Data Analysis and Interpretation Capstone

Report on Net Migration causing factors (population, industry, urbanization) in year 2012

Introduction

This report is examining factors that cause human migration in year 2012. We will examine and check any relationships between predictor variables (FDIs,Health,Facilities,Populations) in regards to target variable which is Net Migration.

The dataset is provided by World Bank.

Data Dictionary

Field	Description
x131_2012	FOREIGN DIRECT INVESTMENT, NET INFLOWS (% OF GDP)
x142_2012	GDP PER CAPITA (CURRENT US$)
x150_2012	HEALTH EXPENDITURE, TOTAL (% OF GDP)
x155_2012	IMPROVED SANITATION FACILITIES (% OF POPULATION WITH ACCESS)
x156_2012	IMPROVED WATER SOURCE (% OF POPULATION WITH ACCESS)
x1_2012	ACCESS TO ELECTRICITY (% OF POPULATION)
x258_2012	RURAL POPULATION (% OF TOTAL POPULATION)
x283_2012	URBAN POPULATION (% OF TOTAL)
x31_2012	AGRICULTURAL LAND (% OF LAND AREA)
x195_2012	NET MIGRATION

Data Preparation

In [1]:
```
import pandas as pd
```

In [2]:
```
df = pd.read_csv("worldbank.csv")
df.head()
```

Out[2]:

	country	x1_2012	x2_2012	x9_2012	x11_2012	x12_2012	x14_2012	x15_2012	x16_2012	x18_2012	...	x244_2013
0	AFGHANISTAN	43.00000	19.480962	1.852684e+10	623.236804	0.677345	8.965464	1.833022e+09	1.600000	0.125552	...	0.0
1	ALBANIA	100.00000	62.086412	1.037147e+10	3575.766967	0.386039	12.381013	1.521033e+09	2.842804	2.730370	...	13.1
2	ALGERIA	100.00000	99.990000	1.480000e+11	3942.202841	0.620760	7.019903	1.417333e+10	4.467196	19.550386	...	2.4
3	AMERICAN SAMOA	59.32891	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN
4	ANDORRA	100.00000	100.000000	NaN	NaN	0.152983	9.237132	2.919443e+08	3.100000	0.000000	...	NaN

5 rows × 163 columns

In [3]:
```
df2 = df[["x131_2012","x142_2012","x150_2012","x155_2012","x156_2012","x1_2012","x258_2012","x283_2012","x31_20
```

In [4]:
```
df2.head()
```

Out[4]:

	x131_2012	x142_2012	x150_2012	x155_2012	x156_2012	x1_2012	x258_2012	x283_2012	x31_2012	x195_2012
0	0.299592	690.842629	8.479199	30.5	51.6	43.00000	74.532	25.468	58.067580	473007.0
1	7.468318	4247.485437	5.627439	92.1	95.4	100.00000	45.670	54.330	43.843066	-91750.0
2	0.717733	5583.616160	6.007278	87.0	84.9	100.00000	31.130	68.870	17.381700	-143268.0
3	NaN	NaN	NaN	62.5	100.0	59.32891	12.587	87.413	24.500000	NaN
4	NaN	39666.369210	8.302019	100.0	100.0	100.00000	13.292	86.708	42.978723	NaN

In [5]:
```
df2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 248 entries, 0 to 247
Data columns (total 10 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   x131_2012  219 non-null    float64
 1   x142_2012  225 non-null    float64
 2   x150_2012  221 non-null    float64
 3   x155_2012  226 non-null    float64
 4   x156_2012  227 non-null    float64
 5   x1_2012    245 non-null    float64
 6   x258_2012  245 non-null    float64
 7   x283_2012  245 non-null    float64
 8   x31_2012   240 non-null    float64
 9   x195_2012  227 non-null    float64
dtypes: float64(10)
memory usage: 19.5 KB
```

In [6]:
```
df2.columns = ["fdi","gdp","health","sanitation","water","electricity","ruralpop","urbanpop","agri","migration"
```

In [7]:
```
df2
```

Out[7]:

	fdi	gdp	health	sanitation	water	electricity	ruralpop	urbanpop	agri	migration
0	0.299592	690.842629	8.479199	30.50000	51.600000	43.000000	74.532000	25.468000	58.067580	473007.0
1	7.468318	4247.485437	5.627439	92.10000	95.400000	100.000000	45.670000	54.330000	43.843066	-91750.0
2	0.717733	5583.616160	6.007278	87.00000	84.900000	100.000000	31.130000	68.870000	17.381700	-143268.0
3	NaN	NaN	NaN	62.50000	100.000000	59.328910	12.587000	87.413000	24.500000	NaN
4	NaN	39666.369210	8.302019	100.00000	100.000000	100.000000	13.292000	86.708000	42.978723	NaN
...	...	...	...	...	...	...	...	...	...	...
243	0.559266	2782.905026	NaN	92.20000	65.100000	97.697830	25.423000	74.577000	43.355482	-43750.0
244	2.537712	10460.113770	9.940811	65.93004	89.487651	84.583874	47.551252	52.448806	37.785216	0.0
245	-0.044394	1289.034078	5.553370	53.30000	54.900000	48.406710	67.126000	32.874000	44.604807	-50000.0
246	6.942853	1686.618024	4.754491	43.20000	63.000000	22.062560	60.413000	39.587000	32.063923	-34490.0
247	3.223668	850.827694	NaN	37.30000	77.500000	40.462560	67.166000	32.834000	41.876696	-219922.0

