



# FINANCIAL CAPABILITY PREDICTOR

SNEHAL D SASE

NIKHIL SUNKU



# CONTENT

- EDA
- MODELS
- WHY ML(MACHINE LEARNING)?
- NEED FOR BIG DATA
- CONCLUSION

# DATASET

```
df = pd.read_csv("BankLoan.csv", delimiter=',')  
# To display the top 5 rows  
df.head()
```

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT
0	100002	1	Cash loans	M	N	Y	0	202500.0	40659
1	100003	0	Cash loans	F	N	N	0	270000.0	129350
2	100004	0	Revolving loans	M	Y	Y	0	67500.0	13500
3	100006	0	Cash loans	F	N	Y	0	135000.0	31268
4	100007	0	Cash loans	M	N	Y	0	121500.0	51300

5 rows × 122 columns

# COLUMNS

```
df.columns
```

```
Index(['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_18', 'FLAG_DOCUMENT_19', 'FLAG_DOCUMENT_20',  
      'FLAG_DOCUMENT_21', 'AMT_REQ_CREDIT_BUREAU_HOUR',  
      'AMT_REQ_CREDIT_BUREAU_DAY', 'AMT_REQ_CREDIT_BUREAU_WEEK',  
      'AMT_REQ_CREDIT_BUREAU_MON', 'AMT_REQ_CREDIT_BUREAU_QRT',  
      'AMT_REQ_CREDIT_BUREAU_YEAR'],  
      dtype='object', length=122)
```

```
df.shape
```

```
(307511, 122)
```

```
print(len(df.columns))
```

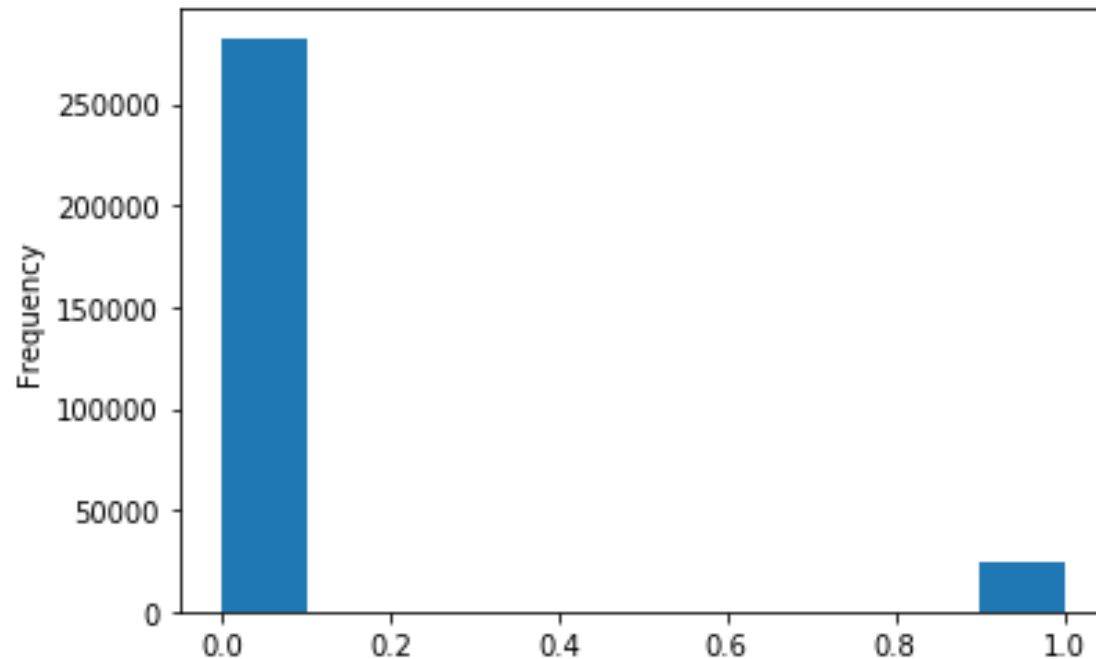
```
122
```

```
print(len(df.index))
```

```
307511
```

