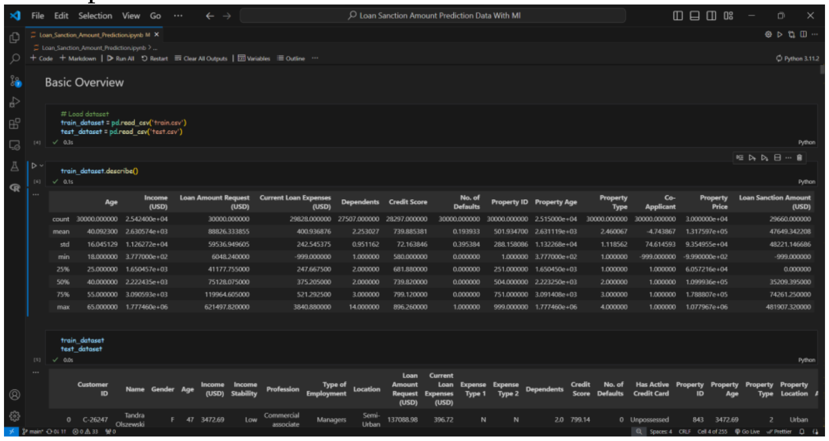
Data Collection and Preprocessing Phase

Date	15 June 2024
Team ID	740185
Project Title	Loan Sanction Amount Prediction Data With Ml
Maximum Marks	6 Marks

Data Exploration and Preprocessing Template

Data Exploration and Preprocessing Template

Dataset variables will be statistically analyzed to identify patterns and outliers, with Python employed for preprocessing tasks like normalization and feature engineering. Data cleaning will address missing values and outliers, ensuring quality for subsequent analysis and modeling, and forming a strong foundation for insights and predictions.

Section	**Description**
Data Overview	Dimension: 8 rows x 131 columns Descriptive stastistics: 

Basic Overview

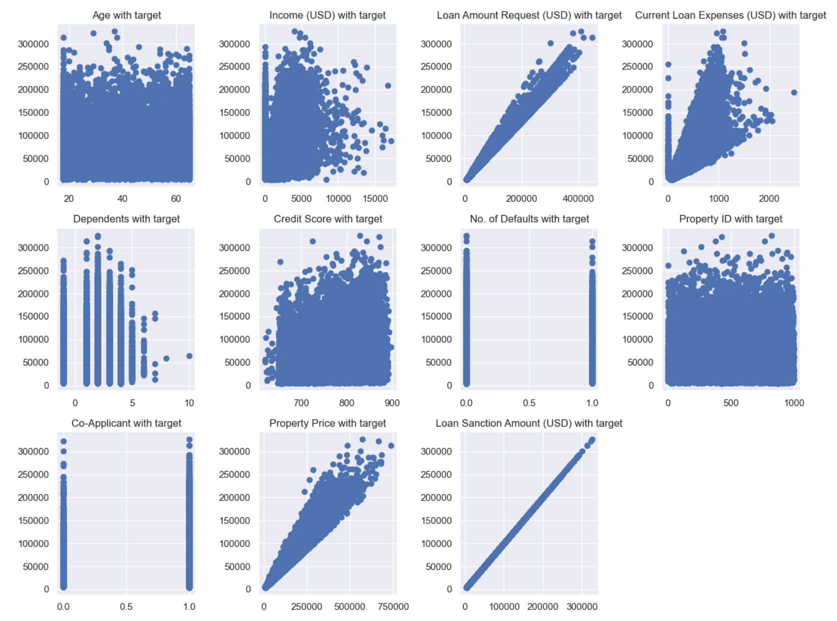
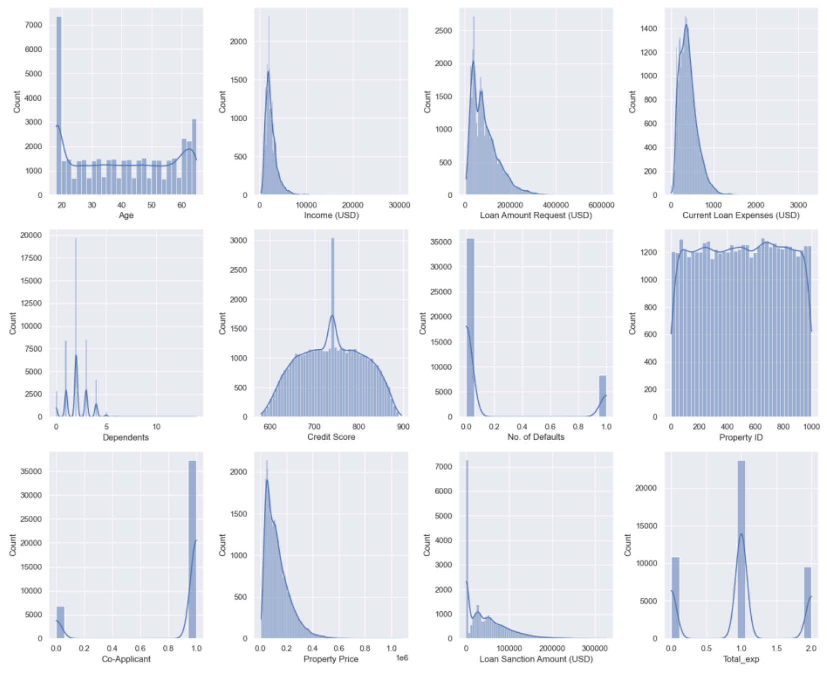
```python
# Load dataset
train_dataset = pd.read_csv('train.csv')
test_dataset = pd.read_csv('test.csv')
```