248 rows × 10 columns

In [8]:
```
df2.to_csv("wbresearch.csv", index=False)
```

In [9]:
```
import seaborn as sns
import matplotlib.pyplot as plt
df = pd.read_csv("wbresearch.csv")
df.head()
```

Out[9]:

	fdi	gdp	health	sanitation	water	electricity	ruralpop	urbanpop	agri	migration
0	0.299592	690.842629	8.479199	30.5	51.6	43.00000	74.532	25.468	58.067580	473007.0
1	7.468318	4247.485437	5.627439	92.1	95.4	100.00000	45.670	54.330	43.843066	-91750.0
2	0.717733	5583.616160	6.007278	87.0	84.9	100.00000	31.130	68.870	17.381700	-143268.0
3	NaN	NaN	NaN	62.5	100.0	59.32891	12.587	87.413	24.500000	NaN
4	NaN	39666.369210	8.302019	100.0	100.0	100.00000	13.292	86.708	42.978723	NaN

Sample

The World Bank dataset has 248 observations with 163 columns. For this research, I chose 9 features to predict the net migration of people based on these factors.

Measures

There are six features measured in percentages. All the data is numeric type.

Analysis

I will use Pearson correlation to check the strength of collinearity with the target variable.

In [10]:
```
df.corr()
```

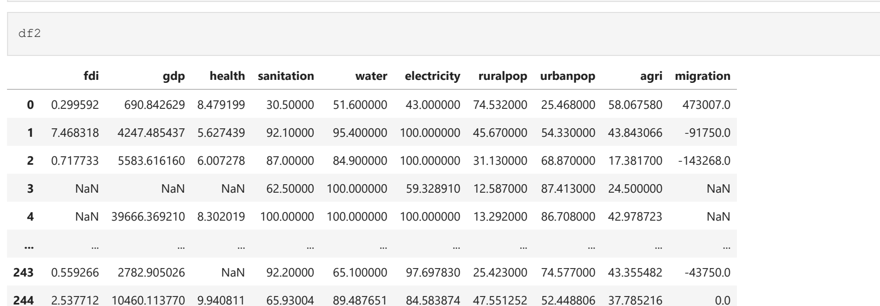
Out[10]:

	fdi	gdp	health	sanitation	water	electricity	ruralpop	urbanpop	agri	migration
fdi	1.000000	0.250559	0.016223	0.023753	0.008043	0.005692	-0.121705	0.121705	-0.013489	0.011989
gdp	0.250559	1.000000	0.334146	0.533179	0.451355	0.411777	-0.508069	0.508069	-0.237129	0.266345
health	0.016223	0.334146	1.000000	0.276713	0.307054	0.165047	-0.271735	0.271734	0.194480	0.332993
sanitation	0.023753	0.533179	0.276713	1.000000	0.804177	0.858107	-0.631365	0.631364	-0.244808	0.202127
water	0.008043	0.451355	0.307054	0.804177	1.000000	0.781008	-0.581884	0.581884	-0.170557	0.134436
electricity	0.005692	0.411777	0.165047	0.858107	0.781008	1.000000	-0.585040	0.585039	-0.149063	0.107798
ruralpop	-0.121705	-0.508069	-0.271735	-0.631365	-0.581884	-0.585040	1.000000	-1.000000	0.197438	-0.228288
urbanpop	0.121705	0.508069	0.271734	0.631364	0.581884	0.585039	-1.000000	1.000000	-0.197438	0.228288
agri	-0.013489	-0.237129	0.194480	-0.244808	-0.170557	-0.149063	0.197438	-0.197438	1.000000	-0.085710
migration	0.011989	0.266345	0.332993	0.202127	0.134436	0.107798	-0.228288	0.228288	-0.085710	1.000000

In [11]:
```
df.corr()["migration"].sort_values()
```

Out[11]:
```
ruralpop      -0.228288
agri          -0.085710
fdi            0.011989
electricity    0.107798
water          0.134436
sanitation     0.202127
urbanpop       0.228288
gdp            0.266345
health         0.332993
migration      1.000000
Name: migration, dtype: float64
```

In [12]:
```
plt.figure(figsize=(16,9))
sns.heatmap(df.corr(),cmap="coolwarm",annot=True,fmt='.2f',linewidths=2)
plt.title("", fontsize=20)
plt.show()
```



Import Libraries

In [13]:
```
------------------------
```

```
  File "<ipython-input-13-8f81c231e55d>", line 1
    ------------------------
                            ^
SyntaxError: invalid syntax
```

In [ ]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
import random

import sklearn

import shap
import statsmodels.api as sm

import datetime
from datetime import datetime, timedelta

import scipy.stats

#import xgboost as xgb
#from xgboost import XGBClassifier, XGBRegressor
#from xgboost import to_graphviz, plot_importance

#from sklearn.experimental import enable_hist_gradient_boosting
#from sklearn.linear_model import ElasticNet, Lasso, LinearRegression, LogisticRegression, Ridge
#from sklearn.ensemble import RandomForestClassifier, RandomForestRegressor, ExtraTreesClassifier, ExtraTreesRe
#from sklearn.ensemble import GradientBoostingClassifier, GradientBoostingRegressor, HistGradientBoostingClass


