# Car Resale Predictor-sprint 1

November 17, 2022

## 1 IMPORTING LIBRARIES

[21]:
```
import pandas as pd
import numpy as np
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import cross_val_score, train_test_split,␣
↪StratifiedShuffleSplit , RandomizedSearchCV
import pickle
import datetime
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score
from sklearn.linear_model import LinearRegression
from sklearn.ensemble import RandomForestRegressor
```

## 2 LOAD DATA

[10]:
```
file = '/Users/rpriyadharshini/Desktop/autos.csv'
```

[11]:
```
import chardet
with open(file, 'rb') as rawdata:
    result = chardet.detect(rawdata.read(100000))
result
```

[11]:
```
{'encoding': 'Windows-1252', 'confidence': 0.73, 'language': ''}
```

[12]:
```
df = pd.read_csv(file, encoding = 'Windows-1252')
df.head()
```

[12]:

| | dateCrawled | name | seller | offerType | \ |
|---|---|---|---|---|---|
| 0 | 2016-03-24 11:52:17 | Golf_3_1.6 | privat | Angebot | |
| 1 | 2016-03-24 10:58:45 | A5_Sportback_2.7_Tdi | privat | Angebot | |
| 2 | 2016-03-14 12:52:21 | Jeep_Grand_Cherokee_"Overland" | privat | Angebot | |
| 3 | 2016-03-17 16:54:04 | GOLF_4_1_4__3TÜRER | privat | Angebot | |
| 4 | 2016-03-31 17:25:20 | Skoda_Fabia_1.4_TDI_PD_Classic | privat | Angebot | |

```
   price abtest vehicleType  yearOfRegistration    gearbox  powerPS  model  \
0    480   test         NaN                1993    manuell        0   golf
1  18300   test       coupe                2011    manuell      190    NaN
2   9800   test         suv                2004  automatik      163  grand
3   1500   test  kleinwagen                2001    manuell       75   golf
4   3600   test  kleinwagen                2008    manuell       69  fabia

   kilometer  monthOfRegistration fuelType       brand notRepairedDamage  \
0     150000                    0   benzin  volkswagen               NaN
1     125000                    5   diesel        audi               ja
2     125000                    8   diesel        jeep               NaN
3     150000                    6   benzin  volkswagen              nein
4      90000                    7   diesel       skoda              nein

           dateCreated  nrOfPictures  postalCode             lastSeen
0  2016-03-24 00:00:00             0       70435  2016-04-07 03:16:57
1  2016-03-24 00:00:00             0       66954  2016-04-07 01:46:50
2  2016-03-14 00:00:00             0       90480  2016-04-05 12:47:46
3  2016-03-17 00:00:00             0       91074  2016-03-17 17:40:17
4  2016-03-31 00:00:00             0       60437  2016-04-06 10:17:21
```

# 3 EXPLORATORY DATA ANALYSIS

## 3.1 SHAPE

[13]:
```
df.shape
```

[13]:
```
(371528, 20)
```

[15]:
```
print(df['fuelType'].unique())
print(df['seller'].unique())
print(df['gearbox'].unique())
```

```
['benzin' 'diesel' nan 'lpg' 'andere' 'hybrid' 'cng' 'elektro']
['privat' 'gewerblich']
['manuell' 'automatik' nan]
```

## 3.2 CHECK NULL VALUES

[16]:
```
df.isnull().sum()
```

[16]:
```
dateCrawled            0
name                   0
seller                 0
offerType              0
price                  0
abtest                 0
```

```
vehicleType            37869
yearOfRegistration         0
gearbox                20209
powerPS                    0
model                  20484
kilometer                  0
monthOfRegistration        0
fuelType               33386
brand                      0
notRepairedDamage      72060
dateCreated                0
nrOfPictures               0
postalCode                 0
lastSeen                   0
dtype: int64
```

