Pràctica 2

Realitzat per: Gerard López Roig i Antonio Àngel Martínez Chamizo

Descripció del dataset. Perquè és important i quina pregunta pretén respondre?

El dataset conté dades que fan referència a vehicles de segona mà de la marca Audi; vehicles que, en el moment de la recol·lecció de dades, hem trobat en venda a Espanya. La informació capturada al conjunt de dades reflexa les propietats quantitatives i qualitatives que fan referència al vehicle en venda, alguns exemples representatius serien: model, preu actual, localització del venedor i fitxa tècnica. S'han ignorat aspectes més subjectius i de difícil anàlisi com l'equipament, les imatges del vehicle o la descripció no estructurada donada per l'anunciant en texte lliure.

La pregunta que volem respondre és, fonamentalment, quins atributs de la fitxa tècnica del vehicle són els millors predictors del seu preu de segona mà.

Integració i selecció de les dades d'interès a analitzar

Pel que fa la integració, totes les nostres dades provenen del mateix procés de web scraping de la pràctica 1. Donat que la informació només té una font, el procés d'integració és innecessari en el nostre cas.

La selecció de les dades a analitzar també va prendre lloc en la pràctica 1, només es varen recollir dades de vehicles de la marca Audi per a conduir aquest anàlisi. La part d'exploració de dades que té lloc en aquesta fase de la neteja de dades continúa en la següent secció.

Neteja de les dades

Les dades contenen zeros o elements buits? Com gestionaries aquests casos?

Veiem que tenim elements buits inesperats en la columna d'eficiència ecològica i, més sovint, en les columnes que corresponen a la fitxa tècnica. Un ~10% dels anuncis no tenen fitxa tècnica, de manera que totes les columnes que capturen les seves dades (e.g potència del motor, consum, etc) queden buides en aquests casos.

En el cas de les variables categòriques de la fitxa tècnica i de l'eficiència ecològica, primer intentem veure quin és el valor més freqüent en aquell model de vehicle (e.g Audi A5) i si no trobem el model entre les dades dels vehicles que tenen fitxa tècnica, ens quedem amb el valor més freqüent de tot el dataset.

En el cas de les variables contínues seguim un procés semblant, però agafant el valor mitjà del model del vehicle, i en el seu defecte, el valor mitjà de la columna per a tot el dataset.

Identificació i tractament de valors extrems

Per detectar els valors extrems en el cas de les variables contínues, generem boxplots amb tal d'identificar-los gràficament. Com podem veure en els boxplots del notebook, gairebé cada variable contínua conté outliers, però cap dels valors observats a la figura sembla no pertanyer a la distribució. Si bé es troben lluny de la mitjana, semblen ser valors legitims, ja que són valors possibles i no infreqüents (sovint es troben acompanyats per altres outliers), com per exemple els vehicles amb un gairebé 500 (models de gamma alta) o un quilometratge de 0 per als quilòmetre zero que no han sortit del concessionari.

Una comú font d'outliers, l'ús intercanviable de ',' i '.' com a separador de millars, es gestiona en el pas de transformació de les dades de string a float.

En els casos categòrics, si bé no trobem outliers per definició estricta, si que trobem valors que són impossibles o que corresponen a 'placeholders'. En aquests casos, el remei és la inspecció visual i tractar-los com hem tractat els NaNs categòrics.

Anàlisi de les dades

Selecció dels grups de dades que es volen analitzar/comparar (planificació dels anàlisis a aplicar)

Agafem els atributs continus de la fitxa tècnica i els preus de venda actual com a label. Per a aquest anàlisi no considerem els atributs categòrics ja que volem fer servir una regressió lineal. Com a afegit, hem considerat la columna del color per a un anàlisi posterior, però no es farà servir en la regressió.

L'atribut 10 no l'agafem, ja que és el preu de venda nova, altament correlacionat amb el preu de venda de segona mà i no és una propietat per si mateixa del vehicle. Hem decidit també deixar enrere l'atribut 28, ja que és equivalent a l'atribut 27 potencia del motor (en kW), però en cavalls de vapor i és aquest darrer el que es fa servir més sovint en documents tècnics.

Els valors seleccionats per a la regressió lineal, són els restants que mostren una major correlació amb el preu i que es troben poc correlacionats entre si. L'any de matricula (2), el quilometratge (4), la potència (27) i l'acceleració (31).