- Using these columns, we are going to train our Machine Learning Model to predict if the customers are able to repay back the loan .It is very important that this column have same amount of data otherwise the prediction will be biased which is not expected ,we have to solve this problem as we can see from the graph data is not evenly distributed.

```
Out[8]: <matplotlib.axes._subplots.AxesSubplot at 0x23aea8875c8>
```



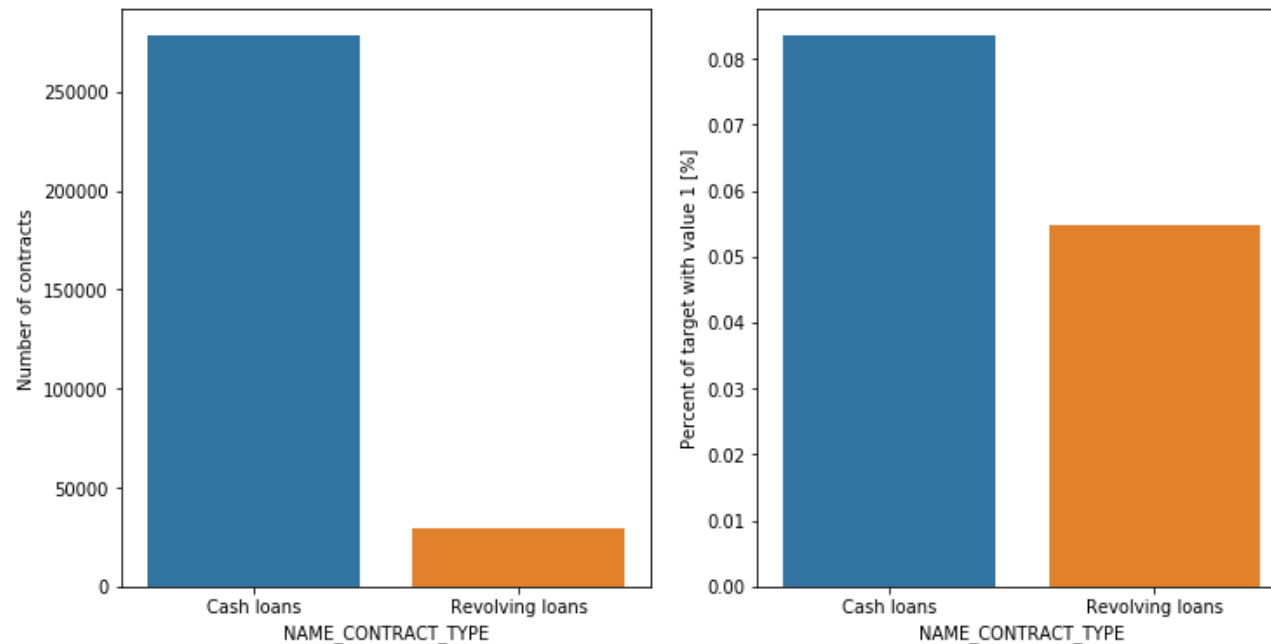
## 1 Categorical Feature

- We are finding out how many categorical features do we have and also how many categories do we need to have.
- This is very important step to find out because when we are going to do one hot encoding, we can find out after encoding how many new columns can we get and which columns we might need to train our machine learning model

```
NAME_CONTRACT_TYPE      2
CODE_GENDER              3
FLAG_OWN_CAR             2
FLAG_OWN_REALTY          2
NAME_TYPE_SUITE          7
NAME_INCOME_TYPE         8
NAME_EDUCATION_TYPE      5
NAME_FAMILY_STATUS       6
NAME_HOUSING_TYPE        6
OCCUPATION_TYPE          18
WEEKDAY_APPR_PROCESS_START 7
ORGANIZATION_TYPE       58
FONDKAPREMONT_MODE      4
HOUSETYPE_MODE           3
WALLSMATERIAL_MODE       7
EMERGENCYSTATE_MODE      2
dtype: int64
```