### 3.3 DESCRIBE THE DATA

[17]:
```
df.describe()
```

[17]:
```
              price  yearOfRegistration        powerPS      kilometer  \
count  3.715280e+05       371528.000000  371528.000000  371528.000000
mean   1.729514e+04         2004.577997     115.549477  125618.688228
std    3.587954e+06           92.866598     192.139578   40112.337051
min    0.000000e+00         1000.000000       0.000000    5000.000000
25%    1.150000e+03         1999.000000      70.000000  125000.000000
50%    2.950000e+03         2003.000000     105.000000  150000.000000
75%    7.200000e+03         2008.000000     150.000000  150000.000000
max    2.147484e+09         9999.000000   20000.000000  150000.000000

       monthOfRegistration  nrOfPictures    postalCode
count        371528.000000      371528.0  371528.00000
mean              5.734445           0.0   50820.66764
std               3.712412           0.0   25799.08247
min               0.000000           0.0    1067.00000
25%               3.000000           0.0   30459.00000
50%               6.000000           0.0   49610.00000
75%               9.000000           0.0   71546.00000
max              12.000000           0.0   99998.00000
```

## 4 PREPROCESSING DATA

[19]:
```
df1 = df.drop(columns='name')
df1.shape
```

[19]:
```
(371528, 19)
```

[24]:
```
df.seller.value_counts()
```

[24]:
```
privat        371525
gewerblich         3
Name: seller, dtype: int64
```

[25]:
```
df=df[df.seller != 'gewerblich']
```

[26]:
```
df=df.drop('seller',1)
```

```
/var/folders/13/mdvvnj0d299_920wwhpnhy5m0000gn/T/ipykernel_93317/1037493778.py:1
: FutureWarning: In a future version of pandas all arguments of DataFrame.drop
except for the argument 'labels' will be keyword-only
  df=df.drop('seller',1)
```

[27]:
```
df.offerType.value_counts()
```

[27]:
```
Angebot    371513
Gesuch         12
Name: offerType, dtype: int64
```

[28]:
```
df=df[df.offerType != 'Gesuch']
```

[29]:
```
df
```

[29]:

| | dateCrawled | name | \ |
|---|---|---|---|
| 0 | 2016-03-24 11:52:17 | Golf_3_1.6 | |
| 1 | 2016-03-24 10:58:45 | A5_Sportback_2.7_Tdi | |
| 2 | 2016-03-14 12:52:21 | Jeep_Grand_Cherokee_"Overland" | |
| 3 | 2016-03-17 16:54:04 | GOLF_4_1_4__3TÜRER | |
| 4 | 2016-03-31 17:25:20 | Skoda_Fabia_1.4_TDI_PD_Classic | |
| … | … | … | |
| 371523 | 2016-03-14 17:48:27 | Suche_t4___vito_ab_6_sitze | |
| 371524 | 2016-03-05 19:56:21 | Smart_smart_leistungssteigerung_100ps | |
| 371525 | 2016-03-19 18:57:12 | Volkswagen_Multivan_T4_TDI_7DC_UY2 | |
| 371526 | 2016-03-20 19:41:08 | VW_Golf_Kombi_1_9l_TDI | |
| 371527 | 2016-03-07 19:39:19 | BMW_M135i_vollausgestattet_NP_52.720____Euro | |

