Predikcia cien ojazdených aut

Semestrálna práca

Anna Dragolovová, Peter Chochula

2020

Obsah

[Definícia problému 2](#_Toc43236349)

[Dataset a preprocessing dát 3](#_Toc43236350)

[Textové dáta 3](#_Toc43236351)

[Vymazanie extrémov a nezmyselných dát 3](#_Toc43236352)

[Rok registrácie auta 3](#_Toc43236353)

[Rozdelenie a škálovanie dát 6](#_Toc43236354)

[Výber features 7](#_Toc43236355)

[Dáta po preprocessingu 7](#_Toc43236356)

[Lineárna regresia 8](#_Toc43236357)

[Neurónová sieť 9](#_Toc43236358)

[Architektúra 9](#_Toc43236359)

[Stupeň hypotézy a podučenie/preučenie 10](#_Toc43236360)

[Analýza chybovosti dát 11](#_Toc43236361)

[Záver 12](#_Toc43236362)

[Referencie 13](#_Toc43236363)

# Definícia problému

Našim cieľom je predpovedať cenu jazdených áut na základe atribútov uvedených v inzerátoch na webovej stránke ebay-kleinanzeigen.de. Problém predikcie ceny spadá do úloh regresie a preto sa v rámci tejto práce zameriame na regresiu pomocou neurónovej siete, ale okrajovo vyskúšame aj lineárne regresie a to konkrétne lineárnu regresiu bez regularizácie s L1 regularizáciou (Lasso regresiu) a s L2 regularizáciou (Ridge regresiu).

Ako východiskový programovací jazyk tohto projektu bude jazyk Python 3 a jeho knižnice, menovite scikit, keras, numpy, pandas, matplotlib a seaborn.

# Dataset a preprocessing dát

Dataset obsahuje 371528 inzerátov, ktoré obsahujú 20 atribútov (19 bez ceny, ktorú chceme stanoviť). Týmito atribútmi sú:

1. DateCrawled - dátum stiahnutia inzerátu
2. Name – meno inzerátu
3. Seller – predajca (osobný alebo dealer)
4. OfferType – typ inzerátu (ponuka versus dopyt)
5. Price – cena auta v eurách
6. Abtest – A/B testovanie
7. VehicleType – typ vozidla (suv, cabrio...)
8. YearOfRegistration – rok prvej registrácie auta (vek auta)
9. Gearbox – typ riadenia
10. PowerPS - výkon auta
11. Model – model auta
12. Kilometer - počet najazdených kilometrov
13. MonthOfRegistration – v akom mesiaci bolo auto registrované
14. FuelType – typ paliva
15. Brand - značka auta
16. NotRepairedDamage - či je auto poškodené
17. DateCreated - dátum vytvorenia inzerátu
18. NrOfPictures - počet fotiek
19. PostalCode - poštové smerovacie číslo

Niektoré atribúty už od prvého pohľadu budú hrať väčšiu rolu pri stanovení ceny, ako iné. Naopak iné dáta nebudú hrať žiadnu rolu a tie budeme môcť s kľudom na duši úplne zanedbať. Na základe osobných skúseností sme preto určili, že názov inzerátu, čas stiahnutia, počet obrázkov, meno predávajúceho, dátum vytvorenia inzerátu, poštové číslo a dátum, kedy bol naposledy videný inzerát pre naše potreby používať nebudeme. Zmysel používať zostávajúce položky podrobíme podrobnejšej analýze neskôr v tejto práci.

## Textové dáta

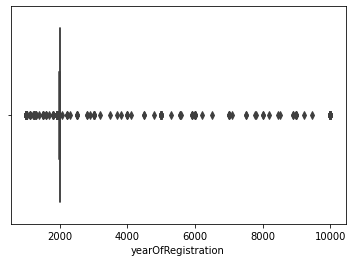
Dataset okrem číselných hodnôt obsahuje aj hodnoty textové. Pre potreby neurónovej siete či lineárnej regresie bolo potrebné preto previesť tieto dáta do ich číselnej reprezentácie. Pre každý atribút bola preto stanovená množina unikátnych hodnôt tak, aby popisovala všetky hodnoty náležajúce danému atribútu a každej hodnote z danej množiny bolo priradené číslo postupne od 1 až do n, kde n reprezentuje veľkosť množiny.

