**Slovenská technická univerzita**

**Fakulta informatiky a informačných technológií**

## OZNAL

**1. priebežná správa**

**Autori: Bc. Martin Grega, Bc. Šimon Zaťko**

**Cvičiaci: Ing. Róbert Móro, PhD.**

# Prieskumná analýza dát

Na náš semestrálny projekt sme si vybrali dáta o ojazdených autách z webu kaggle.com (<https://www.kaggle.com/avikasliwal/used-cars-price-prediction>). Tento dataset sme si vybrali, pretože sa obaja zaujímame o autá a tu máme k dispozícii dostatok atribútov s rôznymi typmi a distribúciou dát. Okrem toho obsahuje aj niekoľko problémov, ktoré je potrebné vyriešiť a tak si osvojíme metódy spracovávanie dát a všetky úlohy dátovej analýzy a strojového učenia.

Dáta obsahujú informácie o názve auta, mieste, kde sa predáva auto, rok výroby, najazdené kilometre, typ pohonnej hmoty pre auto, typ prevodovky, koľko majiteľov malo auto, spotreba auta, zdvihový objem motora, výkon vozidla, počet sedadiel, cena nového vozidla, cena vozidla.

Ukážka prvých troch riadkov dát:

A screenshot of a cell phone

Description automatically generated

## Opis dát

Dáta obsahujú 6019 riadkov záznamov. V 5-tich stĺpcoch z 13 máme nulové hodnoty. Na základe ďalšej analýzy sa rozhodneme, či a ako budeme dané hodnoty dopĺňať. Ak sa pozrieme na unikátne hodnoty v stĺpcoch, tak ich počty sú celkovo nevýrazné, u väčšiny v rozmedzí <2,540>. Tri stĺpce majú výrazne viac unikátnych hodnôt, čo ale vychádza z podstaty ich informácií (názov auta, najazdené kilometre, cena).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Nulové hodnoty* | *Unikátne hodnoty* | *Najčastejšia hodnota* | *Identifikovaný problém* |
| *Name* | 0 | 1876 | Mahindra XUV500 W8 2WD | - |
| *Location* | 0 | 11 | Mumbai | - |
| *Year* | 0 | 22 | 2014 | - |
| *KmDriven* | 0 | 3093 | 60000 | 1 veľmi vzdialený outlier |
| *FuelType* | 0 | 5 | Diesel | - |
| *Transmission* | 0 | 2 | Manual | - |
| *OwnerType* | 0 | 4 | First | - |
| *Mileage* | 2 | 442 | 18.9 kmpl | Rôzne jednotky |
| *Engine* | 36 | 146 | 1197 CC | Označenie na konci |
| *Power* | 36 | 372 | 74 bhp | Označenie na konci |
| *Seats* | 52 | 9 | 5.0 | Typ float, 1 vozidlo má 0 |
| *NewPrice* | 5195 | 540 | 4.78 Lakh | Veľmi veľa nulových hodnôt |
| *Price* | 0 | 1373 | 4.5 | - |

### Problémové dáta

Skontrolovaním počtu hodnôt v jednotlivých stĺpcoch sme identifikovali niekoľko problémov pri vybraných atribútoch. Všetky z nich ale vieme vyriešiť:

* Outlier vieme nahradiť priemernou hodnotou, mediánom alebo priemernou hodnotou vybranej triedy (podľa iného atribútu), prípadne celý záznam vymazať
* Rôzne jednotky vieme skonvertovať na 1 formát
* Označenie na konci vieme odstrániť, čím získame numerický stĺpec
* Vozidlo s 0 sedadlami vieme doplniť podľa iného vozidla s rovnakým typom, ak také neexistuje tak zmažeme záznam
* Počet sedadiel jednoducho konvertujeme z typu float na integer
* Atribút  *New\_Price* obsahuje príliš veľa nulových hodnôt, preto je pre našu úlohu nepotrebný a neskôr ho pravdepodobne odstránime

V niektorých stĺpcoch sa nachádzajú vychýlené hodnoty, ale podľa skontrolovania sme zistili, že dávajú zmysel – napríklad staršie vozidlá majú o mnoho viac najazdených kilometrov ako novšie a to zodpovedá aj inej distribúcii cien týchto vozidiel.

### Rozšírenie dát

Stĺpec názvu vozidla pre jeho veľa unikátnych hodnôt nevieme dobre využiť, ale po extrahovaní prvého slova získame značku vozidla, čím by sme získali menej unikátnych hodnôt, ktoré by nám pomohli v našej úlohe strojového učenia.

### Korelácie atribútov

Z teplotnej mapy (heatmap) vieme zistiť závislosti medzi jednotlivými numerickými stĺpcami:

|  |  |  |
| --- | --- | --- |
| *Atribút1* | *Atribút2* | *Hodnota korelácie (abs)* |
| Engine | Power | 0.866 |
| Power | Price | 0.773 |
| Engine | Price | 0.658 |
| Engine | Seats | 0.393 |
| Year | Price | 0.305 |

## Definovanie úlohy

Zo získaných údajov vieme vytvoriť niekoľko úloh objavovania znalostí, avšak najrozumnejšia bude predikcia cien vozidiel, ktorú budeme realizovať regresiou, pomocou ktorej budeme predikovať výstupné spojité hodnoty na základe vstupných atribútov.

### Opis predikovanej hodnoty *Price*

V rámci definovanej úlohy sa budeme venovať atribútu *Price*. Jeho distribúcia je zošikmená sprava, takže väčšina hodnôt sa sústreďuje v nižších oblastiach a niekoľko vyšších hodnôt je braných ako outliery.

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

*Obr.1: Distribúcia hodnôt atribútu Price*