In [1]:

```
import pandas as pd
import numpy as np
import sklearn
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_absolute_error
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeRegressor
```

In [2]:

```
df = pd.read_csv(r'C:\Users\User\Desktop\Train(DSN hackaton).csv')
```

In [3]:

```
df.dropna(axis=0, thresh=35, subset=None, inplace=True)
```

In [4]:

```
df['form_field6'] = df['form_field6'].fillna(0)
df['form_field8'] = df['form_field8'].fillna(0)
```

In [5]:

```
del df['Applicant_ID']
```

In [6]:

```
df['form_field1'] = df['form_field1'].fillna(df['form_field1'].mean())
df['form_field2'] = df['form_field2'].fillna(df['form_field2'].mean())
df['form_field7'] = df['form_field7'].fillna(df['form_field7'].mean())
df['form_field9'] = df['form_field9'].fillna(df['form_field9'].mean())
df['form_field11'] = df['form_field11'].fillna(df['form_field11'].mean())
df['form_field12'] = df['form_field12'].fillna(df['form_field12'].mean())
df['form_field13'] = df['form_field13'].fillna(df['form_field13'].mean())
df['form_field15'] = df['form_field1'].fillna(df['form_field15'].mean())
df['form_field16'] = df['form_field16'].fillna(df['form_field16'].mean())
df['form_field17'] = df['form_field17'].fillna(df['form_field17'].mean())
df['form_field18'] = df['form_field18'].fillna(df['form_field18'].mean())
df['form_field21'] = df['form_field21'].fillna(df['form_field21'].mean())
df['form_field22'] = df['form_field22'].fillna(df['form_field22'].mean())
df['form_field23'] = df['form_field23'].fillna(df['form_field23'].mean())
df['form_field24'] = df['form_field24'].fillna(df['form_field24'].mean())
df['form_field25'] = df['form_field25'].fillna(df['form_field25'].mean())
df['form_field26'] = df['form_field26'].fillna(df['form_field26'].mean())
df['form_field27'] = df['form_field27'].fillna(df['form_field27'].mean())
df['form_field30'] = df['form_field30'].fillna(0)
df['form_field37'] = df['form_field37'].fillna(df['form_field37'].mean())
df['form_field32'] = df['form_field32'].fillna(df['form_field32'].mean())
df['form_field35'] = df['form_field35'].fillna(df['form_field35'].mean())
df['form_field36'] = df['form_field36'].fillna(df['form_field36'].mean())
df['form_field39'] = df['form_field39'].fillna(df['form_field39'].mean())
df['form_field42'] = df['form_field42'].fillna(df['form_field42'].mean())
df['form_field43'] = df['form_field43'].fillna(df['form_field43'].mean())
df['form_field44'] = df['form_field44'].fillna(df['form_field44'].mean())
df['form_field45'] = df['form_field45'].fillna(df['form_field45'].mean())
df['form_field46'] = df['form_field46'].fillna(df['form_field46'].mean())
df['form_field50'] = df['form_field50'].fillna(df['form_field50'].mean())
```

In [7]:

```
df.tail()
```

Out[7]:

	form_field1	form_field2	form_field3	form_field4	form_field5	form_field6	form_field7	fo
55995	3740.0	0.01730	0.0000	0.0000	0.0	770998.0	9637475.0	4
55996	3360.0	2.01145	0.6252	0.0000	0.0	0.0	927765.0	
55997	3500.0	0.76640	0.0000	0.0000	0.0	118645.0	3662435.0	3
55998	3280.0	0.05235	2.0916	2.2212	0.0	0.0	3458599.0	
55999	3522.0	0.46930	0.0000	0.0000	0.0	98806.0	2053920.0	

5 rows × 51 columns

In [8]:

```
df['form_field48'] = df['form_field48'].fillna(df['form_field48'].mean())
```

In [9]:

```
del df['form_field33']
del df['form_field31']
del df['form_field40']
del df['form_field41']
```

In [10]:

```
df.replace(('yes', 'no'), (1, 0), inplace = True)
```

In [11]:

```
del df['form_field47']
df.head()
```

Out[11]:

	form_field1	form_field2	form_field3	form_field4	form_field5	form_field6	form_field7	form_fi
0	3436.0	0.28505	1.6560	0.0	0.000	0.0	10689720.0	2520
1	3456.0	0.67400	0.2342	0.0	0.000	0.0	898979.0	4975
2	3276.0	0.53845	3.1510	0.0	6.282	0.0	956940.0	
3	3372.0	0.17005	0.5050	0.0	0.000	192166.0	3044703.0	3854
4	3370.0	0.77270	1.1010	0.0	0.000	1556.0	214728.0	2147

