Car Resale Value Prediction

PRE PROCESS THE DATA

Cleaning the Dataset

print(df.seller.value_counts())

```
privat        371525
gewerblich         3
Name: seller, dtype: int64
```

df[df.seller != 'gewerblich']

print(df.offerType.value_counts())

```
Angebot    371516
Gesuch         12
Name: offerType, dtype: int64
```

df[df.offerType != 'Gesuch']

df=df.drop('offerType',axis=1)

print(df.shape)

```
(371528, 18)
```

df=df[(df.powerPS > 50) & (df.powerPS < 900)]
print(df.shape)

```
(309171, 18)
```

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

```
new_df = df.copy()

new_df = new_df.drop_duplicates ([ 'price', 'vehicleType', 'yearOf
Registration',
                                  ,'gearbox', 'powerPS', 'model', 'kilometer',
monthOfRegistration', 'fuelType'
                                  ,'notRepairedDamage'])

new_df.gearbox.replace(('manuell', 'automatik'), ('manual', 'autom
atic'), inplace=True)

new_df.gearbox.replace(('manuell', 'automatik'), ('manual', 'autom
atic'), inplace=True)

new_df.vehicleType.replace(('kleinwagen', 'cabrio', 'kombi', 'ander
e'), ('small car', 'convertible', 'combination', 'others'), inplace=Tru
e)


new_df.notRepairedDamage.replace(('ja', 'nein'), ('Yes',
'No'),inplace=True)

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)

new_df.to_csv("autos_preprocessed.csv")


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_)
```

```
    print(i, ":",mapper[i])
    new_df.loc[:, i + '_labels'] = pd.Series (tr,
index=new_df.index)
```

```
gearbox : LabelEncoder()
notRepairedDamage : LabelEncoder()
model : LabelEncoder()
brand : LabelEncoder()
fuelType : LabelEncoder()
vehicleType : LabelEncoder()
```

```
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')
```