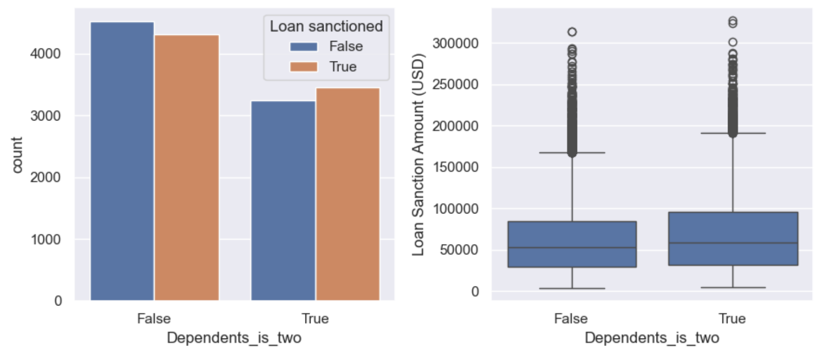
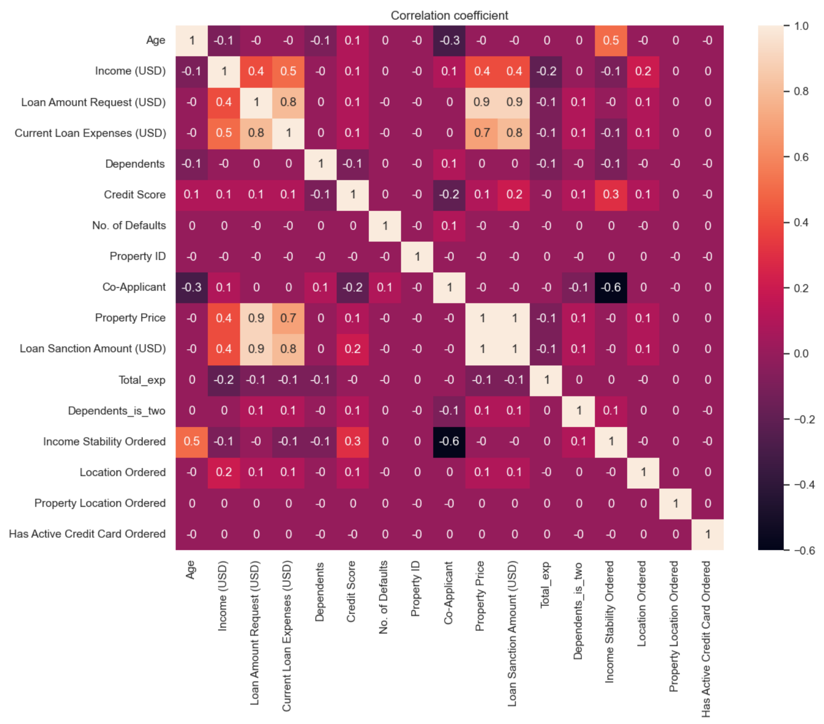
```python
train_dataset.describe()
```

```
             Age        Income  Loan Amount Request  Current Loan Expenses    Dependents   Credit Score   No. of       Property ID  Property Age    Property      Co-           Property      Loan Sanction Amount
                        (USD)   (USD)                (USD)                                                Defaults                                Type          Applicant     Price         (USD)
count  30000.000000  2.542400e+04  30000.000000     29828.000000          27507.000000  28297.000000  30000.000000  30000.000000  2.515000e+04  30000.000000  30000.000000  3.000000e+04  29660.000000
mean      40.092300  2.630574e+03  88826.333855       400.936876              2.253027    739.885381      0.193933    501.934700  2.631119e+03      2.460067     -4.743867  1.317597e+05  47649.342208
std       16.045129  1.126272e+04  59536.949605       242.545375              0.951162     72.163846      0.395384    288.158086  1.132268e+04      1.118562     74.614593  9.354955e+04  48221.146686
min       18.000000  3.777000e+02   6048.240000      -999.000000              1.000000    580.000000      0.000000      1.000000  3.777000e+02      1.000000   -999.000000 -9.990000e+02   -999.000000
25%       25.000000  1.650457e+03  41177.755000       247.667500              2.000000    681.880000      0.000000    251.000000  1.650450e+03      1.000000      1.000000  6.057216e+04      0.000000
50%       40.000000  2.222435e+03  75128.075000       375.205000              2.000000    739.820000      0.000000    504.000000  2.223250e+03      2.000000      1.000000  1.099936e+05  35209.395000
75%       55.000000  3.090593e+03  119964.605000      521.292500              3.000000    799.120000      0.000000    751.000000  3.091408e+03      3.000000      1.000000  1.788807e+05  74261.250000
max       65.000000  1.777460e+06  621497.820000     3840.880000             14.000000    896.260000      1.000000    999.000000  1.777460e+06      4.000000      1.000000  1.077967e+06  481907.320000
```

