# CREDIT_RISK_CUSTOMERS PROJECT

## *DOCUMENTATION*

- **Abstract/Executive Summary :-**

The Project is about Credit Risk of the Customer according to various attributes of the customer. We need to predict about the target Class depending on the features that whether he comes under good class or bad class. There are 1000 data points and 21 features in the dataset of the project. The EDA is done to find the relation of the various features with the target and predictions are done using various Machine Learning Algorithms. After performing all the EDA process and Machine Learning Algorithms, the following were the key findings:-

- The Best Model to predict about the credit risk in this project is Random Forest Random Search CV.
- Generally there are very less number of customers who are risky for granting the credit. The duration for most of the loans are between 10 and 25. The maximum customers have taken credit between 1000 to 5000. Customers between age of 20 and 30 are more in a need of credit.
- Customers who have high savings and more employment duration are very less risky for granting the credit and most of the customers who take credit for luxurious items.
- Higher Loan amount is generally cleared on time. Customer who are in middle age and those have less years of employment are generally prone towards the credit.

## ● Introduction :-

The project is of *Fintech industry*. It is a classification problem.

The following is the problem statement. We need to predict whether the customer is risky or not for providing the credit

***Problem Statement:-Predict if customers are risky or not for credit. There are 21 attributes of the customers a prediction needs to be done whether the customer is risky or not. The Target in the dataset is class, whether the customer comes under good or bad class.***

## ● Data Collection and Preprocessing :-

The data is collected from *Kaggle*. There are 1000 data points and 21 features in the dataset.

The following are the features of the dataset:-***This dataset classifies people described by a set of attributes as good or bad credit risks.***

- checking_status - Status of existing checking account
- duration - Duration in months
- credit_history - credits taken, paid back duly, delays, critical accounts
- purpose - Purpose of the credit
- credit_amount - Amount of credit
- savings_status - Status of savings account/bond
- employment - Present employment, in number of years
- installment_commitment - Installment rate in percentage of disposable income
- personal_status - sex and marital data
- other_parties - Other debtors / guarantors
- residence_since - How many years customer is residing.
- property_magnitude - Type of Property
- age – Age of the customer
- other_payment_plans – Any other Payment Plan
- housing – Customer has own or rented house or has free accomodation
- existing_credits – Number of existing credit
- job – skilled or unskilled worker
- num_dependents - number of dependants
- own_telephone – possession of telephone or not
- foreign_worker – Foreign worker or not
- class – Customer comer under which class for credit – good or bad

There were some inconsistencies in the data type of the features which was dealt and the data was made consistent. There were no null values in the data. Data was Standardised as well as Normalised using Scaling Techniques like Standard Scaler and MinMax Scaler. Encoding was also done so that the Machine Learning Models can give predictions with huge accuracy.

- **Methodology:-**

The problem statement was the classification problem statement so the first ML Algorithm which was used was Logistic Regression. Later the other Algorithms which where used were KNN, Decision Tree Classifier, Support Vector Classifier Random Forest Classifier and Boosting Algorithms. The Algorithms were Hyper Parameter Tuned as well so that the Accuracy of the Model be high. The Performance Metrics used is ACCURACY.

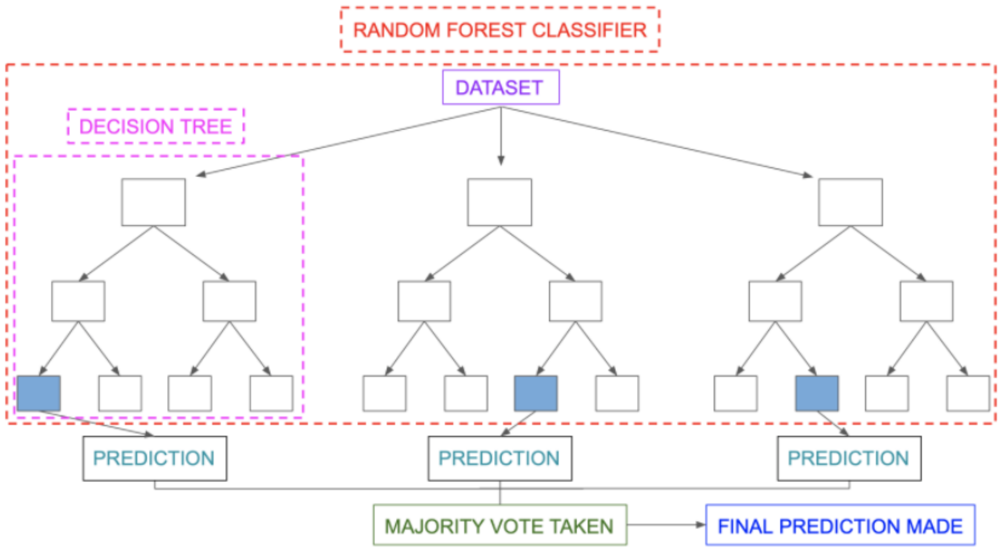
The Result of all the Machine Learning Algorithms are as follows:-

| Model | Training Accuracy | Testing Accuracy |
|---|---|---|
| Logistic Regression | 71.60 | 75.2 |
| KNN | 72.00 | 71.2 |
| SVC | 70.53 | 72.8 |
| Decision Tree | 100.00 | 70.4 |
| Random Forest | 84.90 | 82.4 |
| Cat Boost | 96.80 | 82.4 |
| XG Boost | 100.00 | 78.8 |
| Gardient Boost | 83.60 | 87.4 |

The above results are *Hyper Parameter Tuned*.

From the table we can see that the Best Fit Model is Random Forest Random Search CV.