Comprovació de la normalitat i homogeneïtat de la variància

La normalitat o la seva absència es poden intuir en el gràfic de la distribució de les variables numèriques, en qualsevol cas, executant el test de Shapiro-Wilk veiem que els quatre atributs seleccionats per al model no compleixen la normalitat. De manera semblant, veiem el mateix per la homogeneïtat de la variància, el test de Levene confirma la heteroscedasticitat dels atributs, el qual és d'esperar, ja que és el cas en la majoria de situacions del món real. Tanmateix, encara es pot utilitzar els mínims quadrats ordinaris sense corregir l'heteroscedasticitat o normalitat, ja que la mida mostra és prou gran ($n > 30$). Pel teorema

central del límit i pel fet que la variància de l'estimador de mínims quadrats encara pot ser prou petita per obtenir estimacions precises, donada la elevada quantitat de mostres al dataset.

Aplicació de proves estadístiques per comparar els grups de dades. En funció de les dades i de l'objectiu de l'estudi, aplicar proves de contrast d'hipòtesis, correlacions, regressions, etc. Aplicar almenys tres mètodes d'anàlisi diferents

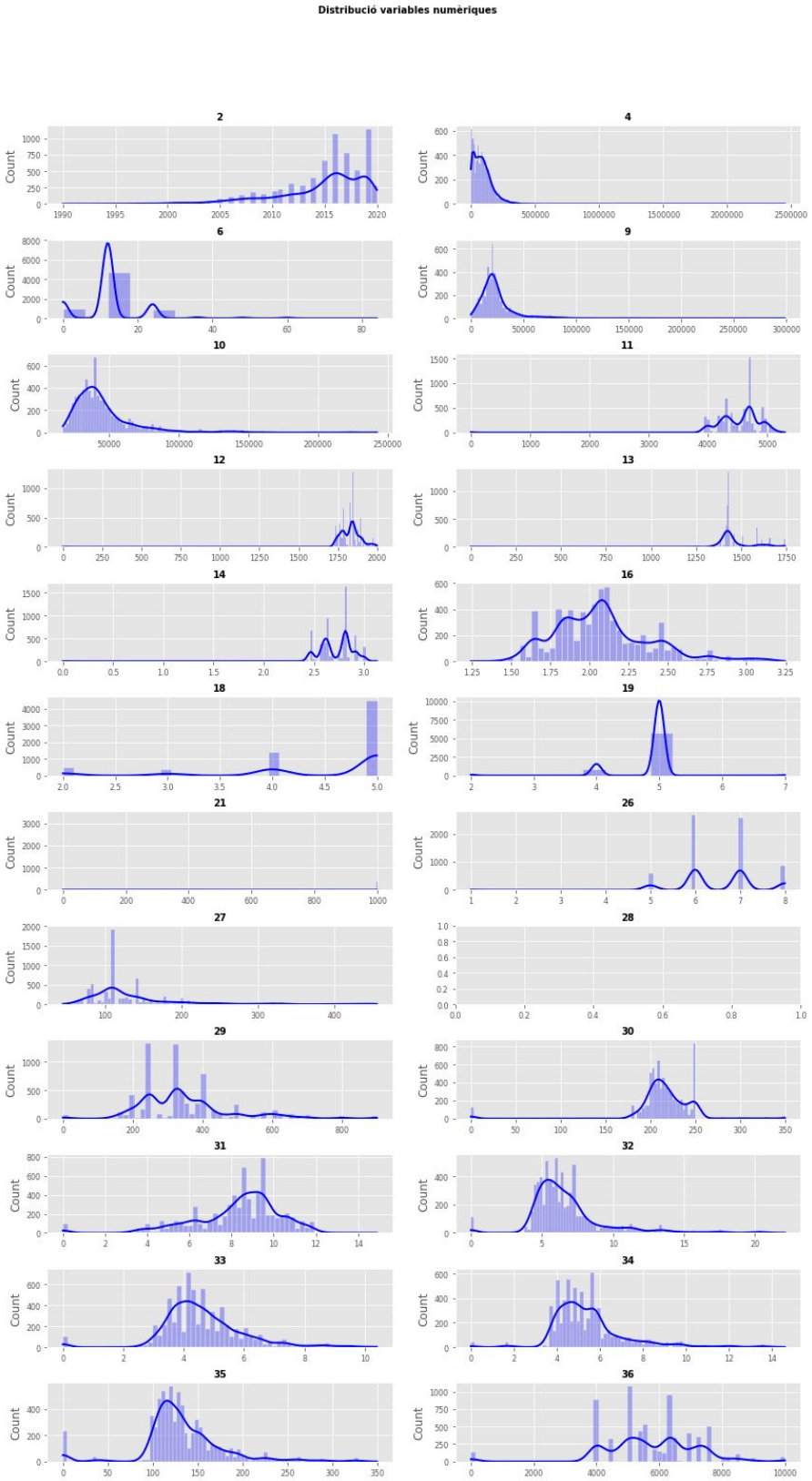
Correlació per a la selecció d'atributs a utilitzar en el model. I hem seleccionat l'any de matricula (2), el quilometratge (4), la potència (27) i l'acceleració (31).