| | offerType | price | abtest | vehicleType | yearOfRegistration | gearbox | \ |
|---|---|---|---|---|---|---|---|
| 0 | Angebot | 480 | test | NaN | 1993 | manuell | |
| 1 | Angebot | 18300 | test | coupe | 2011 | manuell | |
| 2 | Angebot | 9800 | test | suv | 2004 | automatik | |
| 3 | Angebot | 1500 | test | kleinwagen | 2001 | manuell | |
| 4 | Angebot | 3600 | test | kleinwagen | 2008 | manuell | |
| … | … | … | … | … | … | … | |
| 371523 | Angebot | 2200 | test | NaN | 2005 | NaN | |
| 371524 | Angebot | 1199 | test | cabrio | 2000 | automatik | |

```
371525   Angebot   9200     test         bus                1996    manuell
371526   Angebot   3400     test       kombi                2002    manuell
371527   Angebot  28990  control   limousine                2013    manuell

        powerPS        model  kilometer  monthOfRegistration fuelType  \
0             0         golf     150000                    0   benzin
1           190          NaN     125000                    5   diesel
2           163        grand     125000                    8   diesel
3            75         golf     150000                    6   benzin
4            69        fabia      90000                    7   diesel
…           …           …         …                       …       …
371523        0          NaN      20000                    1      NaN
371524      101       fortwo     125000                    3   benzin
371525      102  transporter     150000                    3   diesel
371526      100         golf     150000                    6   diesel
371527      320      m_reihe      50000                    8   benzin

                brand notRepairedDamage          dateCreated  nrOfPictures  \
0           volkswagen               NaN  2016-03-24 00:00:00             0
1                 audi                ja  2016-03-24 00:00:00             0
2                 jeep               NaN  2016-03-14 00:00:00             0
3           volkswagen              nein  2016-03-17 00:00:00             0
4                skoda              nein  2016-03-31 00:00:00             0
…                    …                 …                    …            …
371523  sonstige_autos               NaN  2016-03-14 00:00:00             0
371524           smart              nein  2016-03-05 00:00:00             0
371525      volkswagen              nein  2016-03-19 00:00:00             0
371526      volkswagen               NaN  2016-03-20 00:00:00             0
371527             bmw              nein  2016-03-07 00:00:00             0

        postalCode             lastSeen
0            70435  2016-04-07 03:16:57
1            66954  2016-04-07 01:46:50
2            90480  2016-04-05 12:47:46
3            91074  2016-03-17 17:40:17
4            60437  2016-04-06 10:17:21
…            …                    …
371523       39576  2016-04-06 00:46:52
371524       26135  2016-03-11 18:17:12
371525       87439  2016-04-07 07:15:26
371526       40764  2016-03-24 12:45:21
371527       73326  2016-03-22 03:17:10

[371513 rows x 19 columns]
```

[30]:
```
df=df.drop('offerType',1)
```

```
/var/folders/13/mdvvnj0d299_920wwhpnhy5m0000gn/T/ipykernel_93317/2498620258.py:1
: FutureWarning: In a future version of pandas all arguments of DataFrame.drop
except for the argument 'labels' will be keyword-only
  df=df.drop('offerType',1)
```

[31]:
```
df.shape
```

[31]:
```
(371513, 18)
```

[32]:
```
df = df[(df.powerPS > 50) & (df.powerPS <900)]
```

[33]:
```
df.shape
```

[33]:
```
(319704, 18)
```

[34]:
```
df = df[(df.yearOfRegistration >= 1950) & (df.yearOfRegistration < 2017)]
```

[35]:
```
df.shape
```

[35]:
```
(309166, 18)
```

[36]:
```
df.head()
```

[36]:

| | dateCrawled | name | price | abtest | vehicleType | yearOfRegistration | gearbox | powerPS | model | kilometer | monthOfRegistration | fuelType | brand | notRepairedDamage | dateCreated | nrOfPictures | postalCode | lastSeen |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2016-03-24 10:58:45 | A5_Sportback_2.7_Tdi | 18300 | test | coupe | 2011 | manuell | 190 | NaN | 125000 | 5 | diesel | audi | ja | 2016-03-24 00:00:00 | 0 | 66954 | 2016-04-07 01:46:50 |
| 2 | 2016-03-14 12:52:21 | Jeep_Grand_Cherokee_"Overland" | 9800 | test | suv | 2004 | automatik | 163 | grand | 125000 | 8 | diesel | jeep | NaN | 2016-03-14 00:00:00 | 0 | 90480 | 2016-04-05 12:47:46 |
| 3 | 2016-03-17 16:54:04 | GOLF_4_1_4__3TÜRER | 1500 | test | kleinwagen | 2001 | manuell | 75 | golf | 150000 | 6 | benzin | volkswagen | nein | | | | |
| 4 | 2016-03-31 17:25:20 | Skoda_Fabia_1.4_TDI_PD_Classic | 3600 | test | kleinwagen | 2008 | manuell | 69 | fabia | 90000 | 7 | diesel | skoda | nein | | | | |
| 5 | 2016-04-04 17:36:23 | BMW_316i___e36_Limousine___Bastlerfahrzeug__Ex… | 650 | test | limousine | 1995 | manuell | 102 | 3er | 150000 | 10 | benzin | bmw | ja | | | | |

```
3  2016-03-17 00:00:00            0       91074  2016-03-17 17:40:17
4  2016-03-31 00:00:00            0       60437  2016-04-06 10:17:21
5  2016-04-04 00:00:00            0       33775  2016-04-06 19:17:07
```