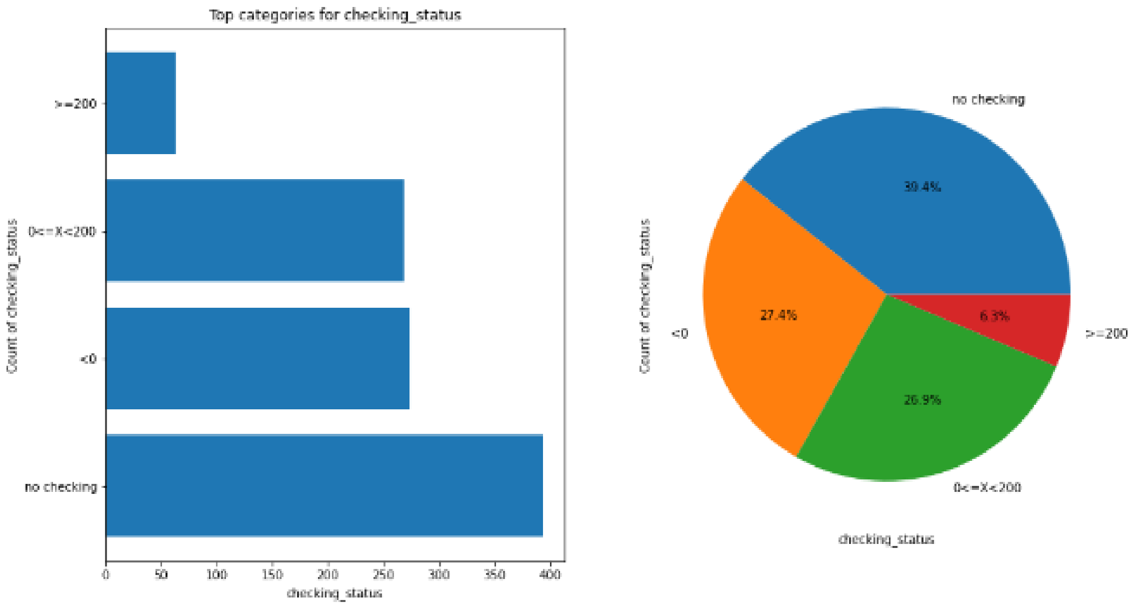
**FLOW CHART:-**

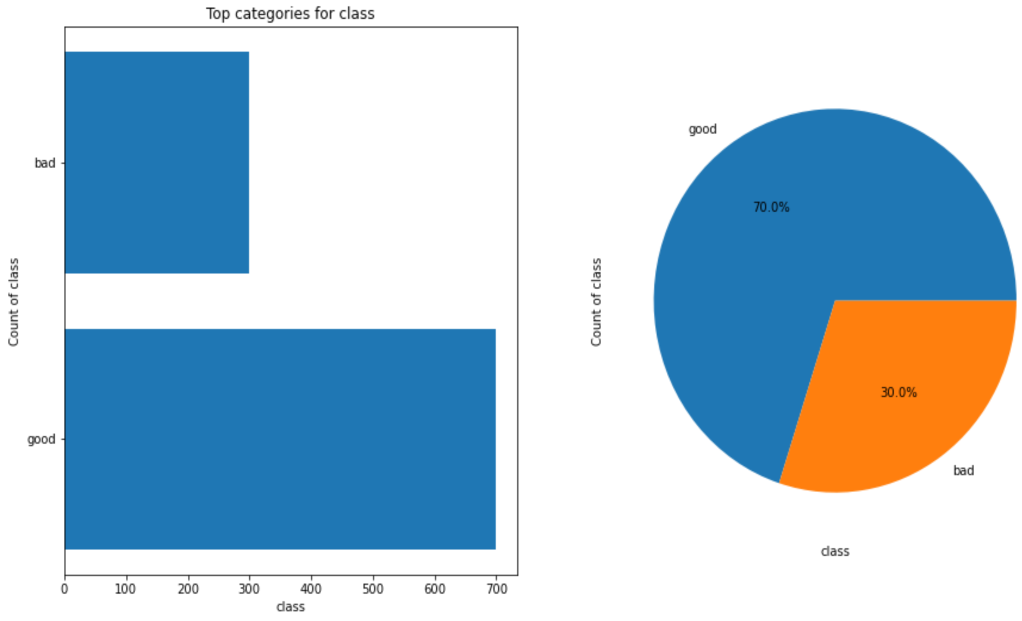


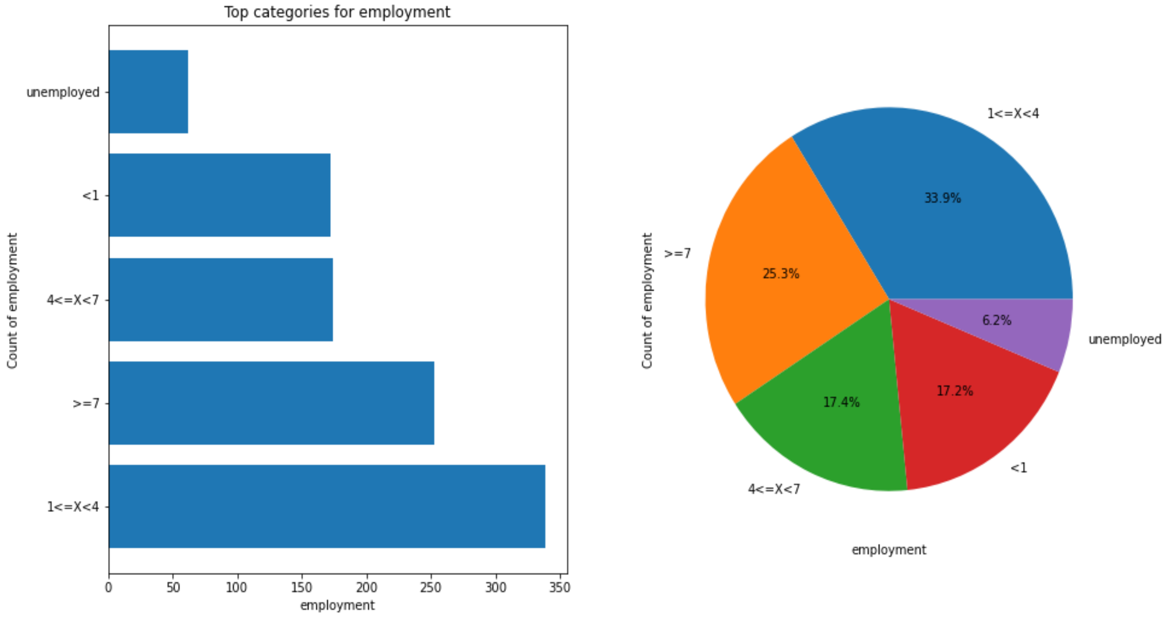
- **Results and Conclusion:-**

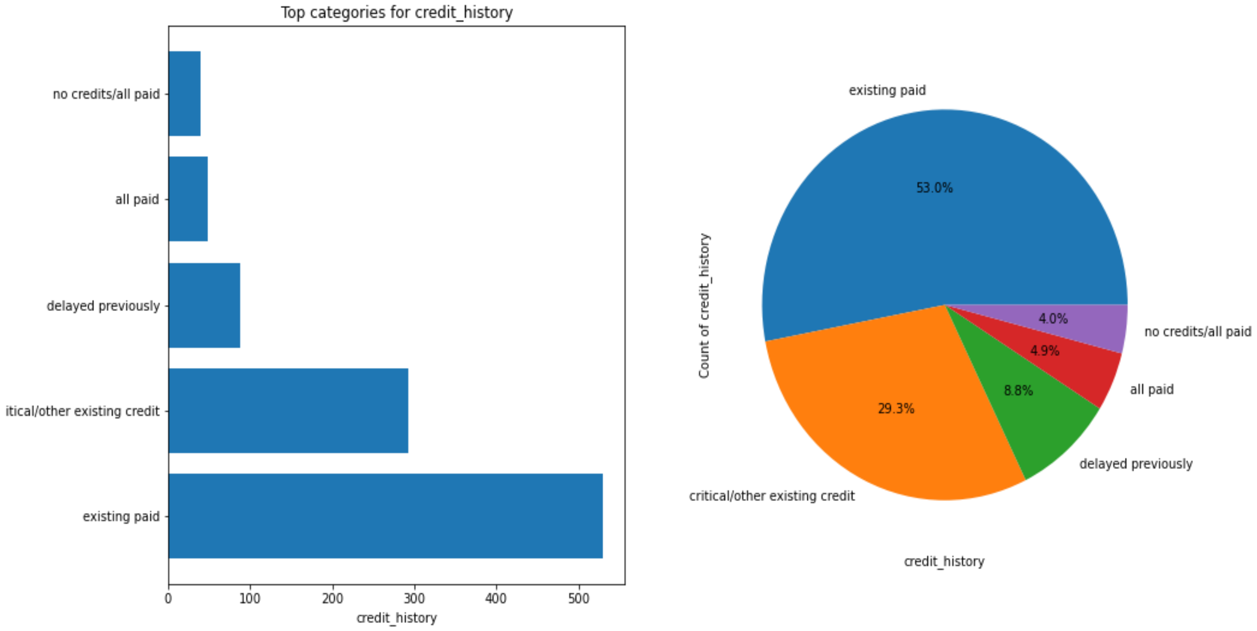
## *EDA RESULTS*

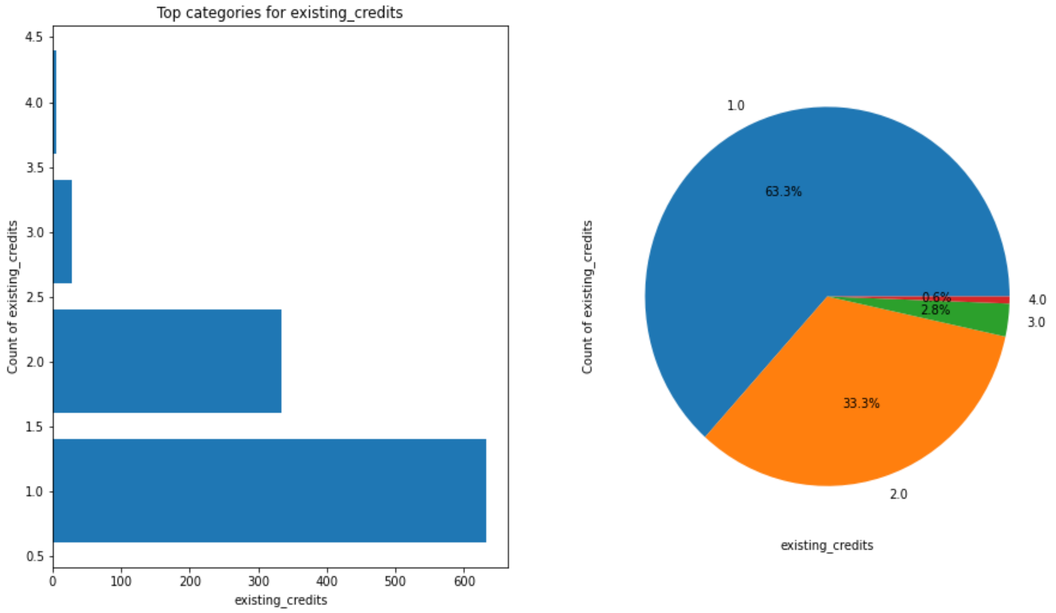
### ***Top categories for categorical features in the data :-***

