Internship Project Report

Loan Amount Analysis & Prediction

At

IC SOLUTIONS

SUBMITTED BY:

NAME: Prajwal M Korishettar

USN: 1BY19IS407

EMAIL ID: prajumk777@gmail.com

INSTRUCTOR: ABHISHEK C

Acknowledgement:

Being part of this internship was a great experience. It is with grateful heart that I thank the coordinators of IC Solutions for giving me an opportunity to intern with their organization. I express my gratitude to my instructor, Abhishek C, for providing me with useful knowledge that can be applied on practical tasks and also for guiding and encouraging me to do my best in this internship.

During the period of my internship, I have received generous help and suggestions from many quarters, for which I would like to remember and thank them all with deep gratitude and great pleasure.

Abstract :

Loan Amount Analysis & Prediction has been of high interest area, it requires noticeable effort and knowledge of the field expert. Considerable number of distinct attributes are examined for the reliable and accurate prediction. Respective performances of different algorithms are compared to find one that best suits the available data set. The final prediction model is evaluated using test data and the accuracy is obtained.

This project involves analysis of the given dataset, which here is one on Land Asset. It includes detailed analysis of the features/factors of the Loan Amount Sanction given in the data set, depending on basis of its Income, Dependents, Credit Score, Property Type & etc. Here specifically, we need to understand how the factors affect the Loan Amount Sanction in the Banks , since the Income , Credit Score & other aspect may vary with different features.

So, by analyzing each of the factors involved, by means of various plots and graphs we study the changes they bring to the final Loan Amount Sanction. We draw conclusions from the obtained plots and models, and the available data list to predict the possible final Loan Amount Sanction.

To do this, we use the concept of Machine Learning. Using Regression approach, Since we have been asked to predict the final Loan Amount Sanction by utilizing all the other specifications, this kind of problem comes under Regression category so we utilize Regression Models or Algorithms to solve this problem.

About the Company :

IC Solution (ICS) is a digital service provider that aims to provide software, designing and marketing solutions to individuals and businesses. At ICS, we believe that service and quality is the key to success.

We provide all kinds of technological and designing solutions from Billing Software to Web Designs or any custom demand that you may have. Experience the service like none other!

Some of our services include:

Development - We develop responsive, functional and super-fast websites. We keep User Experience in mind while creating websites. A website should load quickly and should be accessible even on a small view-port and slow internet connection.

Mobile Application - We offer a wide range of professional Android, iOS & Hybrid app development services for our global clients, from a startup to a large enterprise.

Design - We offer professional Graphic design, Brochure design & Logo design. We are experts in crafting visual content to convey the right message to the customers.

Consultancy - We are here to provide you with expert advice on your design and development requirement.

Videos - We create a polished professional video that impresses your audience.

Index :

Introduction :

Python along with Machine Learning Internship is offered by the collaboration of two companies namely TakeiteasyEngineers and IC Solutions. It originally consisted of 10 days of training which is followed by a month of internship which must be done in order to be able to receive the certificate of completion.
Day 1of the training began by introduction of our mentor Abhishek G sir and started with the basics of Python which were essential for the course of Machine Learning and writing accurate syntax. The basics included mathematical operations, functions and their utilization in programs and other basic topics which were needed for this particular course. Then two important libraries of Python namely Numpy and Pandas was introduced which are used for Data Generation, Data Displaying, Manipulation and Cleaning.
Then we got to know about another python library named Matplotlib which is mainly used for plotting and viewing the data pictorially and deriving suitable conclusions from the parameters given and used in the problem statement and dataset.
This was followed by other libraries named Seaborn and ScikitLearn. Seaborn is also used in Data Visualization while ScikitLearn is used to test and predict the data. It is followed by understanding bout both Supervised and Unsupervised Machine Learning Algorithms which are imported and implemented from ScikitLearn.
Finally, we learnt about the basics of Artificial Neural Networks and how to deploy a Machine Learning Model and use it in Backend Development.