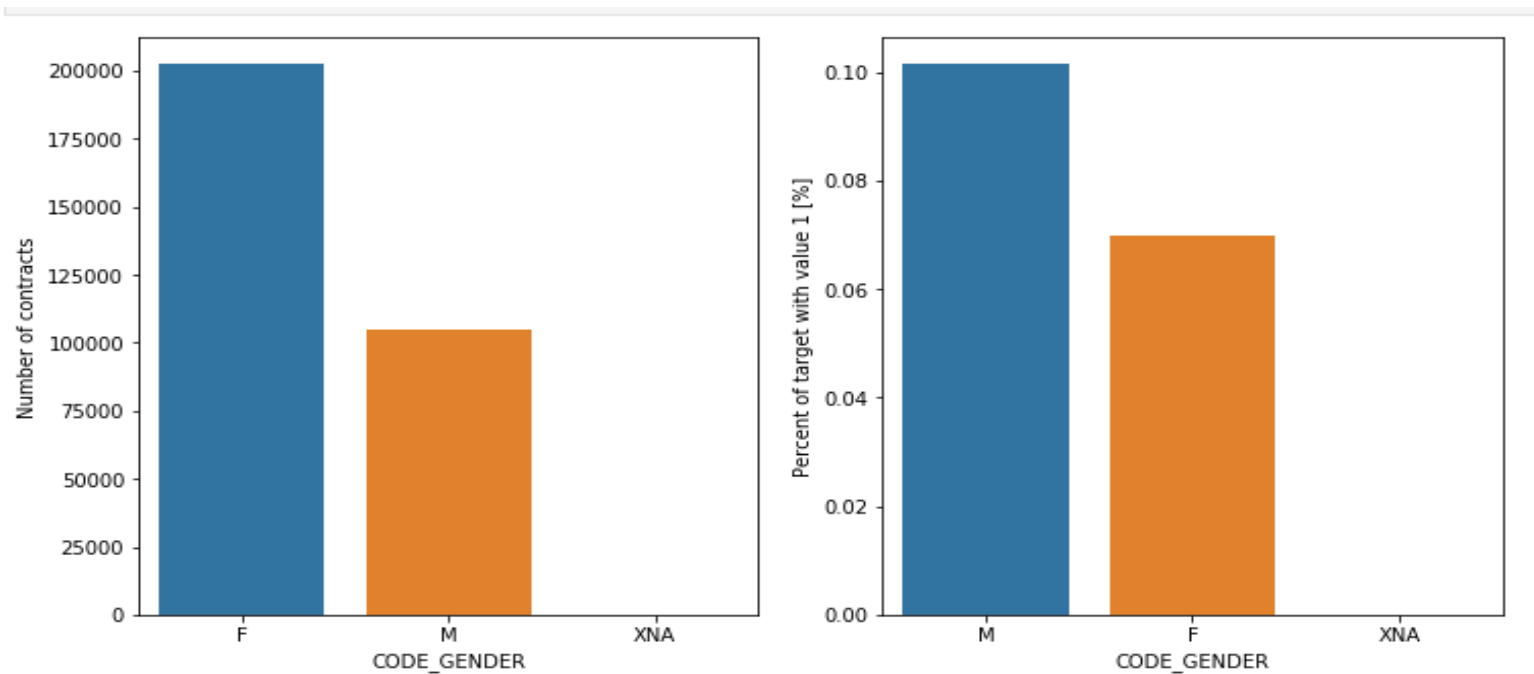
## • Loan Types

- From the historical data that we have ,we are finding out which different type of loan categories we possess. This will help us predict customer who applied for particular category of loan repays it back.
- Here we have two loan types Cash Loans and Revolving Loans.