5 rows × 46 columns

In [12]:

```
features = ['form_field1','form_field2', 'form_field3', 'form_field4','form_field5','form_
field6','form_field7','form_field8','form_field9','form_field10',
'form_field12','form_field13', 'form_field14', 'form_field15', 'form_field16', 'form_field
17', 'form_field18', 'form_field19', 'form_field20', 'form_field21', 'form_field22',
 'form_field23', 'form_field24', 'form_field25', 'form_field26', 'form_field27', 'form_fie
ld28', 'form_field29', 'form_field30', 'form_field32',
 'form_field34', 'form_field35', 'form_field36', 'form_field37', 'form_field38', 'form_fie
ld39', 'form_field42', 'form_field43', 'form_field44', 'form_field45',
 'form_field46', 'form_field48','form_field49','form_field50']
X = df[features]
```

In [13]:

```
y = df['default_status']
```

In [14]:

```
from sklearn.model_selection import train_test_split
train_X, test_X, train_y, test_y = train_test_split(X, y, test_size = 0.3, random_state =
42)
```

In [15]:

```
from sklearn.naive_bayes import GaussianNB
gnb = GaussianNB()
model = gnb.fit(train_X, train_y)
```

In [16]:

```
preds = gnb.predict_proba(test_X)
print(preds)
```

```
[[9.99999999e-01 1.01947028e-09]
 [2.14135192e-01 7.85864808e-01]
 [9.99999987e-01 1.33397231e-08]
 ...
 [1.00000000e+00 6.54331469e-49]
 [5.03578717e-03 9.94964213e-01]
 [2.30219955e-04 9.99769780e-01]]
```

In [17]:

```
model.predict_proba(test_X)[:,1]
```

Out[17]:

```
array([1.01947028e-09, 7.85864808e-01, 1.33397231e-08, ...,
       6.54331469e-49, 9.94964213e-01, 9.99769780e-01])
```

In [ ]:

```

```

In [18]:

```
import pandas as pd
import numpy as np
import sklearn
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_absolute_error
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeRegressor
```

In [19]:

```
df2 = pd.read_csv(r'C:\Users\User\Desktop\Test(DSN hackaton).csv')
```

In [20]:

```
df2['form_field6'] = df2['form_field6'].fillna(0)
df2['form_field8'] = df2['form_field8'].fillna(0)
```

In [21]:

```
df2.head(3)
```

Out[21]:

	Applicant_ID	form_field1	form_field2	form_field3	form_field4	form_field5	form_field6	forr
0	Apcnt_1000032	3236.0	0.34875	10.2006	0.0000	0.0	418564.0	4
1	Apcnt_1000048	3284.0	1.27360	2.9606	9.0198	0.0	0.0	98
2	Apcnt_1000052	NaN	0.27505	0.0600	0.0000	0.0	0.0	

3 rows × 51 columns

In [22]:

```
x = df2.mean()
df2 = df2.fillna(x)
```

In [23]:

```
df2.replace(('yes', 'no'), (1, 0), inplace = True)
```

In [24]:

```
features = ['form_field1','form_field2', 'form_field3', 'form_field4','form_field5','form_
field6','form_field7','form_field8','form_field9','form_field10',
'form_field12','form_field13', 'form_field14', 'form_field15', 'form_field16', 'form_field
17', 'form_field18', 'form_field19', 'form_field20', 'form_field21', 'form_field22',
 'form_field23', 'form_field24', 'form_field25', 'form_field26', 'form_field27', 'form_fie
ld28', 'form_field29', 'form_field30', 'form_field32',
 'form_field34', 'form_field35', 'form_field36', 'form_field37', 'form_field38', 'form_fie
ld39', 'form_field42', 'form_field43', 'form_field44', 'form_field45',
 'form_field46', 'form_field48','form_field49','form_field50']
X_test = df2[features]
```

In [25]:

```
preds_test = gnb.predict_proba(X_test)
print(preds_test)
```

```
[[1.00000000e+00 7.37484248e-34]
 [1.00000000e+00 1.71698085e-20]
 [1.00000000e+00 6.93885809e-34]
 ...
 [1.00000000e+00 8.39624765e-34]
 [1.00000000e+00 2.30280258e-33]
 [1.00000000e+00 5.47758463e-34]]
```

In [26]:

```
model.predict_proba(X_test)[:,1]
```

Out[26]:

```
array([7.37484248e-34, 1.71698085e-20, 6.93885809e-34, ...,
       8.39624765e-34, 2.30280258e-33, 5.47758463e-34])
```

In [27]:

```
Submission = pd.read_csv(r'C:\Users\User\Desktop\SampleSubmission.csv')
Submission['Applicant_ID'] = df2['Applicant_ID']
Submission['default_ID'] = model.predict_proba(X_test)[:,1]
```

In [28]:

```
Submission.to_csv(r'C:\Users\User\Desktop\Submission.csv', index = False)
```