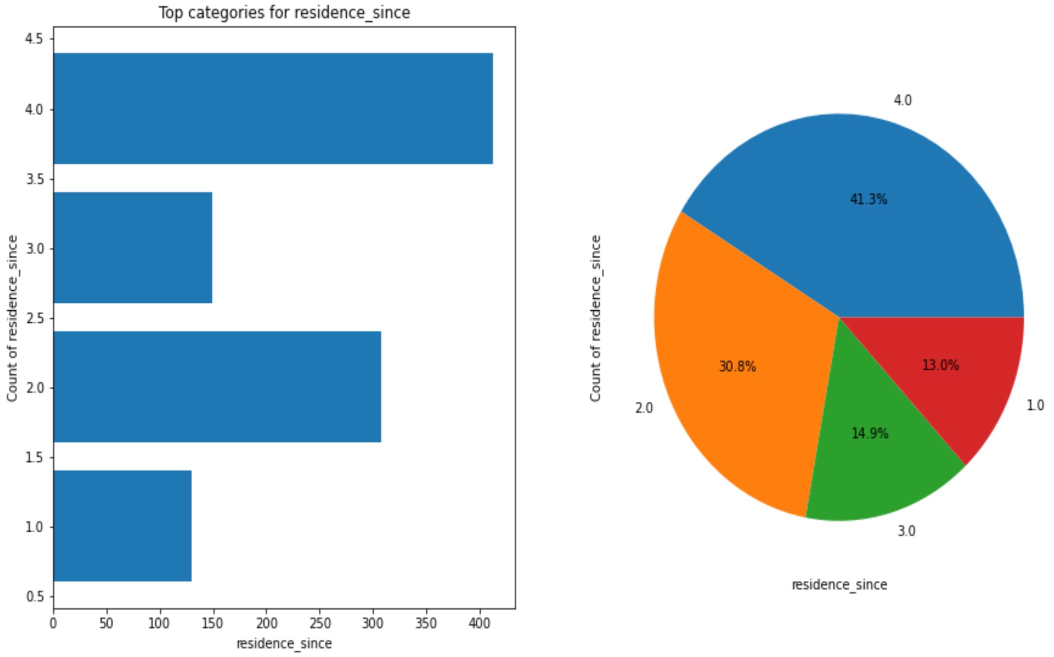


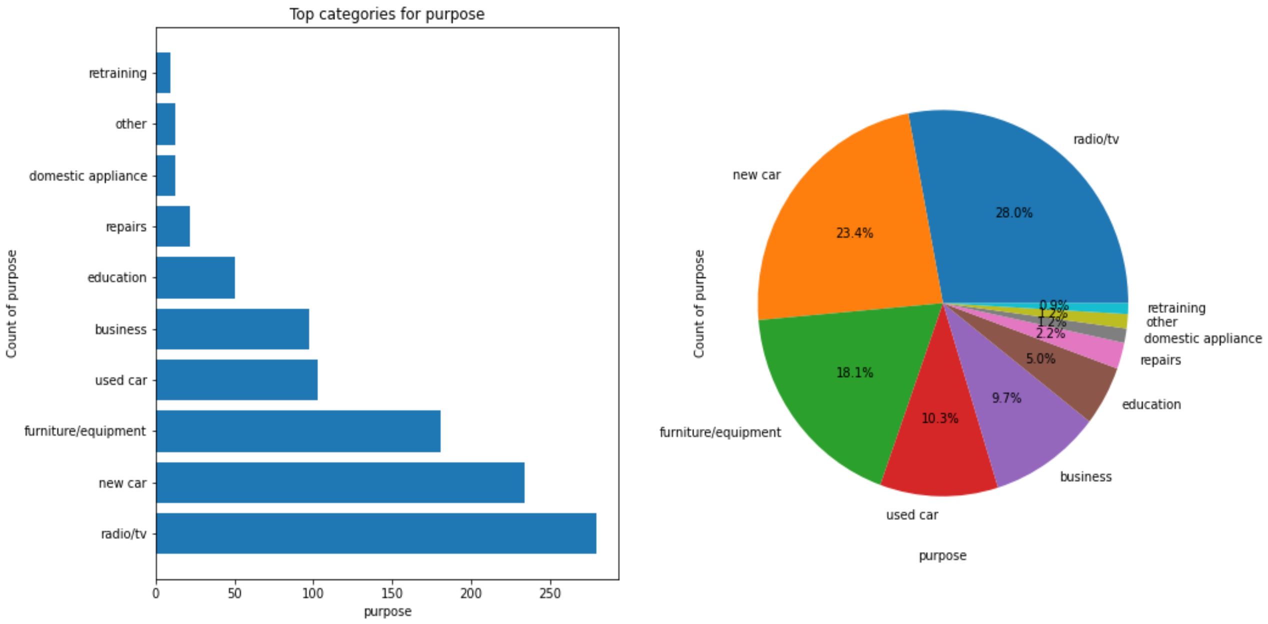


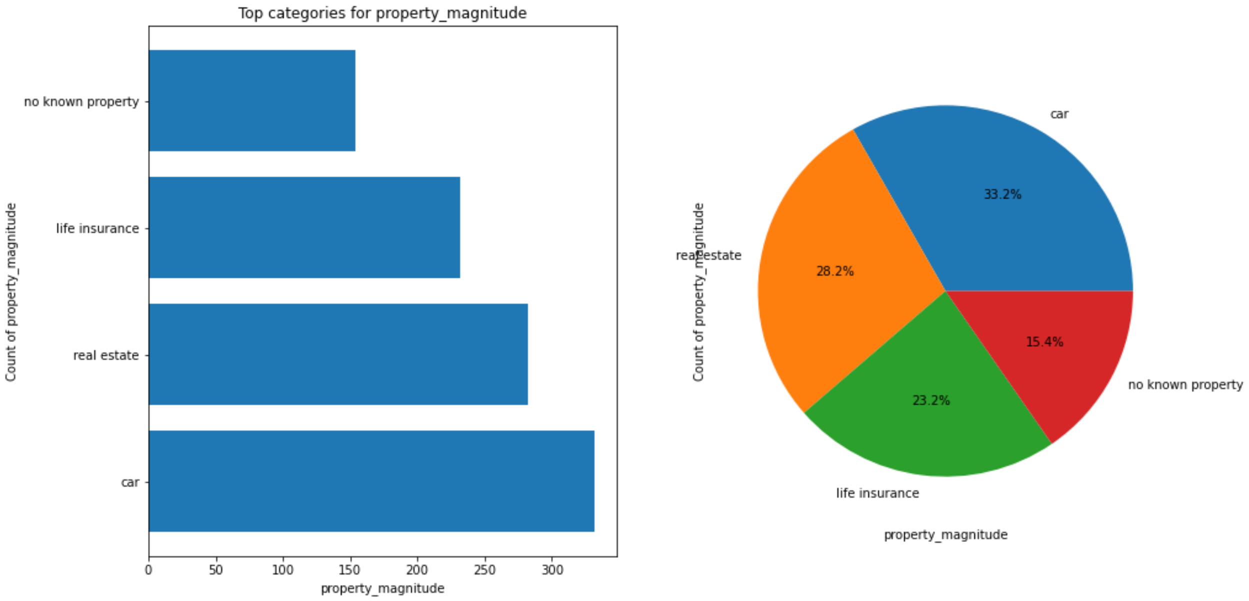

Top categories for employment
unemployed
<1
4<=X<7
>=7
1<=X<4
Count of employment
0
50
100
150
200
250
300
350
employment
1<=X<4
33.9%
>=7
25.3%
6.2%
unemployed
17.4%
17.2%
<1
4<=X<7
Count of employment
employment

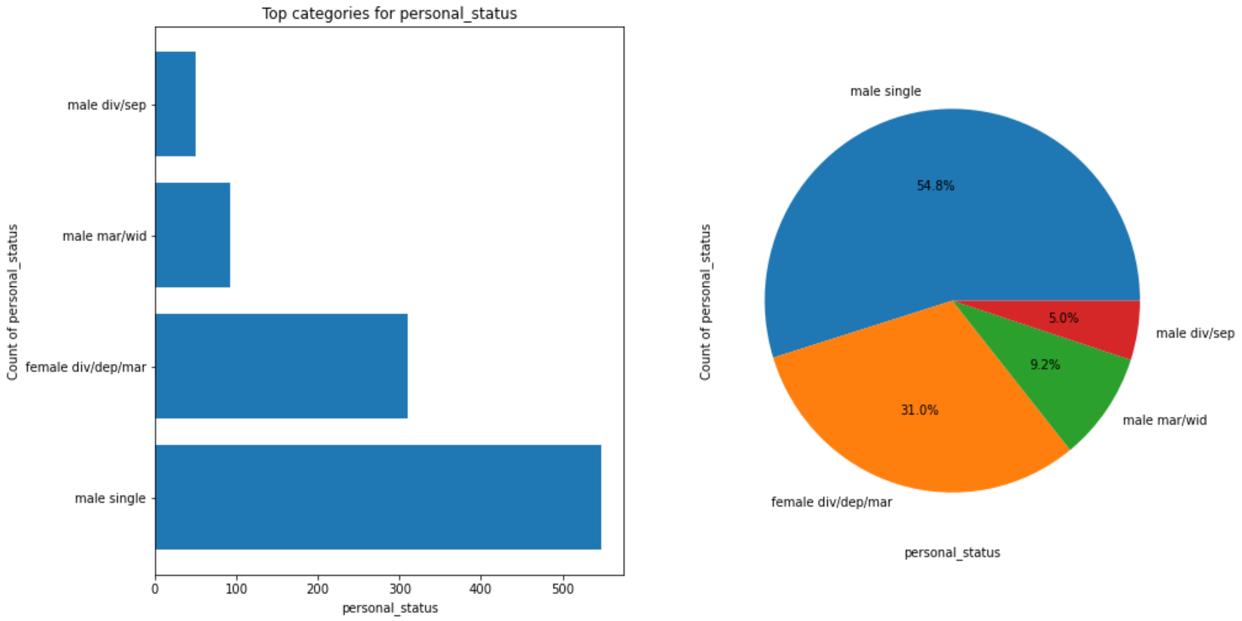

Top categories for credit_history
no credits/all paid
all paid
delayed previously
itical/other existing credit
existing paid
0
100
200
300
400
500
credit_history
existing paid
53.0%
4.0%
no credits/all paid
4.9%
all paid
8.8%
29.3%
delayed previously
critical/other existing credit
Count of credit_history
credit_history

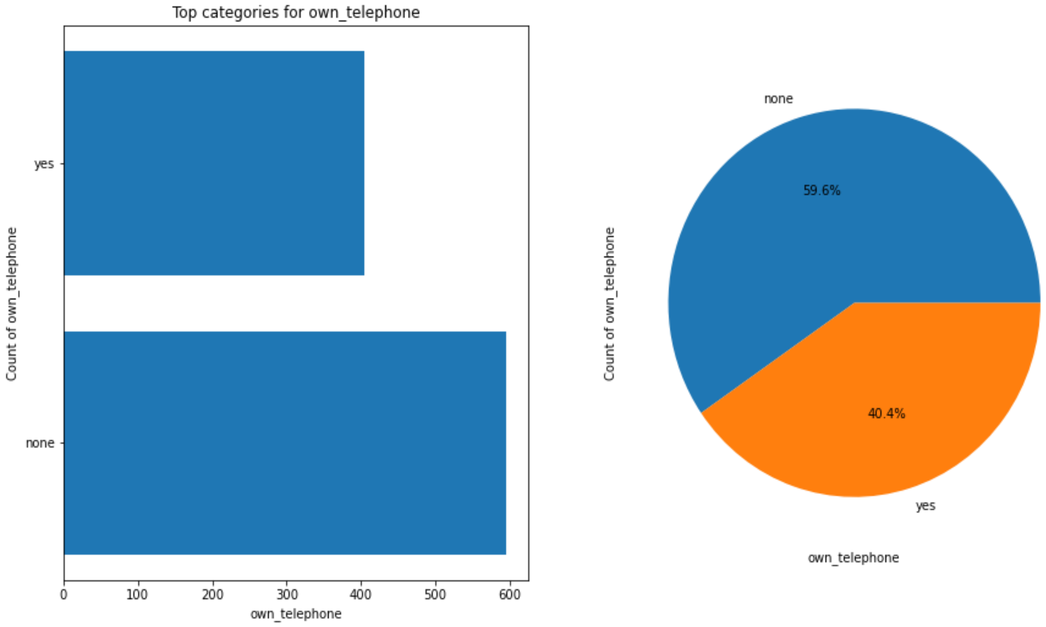

Top categories for existing_credits
Count of existing_credits
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0
100
200
300
400
500
600
existing_credits
1.0
63.3%
0.6%
2.8%
4.0
3.0
33.3%
2.0
existing_credits

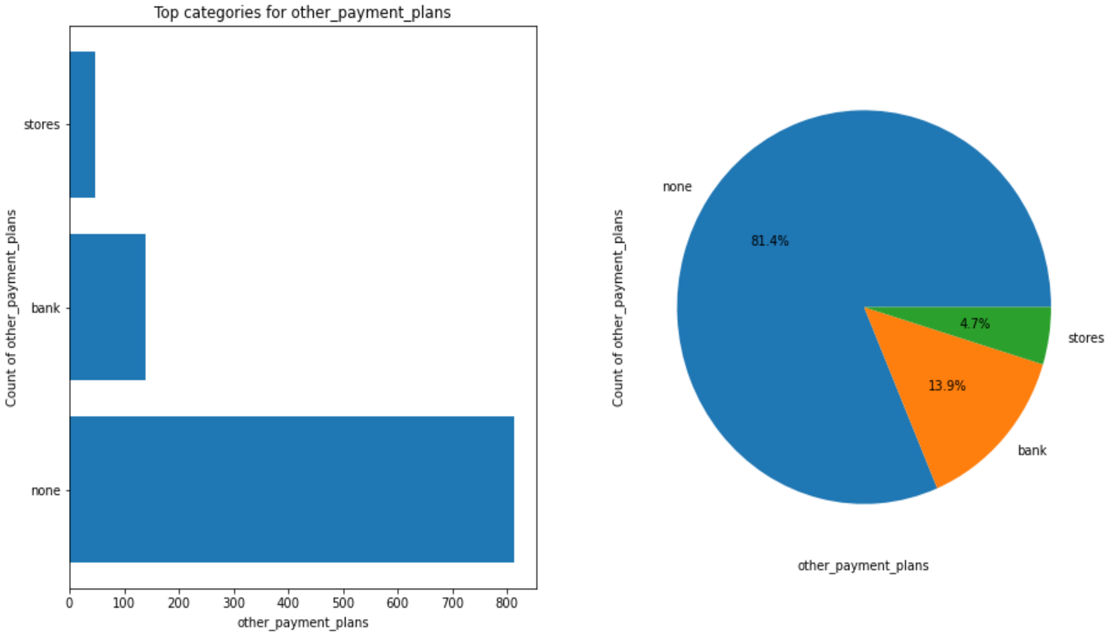

Top categories for residence_since
Count of residence_since
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0
50
100
150
200
250
300
350
400
residence_since
4.0
41.3%
30.8%
13.0%
2.0
14.9%
1.0
3.0
residence_since