## Vymazanie extrémov a nezmyselných dát

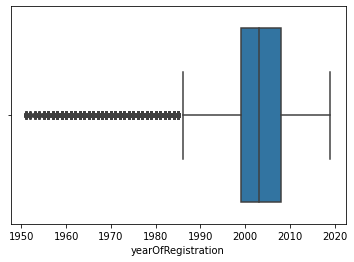
V tejto fáze preprocessingu dát bolo potrebné sa pozrieť bližšie na dáta. Každý zostávajúci atribút bol zobrazený pomocou krabicového grafu a na základe podrobnejšej analýzy boli vymazané nezmyselné či extrémne dáta. Ako príklad z atribútov, na základe ktorých budeme odhadovať cenu, uvedieme rok registrácie auta.

### Rok registrácie auta

Prvotné zobrazenia krabicového grafu vyzeralo nasledovne.



Je vidno, že väčšina áut bolo registrovaných okolo roku 2000, čo je v poriadku. V datasete sa však nachádzajú aj také autá, ktoré boli registrované po roku 2020 či až príliš veľa rokov pred 2020 (stanovili sme minimálny rok 1950). Vzhľadom ku dostatočnému množstvu dát, ako ukážeme neskôr v tejto práci, nebolo potrebné tieto dáta žiadnym spôsobom upravovať, ale stačilo ich odstrániť. Následný graf vyzeral po odstránení dát nasledujúco.

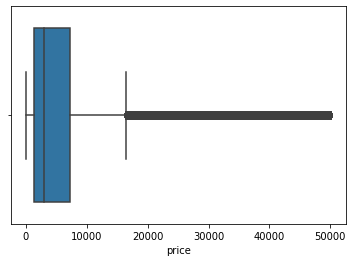


Aj napriek tomu, že nám krabicový graf stále zobrazuje niekoľko extrémnych hodnôt, tieto hodnoty sme sa rozhodli ponechať vzhľadom k tomu, že niektoré inzerované autá mohli byť naozaj registrované neskôr, ako v roku 1985.

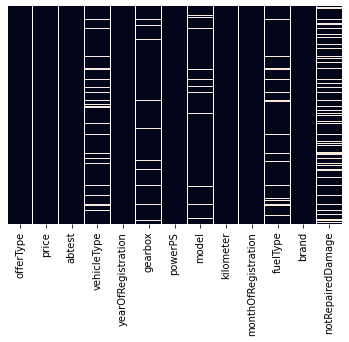
Obdobným spôsobom boli analyzované a upravené aj ďalšie položky vrátane nami predikovanej ceny. Tá mala na začiatku obsahovala širokú škálu hodnôt a to od 0 až po 999999999. Táto distribúcia je zobrazená nižšie na grafe.



Vzhľadom k tomu, že cenu jazdeného auta 999999999 eur považujeme za nezmyselnú, stanovili sme maximálnu hornú cenu na 50 000 a spodnú hranicu na 1. Aj napriek tomu, že nižšie zobrazený krabicový graf má viacero extrémnych hodnôt, rozhodli sme sa ho považovať za dostatočný. Priemerná cena približne 2000 eur na ojazdené auto nám príde reálna, tak ako aj fakt, že niektoré inzerované autá mohli mať naozaj cenovku vyššiu ako 50 000 eur.

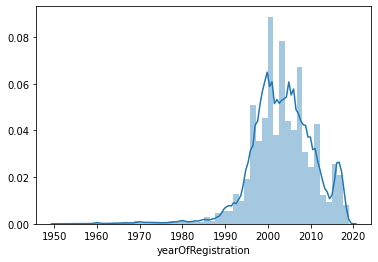


Ďalším nevyhnutným krokom bolo preskúmanie datasetu z pohľadu NaN hodnôt. Pred hocijakými úpravami dataset obsahoval viacero takýchto hodnôt. Na grafe nižšie je NaN hodnota zobrazená bielou vodorovnou čiarou. Tieto dáta boli vymazané.

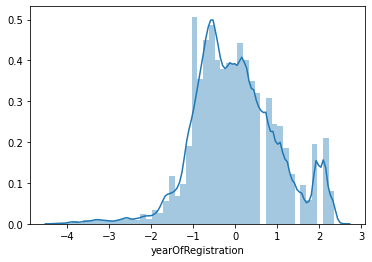


## Rozdelenie a škálovanie dát

Pre efektívnejšie učenie neurónovej siete bolo potrebné preskúmať aj šikmosť dát. Opäť ukážeme v rámci tejto práce jednu ukážku a to na roku registrácie auta.