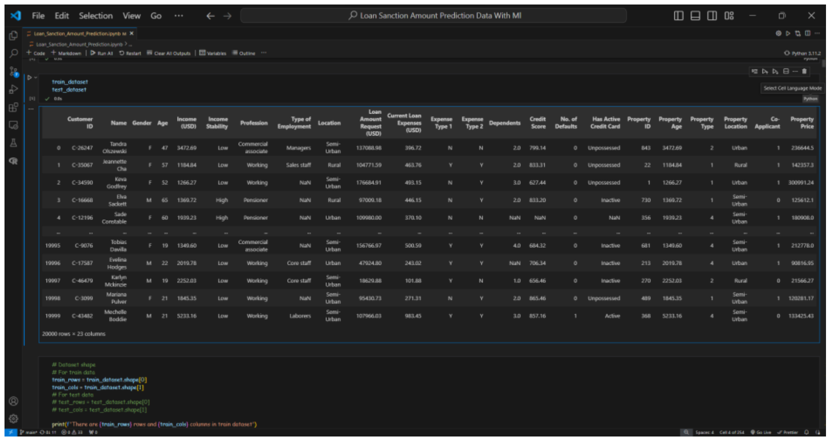
```python
train_dataset
test_dataset
```

```
   Customer  Name       Gender  Age  Income   Income     Profession  Type of     Location  Loan       Current   Expense  Expense  Dependents  Credit  No. of    Has Active   Property  Property  Property  Property
   ID                                (USD)    Stability              Employment            Amount     Loan      Type 1   Type 2              Score   Defaults  Credit Card  ID        Age       Type      Location
                                                                                           Request    Expenses
                                                                                           (USD)      (USD)
0  C-26247   Tandra     F       47   3472.69  Low        Commercial  Managers  Semi-     137088.98  396.72    N        N        2.0         799.14  0         Unpossessed  843       3472.69   2         Urban
             Olszewski                                   associate             Urban
```

Univariate Analysis	
	

Bivariate Analysis	
Multivariate Analysis	
Outliers and Anomalies	-
Data Preprocessing Code Screenshots	

