International Journal of Information & Computation Technology.

ISSN 0974-2239 Volume 4, Number 7 (2014), pp. 753-764

© International Research Publications House

http://www. irphouse.com

clanok: predicting the price of used cars using ML techniques, sameerchand pudaruth, 2014

Použité stĺpce boli: výrobca, počet hláv motoru, rok výroby, najazdené kilometre. Nakoniec použili len 97 záznamov.

Metódy: V tomto prípade, sa pomocou korelácie ukázalo, že vzťah medzi rokom výroby a najazdeným kilometrami je veľmi slabý. Na strane druhej počet hláv motora má silnú koreláciu s cenou. Najvyššia korelácia bola medzi rokom výroby a cenou auta. Na predikciu sa teda použil len rok výroby. Počiatočný koeficient regresie bol 0,814 zlepšenie o 0,005 bolo dosiahnuté zmazaním áut starších ako 20 a mladších ako 4 roky (jednoduchá lineárna regresia). Výrazne zlepšenie dosiahli autori použitím logaritmickej hodnoty (logaritmická regresia) ceny auta namiesto skutočnej. Koeficient vzrástol z 0,814 na 0,851. Výsledkom ale je, že použitie jedného stĺpca je nedostačujúce.

**K-nearest neighbour:** V tomto prípade boli použité 3 atribúty a to: rok výroby, značka auta, počet hláv motora. Dáta boli rozdelené podľa značky auta a zostávajúce 2 stĺpce boli normalizované. Výsledky záležali aj od konzistencie ceny áut. Pri jedno výrobcovi boli dosahované lepšie výsledky predikcie. Najlepšie hodnoty pre *k*boli 1 a 5. Výsledky sa všeobecne zhoršovali so zvyšovaním čísla *k.*

**Rozhodovacie stromy:** Ceny boli rozdelene do kategorií, medzi hranicami kategórii boli značné rozdiely, kvôli chýbajúcim dátam v týchto rozmedziach. Použité stĺpce: výrobca, počet hláv motora, rok. Úspešnosť rozhodovacieho stromu bola okolo 65%. Pri použití náhodného stromu okolo 70%.

**Naive Bayes:** Použité boli rovnaké stĺpce ako pri stromoch a výsledky boli tiež porovnateľne rovnaké.

Problémom pri rozhodovacích stromoch a naive bayes je, že nedokážu pracovať s numerickými triedami. Preto bolo potrebné hodnoty rozdeľovať do kategórií. Takýto postup ale vedie k nepresnostiam. Autor tiež ako jeden z dôvodov slabej úspešnosti uvádza nedostatok dát na trénovanie.

Predicting the Price of Second-hand Cars using Artificial Neural

Networks

Saamiyah Peerun, Nushrah Henna Chummun and Sameerchand Pudaruth

Dáta obsahujú atribúty roku výroby, výrobcu, obsah motoru, typ farby (klasicka/metalicka), typ prevodovky, najazdené kilometre, cenu auta. Dáta obsahovali 200 záznamov. Celkovo autori práce použili 4 metódy predikovania ceny vozidiel. Support vector regression, linear regression, k=nearest neighbour a multilayer perceptron(500 cyklov a learning rate bol 0,05). Porovnanie metód bolo na základe mean absolute error **(stredná odchylka?).** Najhoršia metóda bola knn s chybou 42240. Ostatné metódy dosiahli veľmi podobné hodnotenie s rozpätím 100 bodov medzi sebou.