[37]:
```
df.columns
```

[37]:
```
Index(['dateCrawled', 'name', 'price', 'abtest', 'vehicleType',
       'yearOfRegistration', 'gearbox', 'powerPS', 'model', 'kilometer',
       'monthOfRegistration', 'fuelType', 'brand', 'notRepairedDamage',
       'dateCreated', 'nrOfPictures', 'postalCode', 'lastSeen'],
      dtype='object')
```

[38]:
```
df.drop(['name', 'abtest', 'dateCrawled','nrOfPictures', 'lastSeen',␣
 ↪'postalCode', 'dateCreated'], axis='columns', inplace=True)
```

[39]:
```
df
```

[39]:
```
         price vehicleType  yearOfRegistration    gearbox  powerPS  \
1        18300       coupe                2011    manuell      190
2         9800         suv                2004  automatik      163
3         1500  kleinwagen                2001    manuell       75
4         3600  kleinwagen                2008    manuell       69
5          650   limousine                1995    manuell      102
…          …          …                   …          …        …
371520    3200   limousine                2004    manuell      225
371524    1199      cabrio                2000  automatik      101
371525    9200         bus                1996    manuell      102
371526    3400       kombi                2002    manuell      100
371527   28990   limousine                2013    manuell      320

               model  kilometer  monthOfRegistration fuelType       brand  \
1                NaN     125000                    5   diesel        audi
2              grand     125000                    8   diesel        jeep
3               golf     150000                    6   benzin  volkswagen
4              fabia      90000                    7   diesel       skoda
5                3er     150000                   10   benzin         bmw
…                …          …                    …        …          …
371520          leon     150000                    5   benzin        seat
371524        fortwo     125000                    3   benzin       smart
371525  transporter     150000                    3   diesel  volkswagen
371526          golf     150000                    6   diesel  volkswagen
371527      m_reihe      50000                    8   benzin         bmw

        notRepairedDamage
1                      ja
2                     NaN
3                    nein
```

```
4                    nein
5                      ja
…                     …
371520                 ja
371524               nein
371525               nein
371526                NaN
371527               nein

[309166 rows x 11 columns]
```

[40]:
```
new_df = df.copy()
```

[41]:
```
new_df.columns
```

[41]:
```
Index(['price', 'vehicleType', 'yearOfRegistration', 'gearbox', 'powerPS',
       'model', 'kilometer', 'monthOfRegistration', 'fuelType', 'brand',
       'notRepairedDamage'],
      dtype='object')
```

[42]:
```
new_df.head()
```

[42]:

| | price | vehicleType | yearOfRegistration | gearbox | powerPS | model | kilometer | monthOfRegistration | fuelType | brand | notRepairedDamage |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 18300 | coupe | 2011 | manuell | 190 | NaN | 125000 | 5 | diesel | audi | ja |
| 2 | 9800 | suv | 2004 | automatik | 163 | grand | 125000 | 8 | diesel | jeep | NaN |
| 3 | 1500 | kleinwagen | 2001 | manuell | 75 | golf | 150000 | 6 | benzin | volkswagen | nein |
| 4 | 3600 | kleinwagen | 2008 | manuell | 69 | fabia | 90000 | 7 | diesel | skoda | nein |
| 5 | 650 | limousine | 1995 | manuell | 102 | 3er | 150000 | 10 | benzin | bmw | ja |

[43]:
```
new_df = new_df.drop_duplicates(['price', 'vehicleType',␣
↪'yearOfRegistration','gearbox', 'powerPS', 'model', 'kilometer',␣
↪'monthOfRegistration', 'fuelType' , 'notRepairedDamage'])
```