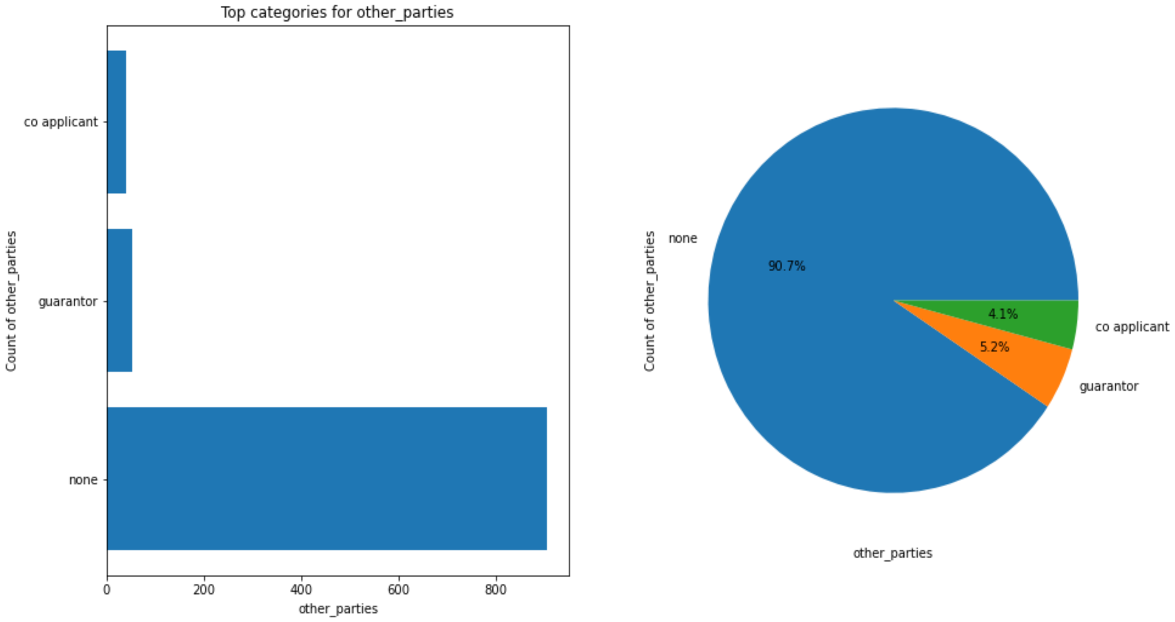

Top categories for purpose
Count of purpose
retraining
other
domestic appliance
repairs
education
business
used car
furniture/equipment
new car
radio/tv
0
50
100
150
200
250
purpose
Count of purpose
radio/tv
28.0%
new car
23.4%
furniture/equipment
18.1%
used car
10.3%
business
9.7%
education
5.0%
repairs
2.2%
domestic appliance
1.2%
other
1.2%
retraining
0.9%
purpose

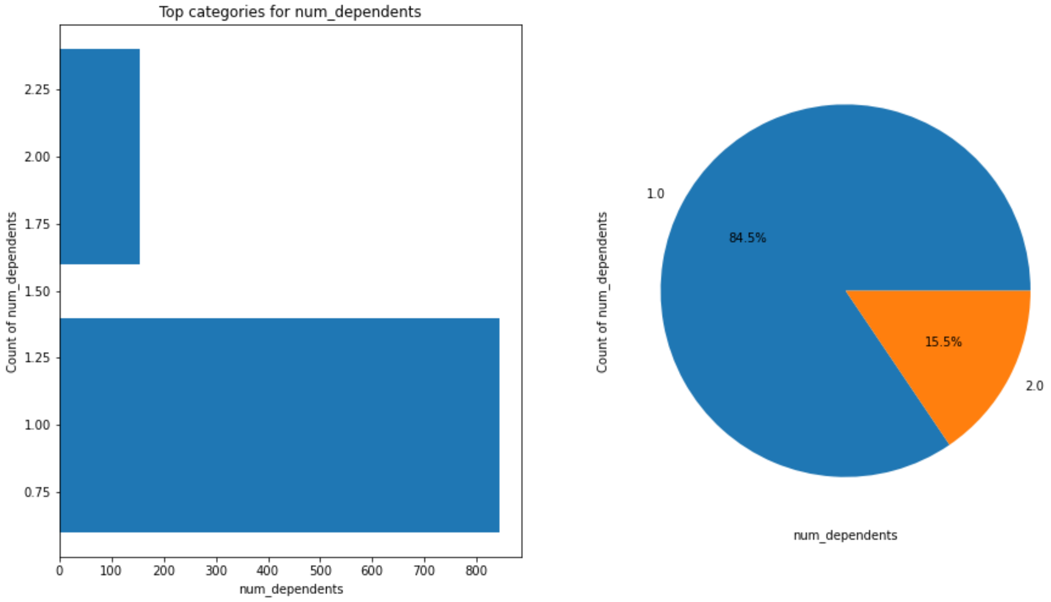

Top categories for property_magnitude
Count of property_magnitude
no known property
life insurance
real estate
car
0
50
100
150
200
250
300
property_magnitude
Count of property_magnitude
car
33.2%
real estate
28.2%
life insurance
23.2%
no known property
15.4%
property_magnitude

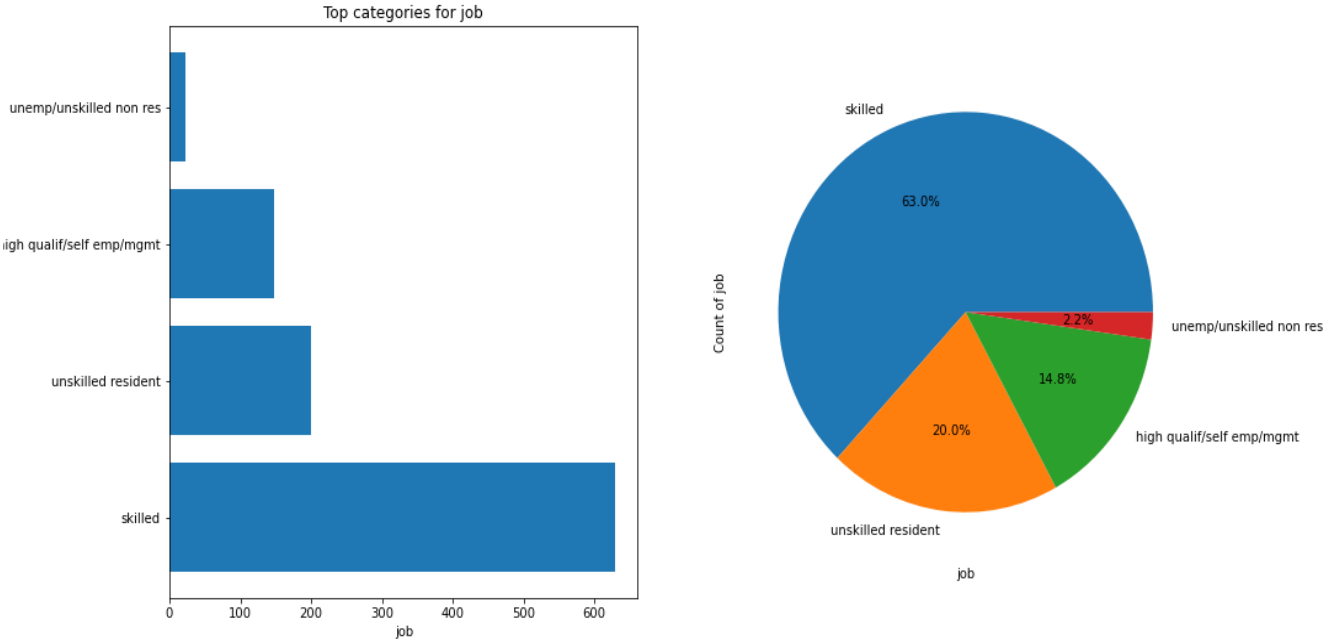

Top categories for personal_status
male div/sep
male mar/wid
female div/dep/mar
male single
Count of personal_status
0
100
200
300
400
500
personal_status
male single
54.8%
31.0%
9.2%
5.0%
male div/sep
male mar/wid
female div/dep/mar
Count of personal_status
personal_status

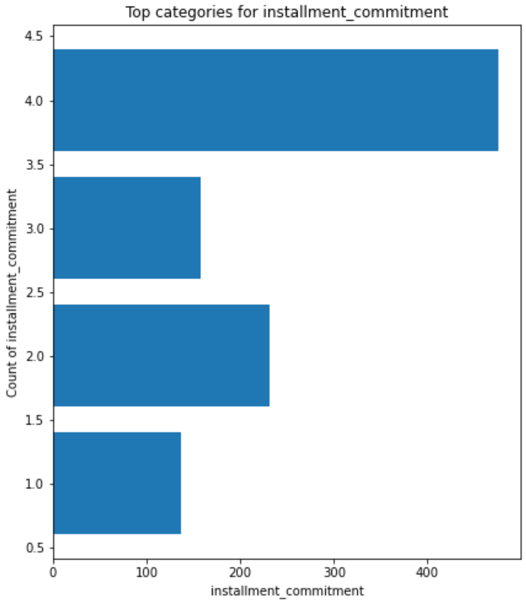

Top categories for own_telephone
yes
none
Count of own_telephone
0
100
200
300
400
500
600
own_telephone
none
59.6%
40.4%
yes
Count of own_telephone
own_telephone

