Chapter II : Project Undertaken

During this internship, our project focused on analyzing the eligibility criteria for loan approval by considering various factors such as credit score, income, desired loan amount, and assets etc. However, a significant inquiry arose: as commerce and data students, we were particularly interested in understanding how an individual's eligibility for a loan can be assessed and the key factors that influence loan approval.

To address these questions, we initiated a comprehensive research study titled ‘From Data to Decision: A Machine Learning Approach to Loan Eligibility.' This study sheds light on the criteria for loan eligibility.

Summary

Banks generate a significant portion of their profits from loans, and with a growing number of loan applicants, the challenge lies in accurately identifying genuine candidates who will repay their loans. Manual assessment can introduce misconceptions and inefficiencies into the selection process. To address this, we are in the process of developing a loan prediction system using machine learning. This system will automatically identify eligible candidates, leading to drastically reduction in the time required for loan approvals.

History of Machine Learning

Machine learning history starts in 1943 with the first mathematical model of neural networks presented in the scientific paper "A logical calculus”.

Then, in 1949, the book ”The Organization of Behavior” by Donald Hebb was published. The book had theories on how behavior relates to neural networks and brain activity.

In 1950 Alan Turing created the Turing Test to determine if a computer has real intelligence. To pass the test, a computer must be able to fool a human into believing

it is also human. He presented the principle in his paper “Computing Machinery and Intelligence”.

The first ever computer learning program was written in 1952. The program was the game of checkers, and the IBM computer improved at the game the more it played, and in 1997, IBM’s Deep Blue shocked the world by beating the world champion at chess.

What is Machine Learning?

Machine learning is a subset of Artificial Intelligence that includes algorithms that parse data, learn from that data, and then apply what they’ve learned to make informed decisions.

Machine Learning is a field of computer science that uses statistical tools to give computer systems the ability to learn from the data without being explicitly programmed.

At its core, machine learning involves the development of algorithms and models that can process and analyze data, identify patterns, and make predictions or decisions based on that data. These algorithms learn by being exposed to large amounts of data, essentially "training" on the data to improve their performance over time.

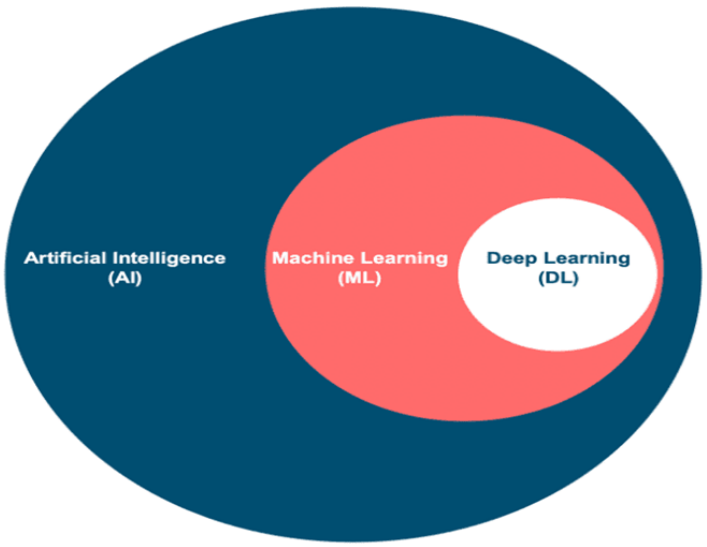
The impact of machine learning extends beyond individual applications. It's reshaping the job market, creating a demand for data scientists, machine learning engineers, and related roles. Additionally, it raises important ethical and societal questions regarding data privacy, algorithmic bias, and the potential consequences of autonomous systems.

One of the most noticeable applications of machine learning in our daily lives is recommendation systems. When you shop online, watch videos on streaming platforms, or use social media, you're likely to encounter recommendations tailored to your interests. These systems analyze your past behavior and preferences, comparing

them to a vast amount of data from other users to suggest products, content, or connections you might like.