[44]:
```
new_df.shape
```

[44]:
```
(285140, 11)
```

[45]:
```
new_df.gearbox.replace(('manuell', 'automatik'),␣
↪('manual','automatic'),inplace=True)
```

```
new_df.fuelType.replace(('benzin', 'andere', 'elektro'), ('petrol', 'others',␣
↪'electric'),inplace=True)
new_df.vehicleType.replace(('kleinwagen', 'cabrio', 'kombi', 'andere'),('small␣
↪car','convertible', 'combination', 'others'), inplace=True)
new_df.notRepairedDamage.replace(('ja', 'nein'), ('Yes', 'No'), inplace =True)
```

[46]:
```
new_df.head()
```

[46]:

| | price | vehicleType | yearOfRegistration | gearbox | powerPS | model | \ |
|---|---|---|---|---|---|---|---|
| 1 | 18300 | coupe | 2011 | manual | 190 | NaN | |
| 2 | 9800 | suv | 2004 | automatic | 163 | grand | |
| 3 | 1500 | small car | 2001 | manual | 75 | golf | |
| 4 | 3600 | small car | 2008 | manual | 69 | fabia | |
| 5 | 650 | limousine | 1995 | manual | 102 | 3er | |

| | kilometer | monthOfRegistration | fuelType | brand | notRepairedDamage |
|---|---|---|---|---|---|
| 1 | 125000 | 5 | diesel | audi | Yes |
| 2 | 125000 | 8 | diesel | jeep | NaN |
| 3 | 150000 | 6 | petrol | volkswagen | No |
| 4 | 90000 | 7 | diesel | skoda | No |
| 5 | 150000 | 10 | petrol | bmw | Yes |

[47]:
```
new_df = new_df[(new_df.price >= 100) & (new_df.price <= 150000)]
new_df['notRepairedDamage'].fillna(value='not-declared', inplace=True)
new_df['fuelType'].fillna(value='not-declared', inplace=True)
new_df['gearbox'].fillna(value='not-declared', inplace=True)
new_df['vehicleType'].fillna(value='not-declared', inplace=True)
new_df['model'].fillna(value='not-declared', inplace=True)
```

[48]:
```
new_df.head()
```

[48]:

| | price | vehicleType | yearOfRegistration | gearbox | powerPS | model | \ |
|---|---|---|---|---|---|---|---|
| 1 | 18300 | coupe | 2011 | manual | 190 | not-declared | |
| 2 | 9800 | suv | 2004 | automatic | 163 | grand | |
| 3 | 1500 | small car | 2001 | manual | 75 | golf | |
| 4 | 3600 | small car | 2008 | manual | 69 | fabia | |
| 5 | 650 | limousine | 1995 | manual | 102 | 3er | |

| | kilometer | monthOfRegistration | fuelType | brand | notRepairedDamage |
|---|---|---|---|---|---|
| 1 | 125000 | 5 | diesel | audi | Yes |
| 2 | 125000 | 8 | diesel | jeep | not-declared |
| 3 | 150000 | 6 | petrol | volkswagen | No |
| 4 | 90000 | 7 | diesel | skoda | No |
| 5 | 150000 | 10 | petrol | bmw | Yes |

[51]:
```
new_df.to_csv("/Users/rpriyadharshini/Desktop/autos_p.csv")
```

[52]:
```
labels = ['gearbox', 'notRepairedDamage', 'model', 'brand', 'fuelType',␣
↪'vehicleType']

mapper = {}
for i in labels:
    mapper[i] = LabelEncoder()
    mapper[i].fit(new_df[i])
    tr = mapper[i].transform(new_df[i])
    np.save(str('classes'+i+'.npy'), mapper[i].classes_)
    new_df.loc[:, i+'_labels'] = pd.Series(tr, index=new_df.index)

labeled = new_df[['price',␣
↪'yearOfRegistration','powerPS','kilometer','monthOfRegistration']
                 +[x+"_labels" for x in labels]]

print(labeled.columns)
```

```
Index(['price', 'yearOfRegistration', 'powerPS', 'kilometer',
       'monthOfRegistration', 'gearbox_labels', 'notRepairedDamage_labels',
       'model_labels', 'brand_labels', 'fuelType_labels',
       'vehicleType_labels'],
      dtype='object')
```

[ ]: