Data Wrangling

The source of my data was https://www.kaggle.com/rikdifos/credit-card-approval-prediction It was not missing any data and had 24 variables, that identified demographics of 5000 customers.

In [1]:
```
import os
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels.api as sm
from statsmodels.graphics.api import abline_plot
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.model_selection import train_test_split
from sklearn import linear_model, preprocessing
from sklearn.model_selection import train_test_split
import warnings
from statsmodels.stats.multicomp import pairwise_tukeyhsd
from statsmodels.stats.multicomp import MultiComparison
cwd = os.getcwd()
print(cwd)
import statsmodels.api as sm

import scipy.stats as stats
import statsmodels.formula.api as smf
from statsmodels.formula.api import ols
from sklearn import preprocessing
from matplotlib import pyplot
from sklearn.metrics import precision_recall_curve
from sklearn.metrics import f1_score
from sklearn.metrics import auc
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report,confusion_matrix,roc_curve,roc_auc_score
from sklearn.metrics import accuracy_score,log_loss
from matplotlib import pyplot
from sklearn.ensemble import RandomForestClassifier
import itertools
from sklearn import svm, datasets
from sklearn.metrics import confusion_matrix
from sklearn.linear_model import LinearRegression
from sklearn import metrics
from tabulate import tabulate
```

```
C:\Users\Tom\Documents\GitHub
```

Reading Data into a Dataframe

In [2]:
```
loans = pd.read_csv("capstone3.csv")
```

In [3]:
```
loans
```

Out[3]:

	Customer_ID	Status_Checking_Acc	Duration_in_Months	Credit_History	Purpose_Credit_Taken	Credit_Amount	Savings_Acc	Years_At_Prese
0	100001	A11	6	A34	A43	1169	A65	
1	100002	A12	48	A32	A43	5951	A61	
2	100003	A14	12	A34	A46	2096	A61	
3	100004	A11	42	A32	A42	7882	A61	
4	100005	A11	24	A33	A40	4870	A61	
...	...	...	...	...	...	...	...	
4995	104996	A14	12	A32	A42	1736	A61	
4996	104997	A11	30	A32	A41	3857	A61	
4997	104998	A14	12	A32	A43	804	A61	
4998	104999	A11	45	A32	A43	1845	A61	
4999	105000	A12	45	A34	A41	4576	A62	

5000 rows × 23 columns

DataTypes

Description

Here are the different variables we have available for each of the 5000 observations. The presentation of those variables is more convenient for human reading and manipulation. The values of the variables are described on the attached file. We will be changing the values to the actual description factors. Out of the 24 variables, 10 of those are numerical, while the other 14 are qualitative, most of them are ordinal, and thiscreates some issues with the dummies variables that we have to add, but we have no missing values and the entire dataset of 5000 observations will be used.

In [4]:
```
loans.dtypes
```

Out[4]:
```
Customer_ID                  int64
Status_Checking_Acc         object
Duration_in_Months           int64
Credit_History              object
Purpose_Credit_Taken        object
Credit_Amount                int64
Savings_Acc                 object
Years_At_Present_Employment object
Inst_Rt_Income               int64
Marital_Status_Gender       object
Other_Debtors_Guarantors    object
Current_Address_Yrs          int64
Property                    object
Age                          int64
Other_Inst_Plans            object
Housing                     object
Num_CC                       int64
Job                         object
Dependents                   int64
Telephone                   object
Foreign_Worker              object
Default_On_Payment           int64
Count                        int64
dtype: object
```

Change of values from identifiers to descriptors.

