

NAME-PAPU SWAIN

PROJECT-HOUSE LOAN DATA ANALYSIS

OBJECTIVE-

Create a model that predicts whether or not an applicant will be able to repay a loan using historical data.

In [1]:

```
import pandas as pd
import numpy
import matplotlib.pyplot as plt
import seaborn as sns
import plotly online
import warnings
warnings.filterwarnings("ignore")
```

Out [2]:

SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	DAYS_BIRTH	DAYS_EMPLOYED	...	FLAG_DOCUMENT_2
0	100002	0	Cash loans	M	N	N	Y	0	202500.0	406597.5	24700.5	...			
1	100003	0	Cash loans	F	N	N	Y	0	270000.0	1293502.5	35689.5	...			
2	100004	0	Revolving loans	M	Y	Y	0	67500.0	135000.0	6750.0	...				
3	100006	0	Cash loans	F	N	N	Y	0	135000.0	312682.5	29686.5	...			
4	100007	0	Cash loans	M	N	Y	0	121500.0	513000.0	21865.5	...				

5 rows × 122 columns

In [3]:

```
>>>data = pd.read_csv("loan_data1.csv")
df.head()
```

Out [4]:

SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	DAYS_BIRTH	DAYS_EMPLOYED	...	FLAG_DOCUMENT_2
count	307511.000000	307511.000000	307511.000000	307511.000000	307499.000000	307233.000000	307511.000000	307511.000000	307511.000000	307511.000000	307511.000000	307511.000000	307511.000000	...	307511.000000
mean	276180.518577	0.080729	0.417052	1.687979e+05	5.900208e+05	27108.737395	5.383962e+05	0.022686	-16036.995067	63815.045904	63815.045904
std	102790.173438	0.272419	0.722121	2.371223e+04	4.024908e+04	14493.737319	5.369446e+04	0.013831	4363.988632	141275.766519	141275.766519
min	110001.200000	0.000000	0.000000	2.566000e+04	4.500000e+04	1615.000000	4.059000e+04	0.000290	-25249.000000	-2760.000000	-2760.000000
25%	169145.500000	0.000000	0.000000	1.125000e+05	5.135310e+05	16524.000000	2.385000e+04	0.010006	-19682.000000	-2760.000000	-2760.000000
50%	276182.000000	0.000000	0.000000	1.471500e+05	5.135310e+05	24903.000000	4.500000e+04	0.018850	-15750.000000	-1213.000000	-1213.000000
75%	367142.500000	0.000000	1.000000	2.025000e+05	8.086500e+05	34596.000000	6.795000e+04	0.028663	-12413.000000	-289.000000	-289.000000
max	452625.000000	1.000000	19.000000	1.170000e+08	4.050000e+06	258025.500000	4.050000e+06	0.072506	-7489.000000	365243.000000	365243.000000

8 rows × 16 columns

In [5]:

```
>>>#Check for null values in the dataset
df.isna().sum()
```

Out [5]:

SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER
SK_ID_CURR	0		
TARGET	0		
NAME_CONTRACT_TYPE	0		
CODE_GENDER	0		

AMT_REQ_CREDIT_BUREAU_YEAR 41519
AMT_REQ_CREDIT_BUREAU_WEEK 41519
AMT_REQ_CREDIT_BUREAU_MON 41519
AMT_REQ_CREDIT_BUREAU_QRT 41519
AMT_REQ_CREDIT_BUREAU_YEAR 41519
Length: 22, dtype: int64

In [6]:

```
>>>df.head()
```

Out [6]:

SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	...	FLAG_DOCUMENT_2	
0	100002	0	Cash loans	M	N	Y	0	202500.0	406597.5	24700.5	...	
1	100003	0	Cash loans	F	N	N	Y	0	270000.0	1293502.5	35689.5	...
2	100004	0	Revolving loans	M	Y	Y	0	67500.0	135000.0	6750.0	...	
3	100006	0	Cash loans	F	N	Y	0	135000.0	312682.5	29686.5	...	
4	100007	0	Cash loans	M	N	Y	0	121500.0	513000.0	21865.5	...	

