Used Car Sales
Predicting Car Prices

Fernando Barros Magariños

Business Goals

- Current situation:
  - Estimations are done by one single person.
  - This person will be on retirement as of next month
  - There is no replacement.
  - Estimations are 30% of the listed price in average
- Goals:
  - Reduce the estimation error to 10% of the listed price in average.
  - Automate the estimation process

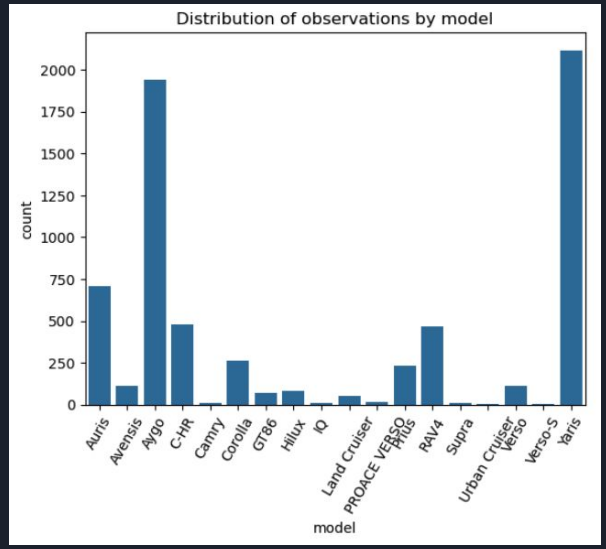
Data

	model	year	price	transmission	mileage	fuelType	tax	mpg	engineSize
0	GT86	2016	16000	Manual	24089	Petrol	265	36.2	2.0
1	GT86	2017	15995	Manual	18615	Petrol	145	36.2	2.0
2	GT86	2015	13998	Manual	27469	Petrol	265	36.2	2.0
3	GT86	2017	18998	Manual	14736	Petrol	150	36.2	2.0
4	GT86	2017	17498	Manual	36284	Petrol	145	36.2	2.0

- 6738 observations of cars
- 8 Features + Target variable (Price)
- Dataset in very good shape

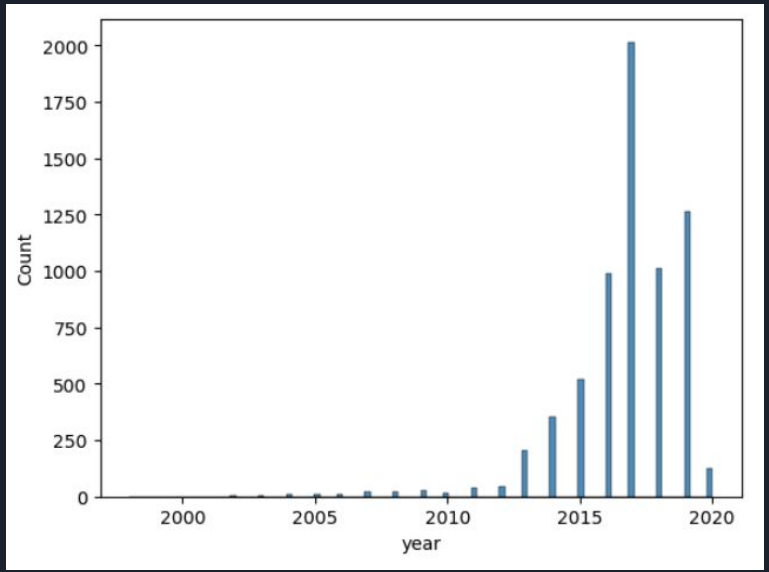
Key Findings

- Big difference in the amount of observations depending on the model



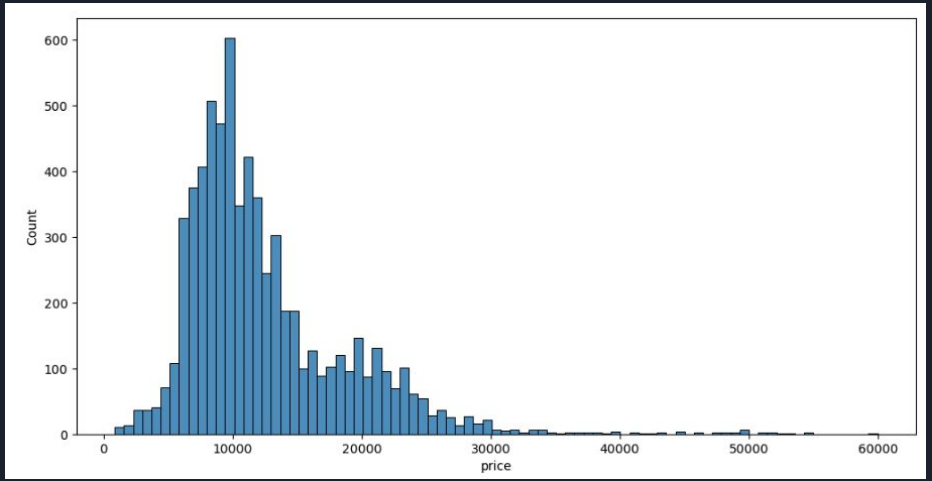
Key Findings II

- Cars dating back from 1998, but mainly from 2015-2020



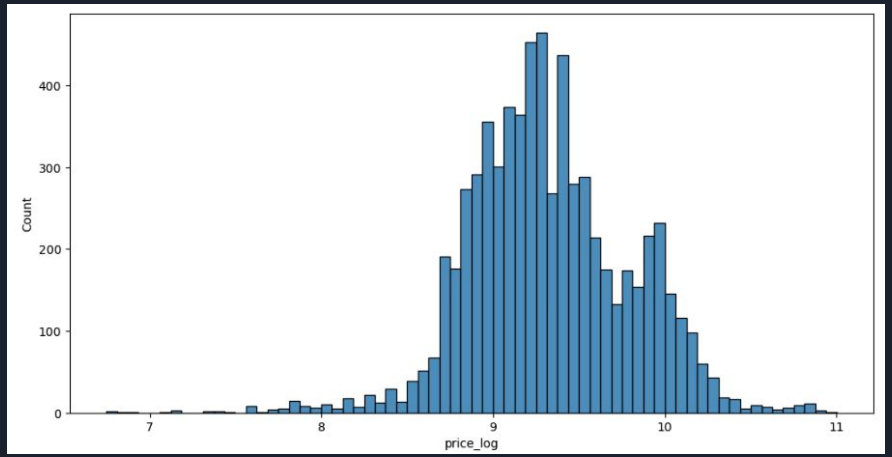
Key Findings IV

The distribution of the price variable has a long tail.



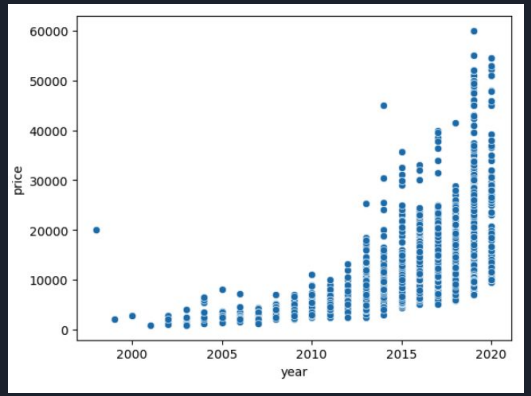
Key Findings IV

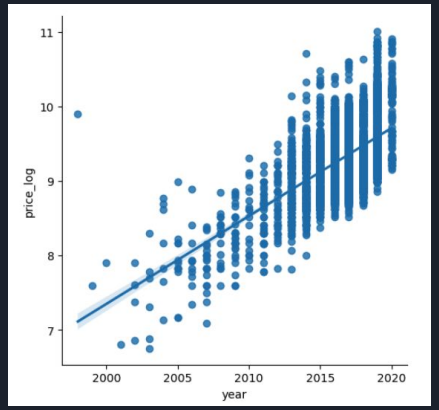
The log transformation of the price normalizes its distribution