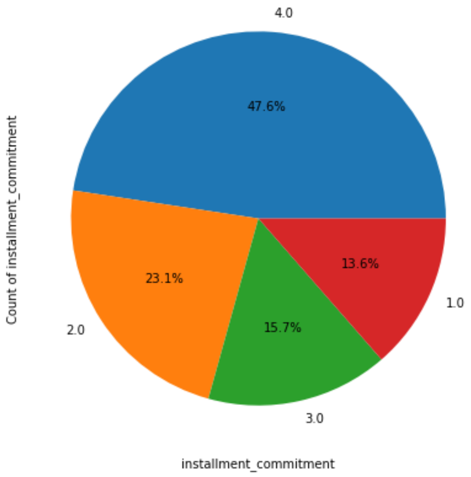

Top categories for other_payment_plans
stores
bank
none
Count of other_payment_plans
0
100
200
300
400
500
600
700
800
other_payment_plans
none
81.4%
4.7%
stores
13.9%
bank
Count of other_payment_plans
other_payment_plans

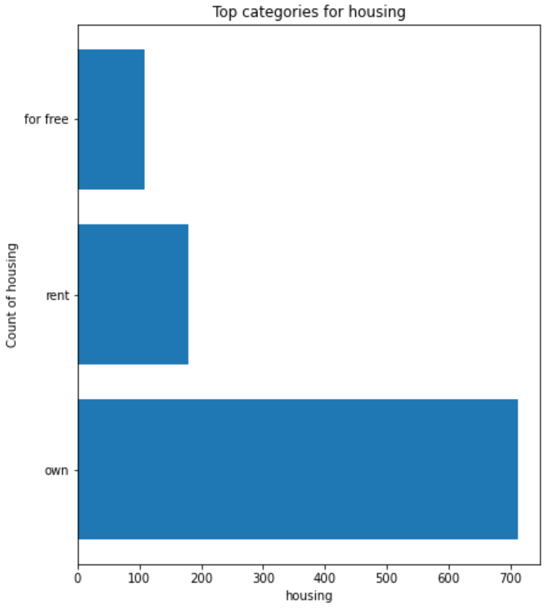

Top categories for other_parties
co applicant
guarantor
none
Count of other_parties
0
200
400
600
800
other_parties
none
90.7%
4.1%
co applicant
5.2%
guarantor
Count of other_parties
other_parties

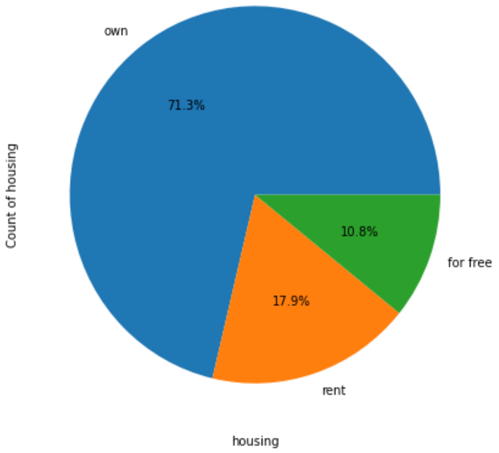

Top categories for num_dependents
2.25
2.00
1.75
1.50
1.25
1.00
0.75
Count of num_dependents
0
100
200
300
400
500
600
700
800
num_dependents
1.0
84.5%
15.5%
2.0
Count of num_dependents
num_dependents

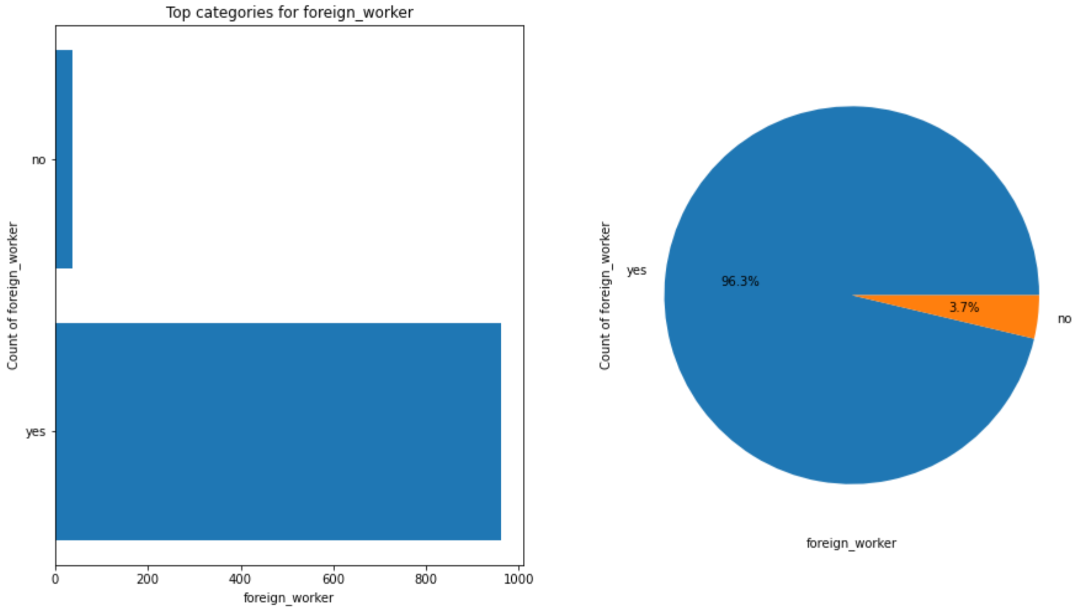

Top categories for job
unemp/unskilled non res
igh qualif/self emp/mgmt
unskilled resident
skilled
0
100
200
300
400
500
600
job
skilled
63.0%
2.2%
unemp/unskilled non res
14.8%
20.0%
high qualif/self emp/mgmt
unskilled resident
Count of job
job

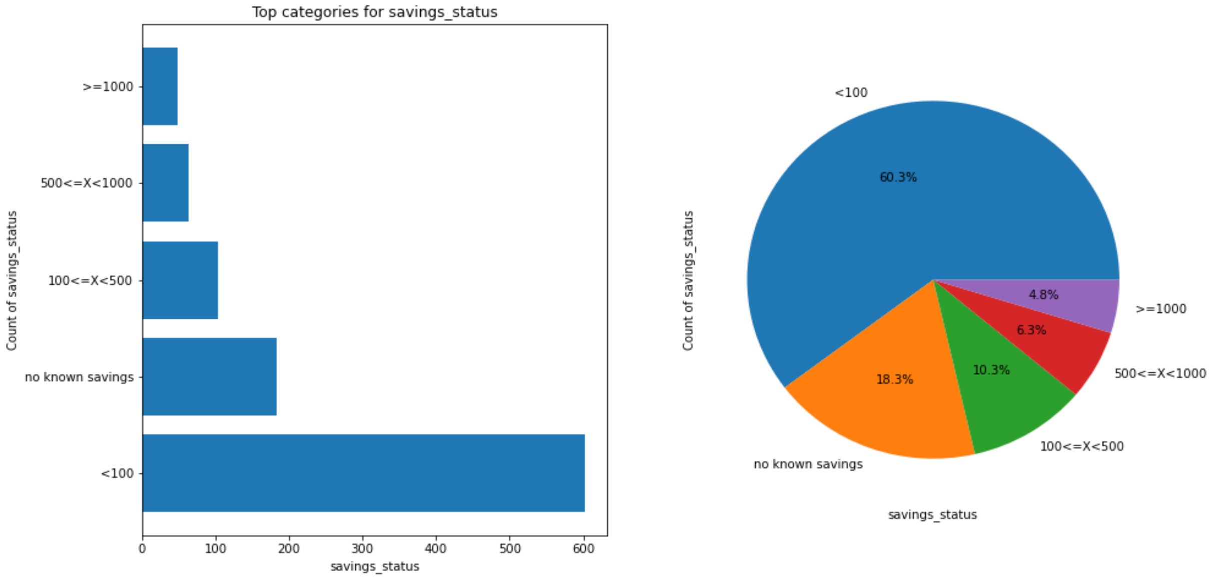

Top categories for installment_commitment
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
Count of installment_commitment
0
100
200
300
400
installment_commitment


4.0
47.6%
2.0
23.1%
3.0
15.7%
1.0
13.6%
Count of installment_commitment
installment_commitment


Top categories for housing
for free
rent
own
Count of housing
0
100
200
300
400
500
600
700
housing