In conclusion, machine learning is a transformative technology that empowers computers to learn from data, adapt to new situations, and make informed decisions. Its applications span various domains, from healthcare to finance, marketing, and scientific research. We encounter machine learning daily through recommendation systems, image and speech recognition, and automation. Its impact on industries, the job market, and society as a whole is profound.

Example- An on-demand music streaming service like Spotify. For Spotify to make a decision about which new songs or artists to recommend to you, machine learning algorithms associate your preferences with other listeners who have a similar musical taste.



(Fig. 1)

Difference between Machine Learning and Deep Learning

Machine Learning(ML)	Deep Learning(DL)
ML can work on a lesser amount of data provided by users.	A lot of unlabeled training data is required to make correct conclusions.
As we provide continues data to our ML model it improves the accuracy but at a certain point the model stops improving and becomes constant.	In DL, as we provide continues data to our ML model it improves the accuracy(always increasing).
ML needs less time to train.	DL needs more time to train.
ML can give good results on large and small data both.	DL gives best results on large data.

(Table no.1)

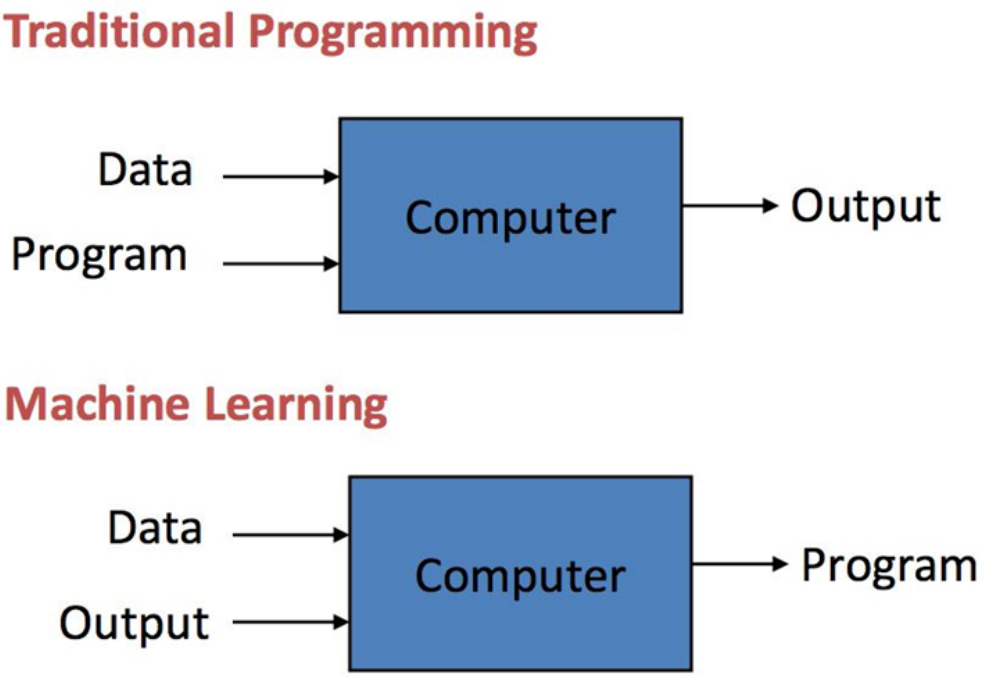
Two Approaches:

1. **Traditional Programming-**

   Traditional programming, often referred to as rule-based, is the conventional approach to instructing a computer to perform specific tasks. In this method, a user gives the data and the program to the computer and the computer gives the output as a result. These instructions are typically in the form of algorithms, which are step-by-step procedures for solving a problem or achieving a goal.

2. **Machine Learning-**

   Machine learning, on the other hand, represents a departure from traditional programming in that it focuses on creating systems capable of learning from data rather than being explicitly programmed for every task. In this method, a user gives the data and the output to the computer and the computer gives the program as a result.



(Fig. 1.1)

Why do we need Machine Learning(ML)?