Key Findings

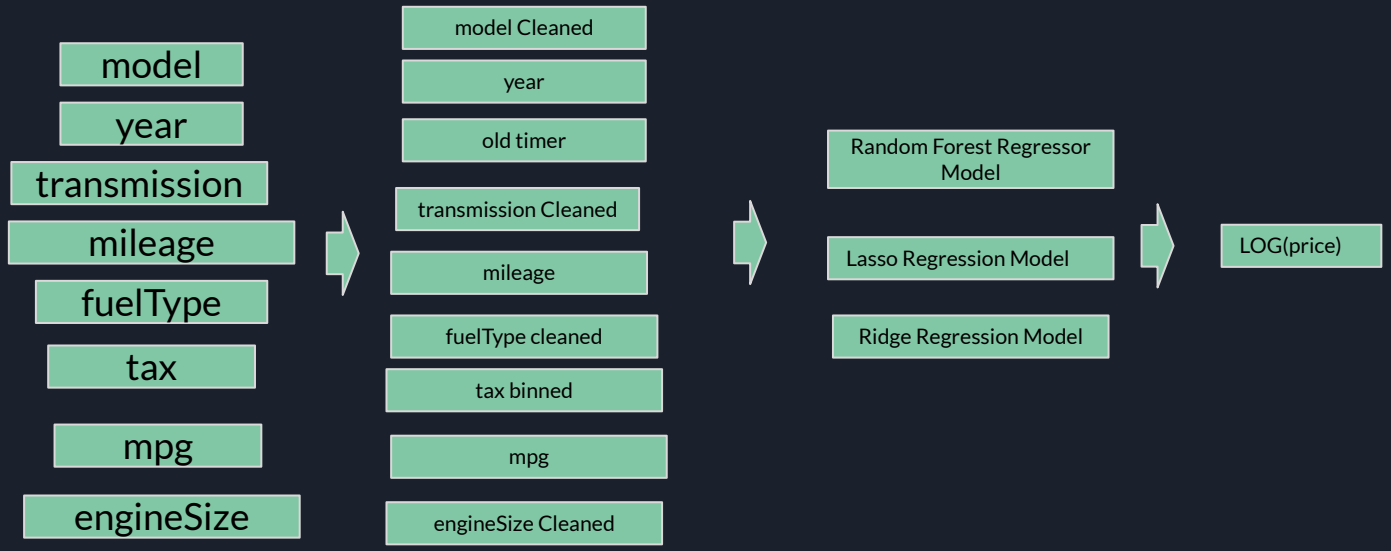
- Example on how the relationship between year and price becomes linear after the log transformation of the price.





Outcomes

We trained 3 models



Outcomes II

- Three metrics: RMSE, R^2, **% of error of the best 80% estimations.**

Metric	**R^2**	**RMSE**	**% of error of the best 80% estimations**
Description	Proportion of the variance in the dependent variable that is predictable from the independent variables	Average magnitude of prediction errors	**% of error of the best 80% estimations**
Range	0-1 (1 is the best)	0-inf. (0 is the best)	0-inf. 0% is the best

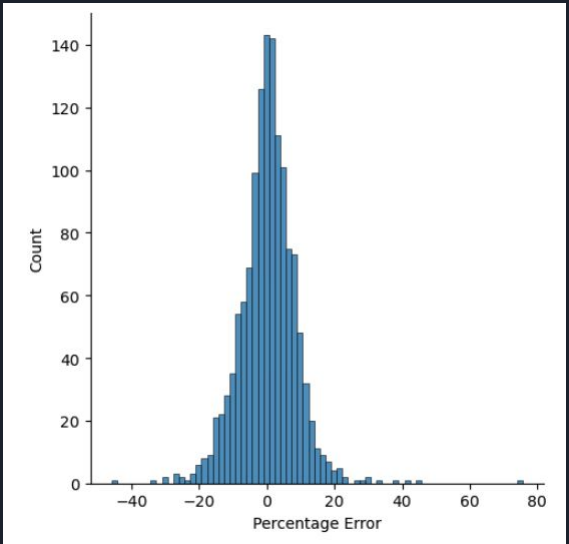
Outcomes III

- Two metrics: accuracy and precision.