5 rows × 122 columns

In [7]:

```
>>>defaulter=df[("TARGET")==1].sum()
payers=df[("TARGET")==0].sum()
default_percent=(defaulters/payers)*100
print("Percentage of default to payers is : ",default_percent)

Percentage of default to payer is  8.781826681345662
```

In [8]:

```
>>>Balance the dataset if the data is imbalanced
```

Out [8]:

df["TARGET"].value_counts()
0 282686 1 24825 Name: TARGET, dtype: int64

In [9]:

```
>>>plt.figure(figsize=(10,7))
sns.countplot(df["TARGET"])
plt.title("Checking whether the dataset is balanced or not")
```

Out [9]:

Text(0.5, 1.0, 'Checking whether the dataset is balanced or not')

In [10]:

```
>>>plt.figure(figsize=(10,7))
sns.heatmap(df.corr(),cmap="bwr")
```

Out [10]:

<AxesSubplot>

In [11]:

```
>>>shuffled_df=pd.sample(frac=1,random_state=25)
fraud_df=shuffled_df[shuffled_df["TARGET"]==1]
nonfraud_df=shuffled_df[shuffled_df["TARGET"]==0]
balanced_df=pd.concat([fraud_df,nonfraud_df])
```

Out [11]:

Text(0.5, 1.0, 'Checking whether the dataset is balanced or not')

In [12]:

```
>>>plt.figure(figsize=(10,7))
sns.countplot(balanced_df["TARGET"])
plt.title("Checking whether the dataset is balanced or not")
```

Out [12]:

Text(0.5, 1.0, 'Checking whether the dataset is balanced or not')

In [13]:

```
>>>plt.figure(figsize=(10,7))
sns.heatmap(balanced_df.corr(),cmap="bwr")
```

Out [13]:

<AxesSubplot>

In [14]:

```
>>>balanced_df.head()
```

Out [14]:

SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	...	FLAG_DOCUMENT_2
122136	241002	1	Cash loans	F	N	0	66600.0	808650.0	31464.0	...	
32365	137520	1	Cash loans	M	Y	Y	0	135000.0	512064.0	25033.5	...
95288	210632	1	Cash loans	M	Y	N	0	180000.0	1078200.0	31522.5	...
243096	381398	1	Cash loans	F	Y	Y	1	117000.0	539100.0	27652.5	...
61628	171473	1	Cash loans	M	N	Y	0	180000.0	900000.0	57649.5	...

5 rows × 122 columns

In [15]:

```
>>>balanced_df.describe(include="object")
```

Out [15]:

	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	NAME_TYPE_SUITE	NAME_INCOME_TYPE	NAME_EDUCATION_TYPE	NAME_FAMILY_STATUS	NAME_HOUSING
count	49650	49650	49650	49650	49470	49650	49650	49650	49650
unique	2	2	2	2	7	6	5	5	5
top	Cash loans	F	N	Y	Unaccompanied	Working	Secondary / secondary special	Married	House / a
freq	45558	30740	33534	34232	40400	27887	36986	30843	35402

In [16]:

```
>>>balanced_df.drop(columns=["WEEKDAY_APPR_PROCESS_START","FLAG_OWN_REALTY","NAME_TYPE_SUITE","FORDKAPROMOT_MODE","EMERGENCYSTATE_MODE"],inplace=True)
```

Out [16]:

NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	NAME_INCOME_TYPE	NAME_EDUCATION_TYPE	NAME_FAMILY_STATUS	NAME_HOUSING_TYPE	OCCUPATION_TYPE	ORGANIZATION_TYPE
count	49650	49650	49650	49650	49650	49650	49650	35402
unique	2	2	2	6	5	5	6	18
top	Cash loans	F	N	Working	Secondary / secondary special	Married	House / apartment	Laborers
freq	45558	30740	33534	27887	36986	30843	43361	10110