Hem fet una regressió lineal amb tal de veure com s'ajusten els atributs seleccionats a un model lineal. El model obtingut té un R2 de 0.747, però en analitzar els seus residuals veiem que el model no s'ajusta a un model lineal.

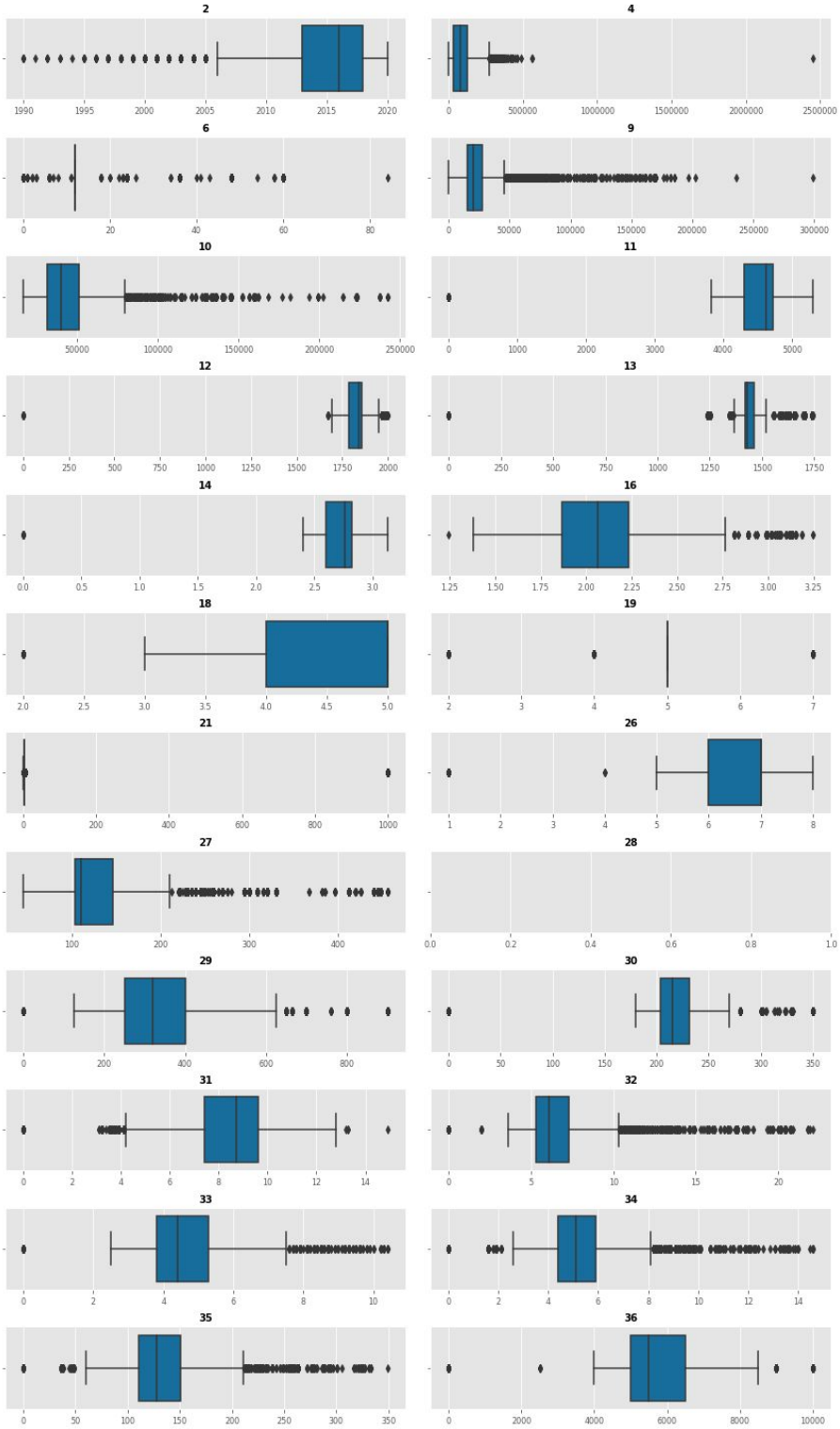
Hem fet un random forest per confirmar la nostra selecció d'atributs, veiem que el random forest retorna el mateix rànquing que hem vist amb el nostre mètode inicial.