Problem Statement and Objective :

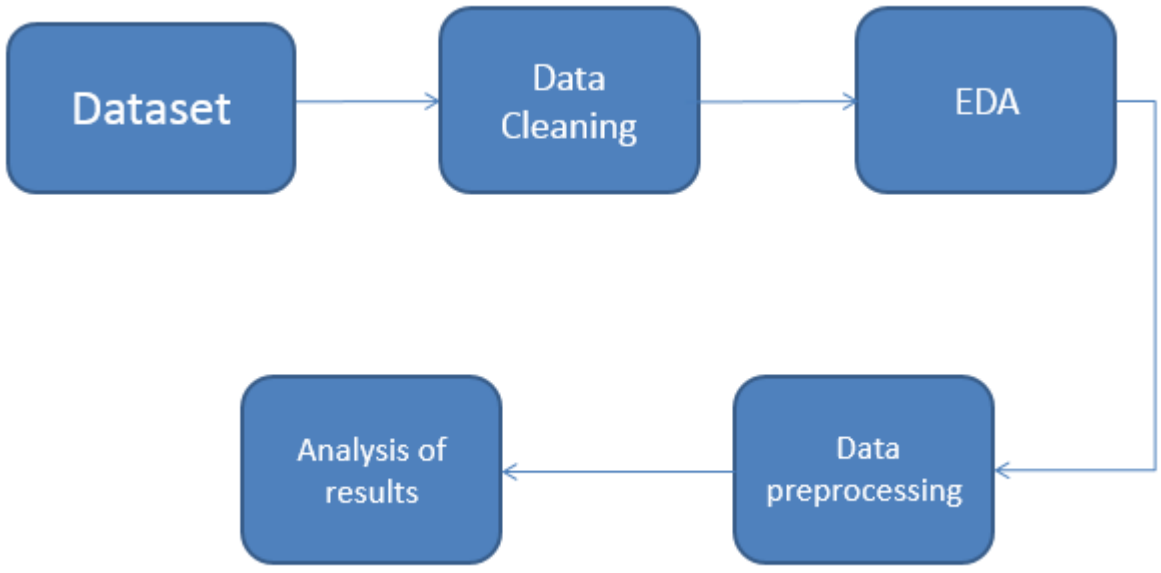
Problem Statement:

Dream Housing Finance company deals in all home loans. They have presence across all urban, semi urban and rural areas. Customer first apply for home loan after that company validates the customer eligibility for loan.

The Company wants to automate the loan eligibility process (real time) based on customer detail provided while filling online application form. These details are Gender, Education, Number of Dependents, Income, Loan Amount, Credit Score and others. To automate this process, they have given a problem to identify the customers segments, those are eligible for loan amount so that they can specifically target these customers.

We'll start by exploratory data analysis , then preprocessing , and finally we'll be testing different models such as Linear regression and decision trees & other models.

Objective :



Requirement Specification :

Hardware Requirements:-

- Dell Inspiron 15 3593 Series
- 10th Generation Intel® Core™ i5-1035G1 Processor (6MB Cache, up to 3.6 GHz)
- 4GB, 1 x 4GB, DDR4, 2666Mhz

Software Requirements:-

- Any Python Compiler available in the internet.
- Numpy, Pandas, ScikitLearn, Seaborn, Matplotlib
- Internet Connection to resolve any doubts or issues
- Anaconda Software as a whole and Jupyter Notebook installed from it