In [17]:

```
>>>from sklearn.preprocessing import LabelEncoder
```

Out [17]:

```
>>>le=LabelEncoder()
balanced_df["NAME_CONTRACT_TYPE"]=le.fit_transform(balanced_df["NAME_CONTRACT_TYPE"])
balanced_df["CODE_GENDER"]=le.fit_transform(balanced_df["CODE_GENDER"])
balanced_df["FLAG_OWN_CAR"]=le.fit_transform(balanced_df["FLAG_OWN_CAR"])
balanced_df["NAME_INCOME_TYPE"]=le.fit_transform(balanced_df["NAME_INCOME_TYPE"])
balanced_df["NAME_EDUCATION_TYPE"]=le.fit_transform(balanced_df["NAME_EDUCATION_TYPE"])
balanced_df["NAME_FAMILY_STATUS"]=le.fit_transform(balanced_df["NAME_FAMILY_STATUS"])
balanced_df["NAME_HOUSING_TYPE"]=le.fit_transform(balanced_df["NAME_HOUSING_TYPE"])
balanced_df["OCCUPATION_TYPE"]=le.fit_transform(balanced_df["OCCUPATION_TYPE"])
balanced_df["ORGANIZATION_TYPE"]=le.fit_transform(balanced_df["ORGANIZATION_TYPE"])
balanced_df["HOUSE_TYPE_MODE"]=le.fit_transform(balanced_df["HOUSE_TYPE_MODE"])
balanced_df["NAME_CONTRACT_TYPE"]=le.fit_transform(balanced_df["NAME_CONTRACT_TYPE"])
balanced_df["WALLSMATERIAL_MODE"]=le.fit_transform(balanced_df["WALLSMATERIAL_MODE"])
```

In [18]:

```
>>>balanced_df.info()
```

Out [18]:

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 49650 entries, 122136 to 61628
dtypes: float64(65), int32(10), int64(42)
memory usage: 43.8 MB
```

In [19]:

```
>>>balanced_df.head()
```

Out [19]:

SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	...	FLAG_DOCUMENT_2
122136	241002	1	0	0	0	0	66600.0	808650.0	31464.0	675000.0	...	
32365	137520	1	0	1	1	0	135000.0	512064.0	25033.5	360000.0	...	
95288	210632	1	0	1	1	0	180000.0	1078200.0	31522.5	900000.0	...	
243096	381398	1	0	1	1	1	117000.0	539100.0	27652.5	450000.0	...	
61628	171473	1	0	1	0	0	180000.0	900000.0	57649.5	900000.0	...	

5 rows × 117 columns

In [20]:

```
>>>x=balanced_df.drop(["TARGET"],axis=1)
y=balanced_df["TARGET"]
```

Out [20]:

SK_ID_CURR	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	NAME_INCOME_TYPE	...
122136	241002	0	0	0	66600.0	808650.0	31464.0	675000.0	2	
32365	137520	0	1	1	135000.0	512064.0	25033.5	360000.0	0	
95288	210632	0	1	1	180000.0	1078200.0	31522.5	900000.0	5	
243096	381398	0	0	1	117000.0	539100.0	27652.5	450000.0	5	
61628	171473	0	1	0	180000.0	900000.0	57649.5	900000.0	5	

5 rows × 116 columns

Out [21]:

```
>>>y.head()
```

Out [21]:

TARGET
122136 1
32365 1
95288 1
243096 1
61628 1

In [22]:

```
>>>from sklearn.model_selection import train_test_split
```

Out [22]:

```
>>>x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.25,random_state=25)
```

In [23]:

```
>>>x_train.shape,x_test.shape
```

Out [23]:

```
((37237, 116), (12413, 116))
```

In [24]:

```
>>>y_train.shape,y_test.shape
```

Out [24]:

```
((37237, 1), (12413, 1))
```

In [25]:

```
>>>from sklearn.preprocessing import StandardScaler
```

Out [25]:

```
>>>scaler=StandardScaler()
x_train=scaler.fit_transform(x_train)
x_test=scaler.transform(x_test)
```

In [26]:

```
>>>Architect the Deep learning Model
```

Out [26]:

```
>>>import tensorflow as tf
```

In [27]:

```
>>>from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense,Dropout
```

Out [27]:

```
>>>model=Sequential()
model.add(Dense(256,activation="relu",input_shape=(x_train.shape[1],)))
model.add(Dropout(0.3))
model.add(Dense(128,activation="relu"))
model.add(Dropout(0.3))
model.add(Dense(64,activation="relu"))
model.add(Dropout(0.3))
model.add(Dense(32,activation="sigmoid"))
```

Out [28]:

```
>>>model.summary()
```

Out [28]:

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 256)	29952
dropout (Dropout)	(None, 256)	0
dense_1 (Dense)	(None, 128)	32896
dropout_1 (Dropout)	(None, 128)	0
dense_2 (Dense)	(None, 64)	8256
dropout_2 (Dropout)	(None, 64)	0
dense_3 (Dense)	(None, 1)	65

Total params: 71,169
Trainable params: 71,169
Non-trainable params: 0

In [29]:

```
>>>model.compile(optimizer="adam",loss="binary_crossentropy",metrics=["accuracy"])
```

Out [29]:

```
>>>model.fit(x_train,y_train,validation_data=(x_test,y_test),epochs=100)
```

Out [30]:

```
>>>Epoch 1/100
116/116 [=====] - 7s 5ms/step - loss: nan - accuracy: 0.5005 - val_loss: nan - val_accuracy: 0.4984
Epoch 2/100
116/116 [=====] - 6s 5ms/step - loss: nan - accuracy: 0.5005 - val_loss: nan - val_accuracy: 0.4984
Epoch 3/100
116/116 [=====] - 6s 5ms/step - loss: nan - accuracy: 0.5005 - val_loss: nan - val_accuracy: 0.4984
Epoch 4/100
116/116 [=====] - 5s 4ms/step - loss: nan - accuracy: 0.5005 - val_loss: nan - val_accuracy: 0.4984
Epoch 5/100
116/116 [=====] - 4s 4ms/step - loss: nan - accuracy: 0.5005 - val_loss: nan - val_accuracy: 0.4984
Epoch 6/100
116/116 [=====] - 4s 4ms/step - loss: nan - accuracy: 0.5005 - val_loss: nan - val_accuracy: 0.4984
Epoch 7/100
116/116 [=====] - 4s 4ms/step - loss: nan - accuracy: 0.5005 - val_loss: nan - val_accuracy: 0.4984
Epoch 8/100
116/116 [=====] - 4s 4ms/step - loss: nan - accuracy: 0.5005 - val_loss: nan - val_accuracy: 0.4984
Epoch 9/100
116/116 [=====] - 4s 4ms/step - loss: nan - accuracy: 0.5005 - val_loss: nan - val_accuracy: 0.4984
Epoch 10/100
116/116 [=====] - 4s 4ms/step - loss: nan - accuracy: 0.5005 - val_loss: nan - val_accuracy: 0.4984
Epoch 11/100
116/116 [=====] - 4s 4ms/step - loss: nan - accuracy: 0.5005 - val_loss: nan - val_accuracy: 0.4984
Epoch 12/100
116/116 [=====] - 4s 4ms/step - loss: nan - accuracy: 0.5005 - val_loss: nan - val_accuracy: 0.4984
Epoch 13/100
116/116 [=====] - 4s 4ms/step - loss: nan - accuracy: 0.5005 - val_loss: nan - val_accuracy: 0.4984
Epoch 14/100
116/116 [=====] - 4s 4ms/step - loss: nan - accuracy: 0.5005 - val_loss: nan - val_accuracy: 0.4984
Epoch 15/100
116/116 [=====] - 4s 4ms/step - loss: nan - accuracy: 0.5005 - val_loss: nan - val_accuracy: 0.4984
Epoch 16/100