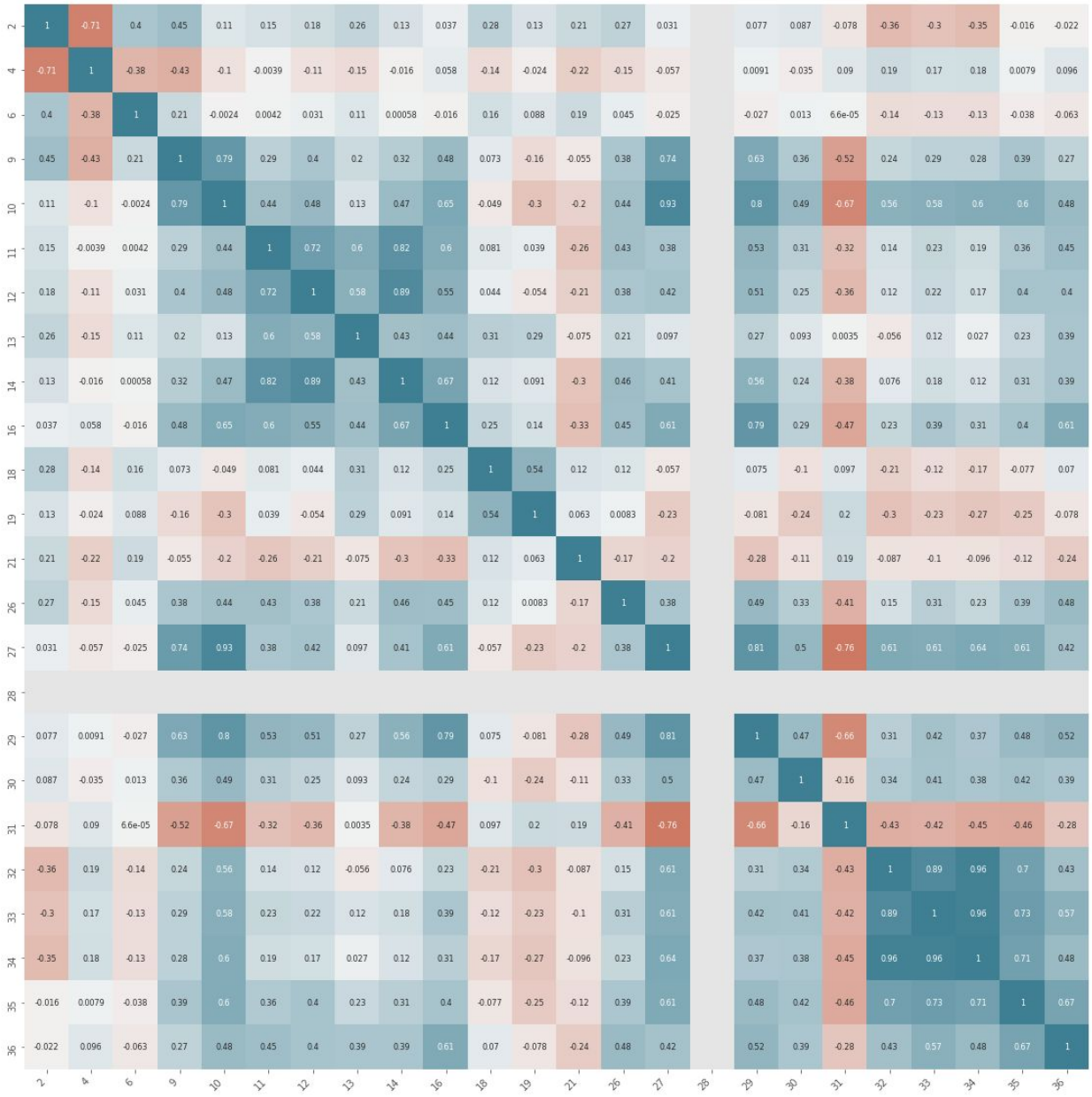
En separar els vehicles en cotxes vermells i no vermells, després de comprovar que tenen distribucions de model semblants, es a dir, que contenen freqüències semblants per a cada model de vehicle i que no hi ha cap model sobrerrepresentat desplaçant la mitjana, hem obtingut amb un t-test que hi ha una diferència significativa a la mitjana de preus entre uns i altres, afavorint la resta de colors sobre els vermells, en contra de la saviesa popular.