In [5]:
```
loans['Status_Checking_Acc'] = loans['Status_Checking_Acc'].replace(['A11', 'A12', 'A13', 'A14' ], ['Neg', 'Low
```

In [6]:
```
loans['Credit_History'] = loans['Credit_History'].replace(['A30', 'A31', 'A32', 'A33', 'A34' ], ['NoCredit', 'P
```

In [7]:
```
loans['Purpose_Credit_Taken'] = loans['Purpose_Credit_Taken'].replace(['A40', 'A41', 'A42', 'A43', 'A44', 'A45'
```

In [8]:
```
loans['Savings_Acc'] = loans['Savings_Acc'].replace(['A61', 'A62', 'A63', 'A64', 'A65'], ['Low', 'LowAvg', 'Avg
```

In [9]:
```
loans['Years_At_Present_Employment'] = loans['Years_At_Present_Employment'].replace(['A71', 'A72', 'A73', 'A74'
```

In [10]:
```
loans['Marital_Status_Gender'] = loans['Marital_Status_Gender'].replace(['A91', 'A92', 'A93', 'A94', 'A95' ], [
```

In [11]:
```
loans['Other_Debtors_Guarantors'] = loans['Other_Debtors_Guarantors'].replace(['A101', 'A102', 'A103'], ['No',
```

In [12]:
```
loans['Property'] = loans['Property'].replace(['A121', 'A122', 'A123', 'A124' ], ['RealEstate', 'LifeInsurance'
```

In [13]:
```
loans['Other_Inst_Plans'] = loans['Other_Inst_Plans'].replace(['A141', 'A142', 'A143'], ['Bank', 'Store', 'None
```

In [14]:
```
loans['Job'] = loans['Job'].replace(['A171', 'A172', 'A173', 'A174' ], ['NoSkills_NoRes', 'NoSkills_Res', 'LowS
```

In [15]:
```
loans['Telephone'] = loans['Telephone'].replace(['A191', 'A192'], ['No', 'Yes'])
```

In [16]:
```
loans['Foreign_Worker'] = loans['Foreign_Worker'].replace(['A201', 'A202'], ['Yes', 'No'])
```

In [17]:
```
loans['Housing'] = loans['Housing'].replace(['A151', 'A152', 'A153'], ['Rent', 'Own', 'Free'])
```

Addition of Response Variable

We are forced to add another variable for the Default customers, as we will need both a numerical and a categorical for the needs of the modeling and the graphing

In [18]:
```
loans['Default_On_Payment2'] = loans['Default_On_Payment']
loans['Default_On_Payment2'] = loans['Default_On_Payment2'].astype(int).astype(str)
```

In [19]:
```
loans['Default_On_Payment2'] = loans['Default_On_Payment2'].replace(['1', '0'], ['Yes', 'No'])
```

In [20]:
```
loans
```

Out[20]:

	Customer_ID	Status_Checking_Acc	Duration_in_Months	Credit_History	Purpose_Credit_Taken	Credit_Amount	Savings_Acc	Years_At_Prese
0	100001	Neg	6	Critical	Radio_TV	1169	NoAcc	
1	100002	Low	48	Current	Radio_TV	5951	Low	
2	100003	NoAcc	12	Critical	Education	2096	Low	
3	100004	Neg	42	Current	Furniture	7882	Low	
4	100005	Neg	24	Delay	NewCar	4870	Low	
...	...	...	...	...	...	...	...	
4995	104996	NoAcc	12	Current	Furniture	1736	Low	
4996	104997	Neg	30	Current	UsedCar	3857	Low	
4997	104998	NoAcc	12	Current	Radio_TV	804	Low	
4998	104999	Neg	45	Current	Radio_TV	1845	Low	
4999	105000	Low	45	Critical	UsedCar	4576	LowAvg	

5000 rows × 24 columns

In [21]:
```
loans.head()
```

Out[21]:

	Customer_ID	Status_Checking_Acc	Duration_in_Months	Credit_History	Purpose_Credit_Taken	Credit_Amount	Savings_Acc	Years_At_Present_E
0	100001	Neg	6	Critical	Radio_TV	1169	NoAcc	
1	100002	Low	48	Current	Radio_TV	5951	Low	
2	100003	NoAcc	12	Critical	Education	2096	Low	
3	100004	Neg	42	Current	Furniture	7882	Low	
4	100005	Neg	24	Delay	NewCar	4870	Low	

5 rows × 24 columns

Saving changes

We created a new data file to work with in the next three phases of the project.

In [22]:
```
loans.to_csv('cleanloans.csv', index=False)
```

In [ ]:
```
```