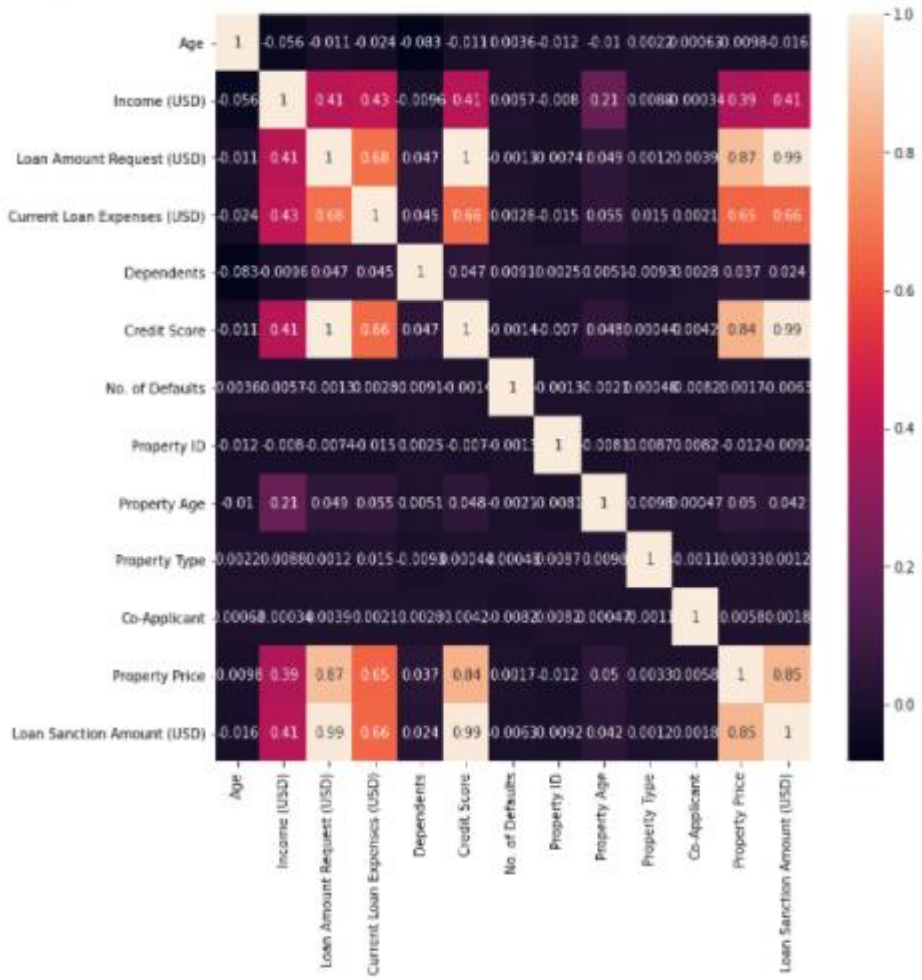
Exploratory Data Analysis (EDA):

After importing the data and data cleaning we go to Exploratory Data analysis part which contains all the analytics:

1) Correlation Matrix- seaborn heatmap -

```
In [40]: corr= dataset.corr()
         plt.figure(figsize=(10,10))
         sns.heatmap(corr,annot =True)

Out[40]: <AxesSubplot:>
```