Loading train Data



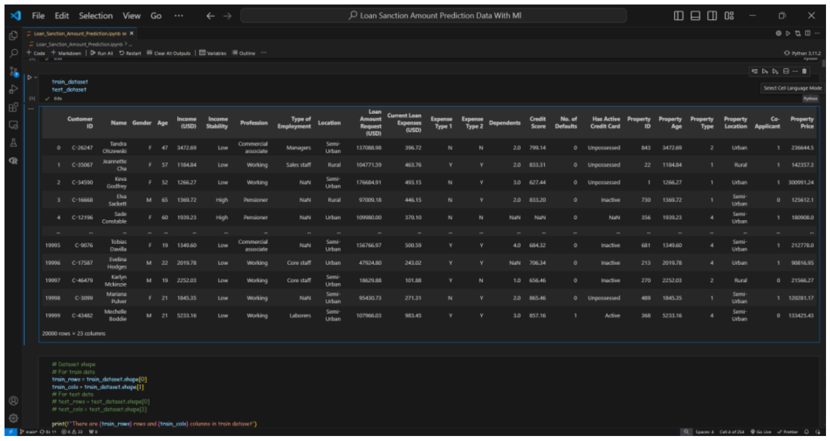
```
train_dataset
test_dataset
```

	Customer ID	Name	Gender	Age	Income (USD)	Income Stability	Profession	Type of Employment	Location	Loan Amount Request (USD)	Current Loan Expenses (USD)	Expense Type 1	Expense Type 2	Dependents	Credit Score	No. of Defaults	Has Active Credit Card	Property ID	Property Age	Property Type	Property Location	Co-Applicant	Property Price
0	C-26247	Tandra Olszewski	F	47	3472.69	Low	Commercial associate	Managers	Semi-Urban	137088.98	396.72	N	N	2.0	799.14	0	Unpossessed	843	3472.69	2	Urban	1	236644.5
1	C-35067	Jeannette Cha	F	57	1184.84	Low	Working	Sales staff	Rural	104771.59	463.76	Y	Y	2.0	833.31	0	Unpossessed	22	1184.84	1	Rural	1	142357.3
2	C-34590	Keva Godfrey	F	52	1266.27	Low	Working	NaN	Semi-Urban	176684.91	493.15	N	Y	3.0	627.44	0	Unpossessed	1	1266.27	1	Urban	1	300991.24
3	C-16668	Elva Sackett	M	65	1369.72	High	Pensioner	NaN	Rural	97009.18	446.15	N	Y	2.0	833.20	0	Inactive	730	1369.72	1	Semi-Urban	0	125612.1
4	C-12196	Sade Constable	F	60	1939.23	High	Pensioner	NaN	Urban	109980.00	370.10	N	N	NaN	NaN	0	NaN	356	1939.23	4	Semi-Urban	1	180908.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
19995	C-9076	Tobias Davilla	F	19	1349.60	Low	Commercial associate	NaN	Semi-Urban	156766.97	500.59	Y	Y	4.0	684.32	0	Inactive	681	1349.60	4	Semi-Urban	1	212778.0
19996	C-17587	Evelina Hodges	M	22	2019.78	Low	Working	Core staff	Urban	47924.80	243.02	Y	Y	NaN	706.34	0	Inactive	213	2019.78	4	Urban	1	90816.95
19997	C-46479	Karlyn Mckinzie	M	19	2252.03	Low	Working	Core staff	Semi-Urban	18629.88	101.88	Y	N	1.0	656.46	0	Inactive	270	2252.03	2	Rural	0	21566.27
19998	C-3099	Mariana Pulver	F	21	1845.35	Low	Working	NaN	Semi-Urban	95430.73	271.31	N	Y	2.0	865.46	0	Unpossessed	489	1845.35	1	Semi-Urban	1	120281.17
19999	C-43482	Mechelle Boddie	M	21	5233.16	Low	Working	Laborers	Semi-Urban	107966.03	983.45	Y	Y	3.0	857.16	1	Active	368	5233.16	4	Semi-Urban	0	133425.43

20000 rows × 23 columns

```
# Dataset shape
# For train data
train_rows = train_dataset.shape[0]
train_cols = train_dataset.shape[1]
# For test data
# test_rows = test_dataset.shape[0]
# test_cols = test_dataset.shape[1]

print(f"There are {train_rows} rows and {train_cols} columns in train dataset")
```

