Hanna Kaimo – Saku Ihalainen – Elyas Addawe

LUOTTORISKIT

Tehtävä n:o 5

Metropolia Ammattikorkeakoulu
Tieto- ja viestintätekniikka
Datan käsittely ja koneoppiminen
28.3.2022

1 Tavoitteet

Analysoitava luottoriski- aineisto sisältää 20 selittävää muuttujaa ja viimeinen muuttuja ilmentää, onko henkilön luottoriski matala vai korkea. Selvitämme tämän koko aineiston perusteella, miten hyvin päätöspuu soveltuu henkilön luottoriskin ennustamiseen. Selvityksen perustamme ositus- validointiin, jossa osa aineistosta on testi- joukossa. Tutkimme myös kokeellisesti, miten päätöspuun koko ja muut parametrit vaikuttavat päätöspuun luokittelu- kykyyn. Painotamme osassa kokeista virhe- tilannetta, jossa henkilö saa luoton, vaikka riski rahojen takaisin saamiseen on korkea.

2 Data

Aineiston lähde: Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: Univ. of California, School of Information and Computer Science

Attribuutit:

Status Of Checking Account: nominaali- muuttuja, vaihteluväli datassa A11- A13
Duration: numeerinen muuttuja, vaihteluväli datassa 4- 72
Credit History: nominaali- muuttuja, vaihteluväli datassa A30- A34
Purpose: nominaali- muuttuja, vaihteluväli datassa (A40- A49) + A410
Amount: numeerinen muuttuja, vaihteluväli datassa 250- 18 484
Bonds: nominaali- muuttuja, vaihteluväli datassa A61- A65
Employment: nominaali- muuttuja, vaihteluväli datassa A71- A75
Installment Rate: numeerinen muuttuja, vaihteluväli datassa 1- 4
Personal Status: nominaali- muuttuja, vaihteluväli datassa A91- A93
Quarantors: nominaali- muuttuja, attribuutin arvo A101, A102 tai A103
Present Residence: numeerinen muuttuja, vaihteluväli 1- 4
Property: nomi naali- muuttuja, vaihteluväli A101- A104
Age: numeerinen muuttuja, vaihteluväli 19- 75
Plans: nominaali- muuttuja, attribuutin arvo A141, A142 ta A143
Housing Type: nominaali- muuttuja, attribuutin arvo A151, A152 tai A153

Credits: numeerinen muuttuja, vaihteluväli 1- 4
Job: nominaali- muuttuja, vaihteluväli A171- A174
Liability Of People: numeerinen muuttuja, attribuutin arvo joko 1 tai 2
Telephone: nominaali- muuttuja, attribuutin arvo joko A191= ei ole tai A192= on
Foreign Worker Or Not: nominaali- muuttuja, attribuutin arvo joko A201= on tai A202= ei ole
Luottoluokitus: numeerinen muuttuja, attribuutin arvo on joko 1= alhainen riski luoton antamiseen tai 2= korkea riski luoton antamiseen

Datan koko on 1000 attribuuttia. Datassa ei ole puuttuvia tai poikkeavia arvoja.

3 Datan valmistelu

Tutustuimme german.doc- tiedostoon. Lisäsimme pilkut aineiston attribuuttien välille välilyöntien sijaan Notepad++ etsi- korvaa- toiminnolla, jolloin attribuutit erottuvat toisistaan. Lisäsimme otsikot attribuuteille, koska Wekassa attribuutit on listattu otsikon mukaan. Muutimme german.data- tiedostomuodon ensin german.csv-tiedostomuodoksi ja sen konverterilla german.arff- tiedostomuodoksi, joka on hyvä tiedostomuoto Wekaakin varten. Tarkistimme Wekassa, että datassa ei ole poikkeavia arvoja. Kävimme datan läpi vielä silmämääräisestikin, koska se oli mahdollista datan koon ollessa suhteellisen pieni. Vaihdoimme luottoluokitus- attribuutin muodon numeerisesta nominaaliksi ja numero- arvoja sisältävien attribuuttien muodon numeeriseksi.