## Gender

From this below column we will find which particular gender is applying for loan and which one among them are paying it back



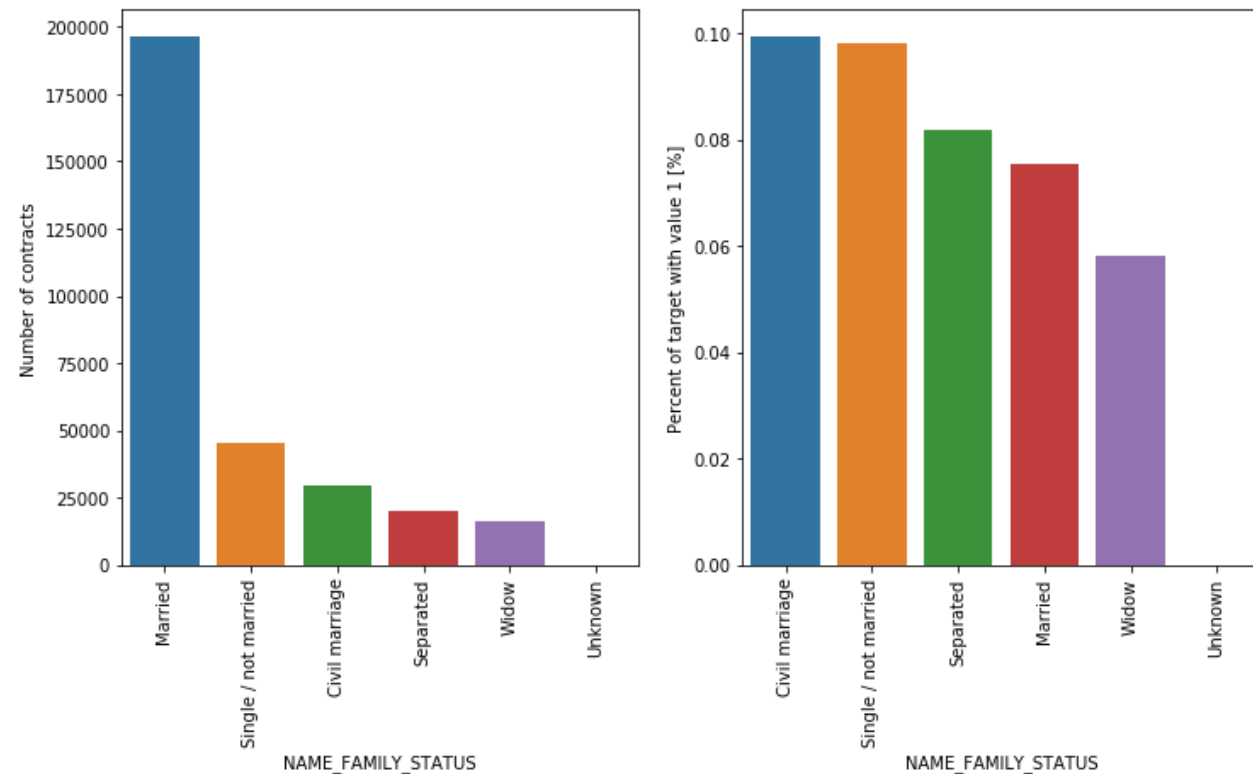


## Status

This column will help us to find how the family status of a person affects its ability of paying back the loan.