Loading test Data



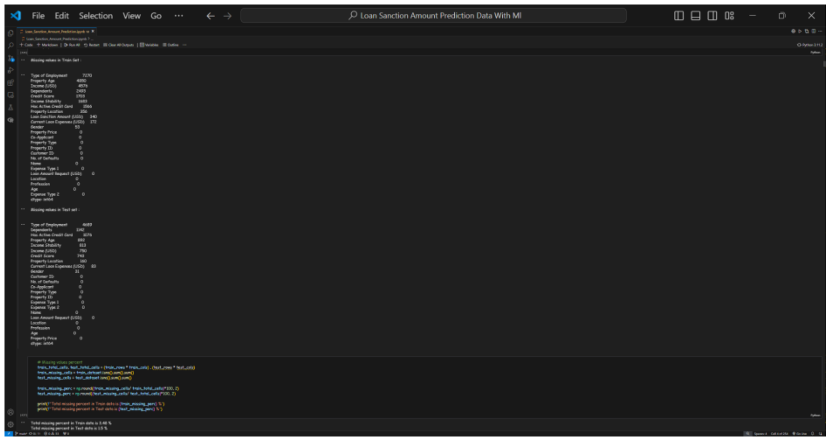
```
train_dataset
test_dataset
```

	Customer ID	Name	Gender	Age	Income (USD)	Income Stability	Profession	Type of Employment	Location	Loan Amount Request (USD)	Current Loan Expenses (USD)	Expense Type 1	Expense Type 2	Dependents	Credit Score	No. of Defaults	Has Active Credit Card	Property ID	Property Age	Property Type	Property Location	Co-Applicant	Property Price
0	C-26247	Tandra Olszewski	F	47	3472.69	Low	Commercial associate	Managers	Semi-Urban	137088.98	396.72	N	N	2.0	799.14	0	Unpossessed	843	3472.69	2	Urban	1	236644.5
1	C-35067	Jeannette Cha	F	57	1184.84	Low	Working	Sales staff	Rural	104771.59	463.76	Y	Y	2.0	833.31	0	Unpossessed	22	1184.84	1	Rural	1	142357.3
2	C-34590	Keva Godfrey	F	52	1266.27	Low	Working	NaN	Semi-Urban	176684.91	493.15	N	Y	3.0	627.44	0	Unpossessed	1	1266.27	1	Urban	1	300991.24
3	C-16668	Elva Sackett	M	65	1369.72	High	Pensioner	NaN	Rural	97009.18	446.15	N	Y	2.0	833.20	0	Inactive	730	1369.72	1	Semi-Urban	0	125612.1
4	C-12196	Sade Constable	F	60	1939.23	High	Pensioner	NaN	Urban	109980.00	370.10	N	N	NaN	NaN	0	NaN	356	1939.23	4	Semi-Urban	1	180908.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
19995	C-9076	Tobias Davilla	F	19	1349.60	Low	Commercial associate	NaN	Semi-Urban	156766.97	500.59	Y	Y	4.0	684.32	0	Inactive	681	1349.60	4	Semi-Urban	1	212778.0
19996	C-17587	Evelina Hodges	M	22	2019.78	Low	Working	Core staff	Urban	47924.80	243.02	Y	Y	NaN	706.34	0	Inactive	213	2019.78	4	Urban	1	90816.95
19997	C-46479	Karlyn Mckinzie	M	19	2252.03	Low	Working	Core staff	Semi-Urban	18629.88	101.88	Y	N	1.0	656.46	0	Inactive	270	2252.03	2	Rural	0	21566.27
19998	C-3099	Mariana Pulver	F	21	1845.35	Low	Working	NaN	Semi-Urban	95430.73	271.31	N	Y	2.0	865.46	0	Unpossessed	489	1845.35	1	Semi-Urban	1	120281.17
19999	C-43482	Mechelle Boddie	M	21	5233.16	Low	Working	Laborers	Semi-Urban	107966.03	983.45	Y	Y	3.0	857.16	1	Active	368	5233.16	4	Semi-Urban	0	133425.43

20000 rows × 23 columns

```
# Dataset shape
# For train data
train_rows = train_dataset.shape[0]
train_cols = train_dataset.shape[1]
# For test data
# test_rows = test_dataset.shape[0]
# test_cols = test_dataset.shape[1]

print(f"There are {train_rows} rows and {train_cols} columns in train dataset")
```