%matplotlib inline
#sets the default autosave frequency in seconds
%autosave 60
sns.set_style('dark')
sns.set(font_scale=1.2)

plt.rc('axes', labelsize=14)
plt.rc('xtick', labelsize=12)
plt.rc('ytick', labelsize=12)


from sklearn.model_selection import cross_val_score, train_test_split, GridSearchCV, RandomizedSearchCV
from sklearn.model_selection import cross_validate, KFold, RepeatedStratifiedKFold
from sklearn.preprocessing import LabelEncoder, StandardScaler, MinMaxScaler, OneHotEncoder
#from sklearn.pipeline import Pipeline
#from sklearn.feature_selection import RFE, RFECV, SelectKBest, f_classif, f_regression, chi2
#from sklearn.inspection import permutation_importance

from sklearn.tree import export_graphviz, plot_tree
from sklearn.metrics import confusion_matrix, classification_report, mean_absolute_error, mean_squared_error,r2
from sklearn.metrics import plot_confusion_matrix, plot_precision_recall_curve, plot_roc_curve, accuracy_score
from sklearn.metrics import auc, f1_score, precision_score, recall_score, roc_auc_score


#from tpot import TPOTClassifier, TPOTRegressor
#from imblearn.under_sampling import RandomUnderSampler
#from imblearn.over_sampling import RandomOverSampler
#from imblearn.over_sampling import SMOTE

import warnings
warnings.filterwarnings('ignore')

# import pickle
# from pickle import dump, load

# Use Folium library to plot values on a map.
#import folium

# Use Feature-Engine library

#import feature_engine.missing_data_imputers as mdi
#from feature_engine.outlier_removers import Winsorizer
#from feature_engine import categorical_encoders as ce


#from pycaret.classification import *
#from pycaret.clustering import *
#from pycaret.regression import *

pd.set_option('display.max_columns',None)
#pd.set_option('display.max_rows',100)
pd.set_option('display.width', 1000)
pd.set_option('display.float_format','{:.2f}'.format)


random.seed(0)
np.random.seed(0)
np.set_printoptions(suppress=True)
```

Exploratory Data Analysis

In [ ]:
```
df = pd.read_csv("worldbank.csv")
```

In [ ]:
```
df
```

In [ ]:
```
df.columns
```

In [ ]:

In [ ]:

In [ ]:
```
df.info(null_counts=True)
```

In [ ]:
```
df.describe(include='all')
```

In [ ]:
```
df.shape
```

In [ ]:
```
df.columns
```

Groupby Function

In [ ]:
```
df.groupby()
```

In [ ]:
```
df.groupby()
```

In [ ]:
```
df.groupby()
```

Pandas-Profiling Reports

In [ ]:
```
profile = ProfileReport(df=df, title='Name of Report', minimal=True)
```

In [ ]:
```
profile.to_notebook_iframe()
```

In [ ]:
```
profile.to_file("your_report.html")
```

Data Visualization

Univariate Data Exploration

In [ ]:
```
df.hist(bins=50, figsize=(20,10))
plt.suptitle('Histogram Feature Distribution', x=0.5, y=1.02, ha='center', fontsize=20)
plt.tight_layout()
plt.show()
```

In [ ]:
```
df.boxplot(figsize=(20,10))
plt.suptitle('BoxPlots Feature Distribution', x=0.5, y=1.02, ha='center', fontsize=20)
plt.tight_layout()
plt.show()
```

In [ ]:
```
fig, ax = plt.subplots(1,2, sharex=False, figsize=(16,5))
fig.suptitle('Main Title')


sns.countplot(x="", data=df, hue=df.income, ax=ax[0])
ax[0].set_title('Title of the first chart')
ax[0].tick_params('x', labelrotation=45)
ax[0].set_xlabel("")
ax[0].set_ylabel("")

sns.countplot(x="", data=df, hue=df.income, ax=ax[1])
ax[1].set_title('Title of the second chart')
ax[1].tick_params('x', labelrotation=45)
ax[1].set_xlabel("")
ax[1].set_ylabel("")


plt.show()
```

In [ ]:
```
#Plot 4 by 2 subplots

fig, ((ax1, ax2), (ax3, ax4), (ax5 ,ax6), (ax7 ,ax8)) = plt.subplots(4,2, sharex=False, figsize=(15,20))
fig.suptitle('Title', y=1)

sns.countplot(x="gender", hue="Churn", data=df, ax=ax1)
ax1.set_title('Title of the first chart')
ax1.tick_params('x', labelrotation=45)
ax1.set_xlabel("")
ax1.set_ylabel("")

sns.countplot(x="SeniorCitizen", hue="Churn", data=df, ax=ax2)
ax2.set_title('Title of the second chart')
ax2.tick_params('x', labelrotation=45)
ax2.set_xlabel("")
ax2.set_ylabel("")

sns.countplot(x="Partner", hue="Churn", data=df, ax=ax3)
#ax3.set_title('Title of the third chart')
#ax3.tick_params('x', labelrotation=45)
# ax1.set_xlabel("")
# ax1.set_ylabel("")