In [17]:

```
plot_stats('NAME_FAMILY_STATUS', True, True)
```

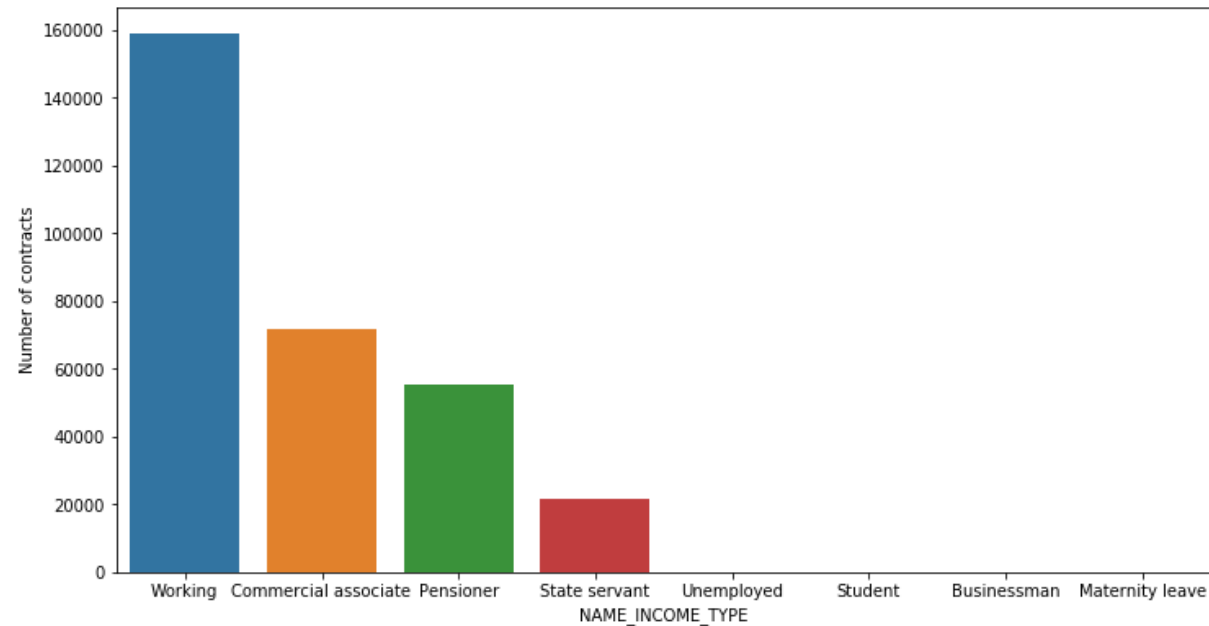


# 1 Income Status

For approval of bank loan its very important to find out the income of a person this column will highly effect ability of loan getting accepted or rejected.

In [20]:

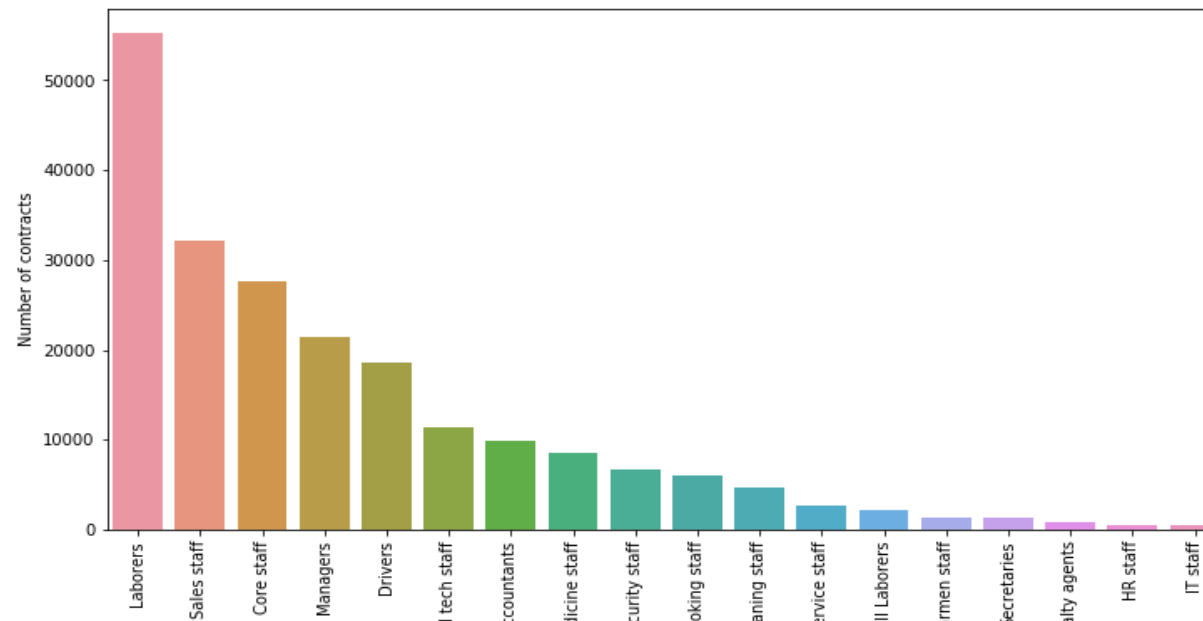
```
plot_stats('NAME_INCOME_TYPE',False,False)
```