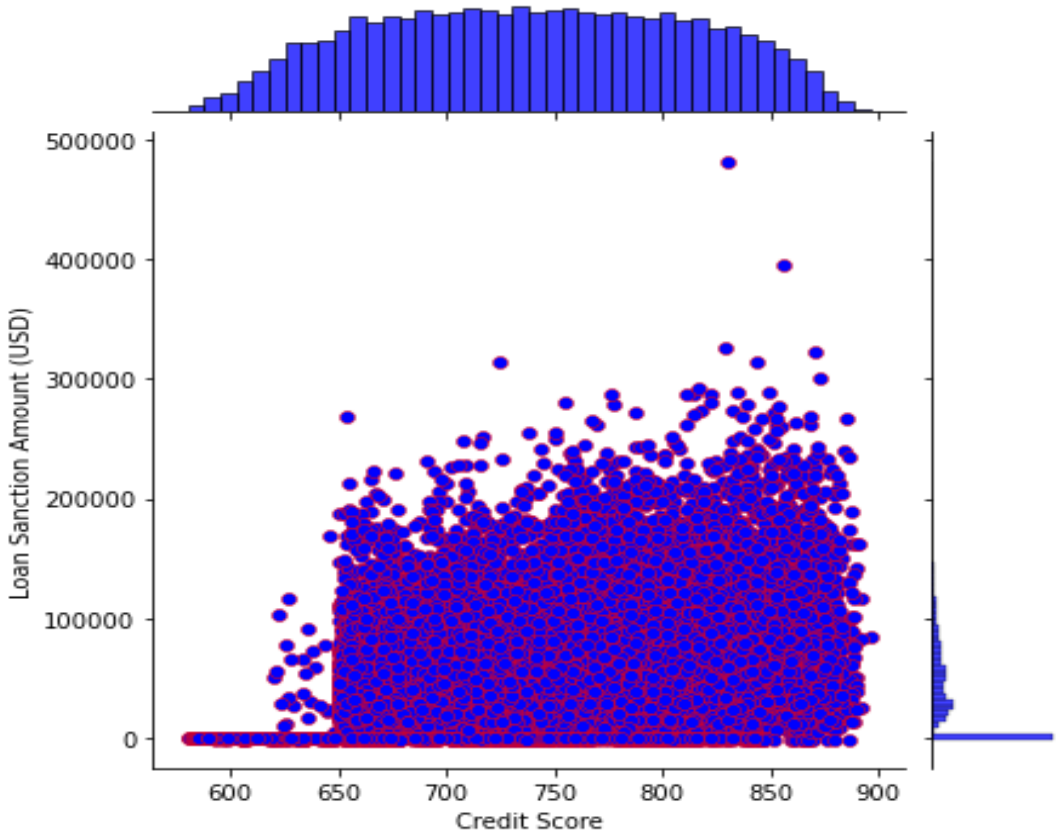


- This plot shows the correlation between the variables.
- The scale on the right will show the correlation coefficient.

- We can see that Credit score and Loan Sanction Amount (USD**)** columns are closely correlated with correlation coefficient of around 0.99.
- Property Price & Property ID are indirectly proportion to each other as the correlation coefficient of around -0.012.
- We can conclude that Current Loan Expenses (USD) and Loan Amount Request (USD) columns are closely correlated with correlation coefficient of around 0.68.

2) Scatterplot of Credit Score and Loan Sanction Amount (USD) -

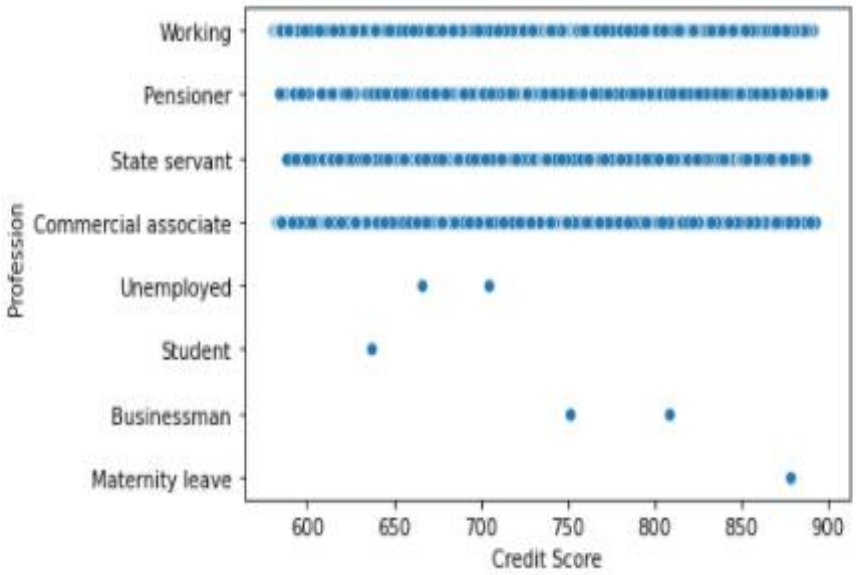
```
In [17]: sns.jointplot(x='Credit Score',y='Loan Sanction Amount (USD)',data=dataset,kind='scatter',color='b',edgecolor='r')
```



- In general, as per plot, between 650 to 880 Credit Score the Loan Sanction Amount is more. This shows that in the data, there are more number of Loan Sanction Amount falling in this range, and hence the high concentration there.
- Only 5 customer got Loan Amount more than 30lakh, remaining 10% customers got zero Loan Amount.
- Only 24 customer got Loan Amount between 630 to 650 credit score.

3) Scatterplot of Credit Score and Profession -

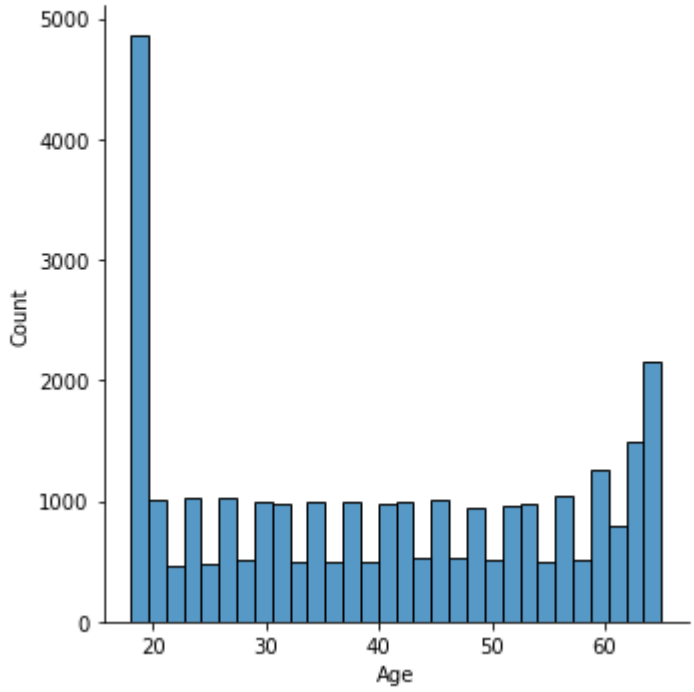
```
In [30]: sns.scatterplot(x=dataset['Credit Score'],y=dataset['Profession'])
Out[30]: <AxesSubplot:xlabel='Credit Score', ylabel='Profession'>
```



