Predikce ceny automobilu[1]

Představte si následující situaci. Právě se vracíte z nákupu ze supermarketu do svého krásného malého domečku na vesničce nedaleko města. Nesete s sebou 3 přeplněné tašky s pečivem, sýry, uzeninami, mraženou pizzou a kyblíkem vanilkové zmrzliny. Se svým nákupem absolvujete 30minutovou jízdu autobusem, kterou následuje téměř 1km pochod do prudkého kopce, než se dostanete do svého milovaného domova. Zmrzlina je z velké části rozteklá, mražená pizza měla na mále a uzeninám doba mimo chlad také zrovna neprospěla. Hladoví a vyčerpaní z nákupu a výšlapu s takto těžkým břemenem si připravíte tousty se šunkou a sýrem, sklenici pomerančového džusu a usednete na koženou pohovku k televizi v obývacím pokoji. Pomyslíte si, „O kolik by byl můj život snazší, kdybych měl/a k vlastní auto…“. S ohledem na to, že jste si teprve rok zpátky brali hypotéku na svůj vysněný dům si však v nejbližší době můžete o vlastním autě nechat zdát.

Po televizních novinách se v televizi objeví reklamní spot, který jako by mluvil přímo k Vám. „Vidíte toto auto? Kolik myslíte, že takové fáro stojí?“ táže se charismatický starší pán z televize. „Dokážete odhadnout ceny vozů lépe než ostatní? Pak můžete být výherce tohoto auta právě Vy! Přihlaste se do naší nové televizní soutěže ještě dnes!“. V tu ránu víte, že tohle je příležitost, kterou si nemůžete nechat ujít. Rychle odešlete přihlášku a necháte se unášet představami svého úžasného budoucího života s autem. Zhruba po 30 minutách odezní prvotní nadšení a narazíte na drobný zádrhel ve Vašem jinak dokonalém plánu. Tím je fakt, že o autech nevíte vůbec nic.

Naštěstí si ještě pamatujete něco málo ze svých vysokoškolských let a hodin statistiky a rozhodnete se své znalosti využít a zvýšit tak své šance v blížící se soutěži. V rámci tréninku najdete na internetových stránkách autobazarů množství dat o vozidlech různých značek, jejich charakteristiky a cenu[2]. Pro analýzu těchto dat využijete všeobecný lineární model[3].

Závislou proměnnou ve vašem modelu bude pochopitelně cena. Tu se pokusíte predikovat pomocí následujících kategoriálních a spojitých regresorů:

Kategorické:
- značka vozidla
- typ paliva
- počet dveří
- karosérie
- umístění motoru (vepředu nebo vzadu)

Spojité:
- délka vozidla
- „počet koní“
- maximální otáčky

U spojitých proměnných předpokládáte vztah „čím víc, tím líp“, nicméně Vás napadne, že příliš dlouhé vozidlo by mohlo být nemotorné, lidi o něj budou mít menší zájem a tím pádem bude i cena takového auta nižší. Příliš krátká auta pak podle Vás vypadají směšně a myslíte si, že by je za velké peníze také nikdo nechtěl. Protože předpokládáte, že regresní přímka by u této proměnné nebyla úplně

[1] *Data a další informace o této zprávě jsou dostupné na adrese https://dostal.vyzkum-psychologie.cz/stat4?i=180.*

[2] Původní data jsou dostupná na stránce https://www.kaggle.com/code/goyalshalini93/car-price-prediction-linear-regression-rfe/input

[3] Data byla analyzována za pomocí software TIBCO Statistica

reprezentativní, přidáte do modelu regresor „délka vozidla na druhou“, který dovolí přímce stát se křivkou s „kopcem“, který zde očekáváte.