sns.countplot(x="Dependents", hue="Churn", data=df, ax=ax4)
#ax4.set_title('Title of the fourth chart')
#ax4.tick_params('x', labelrotation=45)
# ax1.set_xlabel("")
# ax1.set_ylabel("")

sns.countplot(x="PhoneService", hue="Churn", data=df, ax=ax5)
#ax5.set_title('Title of the fourth chart')
#ax5.tick_params('x', labelrotation=45)
# ax1.set_xlabel("")
# ax1.set_ylabel("")

sns.countplot(x="MultipleLines", hue="Churn", data=df, ax=ax6)
#ax6.set_title('Title of the fourth chart')
#ax6.tick_params('x', labelrotation=45)
# ax1.set_xlabel("")
# ax1.set_ylabel("")

sns.countplot(x="InternetService", hue="Churn", data=df, ax=ax7)
#ax7.set_title('Title of the fourth chart')
#ax7.tick_params('x', labelrotation=45)
#ax1.set_xlabel("")
#ax1.set_ylabel("")

sns.countplot(x="OnlineSecurity", hue="Churn", data=df, ax=ax8)
#ax8.set_title('Title of the fourth chart')
#ax8.tick_params('x', labelrotation=45)
#ax1.set_xlabel("")
#ax1.set_ylabel("")


plt.show()
```

In [ ]:
```
fig = plt.figure(figsize=(20,40))

plt.subplot(7,2,1)
plt.title("")
sns.countplot()

plt.subplot(7,2,2)
plt.title("")
sns.countplot()

plt.subplot(7,2,3)
plt.title("")
sns.countplot()

plt.subplot(7,2,4)
plt.title("")
sns.countplot()

plt.subplot(7,2,5)
plt.title("")
sns.barplot()

plt.subplot(7,2,6)
plt.title("")
sns.barplot()

plt.subplot(7,2,7)
plt.title("")
sns.barplot()

plt.subplot(7,2,8)
plt.title("")
sns.barplot()

plt.subplot(7,2,9)
plt.title("")
sns.scatterplot()

plt.subplot(7,2,10)
plt.title("")
sns.scatterplot()

plt.subplot(7,2,11)
plt.title("")
sns.scatterplot()

plt.subplot(7,2,12)
plt.title("")
sns.scatterplot()

plt.subplot(7,2,13)
plt.title("")
sns.relplot()

plt.subplot(7,2,14)
plt.title("")
sns.relplot()

plt.tight_layout()
plt.show()
```

Bivariate Data Exploration

In [ ]:
```
sns.jointplot(x='', y='',data=df, kind='scatter')

sns.jointplot(x='', y='',data=df, kind='scatter')

sns.jointplot(x='', y='',data=df, kind='scatter')

sns.jointplot(x='', y='',data=df, kind='scatter')

sns.jointplot(x='', y='',data=df, kind='kde')

sns.jointplot(x='', y='',data=df, kind='kde')

sns.jointplot(x='', y='',data=df, kind='hex')

sns.jointplot(x='', y='',data=df, kind='hex')

sns.jointplot(x='', y='',data=df, kind='reg')

sns.jointplot(x='', y='',data=df, kind='reg')


plt.show()
```

In [ ]:
```
plt.figure(figsize=(20,20))


g = sns.catplot(x='gender', hue = 'tenure', row = 'division',
           kind='count', data=ratings_df,
           height = 3, aspect = 1)

g.set_xlabels("")
g.set_ylabels("")
#g = (g.set_axis_labels("Tip","Total bill(USD)").set(xlim=(0,10),ylim=(0,100)


g.set(xlim=(0,None))
g.set_xticklabels(rotation=90)

plt.suptitle('', x=0.5, y=1.02, ha='center', fontsize=20)

plt.show()
```

In [ ]:
```
plt.figure(figsize=(20,20))

sns.catplot(x="calories", y="restaurant",

                hue="is_salad", ci=None,

                data=df_calories, color=None, linewidth=3, showfliers = False,

                orient="h", height=20, aspect=1, palette=None,

                kind="box", dodge=True)

plt.xlabel("", size=20)
plt.ylabel("", size=20)
plt.suptitle('', x=0.5, y=1.02, ha='center', fontsize=20)
plt.show()
```

In [ ]:
```
plt.figure(figsize=(20,20))

sns.relplot(x="age", y="eval", hue="gender",
           row="tenure",
           data=ratings_df, height = 3, aspect = 2)

plt.xlabel("", size=20)
plt.ylabel("", size=20)
plt.suptitle('', x=0.5, y=1.02, ha='center', fontsize=20)
plt.show()
```

Time-Series Analysis

In [ ]:
```
timeseries = df[['date','extraction','month', 'day']]
```

In [ ]:
```
timeseries
```

In [ ]:
```
timeseries.info()
```

In [ ]:
```
fig = plt.figure(figsize=(30,10))
sns.lineplot(x=df.month,y=df.amount,data=df, estimator=None)
plt.title("", fontsize=20)
plt.xlabel("", fontsize=20)
plt.ylabel("", fontsize=20)
plt.legend(['',''])
plt.show()
```

In [ ]:
```
fig = plt.figure(figsize=(30,10))
sns.lineplot(x=df.month,y=df.amount,data=df, estimator=None)
plt.title("", fontsize=20)
plt.xlabel("", fontsize=20)
plt.ylabel("", fontsize=20)
plt.legend(['',''])
plt.show()
```

In [ ]:
```
fig = plt.figure(figsize=(30,10))
sns.lineplot(x=df.month,y=df.amount,data=df, estimator=None)
plt.title("", fontsize=20)
plt.xlabel("", fontsize=20)
plt.ylabel("", fontsize=20)
plt.legend(['',''])
plt.show()
```

Regression plot

In [ ]:
```
line_color = {'color': 'red'}
fig , ax = plt.subplots(2,2, figsize=(20,20))