- As per the plot professional like Working, Pensioner, State servant, Commercial associate are given more preferences for the Loan Amount.
- Only 1% of students & Maternity leave applicants got the Loan amount with 640 Credit score.
- Only 2 - 4% of Unemployment , Businessman applicants have got Loan amount Sanction according to the dataset given.
- So we can conclude that only with Profession like Working, Pensioner, State servant, Commercial associate are given first & more preferences for Loan amount with irrespective of the Credit score they have.

4) Displot -

```
In [18]: sns.displot(dataset['Age'],bins=30)
Out[18]: <seaborn.axisgrid.FacetGrid at 0xce8c699ac0>
```

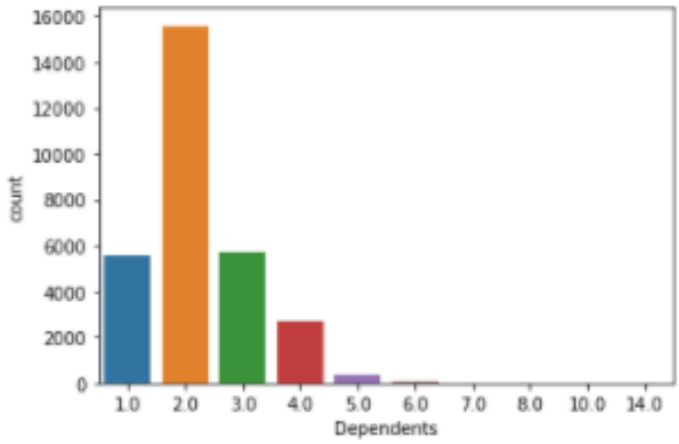


- Here we plotted the Age count using Displot.
- As we can see there are more number of applicants with age 15.
- There are more number of applicants with age 60 to 65.

5) Count Plot –

```
In [14]: sns.countplot(dataset['Dependents'])
```

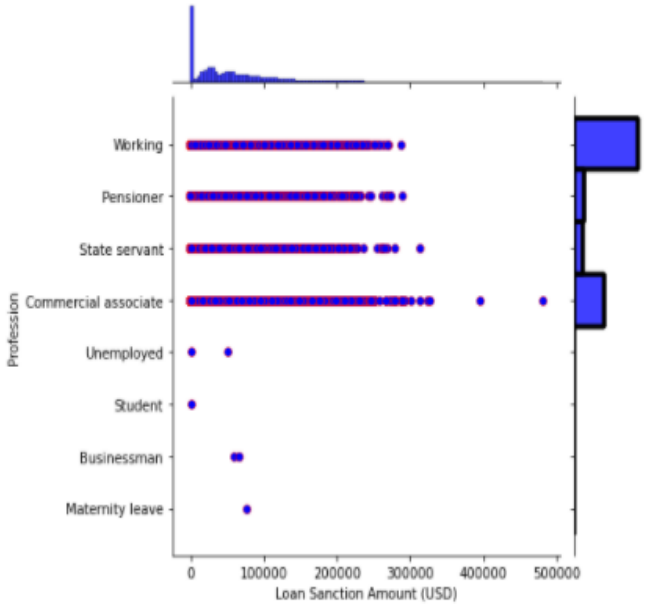
```
Out[14]: <AxesSubplot:xlabel='Dependents', ylabel='count'>
```



- In this plot we have plotted the count of Dependents of the applicants.
- There are 5800 applicants with 1 & 3 Dependents.
- There are 15800 applicants with 2 Dependents.
- There are 2200 applicants with 4 Dependents.
- Remaining 400 applicants with 5 ,6 ,7,8,10 & 14 respectively.

