PROJECT DEVELOPMENT PHASE

DELIVERY OF SPRINT-1

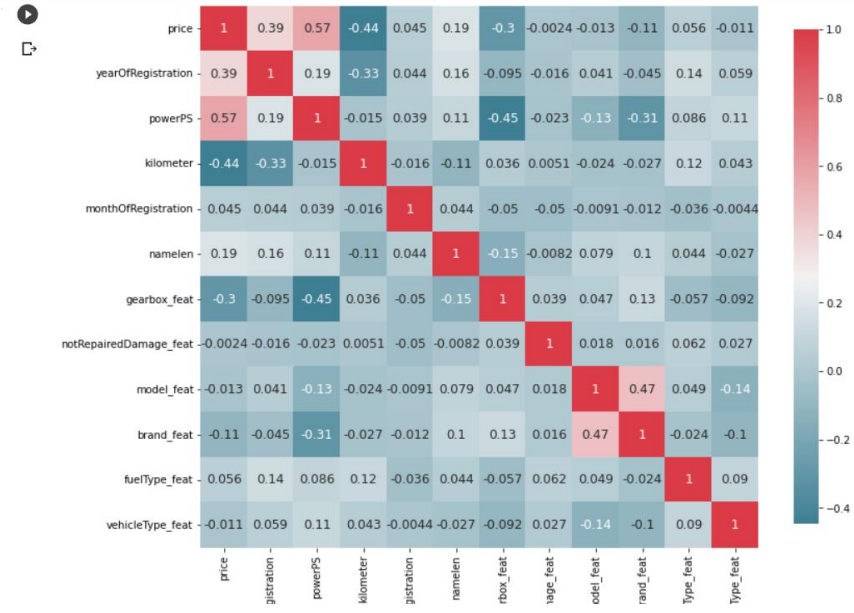
DATE	29 OCTOBER 2022
TEAM ID	PNT2022TMID15486
PROJECT NAME	Car Resale Value Prediction
MAXIMUM MARK	4 Marks

- **Import library and load dataset in python**

```
df = pd.read_csv('/content/drive/MyDrive/Imarticus/autos.csv', sep=',', header=0, encoding='cp1252')
#df = pd.read_csv('autos.csv.gz', sep=',', header=0, compression='gzip',encoding='cp1252')
df.sample(10)
```

	dateCrawled	name	seller	offerType	price	abtest	vehicleType	yearOfRegistration	gearbox	powerPS	model	kilometer	monthOfRegi
35533	2016-04-01 16:52:24	Peugeot_206_5Tuerer_Klima_El_Fenster_2_Hand_8f...	privat	Angebot	999	control	kleinwagen	1999	manuell	75	2_reihe	150000	
104233	2016-03-26 20:58:26	Citroën_C4_Picasso_2.0_HDi_FAP_EGS6_Exclusive	privat	Angebot	9500	control	bus	2008	automatik	136	c4	125000	
81172	2016-04-01 22:53:21	Volkswagen_Passat_Variant_1.9_TDI_DPF_Comfortline	privat	Angebot	6666	test	kombi	2009	manuell	105	passat	150000	
362697	2016-03-09 14:37:44	BMW_E36_Limo	privat	Angebot	2900	test	NaN	2017	NaN	0	andere	150000	
147593	2016-03-21 08:54:07	Ford_Mondeo_an_Bastler	privat	Angebot	250	control	kombi	1999	manuell	0	mondeo	150000	
254916	2016-03-26 12:45:47	Golf_VII_2.0TDI_DSG_Cup	privat	Angebot	22500	control	limousine	2014	automatik	150	golf	40000	
264392	2016-03-27 16:59:13	Peugeot_307_Premium_4Tuerig_Diesel	privat	Angebot	2790	test	NaN	2017	manuell	109	3_reihe	150000	
	2016-03-19												

- **Understanding and Analyzing the dataset by Correlation.**



- **Effective model building**

```
[44] Y = labeled['price']
     X = labeled.drop(['price'], axis='columns', inplace=False)
```

```
[45] from sklearn.model_selection import cross_val_score, train_test_split

     #Split into train and validation
     X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.33, random_state = 3)
     print(X_train.shape, X_test.shape, y_train.shape, y_test.shape)
```

```
     (203769, 7) (100364, 7) (203769,) (100364,)
```

```
[46] from sklearn.ensemble import HistGradientBoostingRegressor
     from sklearn.model_selection import GridSearchCV

     hr = HistGradientBoostingRegressor()

     param_grid = { "loss" : ['squared_error']
                  , "max_leaf_nodes" : [31]
                  , "min_samples_leaf": [20]
                  , "max_depth": [None]
                  , "max_iter":[500]}

     gs = GridSearchCV(estimator=hr, param_grid=param_grid, cv=2, n_jobs=-1, verbose=1)
     gs = gs.fit(X_train, y_train)
     print('Score: %.2f' % gs.score(X_test, y_test))
```

```
     Fitting 2 folds for each of 1 candidates, totalling 2 fits
     Score: 0.78
```

- **Dumping to Pickle file**

```
print(gs.best_score_)
print(gs.best_params_)
```

```
0.7707290246596359
{'loss': 'squared_error', 'max_depth': None, 'max_iter': 500, 'max_leaf_nodes': 31, 'min_samples_leaf': 20}
```

```
[48] import pickle

     pickle.dump(gs,open('histmodel.pkl','wb'))
```