4 Mallinnus

Käytimme datan analysoinnissa kustannuspohjaista algoritmia. Wekassa tämä algoritmi löytyy nimellä Cost Sensitive Classifier. Opetusjoukon tarkkuus on 100 %. Siinä ei ole virheitä.

```
Correctly Classified Instances        1000               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Total Cost                               0
Average Cost                             0
Mean absolute error                      0.001
Root mean squared error                  0.001
Relative absolute error                  0.2375 %
Root relative squared error              0.2178 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     1
                 1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     2
Weighted Avg.    1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000

=== Confusion Matrix ===

   a   b   <-- classified as
 700   0 |   a = 1
   0 300 |   b = 2
```

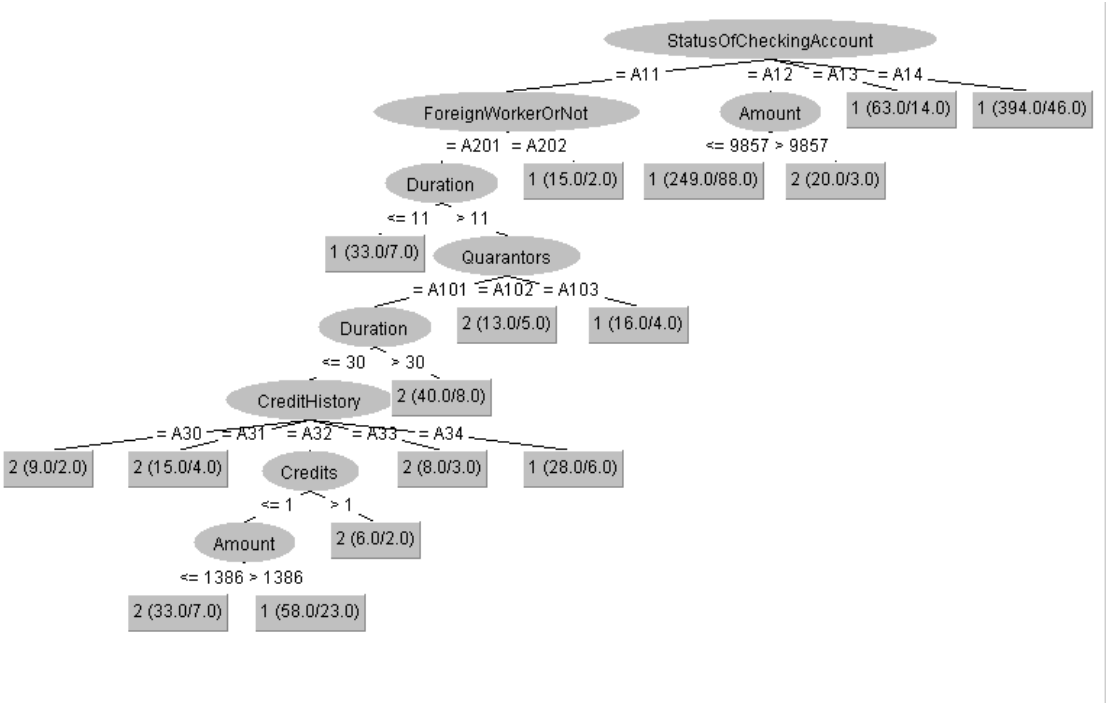
Muutimme Test option arvoksi Split Validation, Percentage Split (66 %, ⅔ opetusjoukoksi ja ⅓ testijoukoksi) ja parametreiksi laitoimme siis J48- päätöspuun ja kustannusmatriisin. Siemen- arvo on 21, attribuuttien määrä datassa. Tämän algoritmin idea on se, että toisin kuin esim. KNN- algoritmissa, tässä algoritmissa on mahdollista painottaa jonkin attribuutin painoarvoa. Ja siksi käytämme sitä. Pyrimme minimoimaan virheistä syntyneet kustannukset. Nyt katsomme, mitä tuloksia saamme ilman painotuksia eri kokoisilla päätöspuilla ja katsomme, mitä tapahtuu, kun painotamme sitä, että henkilölle annetaan luotto, vaikka riski luoton antamiseen henkilölle on korkea. Kustannus on tällöin viisinkertainen. Kustannusmatriisi on silloin tämän näköinen:

0 1
5 0

Kun se ilman painotusta on;
0 1
1 0

Päätöspuun kokoa saa siis muutettua muuttamalla Confidence Factory- parametria. Perustamme tuloksemme testi- joukolle.

Huomaamme heti, että Wekan oletusarvo 0.25 Confidence Factory- parametrille tekee aivan liian ison päätöspuun. Vaihdamme ensimmäiseksi arvoksi 0.1, emme painota aluksi kumpaakaan arvoa ja saamme tämän näköisen päätöspuun:



Tämän testijoukon tarkkuus on 72.65 %. Päätöspuun kokoa vaihtelemalla huomaamme, että päätöspuun koko ei kauhean paljon vaikuta testijoukon tarkkuuteen. Esimerkiksi Confidence Factory arvo 0.02 antaa 73.5 % tarkkuuden ja Confidence Factoryn arvo 0.5: 71.2 %. Puun tasapainoisempi kyllä monilla muilla arvoilla kuin kuvassa näkyvä puu.

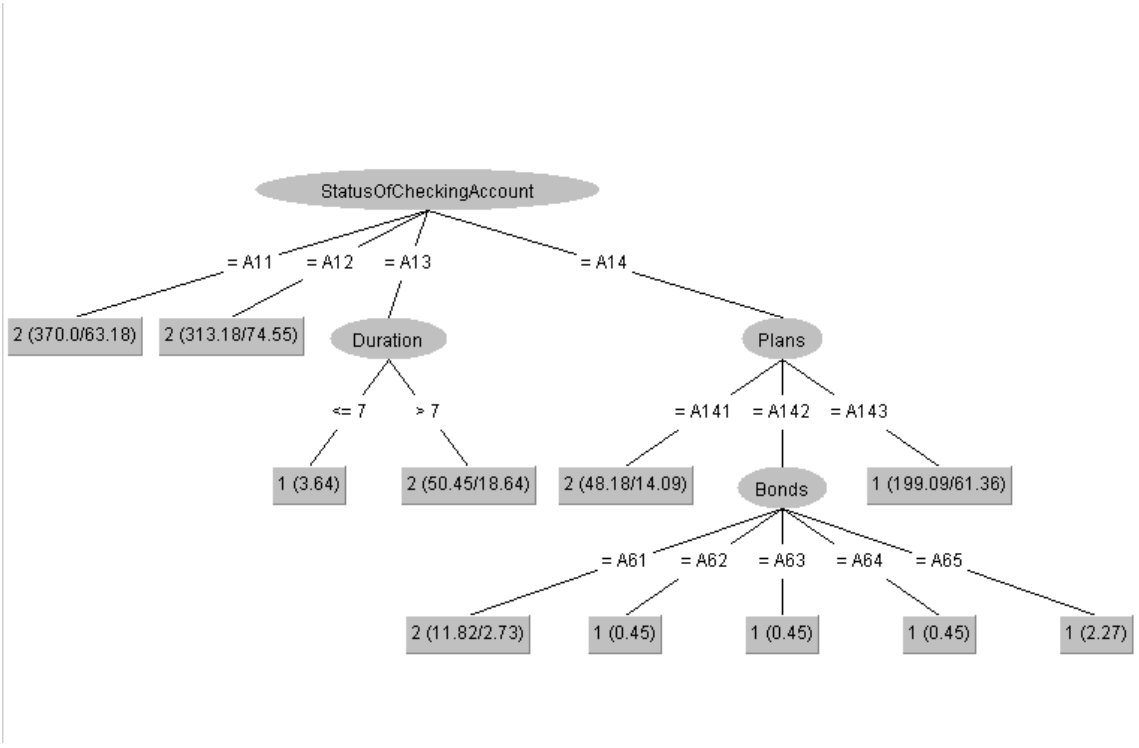
Kokeilemme tehdä päätöspuun Confidence Factory arvolla 0.1 ja vaihdamme luoton antamisen painoarvoksi 5. Päätönpuusta tulee liian iso tulkittavaksi ainakin meidän resursseilla (epäselvä). Ja mikä tärkeä asia selviää: testijoukon tarkkuuskin on vain 57 %. Matrix:sta näkee myös, että nyt paljon arvoja 1 on luokiteltu 2:ksi.

```
=== Confusion Matrix ===

   a   b   <-- classified as
 112 138 |   a = 1
   8  82 |   b = 2
```