116/116 [=====] - 4s 4ms/step - loss: nan - accuracy: 0.5005 - val_loss: nan - val_accuracy: 0.4984
Epoch 17/100
116/116 [=====] - 4s 4ms/step - loss: nan - accuracy: 0.5005 - val_loss: nan - val_accuracy: 0.4984
Epoch 18/100
116/116 [=====] - 4s 4ms/step - loss: nan - accuracy: 0.5005 - val_loss: nan - val_accuracy: 0.4984
Epoch 19/100
116/116 [=====] - 4s 4ms/step - loss: nan - accuracy: 0.5005 - val_loss: nan - val_accuracy: 0.4984
Epoch 20/100
116/116 [=====] - 4s 4ms/step - loss: nan - accuracy: 0.5005 - val_loss: nan - val_accuracy: 0.4984
Epoch 21/100
116/116 [=====] - 4s 4ms/step - loss: nan - accuracy: 0.5005 - val_loss: nan - val_accuracy: 0.4984
Epoch 22/100
116/116 [=====] - 4s 4ms/step - loss: nan - accuracy: 0.5005 - val_loss: nan - val_accuracy: 0.4984
Epoch 23/100
116/116 [=====] - 4s 4ms/step - loss: nan - accuracy: 0.5005 - val_loss: nan - val_accuracy: 0.4984
Epoch 24/100
116/116 [=====] - 4s 4ms/step - loss: nan - accuracy: 0.5005 - val_loss: nan - val_accuracy: 0.4984
Epoch 25/100
116/116 [=====] - 4s 4ms/step - loss: nan - accuracy: 0.5005 - val_loss: nan - val_accuracy: 0.4984
Epoch 26/100
116/116 [=====] - 4s 4ms/step - loss: nan - accuracy: 0.5005 - val_loss: nan - val_accuracy: 0.4984
Epoch 27/100
116/116 [=====] - 4s 4ms/step - loss: nan - accuracy: 0.5005 - val_loss: nan - val_accuracy: 0.4984
Epoch 28/100
116/116 [=====] - 4s 4ms/step - loss: nan - accuracy: 0.5005 - val_loss: nan - val_accuracy: 0.4984
Epoch 29/100
116/116 [=====] - 4s 4ms/step - loss: nan - accuracy: 0.5005 - val_loss: nan - val_accuracy: 0.4984
Epoch 30/100
116/116 [=====] - 4s 4ms/step - loss: nan - accuracy: 0.5005 - val_loss: nan - val_accuracy: 0.4984
Epoch 31/100
116/116 [=====] - 4s 4ms/step - loss: nan - accuracy: 0.5005 - val_loss: nan - val_accuracy: 0.4984
Epoch 32/100
116/116 [=====] - 4s 4ms/step - loss: nan - accuracy: 0.5005 - val_loss: nan - val_accuracy: 0.4984
Epoch 33/100
116/116 [=====] - 4s 4ms/step - loss: nan - accuracy: 0.5005 - val_loss: nan - val_accuracy: 0.4984
Epoch 34/100
116/116 [=====] - 4s 4ms/step - loss: nan - accuracy: 0.5005 - val_loss: nan - val_accuracy: 0.4984
Epoch 35/100
116/116 [=====] - 4s 4ms/step - loss: nan - accuracy: 0.5005 - val_loss: nan - val_accuracy: 0.4984
Epoch 36/100
116/116 [=====] - 4s 4ms/step - loss: nan - accuracy: 0.5005 - val_loss: nan - val_accuracy: 0.4984
Epoch 37/100
116/116 [=====] - 4s 4ms/step - loss: nan - accuracy: 0.5005 - val_loss: nan - val_accuracy: 0.4984
Epoch 38/100
116/116 [=====] - 4s 4ms/step - loss: nan - accuracy: 0.5005 - val_loss: nan - val_accuracy: 0.4984
Epoch 39/100
116/116 [=====] - 4s 4ms/step - loss: nan - accuracy: 0.5005 - val_loss: nan - val_accuracy: 0.4984
Epoch 40/100
116/116 [=====] - 4s 4ms/step - loss: nan - accuracy: 0.5005 - val_loss: nan - val_accuracy: 0.4
```