***Tabulka č. 1** - Koeficient determinace a test podmodelu*

Model	koeficient determinace	testová statistika F	p-hodnota
	0,86	32,33	<0,001

Jak lze vidět v Tabulce č. 1, Vašim modelem se Vám podařilo vysvětlit 86% rozptylu závislé proměnné, tedy ceny vozidla. Ačkoli stále někde létá 14% nevysvětleného rozptylu, může tento model nabídnout stále poměrně slušné indicie, na co při odhadu ceny koukat. V tabulce číslo 2 vidíme, že nejlepší rychlý odhad uděláme, pokud budeme koukat na značku vozidla, počet koní a na to, zda má auto motor vepředu nebo vzadu.

***Tabulka č. 2** - Test statistické významnosti a míra účinku*

Regresor	SS	Testová statistika F	p-hodnota	Míra účinku
Značka	2850476246,21	12,31	0,00	0,60
Palivo	3303559,55	0,30	0,58	0,00
Počet dveří	6079053,21	0,55	0,46	0,00
Karosérie	47754175,32	1,08	0,37	0,02
Umístění motoru	528702498,65	47,95	0,00	0,22
Délka vozidla	33014384,99	2,99	0,09	0,02
Koňských sil	810509496,71	73,51	0,00	0,30
Max. otáček	20713965,12	1,88	0,17	0,01
Délka vozidla^2	38640986,13	3,50	0,06	0,02

V tabulce č. 3 pak vidíte jednotlivé značky, a jak se promítají do výsledné ceny spolu s konfidenčními intervaly. Dále z ní vyčteme, že vozy s motorem vzadu jsou výrazně dražší než ty, s motorem vepředu a že za každou koňskou sílu navíc roste hodnota vozu asi o 80 amerických dolarů.

Tabulka č. 3 – *Vztah ceny vozidla a značky, umístění motoru a počtu koňských sil*

	Regresor	Regresní koeficient	Waldova statistika T	p-hodnota	Konfidenční interval -95,00 %	Konfidenční interval +95,00 %
Značka	alfa-romeo	-3842,0	-1,274	0,205	-9796,5	2112,5
Značka	audi	-1132,3	-0,500	0,617	-5597,8	3333,2
Značka	bmw	6171,2	3,016	0,003	2132,2	10210,2
Značka	chevrolet	-4985,1	-1,945	0,053	-10045,3	75,1
Značka	dodge	-4821,9	-2,535	0,012	-8576,7	-1067,1
Značka	honda	-3234,7	-1,929	0,055	-6544,5	75,2
Značka	isuzu	-7636,7	-3,442	0,001	-12016,0	-3257,5
Značka	jaguar	12172,4	4,886	0,000	7255,2	17089,6
Značka	mazda	-2460,4	-1,699	0,091	-5319,4	398,6
Značka	buick	15010,4	7,871	0,000	11246,2	18774,6
Značka	mercury	-2396,3	-0,657	0,512	-9595,8	4803,3
Značka	mitsubishi	-4896,0	-3,098	0,002	-8015,9	-1776,2
Značka	nissan	-3165,7	-2,142	0,034	-6082,9	-248,5
Značka	peugeot	871,0	0,536	0,593	-2339,2	4081,1
Značka	plymouth	-4671,7	-2,515	0,013	-8338,8	-1004,7
Značka	porsche	-10802,3	-3,776	0,000	-16448,6	-5156,0
Značka	renault	-3755,1	-1,395	0,165	-9067,9	1557,7
Značka	saab	-426,8	-0,232	0,817	-4053,2	3199,6
Značka	subaru	-5758,5	-3,809	0,000	-8742,8	-2774,2
Značka	toyota	-4418,0	-3,310	0,001	-7053,0	-1783,1
Značka	volkswagen	-4067,4	-2,740	0,007	-6996,9	-1137,8
Značka	volvo	0,0				
Umístění motoru	Vepředu	-22844,5	-6,925	0,000	-29356,3	-16332,7
Umístění motoru	Vzadu	0,0				
Koňských sil		80,3	8,574	0,000	61,8	98,8

Nám už nyní nezbývá, než popřát Vám co nejvíce štěstí do další přípravy a do samotné soutěže. Nepochybujeme, že auto bude brzy Vaše.