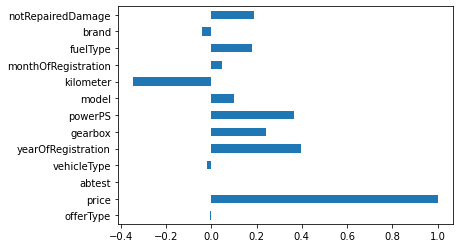
Z grafu je vidno, že dáta reprezentujúce rok registrácie majú zápornú šikmosť. V rámci projektu sme sa túto nerovnomernosť snažili odstrániť. Použili sme tri prístupy a to transformáciu pomocou odmocniny, logaritmickú transformáciu a box cox transformáciu a porovnali, ktorý dáva lepšie výsledky všeobecne na dátach. Po tejto analýze sme zvolili za východiskovú transformáciu box cox a transformovali pomocou nej všetky dáta. Tento algoritmus zároveň dáta automaticky aj škáluje a preto sme už neskôr v tejto práci nemuseli dáta škálovať. Rok registrácie mal po transformovaní nasledujúce rozloženie.



Na koniec je potrebné dodať, že cena nebola žiadnym spôsobom transformovaná ani normalizovaná.

## Výber features

Posledným krokom v preprocesingu dát bol výber atribútov, ktoré budú použité pri učení neurónovej siete. Ako už je vidno z grafu nižšie, ktorý zobrazuje závislosť ceny na jednotlivých atribútoch, niektoré atribúty majú menší vplyv na cenu, ako iné. Jedná sa napríklad o abtest či offertype.



Rozhodli sme sa preto na základe článku od [Abhini Shetye](https://towardsdatascience.com/@abhini.shetye?source=post_page-----e3690ad8504b----------------------) (Shetye, 2020) popisujúcom výber atribútov vyskúšať backward elimination a RFE pre výber atribútov. Oba prístupy vyprodukovali podobné výsledky s tým rozdielom, že backward elimination vybralo aj mesiac registrácie ako dôležitý atribút, kdežto RFE nie. Na základe osobných znalosti sme sa rozhodli túto feature ďalej nepoužívať a obmedzili sme sa len na dáta popisujúce typ vozidla, rok registrácie, typ riadenia, výkon, počet kilometrov, typ paliva, značku a či je poškodené alebo nie. Tento prístup prekvapujúco vyradil atribút model.

## Dáta po preprocessingu

Náš dataset sa po preprocessingu zmenšil celkovo na 232359 inzerátov a 8 atribútov.

# Lineárna regresia

Po spracovaní dát, počas toho, ako sa učila neurónová sieť, sme sa rozhodli vyskúšať riešiť problém aj pomocou lineárne regresie.

|  |  |  |
| --- | --- | --- |
| Algoritmus | Mean absolute error | R squared |
| Linear regression | 2872.2839137529763 | 0.6428558735658016 |
| Lasso regression | 2872.2810739171523 | 0.6428558504235524 |
| Ridge regression | 2872.283527029996 | 0.642855870603666 |

Z výsledkov je zrejmé, že všetky tri prístupy dosahujú takmer rovnakých výsledkov. U Lasso a Ridge regresie bolo zároveň potrebné stanoviť aj optimálny parameter regularizácie lambda. Tento parameter sme stanovili pomocou algoritmu Grid Search. Z množiny čísel 0,001, 0,01, ... 100, 1000) bola pre Lasso regresiu vybraná hodnota 0,01 a pre Ridge regresiu hodnota 10. Algoritmus Lasso si zároveň sám riadi, ktoré atribúty použije. Z našich predspracovaných dát nevyradil dodatočne už žiadny ďalší atribút.