#BMI

ax1 = sns.regplot(x=X_test.bmi, y=lr_pred, line_kws=line_color, ax=ax[0,0])
ax1.set_xlabel("x")
ax1.set_ylabel("y")


#s5

ax2 = sns.regplot(x=X_test.s5, y=lr_pred, line_kws=line_color, ax=ax[0,1])
ax2.set_xlabel("x")
ax2.set_ylabel("y")


#bp

ax3 = sns.regplot(x=X_test.bp, y=lr_pred, line_kws=line_color, ax=ax[1,0])
ax3.set_xlabel("x")
ax3.set_ylabel("y")

#s4

ax4 = sns.regplot(x=X_test.s4, y=lr_pred, line_kws=line_color, ax=ax[1,1])
ax4.set_xlabel("x")
ax4.set_ylabel("y")

plt.show()
```

FacetGrid

In [ ]:
```
g = sns.FacetGrid(data=df, col="column_name", height=3, aspect=1)
g.map(plt.scatter, "numeric", "numeric")
g.add_legend()
plt.show()
```

Pairplots

In [ ]:
```
sns.pairplot(df)
plt.suptitle('Pairplots of features', x=0.5, y=1.02, ha='center', fontsize=20)
plt.show()
```

Data Preprocessing

Feature Engineering

In [ ]:

In [ ]:

In [ ]:

Drop unwanted features

In [ ]:
```
df.columns
```

In [ ]:
```
df.drop()
```

In [ ]:
```
df
```

Treat Missing Values

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

In [ ]:
```
df[''] = df[''].replace(np.nan,df.mean())
```

In [ ]:
```
#imputer = mdi.MeanMedianImputer(imputation_method='median',variables=None)
```

In [ ]:
```
#imputer.fit(df)
```

In [ ]:
```
#df = imputer.transform(df)
```

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

Replacing values

In [ ]:
```
df.replace()
```

In [ ]:

Rounding Values

In [ ]:
```
###pandas.DataFrame.round
df[['internetuserate']] = df[['internetuserate']].round(decimals=0)
```

Treat Duplicate Values

In [ ]:
```
df.duplicated(keep='first').sum()
```

In [ ]:
```
df[df.duplicated(keep=False)] #Check duplicate values
```

In [ ]:
```
df.drop_duplicates(ignore_index=True, inplace=True)
```

In [ ]:

Treat Outliers

In [ ]:
```
df.columns
```

In [ ]:
```
df
```

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

In [ ]:
```
#windsorizer = Winsorizer(distribution='skewed',tail='both',fold=1.5, variables=[])
```

In [ ]:
```
#windsorizer.fit(df)
```

In [ ]:
```
#df2 = windsorizer.transform(df)
```

In [ ]:
```
#df2
```

In [ ]:
```
#df2.describe()
```

In [ ]:
```
#windsorizer.left_tail_caps_
```

In [ ]:
```
#windsorizer.right_tail_caps_
```

Treat Data Types

In [ ]:
```
df.info()
```

In [ ]:
```
df["breastcancerper100th"] = df["breastcancerper100th"].astype('int')
```

In [ ]:
```
df.info()
```

Perform One-Hot Encoding

In [ ]:
```
pd.get_dummies(df_category)
```

In [ ]:
```
df["has_gas"] = pd.get_dummies(data=df["has_gas"],drop_first=True)
```

Visualize pairplots for multi-classification

In [ ]:
```
#Must convert all strings to numbers
sns.pairplot(df2,vars=df2.columns[:-1],hue='logo')
plt.show()
```

In [ ]:

Create and save processed dataset

In [ ]:
```
df.to_csv("",index=False)
```

Train Test Split

```python
X = df.iloc[:,:]
y = df.iloc[:,:]
```

```python
df.shape
```

Treat Imbalance Data

```python
y.value_counts()
```

```python
ros = RandomOverSampler(sampling_strategy='all',random_state=0)
```

```python
new_X, new_y = ros.fit_resample(X, y)
```

```python
new_y[].value_counts()
```

```python
new_X
```

Train Test Split Cont'd

```python
X.values, y.values
```

```python
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)
```

```python
X_train.shape, X_test.shape, y_train.shape, y_test.shape
```

Train Test Split to create Train, Validation and Test Set

```python
#Set test set size
X_df, X_test, y_df, y_test = train_test_split(X.values, y.values, test_size=0.2, random_state=0)
```

```python
X_df.shape, X_test.shape, y_df.shape, y_test.shape
```

```python
X_train, X_val = train_test_split(X_df, test_size=0.2, random_state=0)
```

```python
X_train.shape, X_val.shape
```

Feature Selection

Using SelectKBest

```python
X_new = SelectKBest(f_regression, k=10).fit_transform(X_train,y_train)
```

