Извештај за задатак 2 – вишеструка регресија

Даница Газдић
SV 12/2020

Милош Обрадовић
SV 55/2020

1. Проблем

У овом извештају се решава проблем вишеструке регресије над проблемом предикције цене аутомобила. Задатак је решен употребом алгоритма К-најближих суседа уз претпроцесирање које је обухватало *z-score* нормализацију нумеричких обележја и *label* и *one-hot encoding* за категоричка обележја.

2. Решење

Прво су из тренинг скупа избачени аутлајери који нису имали смисла (редови који су за годину производње имали вредност мању од 1900.), такође избачени су и дупликати из тренинг скупа. Извршена је *z-score* нормализација нумеричких обележја. Што се тиче категоричких обележја, пробали смо *label encoding* за све вредности, али с обзиром да су номиналне (немају никакву квантитативну вредност на основу које би се могле рангирати), много боље резултате нам је дао *one-hot encoding.* Ипак, за бинарно обележје 'мењач' је остављен *label encoding* јер су свакако могуће вредности само 0 и 1, а на тај начин смо добили боље предикције. Избачена је колона 'град' јер је енкодовањем уносила велики број нових колона које су кариле добијене резултате и беспотребно повећавале димензионалност проблема. Као што је већ поменуто за предикцију је коришћен модел КНН.

У првој итерацији аутлајери нису избацивани (осим већ поменутих вредности за годину производње) и то решење није дало задовољавајуће резултате (*RMSE* = 95072.2536). Затим смо покренули имплементирано решење кроз две *for* петље, једна је итерирала кроз могуће горње границе цене након које се редови не би узимали у обзир, а друга петља је итерирала кроз могуће вредности параметра *k.* Најмању грешку је дао пар *k*=10, максимална цена=63000. Овако имплементирано решење је дало крајњу грешку *RMSE* = 2670.04874.