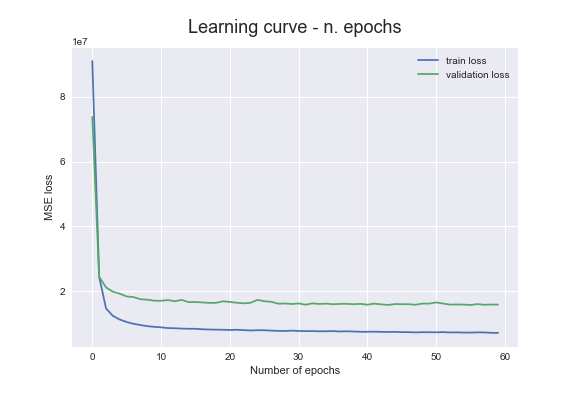
Mean absolute error vo výsledku určuje, o koľko eur sa v priemere odchyľovala predikovaná cena jednotlivými algoritmami od reálnych cien. Ako je z výsledkov zrejmé, priemerná odchýlka je pomerne vysoká a preto si sľubujeme od neurónovej siete značné zníženie tejto odchýlky.

# Neurónová sieť

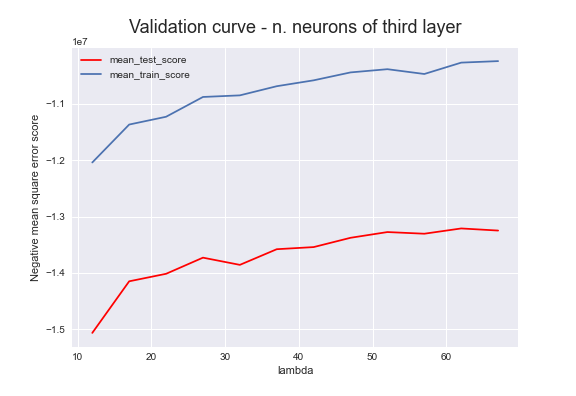
Naše neurónová sieť má na vstupe 8 vybraných, upravených atribútov, ktoré sú spomenuté vyššie. Cieľom neurónovej siete je minimalizovať mean squared error. Inicializácia váh sa riadi náhodnou veličinou z normálnym rozdelením. Vzhľadom k tomu, že chceme predikovať cenu, budeme predikovať nenulové kladné hodnoty. Za týmto účelom sme využili aktivačnú funkciu Relu (Recfied linear unit). Dáta boli vždy náhodne rozdelené na trénovacie (60%), validačné (20%) a testovacie (20%)

## Architektúra

Najskôr sme vytvorili jednoduchú neurónovú sieť s jednou skrytou vrstvou a skontrolovali priebeh učenia. Učenie prebiehalo tak, ako malo. Na grafe nižšie je zobrazené, že na správne naučenie siete nám stačí približne 20 epoch. Ďalšie pridané epochy už nemajú veľký vplyv.



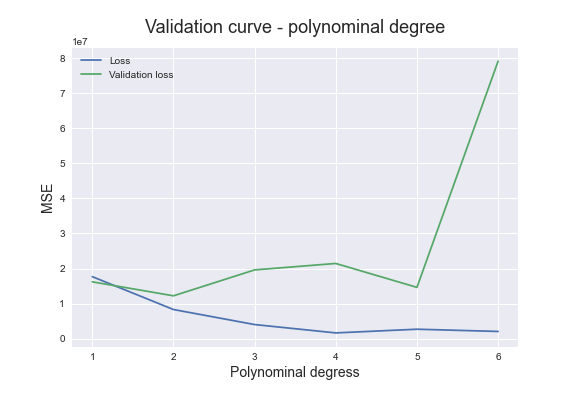
Po určení epoch sme sa za pomoci Grid Search zamerali na určenie vhodného počtu neurónov. Na začiatok sme si stanovili počet neurónov od 13 do 70. Nakoľko algoritmus vyhodnotil, že 69 neurónov je optimálnych, usúdili sme, že model je málo komplexný a z toho dôvodu sme pridali ďalšiu skrytú vrstvu, kde sa optimálny počet neurónov opäť pohyboval na číslach 69 na prvej skrytej vrstve a 54 na druhej. Tento model s dvoma skrytými vrstvami predikoval výslednú cenu s odchýlkou 2473 eur. Toto nebol uspokojivý výsledok a preto sme pridali ešte tretiu skrytú vrstvu s 67 neurónmi, ktorá už predikovala výsledok s priemernou odchýlkou 1909 eur. Vývoj prevrátenej hodnoty MSE v závislosti na počte neurónov na tretej vrstve je zobrazený na nasledujúcom grafe.