Representació dels resultats a partir de taules i gràfiques

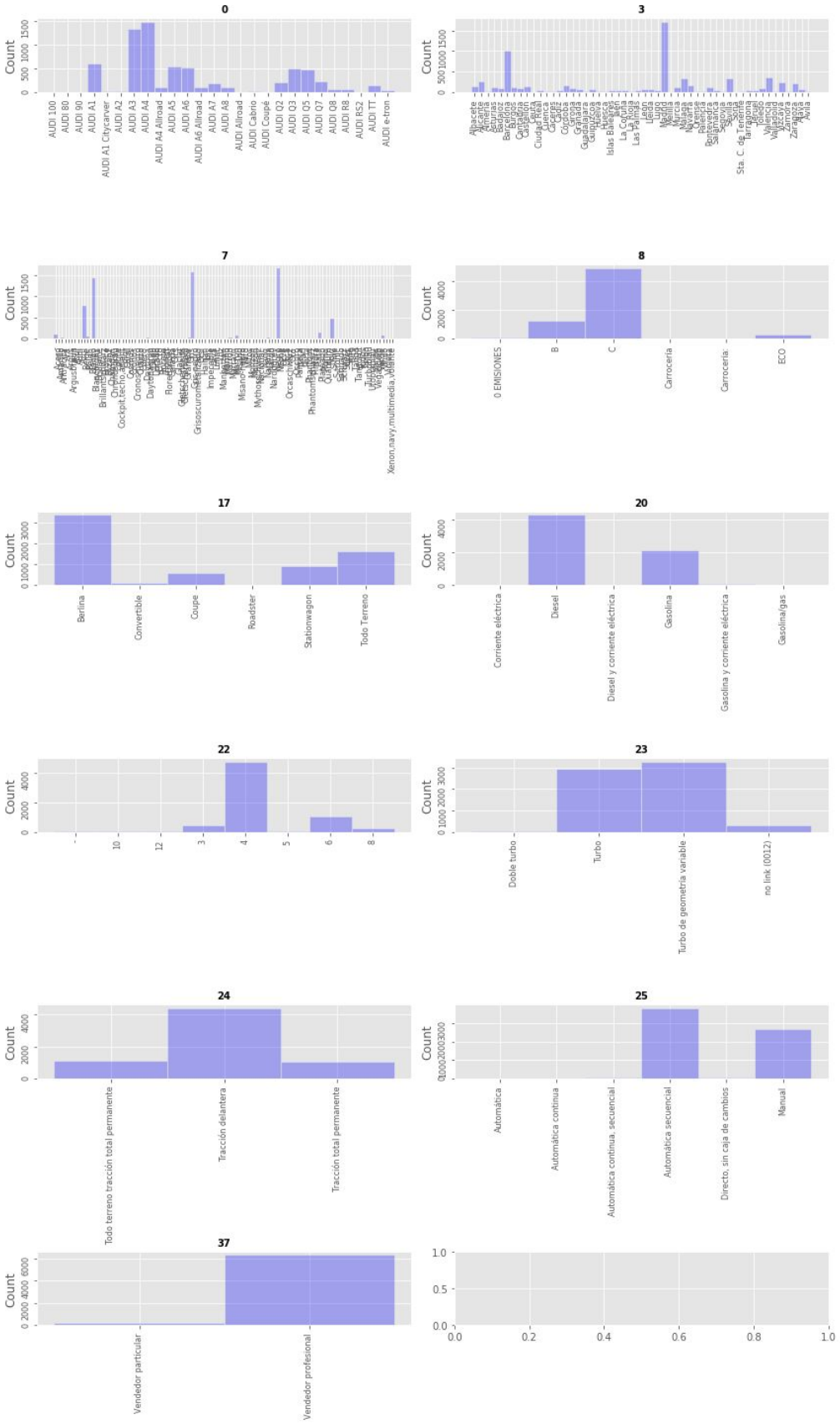


Distribució variables numèriques




2 1 -0.71 0.4 0.45 0.11 0.15 0.18 0.26 0.13 0.037 0.28 0.13 0.21 0.27 0.031 0.077 0.087 -0.078 -0.36 -0.3 -0.35 -0.016 -0.022
4 -0.71 1 -0.38 -0.43 -0.1 -0.0039 -0.11 -0.15 -0.016 0.058 -0.14 -0.024 -0.22 -0.15 -0.057 0.0091 -0.035 0.09 0.19 0.17 0.18 0.0079 0.096
6 0.4 -0.38 1 0.21 -0.0024 0.0042 0.031 0.11 0.00058 -0.016 0.16 0.088 0.19 0.045 -0.025 -0.027 0.013 6.6e-05 -0.14 -0.13 -0.13 -0.038 -0.063
9 0.45 -0.43 0.21 1 0.79 0.29 0.4 0.2 0.32 0.48 0.073 -0.16 -0.055 0.38 0.74 0.63 0.36 -0.52 0.24 0.29 0.28 0.39 0.27
10 0.11 -0.1 -0.0024 0.79 1 0.44 0.48 0.13 0.47 0.65 -0.049 -0.3 -0.2 0.44 0.93 0.8 0.49 -0.67 0.56 0.58 0.6 0.6 0.48
11 0.15 -0.0039 0.0042 0.29 0.44 1 0.72 0.6 0.82 0.6 0.081 0.039 -0.26 0.43 0.38 0.53 0.31 -0.32 0.14 0.23 0.19 0.36 0.45
12 0.18 -0.11 0.031 0.4 0.48 0.72 1 0.58 0.89 0.55 0.044 -0.054 -0.21 0.38 0.42 0.51 0.25 -0.36 0.12 0.22 0.17 0.4 0.4
13 0.26 -0.15 0.11 0.2 0.13 0.6 0.58 1 0.43 0.44 0.31 0.29 -0.075 0.21 0.097 0.27 0.093 0.0035 -0.056 0.12 0.027 0.23 0.39
14 0.13 -0.016 0.00058 0.32 0.47 0.82 0.89 0.43 1 0.67 0.12 0.091 -0.3 0.46 0.41 0.56 0.24 -0.38 0.076 0.18 0.12 0.31 0.39
16 0.037 0.058 -0.016 0.48 0.65 0.6 0.55 0.44 0.67 1 0.25 0.14 -0.33 0.45 0.61 0.79 0.29 -0.47 0.23 0.39 0.31 0.4 0.61
18 0.28 -0.14 0.16 0.073 -0.049 0.081 0.044 0.31 0.12 0.25 1 0.54 0.12 0.12 -0.057 0.075 -0.1 0.097 -0.21 -0.12 -0.17 -0.077 0.07