```python
X_new[0:5]
```

Univariate Feature Selection

```python
select_feature = SelectKBest(chi2, k=10).fit(X_train,y_train)
```

```python
select_feature.scores_
```

Recursive Feature Elimination

```python
rfe = RFE(estimator=XGBRegressor(),n_features_to_select=10,verbose=1, step=1)
```

```python
rfe.fit(X_train,y_train)
```

```python
selected_rfe_features = pd.DataFrame({'Feature':list(X_train.columns),'Ranking':rfe.ranking_})
```

```python
selected_rfe_features
```

Recursive Feature Elimnation with Cross Validation

```python
rfecv = RFECV(estimator=XGBRegressor(), cv=5,scoring="neg_mean_squared_error",verbose=1, step=1)
```

```python
rfecv.fit(X_train,y_train)
```

```python
print("Optimal no of features:", rfecv.n_features_)
```

```python
print("Best features:", rfecv.support_)
```

Feature Scaling

```python
X_train
```

```python
encoder = LabelEncoder()
```

```python
scaler = StandardScaler()
```

```python
minmax = MinMaxScaler()
```

```python
ohe = OneHotEncoder()
```

```python
X_train_scaled = minmax.fit_transform(X_train)
```

```python
X_test_scaled = minmax.transform(X_test)
```

```python
X_train_scaled
```

```python
X_test_scaled
```

Model Training

Using PyCaret

```python
exp_reg = setup(data = df, target = '', session_id=0, normalize=True)
```

```python
compare_models(exclude=['catboost','lightgbm','lda','qda','mlp','ada','nb','ridge','rbfsvm','svm'],fold=5) #For
```

```python
compare_models(exclude=['omp','br','ard','par','ransac','tr','huber','kr','svm','knn','dt','rf',
                        'et','ada','gbr','mlp','xgboost','lightgbm','catboost'],fold=5) # For Regressor
```

```python
model_selected = create_model('catboost')
```

```python
print(model_selected)
```

```python
tuned_model = tune_model(catboost, optimize='mse')
```

```python
print(tuned_model)
```

```python
plot_model(tuned_model)
```

```python
plot_model(tuned_model, plot = 'error')
```

```python
plot_model(tuned_model, plot='feature')
```

```python
interpret_model(tuned_model)
```

```python
evaluate_model(tuned_model)
```

```python
predict_model(tuned_model)
```

```python
final_model = finalize_model(tuned_model)
```

```python
unseen_predictions = predict_model(final_model, data=data_unseen)
unseen_predictions.head()
```

Using Regression or Classification Models

```python
reg_model
```

```python
classi_model
```

K-Fold Cross-Validation (Generalization Performance)

```python
lasso = Lasso(random_state=0)
```

```python
kf = KFold(n_splits=5, shuffle=True, random_state=0)
```

```python
lasso_cv = cross_validate(estimator=lasso, X=X_train_scaled, y=y_train, scoring="neg_root_mean_squared_error",
                          cv=kf, n_jobs=-1,return_train_score=True)
```

```python
lasso_cv
```

```python
np.mean(lasso_cv["train_score"]), np.std(lasso_cv["train_score"])
```

```python
np.mean(lasso_cv["test_score"]), np.std(lasso_cv["test_score"])
```

Using TPOT

```python
# tpot = TPOTClassifier(generations=3,population_size=10,scoring='accuracy', cv=5, verbosity=2, random_state=0,
#                       early_stop=1)
```

```python
# tpot.fit(X_train,y_train)
```

```python
# tpot.score(X_test, y_test)
```

Using XGBoost (Scikit-Learn)

Using RandomSearchCV

```python
model = XGBRegressor(random_state=0, n_estimators=100, objective='reg:squarederror')
```

```python
model = XGBClassifier(random_state=0, n_estimators=100, objective='softmax:multi')
```

```python
parameters = {'max_depth': np.arange(3,10,1),
              'eta': np.arange(0.05,0.3,0.05),
              'n_estimators':np.arange(100,1000,100),
              'min_child_weight': np.arange(1,4,1),
              'gamma':np.arange(0,10,2),
              'subsample':np.arange(0.5,0.9,0.1),
              'colsample_bytree':np.arange(0.5,0.9,0.1),
              'reg_alpha':np.arange(0,1,0.1),
              'reg_lambda':np.arange(0,1,0.1)
             }
```

```python
randm = RandomizedSearchCV(estimator=model, param_distributions = parameters, cv = 5, n_iter = 50,
                           n_jobs=-1, scoring='')
```

```python
randm.fit(X, y)
```

```python
randm.best_estimator_
```

```python
randm.best_score_
```

```python
randm.best_params_
```

Final Model

```python
xgbmodel = XGBRegressor(random_state=0, n_estimators=100, objective='reg:squarederror')
```

```python
xgbmodel = XGBClassifier(random_state=0, n_estimators=100, objective='binary:logistic')
```

```python
xgbmodel = XGBClassifier(random_state=0, n_estimators=100, objective='softmax:multi')
```

```python
xgbmodel.fit(X_train_scaled,y_train,eval_set=[(X_test_scaled,y_test)],eval_metric='rmse',early_stopping_rounds=
```

```python
xgbmodel.fit(X_train_scaled,y_train,eval_set=[(X_test_scaled,y_test)],eval_metric='error',early_stopping_rounds
```