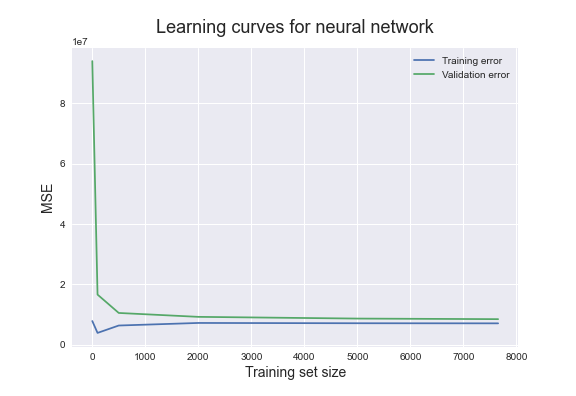
Počet neurónov na jednotlivých vrstvách sme vždy počítali zvlášť. Sme si vedomí, že najvhodnejšiu kombináciu neurónov by bolo optimálne hľadať pomocou rôznych kombinácii na jednotlivých vrstvách, ale kvôli časovej náročnosti sme od toho upustili. Pre ďalšie ladenie neurónovej siete sme už nechceli pridávať nové vrstvy a preto sme a zamerali na optimalizovanie stupňa hypotézy.

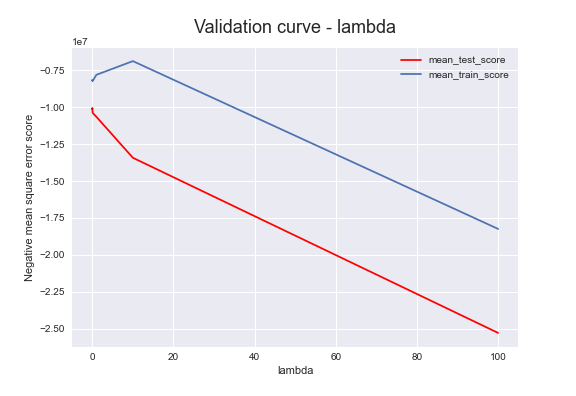
## Stupeň hypotézy a podučenie/preučenie

Pre správne určenie stupňa hypotézy v závislosti na MSE sme si vykreslili nižšie uvedený graf. Z neho vyplýva, že vhodný stupeň hypotézy je 2. Táto zmena znížila opäť priemernú odchýlku a to na 1622 eur. Vzhľadom k tomu, že 1622 eur je stále relatívne veľké číslo, vyskúšali sme do siete pridať aj 4 skrytú vrstvu, ale tá už výraznejšie nezmenila odchýlku.



Z nižšie uvedenej učiacej krivky je zrejmé, že model nie je preučený a väčšie množstvo dát nebude mať významný dopad na proces učenie. Môžeme zároveň predpokladať, že regularizácia nie je potrebná. To potvrdzuje krivka závislosti MSE na regularizačnom parametre lambda.





## Analýza chybovosti dát

Z 900 testovaných má 285 odchýlku väčšiu ako je priemer. Medián rozdielu predikovanej a skutočnej hodnoty je 960. Analýzu vstupných parametrov nám komplikuje cox box škálovanie, ale je patrné, že model zle predikuje autá drahšie ako 2000 eur, nakoľko všetky tieto autá mali odchýlku väčšiu ako bola priemerná odchýlka. Naopak medián u aut s nižšou než priemernou odchýlkou je 567 eur.

# Záver

V tejto práci sme sa zamerali na predikciu cien áut podľa atribútov nachádzajúcich sa na online bazáre s ojazdenými autami. Pre dosiahnutie nami stanoveného cieľa sme okrem tvorby neurónovej siete vyskúšali aj lineárnu regresiu. Po preprocessingu dát a výberu vhodných atribútov naša neurónová sieť pozostávala z troch skrytých vrstiev a stupňom hypotézy 2. Výsledkom bola predikcie s priemernou odchýlkou 1622 eur pričom táto odchýlka je vyššia u áut, ktoré sú drahšie. Je to spôsobené nerovnomerným rozložením cien inzerátov.

# Referencie

Leka, O., 2020. *Used Cars Database*. [online] Kaggle.com. Available at: <https://www.kaggle.com/orgesleka/used-cars-database> [Accessed 13 June 2020].

[Shetye](https://towardsdatascience.com/@abhini.shetye?source=post_page-----e3690ad8504b----------------------) , A., 2019. *Feature selection with sklearn and pandas*. [online] medium.com. Available at: <<https://towardsdatascience.com/feature-selection-with-pandas-e3690ad8504b>> [Accessed 13 June 2020].