19 0.13 -0.024 0.088 -0.16 -0.3 0.039 -0.054 0.29 0.091 0.14 0.54 1 0.063 0.0083 -0.23 -0.081 -0.24 0.2 -0.3 -0.23 -0.27 -0.25 -0.078
21 0.21 -0.22 0.19 -0.055 -0.2 -0.26 -0.21 -0.075 -0.3 -0.33 0.12 0.063 1 -0.17 -0.2 -0.28 -0.11 0.19 -0.087 -0.1 -0.096 -0.12 -0.24
26 0.27 -0.15 0.045 0.38 0.44 0.43 0.38 0.21 0.46 0.45 0.12 0.0083 -0.17 1 0.38 0.49 0.33 -0.41 0.15 0.31 0.23 0.39 0.48
27 0.031 -0.057 -0.025 0.74 0.93 0.38 0.42 0.097 0.41 0.61 -0.057 -0.23 -0.2 0.38 1 0.81 0.5 -0.76 0.61 0.61 0.64 0.61 0.42
28
29 0.077 0.0091 -0.027 0.63 0.8 0.53 0.51 0.27 0.56 0.79 0.075 -0.081 -0.28 0.49 0.81 1 0.47 -0.66 0.31 0.42 0.37 0.48 0.52
30 0.087 -0.035 0.013 0.36 0.49 0.31 0.25 0.093 0.24 0.29 -0.1 -0.24 -0.11 0.33 0.5 0.47 1 -0.16 0.34 0.41 0.38 0.42 0.39
31 -0.078 0.09 6.6e-05 -0.52 -0.67 -0.32 -0.36 0.0035 -0.38 -0.47 0.097 0.2 0.19 -0.41 -0.76 -0.66 -0.16 1 -0.43 -0.42 -0.45 -0.46 -0.28
32 -0.36 0.19 -0.14 0.24 0.56 0.14 0.12 -0.056 0.076 0.23 -0.21 -0.3 -0.087 0.15 0.61 0.31 0.34 -0.43 1 0.89 0.96 0.7 0.43
33 -0.3 0.17 -0.13 0.29 0.58 0.23 0.22 0.12 0.18 0.39 -0.12 -0.23 -0.1 0.31 0.61 0.42 0.41 -0.42 0.89 1 0.96 0.73 0.57
34 -0.35 0.18 -0.13 0.28 0.6 0.19 0.17 0.027 0.12 0.31 -0.17 -0.27 -0.096 0.23 0.64 0.37 0.38 -0.45 0.96 0.96 1 0.71 0.48
35 -0.016 0.0079 -0.038 0.39 0.6 0.36 0.4 0.23 0.31 0.4 -0.077 -0.25 -0.12 0.39 0.61 0.48 0.42 -0.46 0.7 0.73 0.71 1 0.67
36 -0.022 0.096 -0.063 0.27 0.48 0.45 0.4 0.39 0.39 0.61 0.07 -0.078 -0.24 0.48 0.42 0.52 0.39 -0.28 0.43 0.57 0.48 0.67 1
2 4 6 9 10 11 12 13 14 16 18 19 21 26 27 28 29 30 31 32 33 34 35 36