own
71.3%
rent
17.9%
for free
10.8%
Count of housing
housing

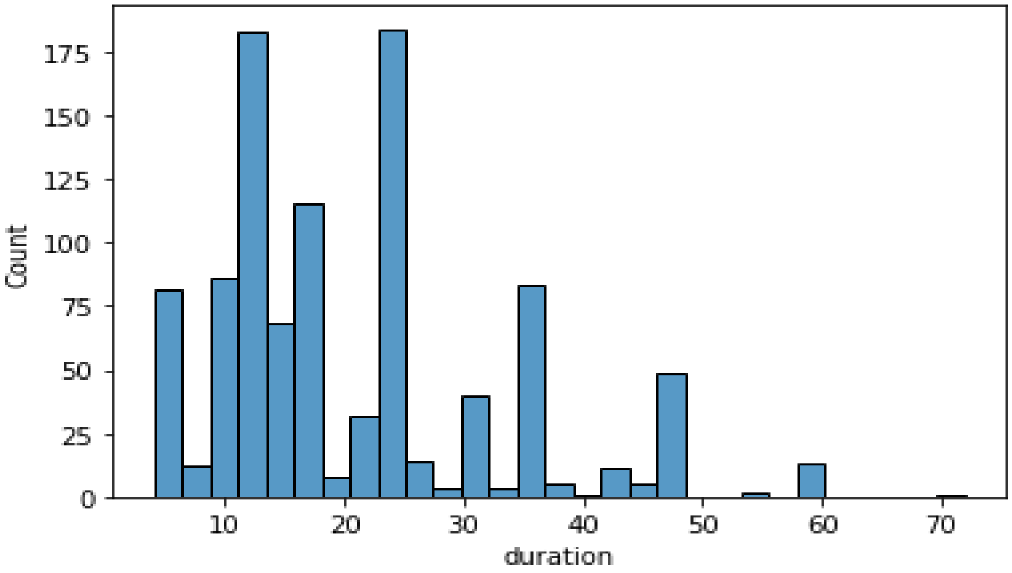


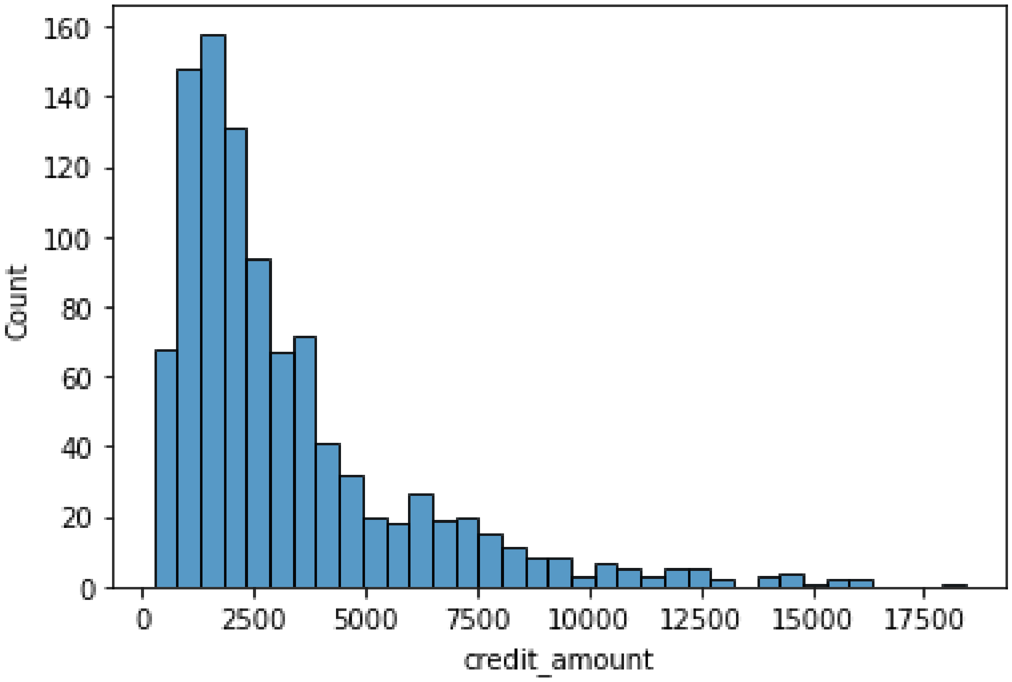


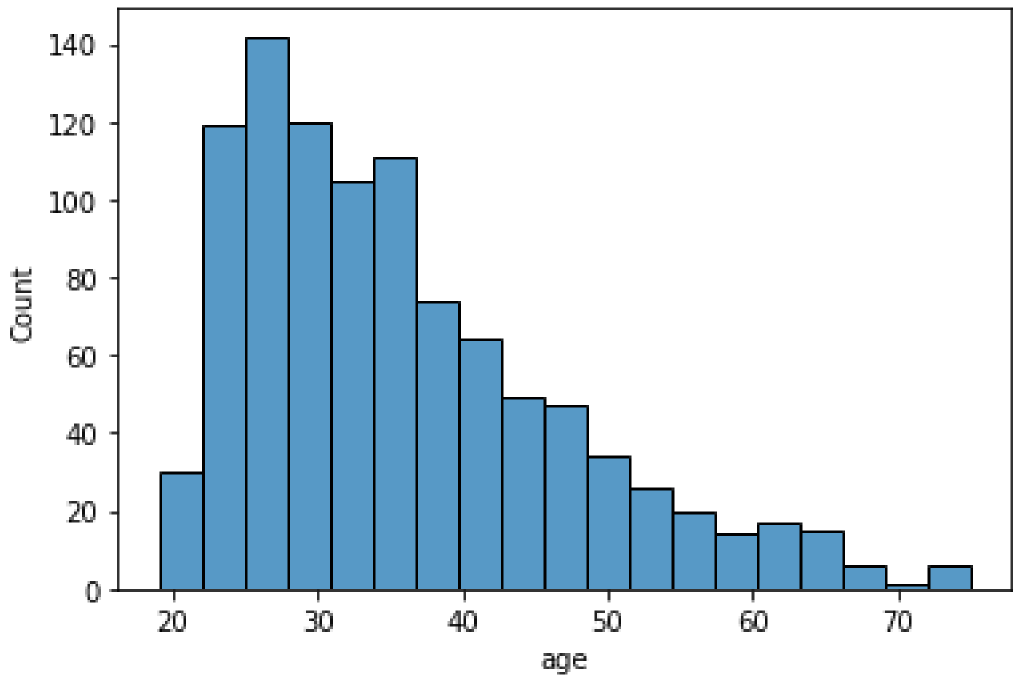
- ✓ **Inference from above charts**
- ➢ Almost 40% people have checking_status as no checking, which means there will be no account checking for them if they are availing for credit. 27.5% people have less than 0 in their account checking. 27% have between 0 to 200 and only 6% people have more than 200 in their account checking.
- ➢ People who are existing paid or critical/another existing credit are more prone towards credit.
- ➢ People are taking more credit for purchasing Radio, TV, New Car, Furnitures and equipments.
- ➢ Customers who have less savings are more prone towards credit as they don’t have liquidity.

- Customers who have more tenure of employment are more prone towards credit which shows they need more money.
- Almost 50% customers have installment commitment of 4
- Male single are in need of more credit which is surprising as bachelor male have less expenses.
- In more than 90% cases there is no involvement of other parties in loan cases.
- Customers who have more tenure of residence are more prone towards credit.
- Customers who are in possession of Car and Real Estate are more prone towards credit
- Customers who are owning house are more inclined towards credit
- Maximum customers are having only one existing loan credit
- 70% customers have good credit and 30% have bad credit