6) Jointplot of Profession and Loan Sanction Amount -

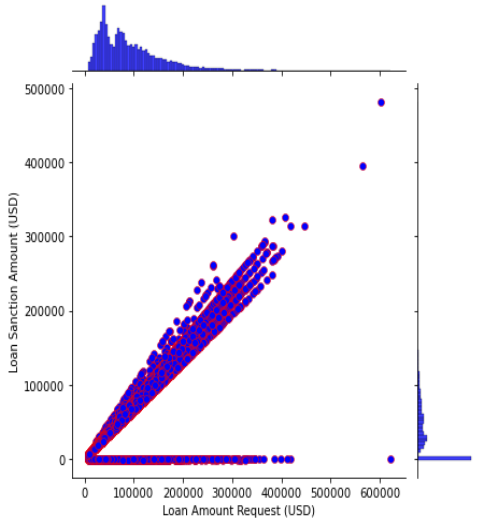
```
In [16]: sns.jointplot(y='Profession',x='Loan Sanction Amount (USD)',data=dataset,kind='scatter',color='b',edgecolor='r')
Out[16]: <seaborn.axisgrid.JointGrid at 0xa119a3e2e0>
```



- This plot shows the percentage of Loan Amount Sanction for each Profession.
- High number of applicants of profession like Workers & Commercial associate got loan amount between 0 to 30.3lak. And also 2% of applicants of commercial associate got 30.8lakh & 40.8lakh respectively according to their credit score.
- Only few applicants got the Loan amount with profession like Student, Unemployment, Businessman & Maternity leave.
- So we can conclude that profession like Commercial associate , Working persons, Pensioner have high changes of getting Loan amount.

7) Jointplot of Loan Amount Requested & Loan Sanction Amount -

```
In [18]: sns.jointplot(x='Loan Amount Request (USD)',y='Loan Sanction Amount (USD)',data=dataset,kind='scatter',color='b',edgecolor='r')
Out[18]: <seaborn.axisgrid.JointGrid at 0xa119d548b0>
```

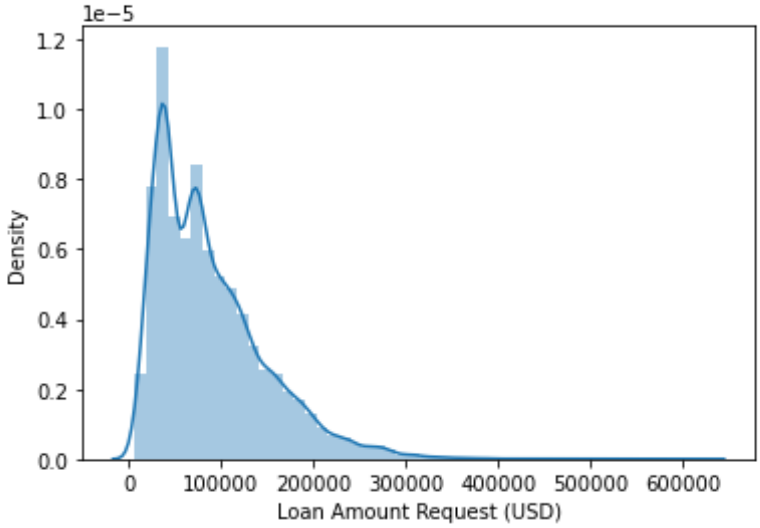


- It looks quite similar to the plot of Credit Score and Loan Sanction Amount (USD).
- There are more number of applicants requesting loan between 2 to 30.5lakh. And got sanction of loan of 2 to 20.8lakh.
- Only 2 applicant has got more sanction amount of 40lakh & 49lakh.