## Occupation of a Person

The feature will also help us to find out repayment capability whether the individual is able to repay the loan, it totally depends on the several factors among one of them we consider is occupation, for example if a person is working as CEO of a company, he is very likely to repay the loan, so the acceptance of his loan application has greater chances.

```
In [21]: plot_stats('OCCUPATION_TYPE', True, False)
```

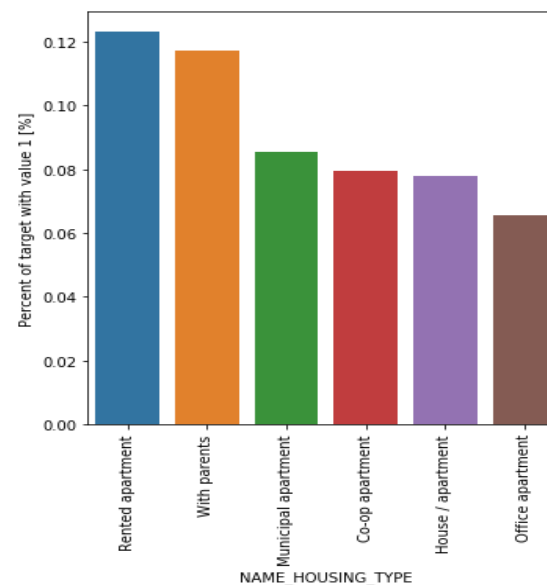
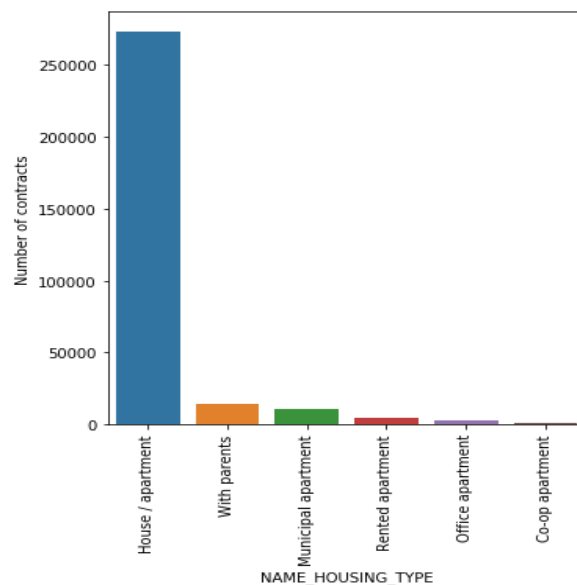


## House Type

This feature will also help up in deciding the ability of person to repay the loan. For example, if a person has a rental apartment this means he is incurring an expense monthly already which might affect his/her power of repaying the loan.