***<u>Top categories for continuous features in the data :-</u>***

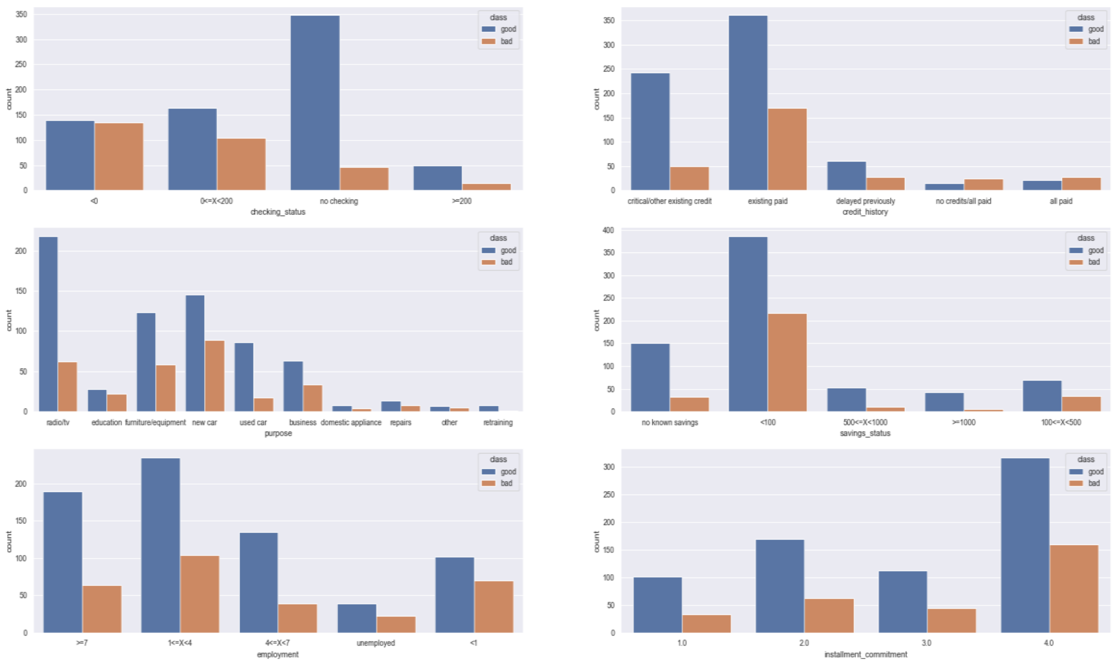


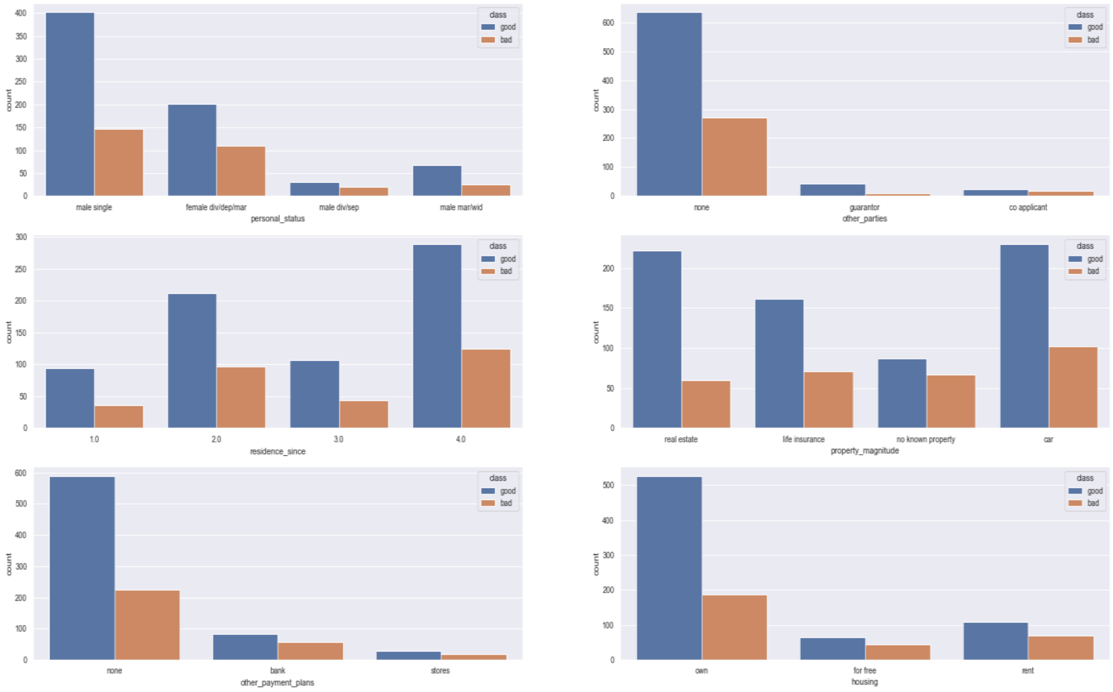


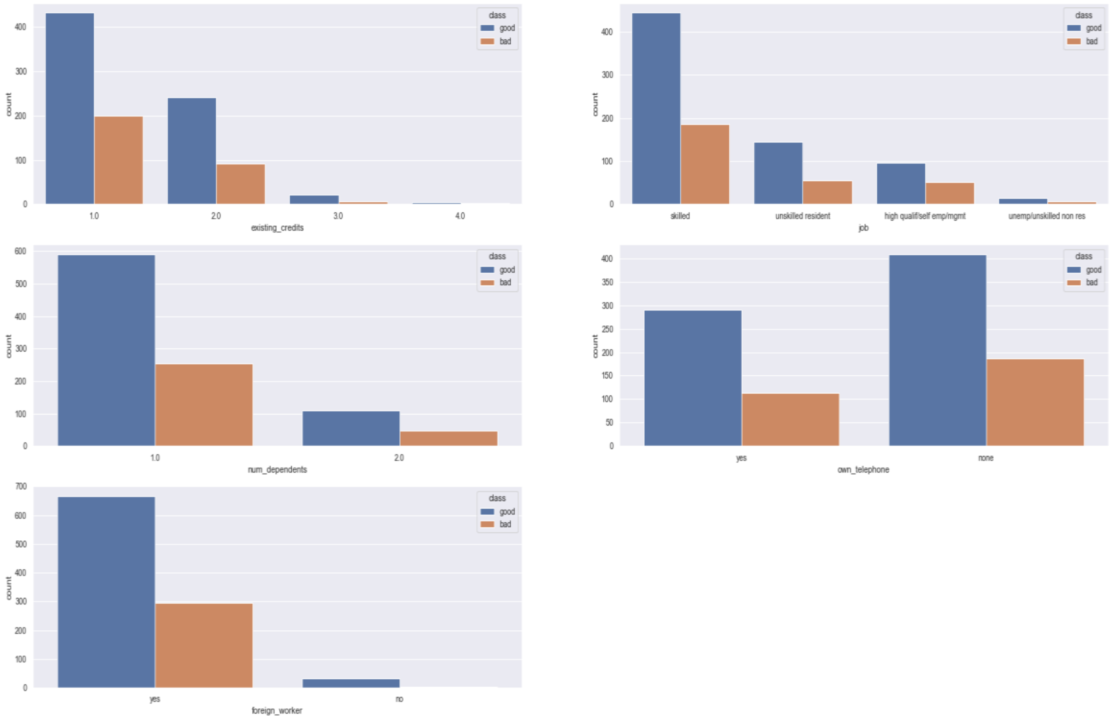


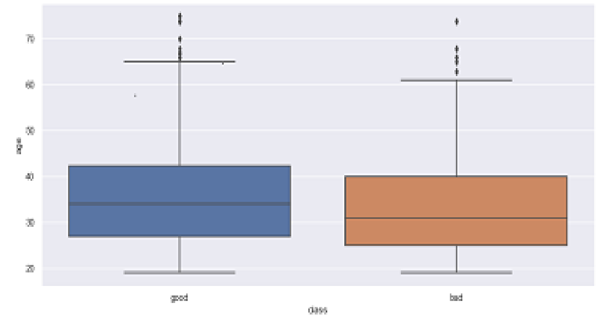
- ✓ **Inference from above charts**
- ➢ The duration for most of the loans are between 10 and 25
- ➢ The maximum customers have taken credit between 1000 to 5000
- ➢ Customers between age of 20 and 30 are more in a need of credit

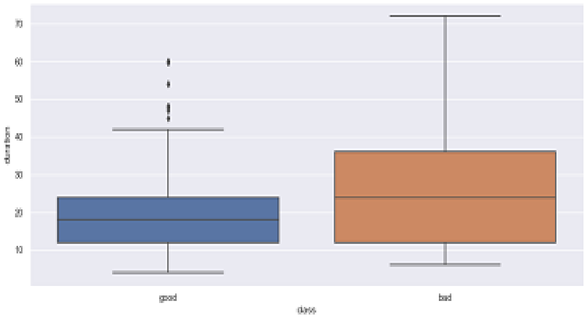
## ***Relation of categorical and continuous columns with the Target:-***

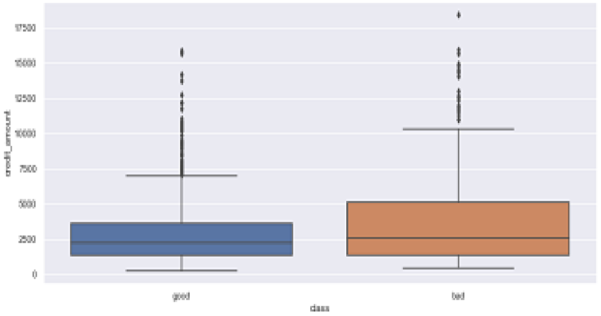



count
class
good
bad
male single
female div/dep/mar
male div/sep
male mar/wid
personal_status
none
guarantor
co applicant
other_parties
1.0
2.0
3.0
4.0
residence_since
real estate
life insurance
no known property
car
property_magnitude
none
bank
stores
other_payment_plans
own
for free
rent
housing


count
class
good
bad
1.0
2.0
3.0
4.0
existing_credits
skilled
unskilled resident
high qualif/self emp/mgmt
unemp/unskilled non res
job
1.0
2.0
num_dependents
yes
none
own_telephone
yes
no
foreign_worker







### ✓ Inference from above charts

- We can see that number of people in class good is 700 and in class bad is 300.
- Generally for all the checking status customers who are not risky for credit are more than number of risky customers but most of those who have no checking as checking status are very less risky customer and have class good and credit can be granted to them without any second thought.
- Most of the customers having credit history as existing paid are not risky for credit. Same is the case for critical/other existing credit. In customers who have credit history as no credits/all paid & all paid there are more number of customers who are in bad class, hence they are risky but a small credit can be given to them because they will pay back the money later and that credit will not go in NPA because they are in class of no credits or all paid which clearly shows they return the money but not in time.
- Most of the customers who take credit for radio, TV, furniture, new car and retraining generally pay of the loan and they are not risky customers.
- Customers who are having savings_status between 500 to 1000 are generally very less risky customers.
- Customers who are employed between 1 and 4 are generally less risky. Same is the case for employment greater than 7.
- Most of the customers who are having installment commitment of 4 are less risky.
- Customers who are residing for a longer period of time are less risky.
- Customers who own the house are generally less risky as compared to those who does not posses a property.
- Customer who posses car and also investment in real estate are less risky.
- Individual loan account customers are less risky than loan accounts where co-applicant and guarantor are involved.
- Customers who have existing credit of less than 3 are more risky to customers who have more existing credits.
- Customers who are skilled are less risky compared to others.

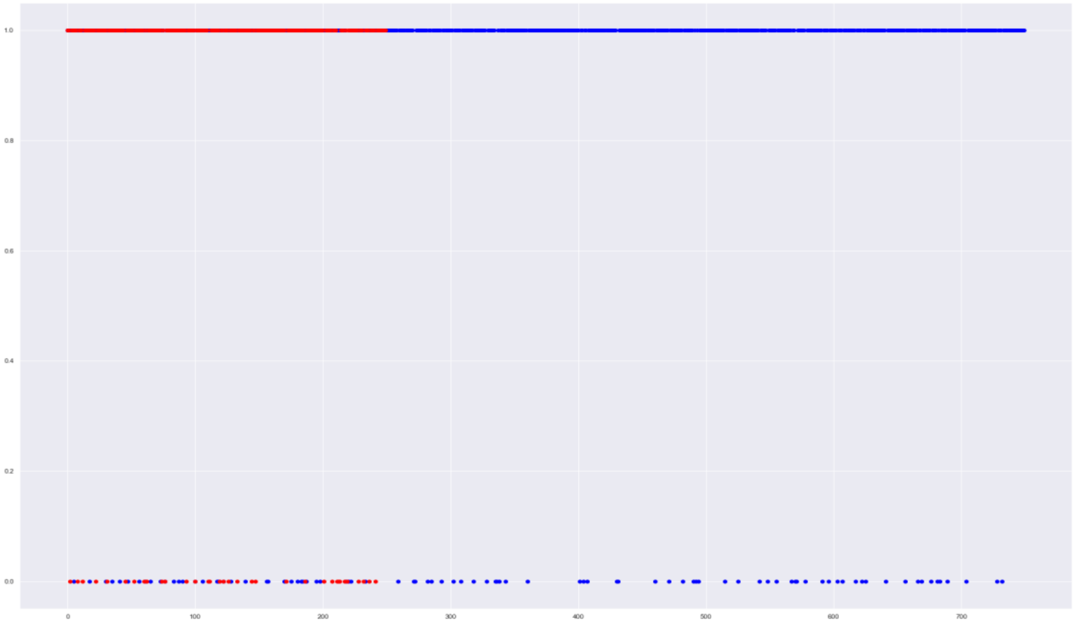
- Customers whose age is between 25 to 45 are who are need of credit. Customers who are above 40 years of age are very less risky compared to other age groups.
- Customers where duration of credit is less are less risky. And as the duration of credit increases the risk of the customer also increases.
- Customers who take credit more than 3500 are very risky compared to customers who take less credit amount.

## ***MACHINE LEARNING RESULTS:-***

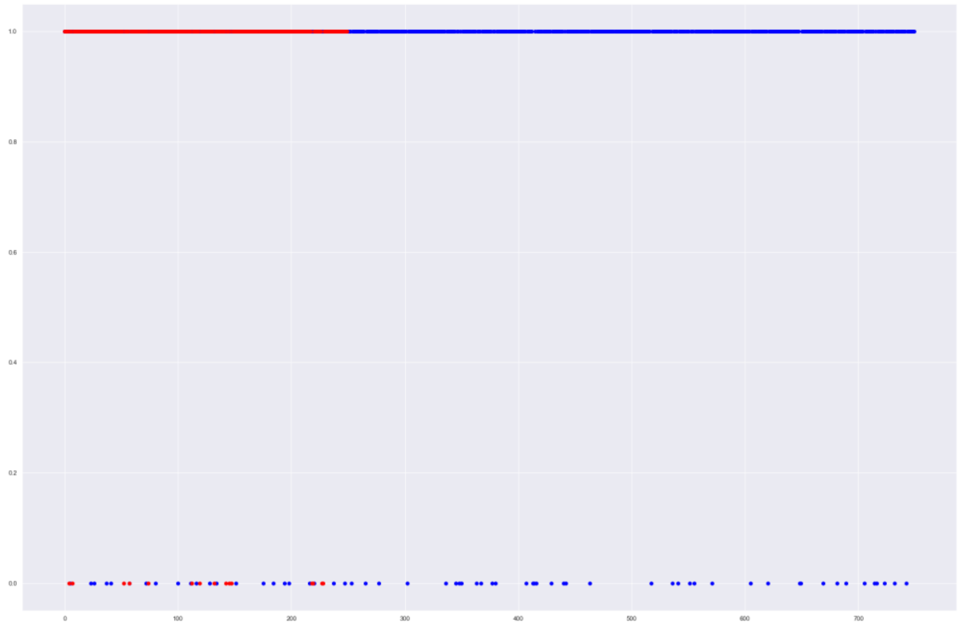
Below are the scatter plots of all the models involved in the prediction of the customer's risk towards the credit.