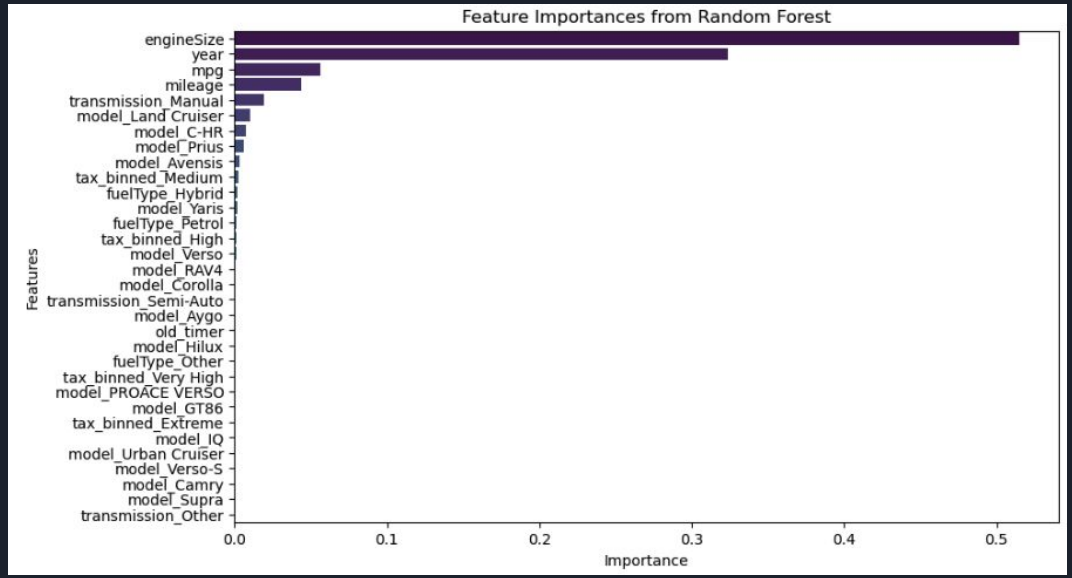
Metric	**R^2**	**RMSE**	**% of error of the best 90% estimations**
Linear Regression	0.96	1265.37	
Ridge Regression	0.96	1261.11	
Random Forest	**0.97**	**1105.25**	80% of the estimations are off by 9.37% or less

Outcomes

- Distribution of the % error compared to the listed price.



Outcomes IV


Feature Importances from Random Forest
engineSize
year
mpg
mileage
transmission_Manual
model_Land Cruiser
model_C-HR
model_Prius
model_Avensis
tax_binned_Medium
fuelType_Hybrid
model_Yaris
fuelType_Petrol
tax_binned_High
model_Verso
model_RAV4
model_Corolla
transmission_Semi-Auto
model_Aygo
old_timer
model_Hilux
fuelType_Other
tax_binned_Very High
model_PROACE VERSO
model_GT86
tax_binned_Extreme
model_IQ
model_Urban Cruiser
model_Verso-S
model_Camry
model_Supra
transmission_Other
Features
0.0
0.1
0.2
0.3
0.4
0.5
Importance

Recommendation

- Validate the model against the experts
- Fix eventual errors
- Deploy the model to start assisting the new members
- Collect more data and retrain the model regularly
- Monitor the model to identify drops in performance

THANKS!!