8) Displot of Loan Amount Request -

```
In [30]: sns.distplot(dataset['Loan Amount Request (USD)'])
         plt.show()
         dataset['Loan Amount Request (USD)'].plot.box(figsize=(16,5))
         plt.show()
```

```
C:\ProgramData\Anaconda3\lib\site-packages\seaborn\distributions.py:2557
d will be removed in a future version. Please adapt your code to use eit
xibility) or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```



- In general, as per plot, there is a increase in applicant of Loan Amount between 0 to 15lakh.
- As per plot there is a less count of request between 25lakh to 65lakh.
- There is a high request of applicants for Loan Amount of 2lakh to 5lakh as we can see in the above graph.

Preparing Machine Learning Model :

The Machine Learning Algorithms I used are:-

- Linear Regression Mode
- Decision Tree Regression Model
- Support Vector Regression Model
- Random Forest Regression Model
- Bayesian Regression Model

Train Test Split

```
In [45]: x=dataset.drop(columns=['Loan Sanction Amount (USD)'],axis=1)
         y=dataset['Loan Sanction Amount (USD)']

In [46]: from sklearn.model_selection import train_test_split

In [47]: x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.3,random_state=101)
```

- Here the data is prepared for training and testing.
- This data is now trained for different Regression algorithms and after testing the models the accuracy of prediction is noted.

Decision Tree model

```
In [65]: from sklearn import tree
         clf = tree.DecisionTreeRegressor()
         clf.fit(x_train, y_train)
         y_pred = clf.predict(x_test)

In [66]: from sklearn.metrics import r2_score

In [67]: r2_score(y_test,y_pred)
Out[67]: 0.5235955347794101
```

R2 Score of Decision Tree Regression is: 52.3595

Random Forest Model

In [63]:
```
from sklearn.ensemble import RandomForestRegressor
model = RandomForestRegressor()
model.fit(x_train,y_train)
pred =model.predict(x_test)

from sklearn.metrics import r2_score
r2_score = r2_score(y_test,pred)
```

In [64]:
```
r2_score
```

Out[64]:
```
0.7485246077843142
```

R2 Score of Random Forest Regression is : 74.8524

Linear Regression Model

```
In [102]: from sklearn.metrics import mean_absolute_error,mean_squared_error,r2_score,accuracy_score

In [103]: r2_score(y_test,predict)
Out[103]: 0.597649810101416

In [61]: mean_absolute_error(y_test,predict)
Out[61]: 21731.54937286625

In [62]: mean_squared_error(y_test,predict)
Out[62]: 988025195.9301953
```

R2 Score of Linear Regression Model is : 59.7649

Support Vector Regression

```
In [117]: from sklearn.svm import SVR
          model = SVR()
          model.fit(x_train,y_train)
          pred3 = model.predict(x_test)
          accuracy =r2_score(y_test,pred3)

In [118]: accuracy

Out[118]: 0.0002878463167897971
```

Accuracy of Support Vector Regression is : 0.02878

Bayesian Regression Model

```
In [70]: from sklearn import linear_model
         reg = linear_model.BayesianRidge()
         reg.fit(x_train, y_train)
         predic = reg.predict(x_test)
         accuracy = r2_score(y_test,predic)

In [71]: accuracy

Out[71]: 0.5676246826817665
```

Accuracy of Support Vector Regression is : 56.7624

Machine Learning Model Chart

Serial Number	ML Algorithm Used	Accuracy Score (approx. 4 decimal places)
01	Support Vector Regression	**0.02878**
02	Random Forest Regression	**74.8524**
03	Linear Regression	**59.7649**
04	Decision Tree Regression	**52.3595**
05	Bayesian Regression Model	**56.7624**

Conclusion :

Loan Amount Analysis & Prediction involves a high number of attributes that should be considered for accurate prediction. A major step in the prediction is processing of the data. In this project a few machine learning algorithms were used to process this data. The data is first analyzed carefully considering each factor with the other possible factor, and the analytics done with the help of matplotlib or seaborn are recorded. The codes, graph/diagram and the insights gained from these helps in the prediction.

Then the data was prepared for training and testing. Data cleaning is one of the processes that increases prediction performance so it is important to set the data ready for further processes. The type of task, regression or classification, is identified and data is trained on multiple algorithms like Linear Regression, Support Vector Regression, etc. and the best algorithm that fits the data is identified with the help of a metric like accuracy score.

All the outcomes of the different algorithms used are noted, compared and the better one is taken i.e. Random Forest Regression.

Bibliography

- stackoverflow.com
- https://seaborn.pydata.org/generated/seaborn.heatmap.html
- https://towardsdatascience.com/
- https://pbpython.com/pages/resources.html
- https://deepnote.com/@rhishab-mukherjee/Loan-Prediction-Project-TermPaper-VPSOpiywSu6FZeN2fK8fug