Vaikka vaihdamme Confidence Factoryn arvoa, jää testijoukon tarkkuus alhaiseksi, kun painotamme luoton antoa arvolla 5. Ja matriisin arvot ovat näyttämämme kaltaisia.

```
=== Summary ===

Correctly Classified Instances         196               57.6471 %
Incorrectly Classified Instances       144               42.3529 %
Kappa statistic                          0.2523
Total Cost                             144
Average Cost                             0.4235
Mean absolute error                      0.4263
Root mean squared error                  0.5419
Relative absolute error                102.7906 %
Root relative squared error            121.909  %
Total Number of Instances              340

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,456    0,089    0,934      0,456    0,613      0,338    0,694     0,848     1
               0,911    0,544    0,376      0,911    0,532      0,338    0,694     0,387     2
Weighted Avg.  0,576    0,209    0,787      0,576    0,592      0,338    0,694     0,726

=== Confusion Matrix ===

   a   b   <-- classified as
 114 136 |   a = 1
   8  82 |   b = 2
```

Päätöspuu luoton antamisen virheellisyyden painotuksella 5. Confidence Factoryn arvolla 0.01:



5 Arviointi

Testi- joukosta laskettu kustannus oli n. 100 ilman arvon 1 painotusta ja kun painotimme luoton antamista virheellisesti arvolla 5, testi- joukosta laskettu kustannus oli paljon korkeampi, pääsääntöisesti yli 140. Emme saaneet niin tarkkoja tuloksia kuin esimerkiksi harjoituksessa 4. Arvon 1 Painotus arvolla 5 ei anna myöskään kauhean tarkkoja tuloksia. Arvojen 1 luokittelu virheellisesti arvoksi 2 nousi, jos painotus oli 5 arvolle 1. Kokeilimme mielenkiinnosta vaihtaa painotuksen arvoa esim. arvoon 2 ja arvoon 3. Tarkkuus nousi hieman ja väärin Matrixissa luokitellut arvot vähenivät aika paljon ja jakautuivat tasaisemmin. Päätöspuu oli myös epätasainen monilla arvoilla, siis vasen tai oikea puoli oli dramaattisesti isompi. Tottakai kun painotimme arvoa 1, arvon 2 luokittelujen määrä virheellisesti arvoksi 1 laski todella paljon.

6 Täytäntöönpano

Päätöspuu ei sovellu hyvin henkilön luottoriskin ennustamiseen. Emme saaneet testijoukon tarkkuutta nousemaan edes yli 80 %. Jos painotimme luoton antamista potentiaaliselle maksukyvyttömälle henkilölle tilanne vain paheni. Tarkkuus jäi todella alhaiseksi. Kun tarkastelemme päätöspuita mielestäni myös niiden rakenne on sellainen, että siinä esimerkiksi epäolennaisia attribuutteja on korkealla (lähellä puun juurta), mutta tämä on vain mutu- tuntuma, koska juuri opimme tekemään ylipäätään päätöspuita Wekalla. Toimenpidesuositus olisi etsiä vaihtoehtoinen luottoriskien ennustustapa tai ehkä muokata dataa, tarkasteltavia attribuutteja tms.