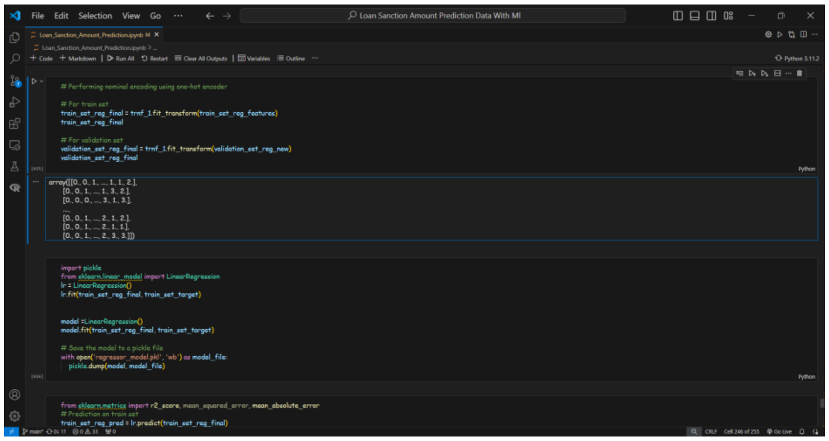
Handling Missing Data
In train and test



```
Missing values in Train Set :

Type of Employment             7270
Property Age                   4850
Income (USD)                   4576
Dependents                     2493
Credit Score                   1703
Income Stability               1683
Has Active Credit Card         1566
Property Location               356
Loan Sanction Amount (USD)      340
Current Loan Expenses (USD)     172
Gender                           53
Property Price                    0
Co-Applicant                      0
Property Type                     0
Property ID                       0
Customer ID                       0
No. of Defaults                   0
Name                              0
Expense Type 1                    0
Loan Amount Request (USD)         0
Location                          0
Profession                        0
Age                               0
Expense Type 2                    0
dtype: int64

Missing values in Test set :

Type of Employment             4689
Dependents                     1142
Has Active Credit Card         1076
Property Age                    882
Income Stability                813
Income (USD)                    750
Credit Score                    743
Property Location               160
Current Loan Expenses (USD)      83
Gender                           31
Customer ID                       0
No. of Defaults                   0
Co-Applicant                      0
Property Type                     0
Property ID                       0
Expense Type 1                    0
Expense Type 2                    0
Name                              0
Loan Amount Request (USD)         0
Location                          0
Profession                        0
Age                               0
Property Price                    0
dtype: int64
```

```
# Missing values percent
train_total_cells, test_total_cells = (train_rows * train_cols) , (test_rows * test_cols)
train_missing_cells = train_dataset.isna().sum().sum()
test_missing_cells = test_dataset.isna().sum().sum()

train_missing_perc = np.round((train_missing_cells/ train_total_cells)*100, 2)
test_missing_perc = np.round((test_missing_cells/ test_total_cells)*100, 2)

print(f"Total missing percent in Train data is {train_missing_perc} %")
print(f"Total missing percent in Test data is {test_missing_perc} %")
```

```
Total missing percent in Train data is 3.48 %
Total missing percent in Test data is 1.5 %
```

Data Transformation	
Feature Engineering	-
Save Processed Data	-

```
# Performing nominal encoding using one-hot encoder

# For train set
train_set_reg_final = trnf_1.fit_transform(train_set_reg_features)
train_set_reg_final

# For validation set
validation_set_reg_final = trnf_1.fit_transform(validation_set_reg_new)
validation_set_reg_final
```

```
array([[0., 0., 1., ..., 1., 1., 2.],
       [0., 0., 1., ..., 1., 3., 2.],
       [0., 0., 0., ..., 3., 1., 3.],
       ...,
       [0., 0., 1., ..., 2., 1., 2.],
       [0., 0., 1., ..., 2., 1., 1.],
       [0., 0., 1., ..., 2., 3., 3.]])
```

```
import pickle
from sklearn.linear_model import LinearRegression
lr = LinearRegression()
lr.fit(train_set_reg_final, train_set_target)


model =LinearRegression()
model.fit(train_set_reg_final, train_set_target)

# Save the model to a pickle file
with open('regressor_model.pkl', 'wb') as model_file:
    pickle.dump(model, model_file)
```

```
from sklearn.metrics import r2_score, mean_squared_error, mean_absolute_error
# Prediction on train set
train_set_reg_pred = lr.predict(train_set_reg_final)
```