasterovanje i parametri: Prilikom klasterovanja korišćen je Kohonenov algoritam jer je prilagođen neprekidnim vrednostima atributa. Klasterovanje je izvršeno sa 10, 20 ,30 i 40 prolaza sa maksimalnim brojem klastera 5. U sva četiri slučaja dobijeni su dosta slični rezultati, obzirom da je potrebno pronaći samo one klastere kod kojih je odstupanje kursa sa početka I sa kraja meseca najveće/najmanje. Promene su uočljive u veličini klastera, ali one su posledica prelaska elemenata sa srednjom vrednošću razlike iz jednog klastera u drugi. Sa promenom broja klastera na 3 i testom sa 10, 20 ,30 i 40 rezultati su isti. Sa brojem preko 20 iteracija uvek dobijamo iste klastere. Meseci sa najvećim I najmanjim odstupanjem uvek se javljaju u posebnom klasteru. Kao zvanično rešenje može se uzeti bilo koje od rezultata klasterovanja.

Zaključak: Potrebno je pronaći klastere koji imaju vrednost very-high,high,low,very-low na atributu RAZLIKA_PRODAJNI. Pošto je od kursa sa kraja meseca oduzet kurs sa početka meseca vrednosti very-high i high označavaju rast kursa, a vrednosti low i very-low opadanje kursa. Na osnovu rezultata (videti mesec_odnos_pocetak_kraj.vis) vidi se da kurs za sve mesece u većini slučajeva raste i da je razlika sa početka i kraja meseca u interval [0, 2] dinara. Slično važi za pad kursa sa intervalom [-2, 0]. Drastični padovi kursa dešavali su se u junu, a skokovi u novembru i januaru. Ne možese uočiti jaka zakonitost, ali bi se moglo reći da u januaru, februaru i novembru kurs ne opada, dok je najveća mogućnost za opadanje u junu.

4.3. Da li se na početku i kraju meseca kurs menja na odredjen način?

Za rešavanje ovog zadatka korišćena je razlika kursa sa kraja i početka prve/poslednje sedmice svakog meseca svake godine.



1. Table source – tabela KURSNA_LISTA
2. Where condition - izdvaja prvu nedelju svakog meseca svake godine. Prosleđuje atribut DATUM_PRIMENE kao i kupovni i prodajni kurs za te datume.
3. Group by – iz prethodno dobijene tabele izdvaja poslednji dan prve sedmice za svaki mesec u svakoj godini, takav da je tada došlo do promene kursa. Dobijamo atribut MAX_DAY. Grupiše se po atributima GODINA i MESEC.
4. Group by – iz prethodno dobijene tabele izdvaja prvi dan prve sedmice za svaki mesec u svakoj godini, takav da je tada došlo do promene kursa. Dobijamo atribut MIN_DAY. Grupiše se po atributima GODINA i MESEC.
5. Table join – spaja tabelu koja sadrži MAX_DAY atribut sa tabelom KURSNA_LISTA. Dobijamo vrednost kupovnog kursa poslednjeg dana prve sedmice svakog meseca svake godine. Spajanje se vrši preko atributa DATUM_PRIMENE. (Atributi PRODAJNI_KURS i KUPOVNI_KURS su jako korelisani pa se može koristiti bilo koji od njih.)

6. Table join – spaja tabelu koja sadrži MIN_DAY atribut sa tabelom KURSNA_LISTA. Dobijamo vrednost kupovnog kursa prvog dana prve sedmice svakog meseca svake godine. Spajanje se vrši preko atributa DATUM_PRIMENE. (Atributi PRODAJNI_KURS i KUPOVNI_KURS su jako korelisani pa se može koristiti bilo koji od njih.)
7. Table join – spaja prethodne dve tabele po atributim GODINA i MESEC uz dodavanje novog atributa RAZLIKA koji predstavlja razliku kupovnog kursa poslednjeg i prvog dana prve sedmice svakog meseca svake godine pomnoženu sa 100 radi lepše intervalne podele.
8. Clusterer – klasterovanje datih podataka. Klasterovanje se vrši na osnovu atributa RAZLIKA Kohonenovim algoritmom u 10 iteracija. Broj klastera nije bitan. Prilikom klasterovanja atributi GODINA i MESEC nisu korišćen jer klastere ne treba formirati prema njihovim vrednostima.
9. Visualizer – vizualizacija podataka.

Klasterovanje i parametri: Prilikom klasterovanja korišćen je Kohonenov algoritam jer je prilagođen neprekidnim vrednostima atributa. Klasterovanje je izvršeno sa 10, 20 ,30 i 40 prolaza sa neograničenim brojem klastera. U sva četiri slučaja dobijeno je 9 klastera sa relativno sličnom raspodelom vrednosti. Obzirom da je potrebno analizirati vrednost atributa RAZLIKA u smislu da li je veća ili manja od nule raspodela po klasterima ne igra veliku ulogu. Štaviše, sa povećanjem broja iteracija još je uočljivije da se ne može uočiti jaka pravilnost o padu i rastu kursa na početku I na kraju meseca.

Zaključak: Posmatranjem atributa RAZLIKA vidi se da je njegova vrednost u većini slučajeva u intervalu [0, 50] ili u nekom iznad njega (videti pocetak_meseca.vis). Jedna trećina instanci ima vrednost manju od 0 na ovom atribut.. Ovo znači da je kurs na kraju prve nedelje veći nego na početku prve nedelje svakog meseca svake godine u dve trećine slučajeva. Može se zaključiti da je veća šansa da početkom meseca kurs raste.

Šema je simetrična i za kraj meseca. Promena: Where condition - izdvaja poslednju nedelju svakog meseca svake godine. Prosleđuje atribut DATUM_PRIMENE kao i kupovni i prodajni kurs za te datume.

Zaključak: Posmatranjem atributa RAZLIKA vidi se da je njegova vrednost u većini slučajeva u intervalu [0, 50] ili u nekom iznad njega (videti kraj_meseca.vis). Preko jedne trećina instanci ima vrednost manju od 0 na ovom atribut.. Ovo znači da je kurs na kraju poslednje nedelje veći nego na početku poslednje nedelje svakog meseca svake godine u nešto manje od dve trećine slučajeva. Može se zaključiti da je veća šansa da krajem meseca kurs raste.

Zaključak: Poređenjem ovih rezultata vidi se da je verovatnoća rasta kursa veća na početku nego na kraju meseca. Stroga zavisnost o padu i rastu kursa ne postoji. Jedini zaključak je da kurs početkom i krajem meseca ima veću verovatnoću porasta nego opadanja.