ML harnesses patterns within its training data to provide precise predictions and decisions. Machine learning models progressively improve over time as they process new data.

ML is important because it gives enterprises a view of trends in customer behavior and business operational patterns, as well as supports the development of new products.

ML is a powerful and versatile field of artificial intelligence that has become increasingly important and prevalent for a variety of reasons. Here are some key reasons:

1. **Automation and Efficiency:** ML allows computers to learn from data and make predictions or decisions without being explicitly programmed. This automation can lead to increased efficiency and productivity in various industries. ML algorithms can automate repetitive tasks, analyze large

datasets, and make decisions faster and more accurately than humans in many cases.

2. **Pattern Recognition:** ML excels at recognizing complex patterns and trends within data that might be difficult or impossible for humans to discover.
3. **Personalization:** ML is widely used in recommendation systems (e.g., Netflix, Amazon) to provide personalized content and product recommendations to users. By analyzing user behavior and preferences, ML algorithms can offer a tailored experience to individual users.
4. **Data-driven Decision Making:** ML enables organizations to make data-driven decisions based on insights extracted from their data. It can identify trends, correlations, and outliers in data that can inform strategic decisions and improve business processes.

Types of Machine Learning:

ML can be categorized into several types based on the learning approach and the nature of the tasks they are designed for. Some common types of Machine Learning include:

1. **Supervised Machine Learning:**

   Supervised learning is a type of ML in which machines are trained using well "labeled" training data, and on the basis of that data, machines predict the output.

   If data has both input as well as output column and the task is to find relationship between both of them so, for every new input our model predicts the output is called supervised ML.

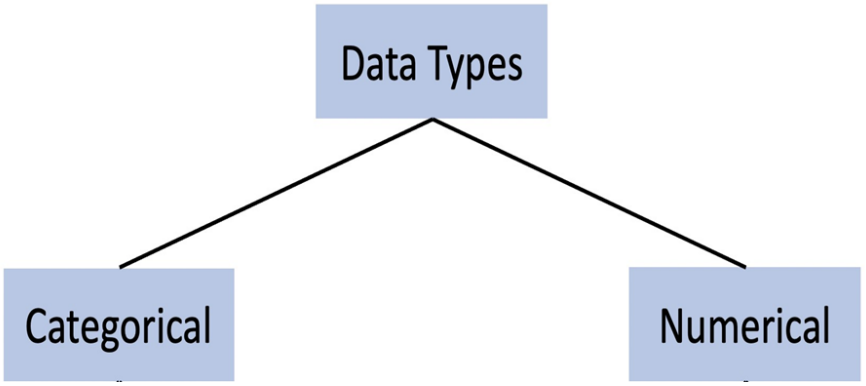
   Through iterative optimization of an objective function, supervised learning algorithms learn a function that can be used to predict the output associated with new inputs. An optimal function will allow the algorithm to correctly determine the output for inputs that were not a part of the training data. An

algorithm that improves the accuracy of its outputs or predictions over time is said to have learned to perform that task.

Example- (Students)

IQ	CGPA	Placement
87	7.1	Yes
111	8.9	Yes
56	5.3	No

Supervised learning algorithms includes-



A. **Regression:**

If we are working in supervised ML and we have input as well as output column, the output column is numerical and that supervised ML problem is called regression.

Example- (Students)

IQ	CGPA	Package
89	8.6	6L
42	5.1	3L
111	9.3	10L

B. **Classification:**

If we are working in supervised ML and we have input as well as output column, the output column is categorical and that supervised ML problem is called classification.

Example- (Email Spam Prediction)

Email	**Words**	**Prediction**
College Notes	Notes, Theory etc	Not Spam
Lottery	Free, 100% off etc	Spam
LinkedIn	Connection Request	Not Spam

2. **Unsupervised Machine Learning:**

In Unsupervised ML, the machine uses unlabeled data and learns on itself without any supervision. The machine tries to find a pattern in the unlabeled data and gives a response.