```python
xgbmodel.fit(X_train_scaled,y_train,eval_set=[(X_test_scaled,y_test)],eval_metric='mlogloss',early_stopping_rou
```

```python
y_pred = xgbmodel.predict(X_test_scaled)
```

```python
y_pred
```

Model Evaluation

```python
cm = confusion_matrix(y_test,y_pred)
cm
```

```python
fig , ax = plt.subplots(figsize=(10,5))
sns.heatmap(cm, annot=True,fmt='.4g',linewidths=2, cmap='viridis')
plt.ylabel('True label')
plt.xlabel('Predicted label')
plt.show()
```

```python
fig , ax = plt.subplots(figsize=(10,5))
plot_confusion_matrix(xgbmodel,X_test_scaled,y_test,values_format='.4g',ax=ax)
plt.show()
```

```python
print(classification_report(y_test,y_pred))
```

```python
plot_roc_curve(xgbmodel,X_test,y_test)
plt.show()
```

```python
mse = mean_squared_error(y_test,y_pred)
mse
```

```python
rmse = np.sqrt(mse)
rmse
```

```python
r2score = r2_score(y_test,y_pred)
r2score
```

```python
fig, ax = plt.subplots(figsize=(10,8))
sns.regplot(x=y_test, y=y_pred, ax=ax)
plt.title("Plot to compare actual vs predicted", fontsize=20)
plt.ylabel("Predicted")
plt.xlabel("Actual")
plt.show()
```

Plot Feature Importances

```python
rf.feature_importances_
```

```python
feat_importances = pd.Series(rf.feature_importances_, index=X.columns)
```

```python
feat_importances
```

```python
feat_importances.nlargest(10).plot(kind='barh', figsize=(10,10))
plt.title('Feature Importances')
plt.show()
```

The permutation based importance

```python
perm_importance = permutation_importance(rf,X_test,y_test, random_state=0, scoring='neg_mean_squared_error')
```

```python
sorted_idx = perm_importance.importances_mean.argsort()
plt.figure(figsize=(10,10))
plt.title("Permutation-based Importance")
plt.barh(X.columns[sorted_idx], perm_importance.importances_mean[sorted_idx])
plt.xlabel("Permutation Importance")
plt.show()
```

Compute Importance from SHAP Values

```python
explainer = shap.TreeExplainer(rf)
```

```python
shap_values = explainer.shap_values(X_test)
```

```python
shap.summary_plot(shap_values, X_test, plot_type="bar")
```

```python
shap.summary_plot(shap_values, X_test)
```

Available importance_types = ['weight', 'gain', 'cover', 'total_gain', 'total_cover']

```python
X.columns
```

```python
xgbmodel.get_booster().feature_names = [X.columns]
```

```python
fig, ax = plt.subplots(figsize=(20,10))
xgb.plot_importance(xgbmodel.get_booster(),ax=ax)
plt.show()
```

```python
xgb.to_graphviz(xgbmodel,num_trees=100)
```

Example:

f = 'gain'

XGBClassifier.get_booster().get_score(importance_type= f)

Plot Tree

```python
X.columns
```

```python
plt.figure(figsize=(40,25))
plot_tree(treeclf, feature_names=X.columns,class_names=['0','1'], fontsize=14, filled=True)
plt.show()
```

Cross-Validation

```python
cv = cross_val_score(xgbmodel,X,y,cv=5,verbose=1,scoring='')
```

```python
cv.mean()
```

Using XGBoost (API)

```python
dtrain = xgb.DMatrix(data=X_train,label=y_train)
dtest = xgb.DMatrix(data=X_test,label=y_test)
```

```python
params = {'n_estimators':,
          'learning_rate':,
          'max_depth':,
          'objective': '',
          'num_class': ,
          'seed': 0,
          'eval_metric':''}
```

```python
xgbmodel = xgb.train(params=params,dtrain=dtrain,num_boost_round=100,evals=[(dtest,"Test")],
                     early_stopping_rounds=10)
```

```python
y_pred = xgbmodel.predict(dtest)
```

```python
y_pred
```

Cross-Validation (API)

```python
cv = xgb.cv(params=params,
            dtrain=dtrain,
            num_boost_round=100,
            nfold=5,
            stratified=False,
            folds=None,
            metrics=('merror'),
            obj=None,
            feval=None,
            maximize=False,
            early_stopping_rounds=10,
            fpreproc=None,
            as_pandas=True,
            verbose_eval=None,
            show_stdv=True,
            seed=0,
            callbacks=None,
            shuffle=True,)
```

```python
cv
```

```python
cv['test-merror-mean'].min()
```

Model Evaluation (Classification)

```python
cm = confusion_matrix(y_test,y_pred)
cm
```

```python
fig , ax = plt.subplots(figsize=(10,5))
sns.heatmap(cm, annot=True,fmt='.4g',linewidths=2, cmap='viridis')
plt.ylabel('True label')
plt.xlabel('Predicted label')
plt.show()
```

```python
fig , ax = plt.subplots(figsize=(10,5))
plot_confusion_matrix(xgbmodel,X_test_scaled,y_test,values_format='.4g',ax=ax)
plt.show()
```

```python
print(classification_report(y_test,y_pred))
```

```python
plot_roc_curve(xgbmodel,X_test,y_test)
plt.show()
```

Model Evaluation (Regression)