In [24]:


```
plot_stats('NAME_HOUSING_TYPE', True)
```

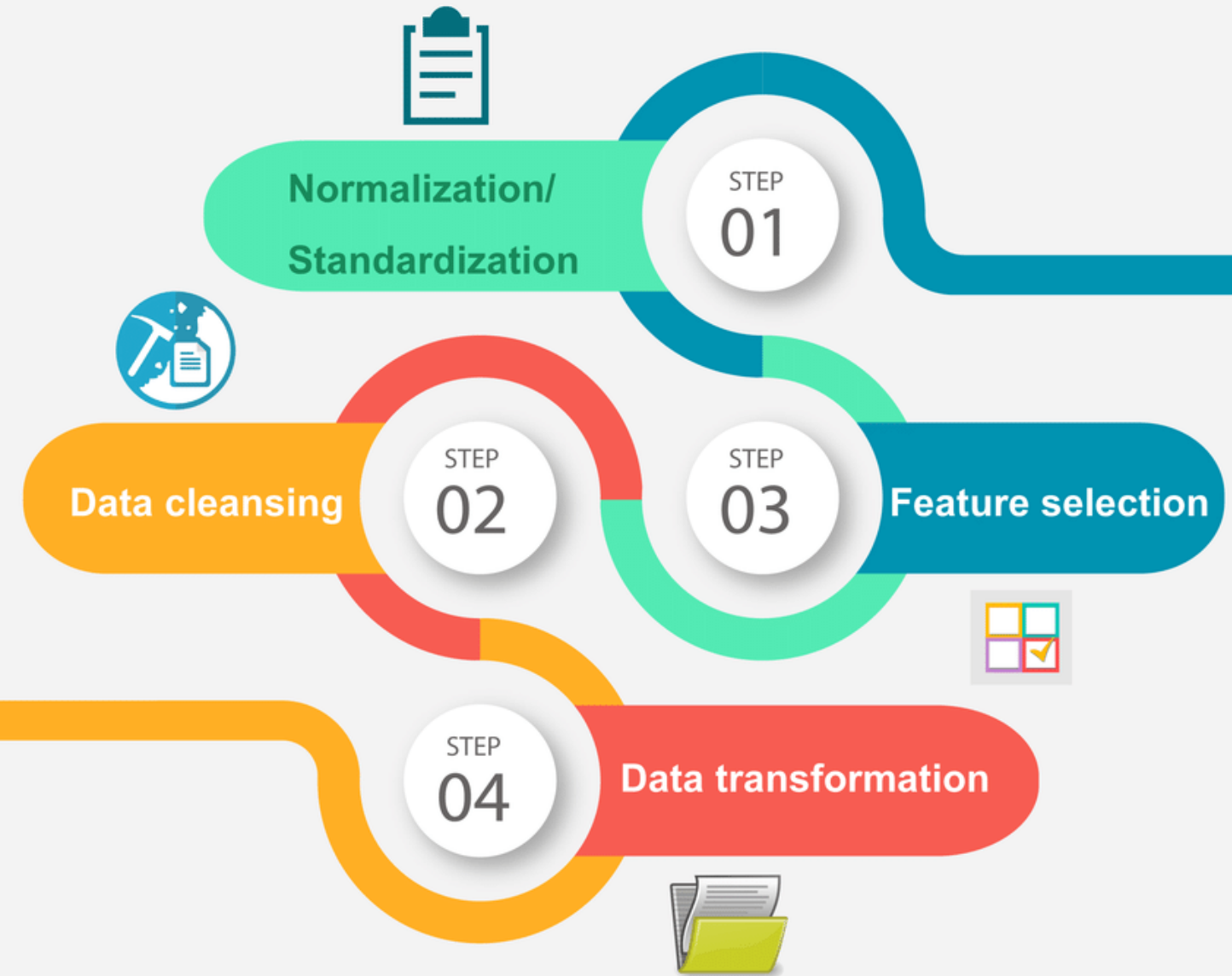




## Models

In this project we use three models to predict the ability of a customer to repay the loan. The problem we are working on is binary classification problem. We are going to use following three models.

1. Decision Tree
  2. Support Vector Machine
  3. Naïve Bayes
- 



## Why ML ?

- As the input data is large it might contain unstructured data
- So, we Use ML to Preprocess and Organize it properly

# Need for Big data

- As our data set has 307511 rows and 122 columns and each column would be occupying a lot of space. The traditional data processing will slow down our process a lot because one CPU core cannot process this large amount of data.
- We need to process data in parallel, Here we are going to use most famous big data processing framework which is Pyspark. Using Pyspark we will be able to process data in multi-threaded manner.
- The use of Pyspark would ease the process of parallel data processing as we will be utilizing the multiple CPU cores in parallel to execute different chunk of data and using Pyspark would optimize processing and enhance performance.

# CONCLUSION

- THE GOAL OF THIS PROJECT IS TO DEVELOP A BIG DATA ARCHITECTURE THAT WILL PREDICT A CUSTOMER THAT IS APPLYING FOR BANK LOAN IS CAPABLE ENOUGH TO REPAY THE LOAN OR NOT.
- EDA GIVES A CLEAR PICTURE ABOUT THE PAYING BACK INSIGHTS OF THE CUSTOMERS, WHICH WILL IN TURN HELP FINANCIAL INSTITUTIONS TO MAKE DECISION ABOUT CREDIT.