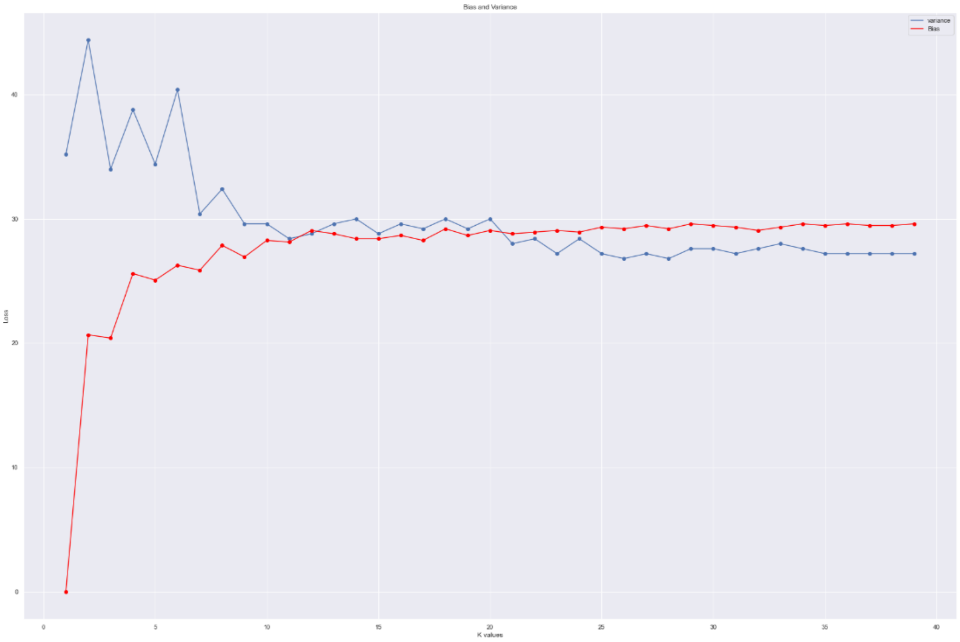
The Blue colour is for Training Data and Red colour is for Test Data.

### LOGISTIC REGRESSION

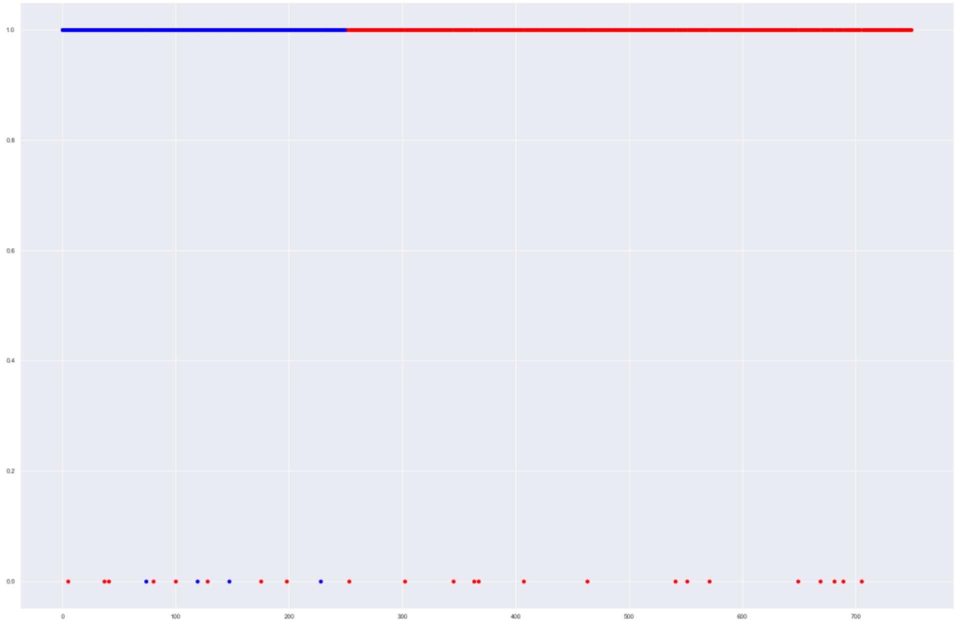


# KNN & KNN BV Trade Off

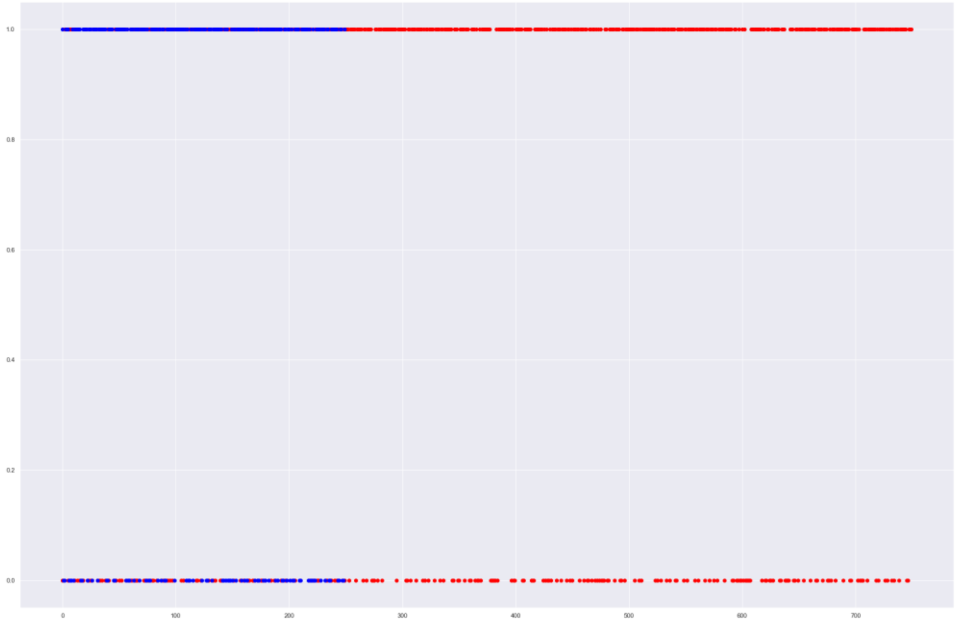




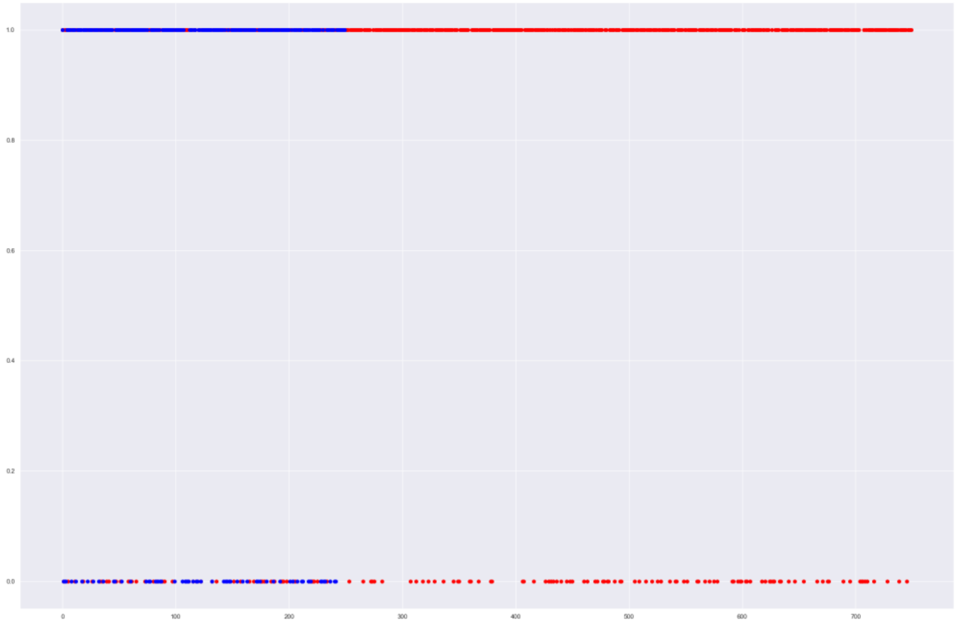
## SVM



## Decision Tree



## Random Forest



***The Best Prediction for the Problem Statement is Random Forest Random Search CV.***

● **References:-**

- Random Forest Flow Chart:- https://www.google.com/search?q=Random+forest+random+search+cv+flow+chart&rlz=1C1CHBD_enIN1056IN1056&sxsrf=APwXEdeHtBsPdqdn3BMvLplpTTtGrzoYJA:1686565153893&source=lnms&tbm=isch&sa=X&ved=2ahUKEwjhro3nwL3_AhUIyDgGHXBODfEQ_AUoAXoECAIQAw&biw=1280&bih=512&dpr=1.5#imgrc=hPgBAWOXUzu0VM

- Dataset Source:- https://www.kaggle.com/datasets/ppb00x/credit-risk-customers

- https://www.google.com/search?q=credit+risk&tbm=isch&ved=2ahUKEwjH9PjXmL7_AhVSLLcAHfV7CPgQ2-cCegQIABAA&oq=credit+risk&gs_lcp=CgNpbWcQAzIECCMQJzIFCAAQgAQyBQgAEIAEMgUIABCABDIFCAAQgAQyBQgAEIAEMgUIABCABDIFCAAQgAQyBQgAEIAEMgUIABCABDoHCAAQGBCABFDYBVi3FGCbGWgAcAB4AIABvAKIAfkQkgEHMC42LjQuMZgBAKABAaoBC2d3cy13aXotaW1nwAEB&sclient=img&ei=SE2HZMewGdLY3LUP9fehwA8&bih=569&biw=1280&rlz=1C1CHBD_enIN1056IN1056#imgrc=PqVdVup3nHUU8M