```python
mse = mean_squared_error(y_test,y_pred)
mse
```

```python
rmse = np.sqrt(mse)
rmse
```

```python
r2score = r2_score(y_test,y_pred)
r2score
```

```python
fig, ax = plt.subplots(figsize=(10,8))
sns.regplot(x=y_test, y=y_pred, ax=ax)
plt.title("Plot to compare actual vs predicted")
plt.ylabel("Predicted")
plt.xlabel("Actual")
plt.show()
```

Table Formatted View

```python
table = X_test.copy()
```

```python
table["True Value"] = y_test.copy()
```

```python
table["Predicted"] = np.round(lr_pred,2)
```

```python
table
```

Cross-Validation

```python
cv = cross_val_score(xgbmodel,X,y,cv=5,verbose=1,scoring='accuracy')
```

```python
cv.mean()
```

Feature Selection

```python
df.columns
```

```python
df2 = df[['']]
```

```python
df2
```

```python
X = df2.iloc[:,0:7]
y = df2.iloc[:,7]
```

```python
X.values, y.values
```

```python
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)
```

```python
xgbmodel2 = XGBClassifier(random_state=0, n_estimators=100, objective='binary:logistic')
```

```python
xgbmodel2 = XGBRegressor(random_state=0, n_estimators=100, objective='reg:squarederror')
```

```python
xgbmodel2.fit(X_train,y_train,eval_set=[(X_test,y_test)],eval_metric='error',early_stopping_rounds=10)
```

```python
xgbmodel2.fit(X_train,y_train,eval_set=[(X_test,y_test)],eval_metric='rmse',early_stopping_rounds=10)
```

```python
y_pred = xgbmodel2.predict(X_test)
```

```python
y_pred
```

```python
fig , ax = plt.subplots(figsize=(5,5))
plot_confusion_matrix(xgbmodel2,X_test,y_test,values_format='4g',ax=ax)
plt.show()
```

```python
print(classification_report(y_test,y_pred))
```

```python
plot_roc_curve(xgbmodel2,X_test,y_test)
plt.show()
```

Model Prediction

```python
testdata = pd.read_csv()
```

```python
answer = xgbmodel.predict(testdata)
```

```python
answer
```

Model Tuning

Using RandomSearchCV

```python
model = XGBClassifier(random_state=0, n_estimators=100, objective='binary:logistic')
```

```python
model = XGBRegressor(random_state=0, n_estimators=100, objective='reg:squarederror')
```

```python
parameters = {'max_depth': np.arange(3,10,1),
              'learning rate': np.arange(0.05,0.3,0.03),
              'n_estimators':np.arange(100,1000,100),
              'min_child_weight': np.arange(1,4,1),
              'gamma':np.arange(0,50,2),
              'subsample':np.arange(0.5,0.9,0.1),
              'colsample_bytree':np.arange(0.5,0.9,0.1)
             }
```

```python
randm = RandomizedSearchCV(estimator=model, param_distributions = parameters, cv = 5, n_iter = 50,
                           n_jobs=-1, scoring='')
```

```python
randm.fit(X, y)
```

```python
randm.best_estimator_
```

```python
randm.best_score_
```

```python
randm.best_params_
```

Using GridSearchCV

```python
model = XGBClassifier(random_state=0, n_estimators=100, objective='binary:logistic')
```

```python
parameters = {'max_depth': np.arange(3,10,1),
              'learning rate': np.arange(0.05,0.3,0.03),
              'n_estimators':np.arange(100,1000,100),
              'min_child_weight': np.arange(1,4,1),
              'gamma':np.arange(0,50,2),
              'subsample':np.arange(0.5,0.9,0.1),
              'colsample_bytree':np.arange(0.5,0.9,0.1)
             }
```

```python
grids = GridSearchCV(estimator=model,param_grid=parameters,scoring='accuracy',
                     n_jobs=-1,cv=5,verbose=1,return_train_score=True)
```

```python
grids.fit(X,y)
```

```python
grids.best_estimator_
```

Final Model

```python
xgbnew = XGBClassifier(random_state=0, n_estimators=, objective='binary:logistic',max_depth=,
                       gamma=, min_child_weight=,learning_rate=,subsample=,colsample_bytree=)
```

```python
xgbnew = XGBRegressor(random_state=0, n_estimators=, objective='binary:logistic',max_depth=,
                      gamma=, min_child_weight=,learning_rate=,subsample=,colsample_bytree=)
```

```python
xgbnew.fit(X_train,y_train,eval_set=[(X_test,y_test)],eval_metric='error',early_stopping_rounds=10)
```

```python
y_pred = xgbnew.predict(X_test)
```

```python
y_pred
```

```python
fig , ax = plt.subplots(figsize=(5,5))
plot_confusion_matrix(xgbnew,X_test,y_test,values_format='4g',ax=ax)
plt.show()
```

```python
print(classification_report(y_test,y_pred))
```

```python
plot_roc_curve(xgbnew,X_test,y_test)
plt.show()
```

Save the Model

```python
filename = 'model.sav'
dump(xgbnew,open(filename,'wb'))
```

Load the Model

```python
loaded_model = load(open(filename,'rb'))
```

```python
loaded_model
```

Python code done by Dennis Lam