Distribució variables categòriques

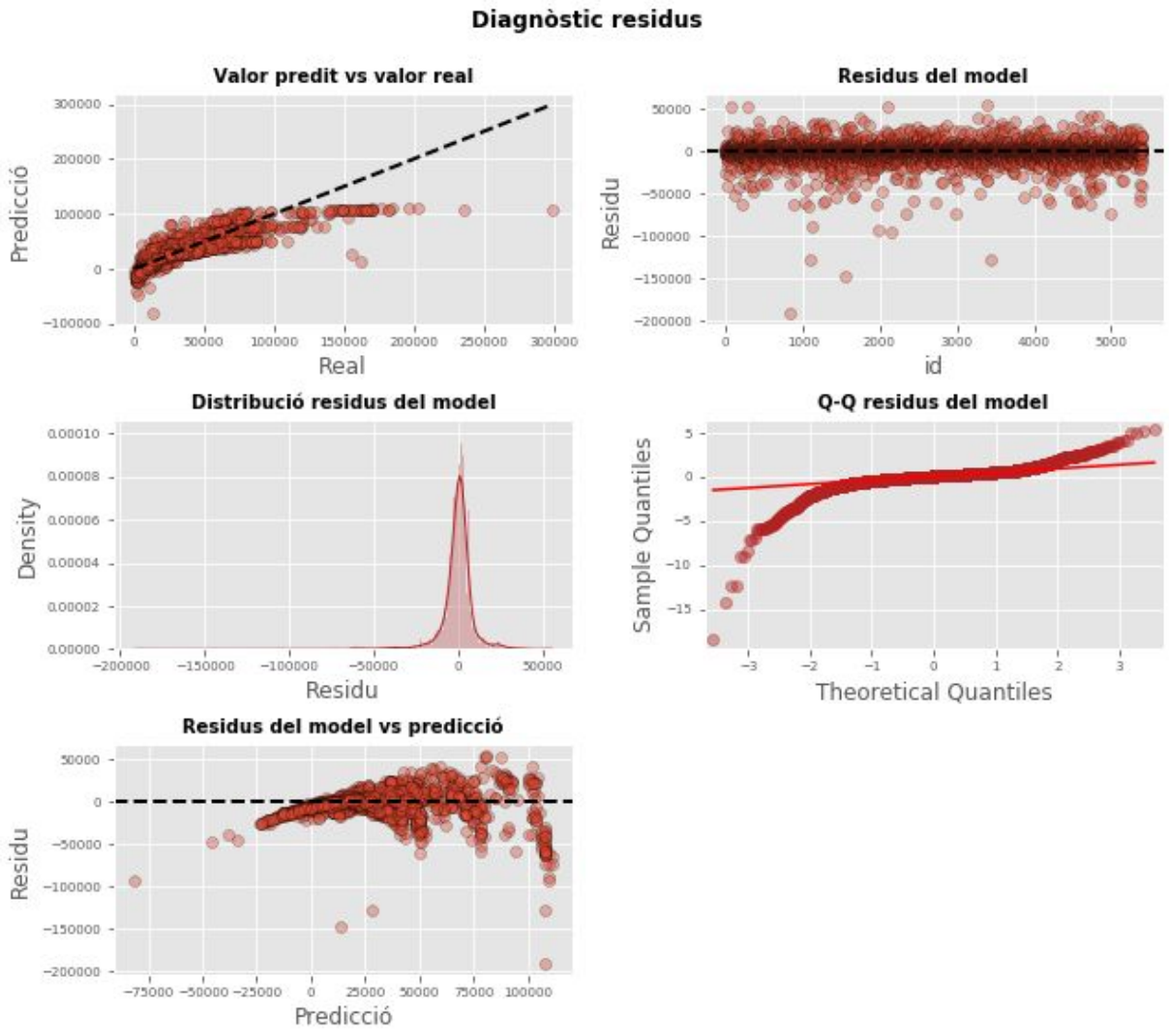


OLS Regression Results

Dep. Variable:	y	R-squared:	0.747
Model:	OLS	Adj. R-squared:	0.747
Method:	Least Squares	F-statistic:	3991.
Date:	Tue, 05 Jan 2021	Prob (F-statistic):	0.00
Time:	21:55:38	Log-Likelihood:	-57654.
No. Observations:	5404	AIC:	1.153e+05
Df Residuals:	5399	BIC:	1.154e+05
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-3.002e+06	8.82e+04	-34.023	0.000	-3.17e+06	-2.83e+06
2	1479.9457	43.668	33.891	0.000	1394.338	1565.553
4	-0.0429	0.003	-16.790	0.000	-0.048	-0.038
27	258.9752	3.225	80.297	0.000	252.652	265.298
31	1621.0904	104.958	15.445	0.000	1415.330	1826.851

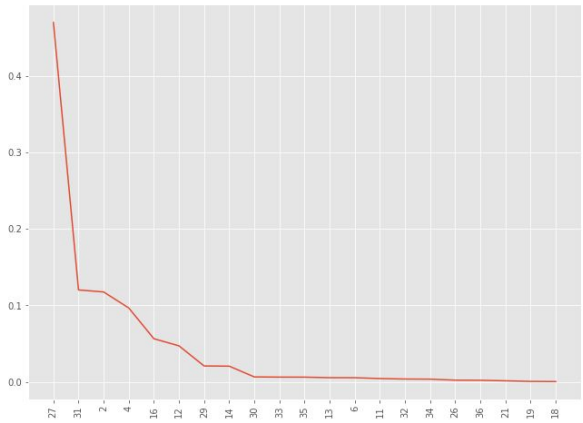
Omnibus:	4846.062	Durbin-Watson:	1.996
Prob(Omnibus):	0.000	Jarque-Bera (JB):	519212.829
Skew:	3.841	Prob(JB):	0.00
Kurtosis:	50.402	Cond. No.	7.41e+07

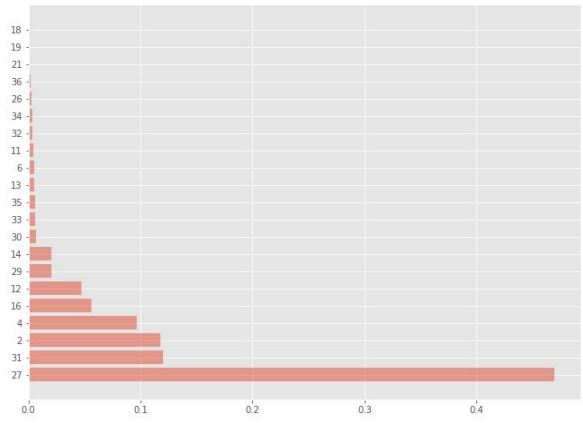


Random Forest Regression features importance

	index	feature	importance
0	12	27	0.469744
1	15	31	0.120269
2	0	2	0.117632
3	1	4	0.096727
4	7	16	0.056534

5	4	12	0.047225
6	13	29	0.020953
7	6	14	0.020683
8	14	30	0.006617
9	17	33	0.006397
10	19	35	0.006350
11	5	13	0.005620
12	2	6	0.005605
13	3	11	0.004501
14	16	32	0.003875
15	18	34	0.003712
16	11	26	0.002395
17	20	36	0.002276
18	10	21	0.001560
19	9	19	0.000772
20	8	18	0.000554





Resolució del problema. A partir dels resultats obtinguts, quines són les conclusions? Els resultats permeten respondre al problema?

A partir dels resultats podem concloure que la predicció del preu de segona mà a partir de les dades de la fitxa tècnica no s'ajusta bé a un model lineal i que el millor predictor del preu de segona mà, ignorant el preu del vehicle nou, és la potència del motor, seguit de l'acceleració. Tant en correlació directa com en la importància de l'atribut dintre del random forest. També hem observat que el color del cotxe, en aquest cas vermell, té un efecte significatiu en el preu, de manera que seria interessant l'estudi d'altres atributs categòrics de la mateixa base de dades i el seu efecte en el preu.

Contribucions	Signa
Recerca Previa	A.A.M.C, G.L.R
Redacció de les respostes	A.A.M.C, G.L.R
Desenvolupament del codi	A.A.M.C, G.L.R