If data has only an input column and no output column and the model predicts the output by its own, this is called unsupervised learning.

Unsupervised learning algorithms includes-

A. **Clustering:**

It divides the overall data into small small groups known as clusters. Clustering in unsupervised learning is like sorting objects into different groups based on their similarities, without knowing in advance what those groups should be.

Example- Let's say we have a bunch of fruits, and you start by putting all the red apples in one basket, all the yellow bananas in another, and

all the green pears in yet another. We continue to create baskets for different types of fruits based on their similarities. This process of grouping similar items without prior knowledge of the categories is what clustering algorithms do in unsupervised learning.

B. Dimensionality Reduction:

By reducing the input columns, we combine or remove the columns that are not useful according to our predictions or the columns that do not have an effect on our output. This process is called dimensionality reduction.

Example- Let's consider a real estate dataset with various features, including the number of rooms and the number of master rooms in a house, as well as the square footage area.

If you believe that the number of rooms and the number of master rooms provide similar information and can be combined into a single feature, you can perform dimensionality reduction.

You could create a new feature, such as "total room area," by multiplying the number of rooms by the square footage area. This single feature captures the essence of the two original features, reducing the dimensionality of the dataset while preserving relevant information.

C. Anomaly Detection:

Anomaly detection in unsupervised learning is the process of identifying unusual or rare data points that deviate significantly from the norm or expected behavior within a dataset. Anomalies, often referred to as outliers, can represent events or observations that are different in some way, which may be due to errors, fraud, or interesting

phenomena. The goal is to automatically detect such anomalies without the need for labeled data (unlike unsupervised learning).

Example- In the stock market, anomaly detection can be used to identify unusual trading behavior, market manipulation, or news-driven price movements. Traders and investors rely on these techniques to gain insights into market dynamics and to make informed decisions. It can also be helpful in automated trading systems for executing orders based on anomalies, thus potentially capitalizing on market inefficiencies.

D. **Association Rule Learning:**

Association rule learning is a ML technique used to discover interesting relationships, patterns, or associations in large datasets. It is a form of unsupervised learning, where the goal is to identify rules that describe the co-occurrence of items in transactions or events. Association rule learning is a type of unsupervised learning technique that checks for the dependency of one data item on another data item and maps accordingly so that it can be more profitable.

Example- Walmart.

3. **Semi-Supervised Learning:**

Semi-supervised learning is a type of ML that falls in between supervised and unsupervised learning. It is a method that uses a small amount of labeled data and a large amount of unlabeled data to train a model. The goal of semi-supervised learning is to learn a function that can accurately predict the output variable based on the input variables, similar to supervised learning.

Example- Google photos, after analyzing all the photos google photos identifies the people in the photos and make there separate individual groups and ask the user to provide the name of the person after giving the name of the

person google photos automatically write the name of that person in all of his/her photos.

4. **Reinforcement Learning:**

   Reinforcement learning is a ML training method based on rewarding desired behaviors and punishing undesired ones.There is no data provided, the model learns from scratch. It improves by making mistakes.

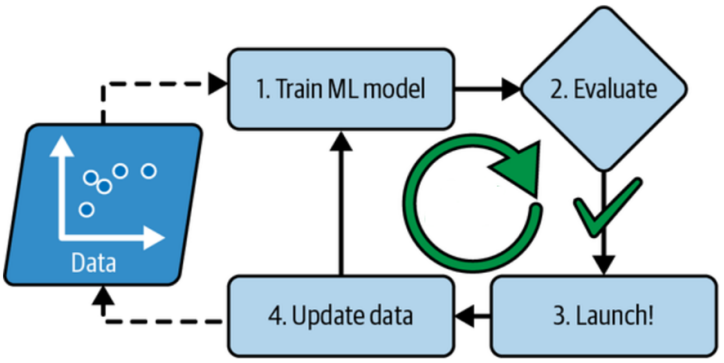
   Reinforcement learning uses algorithms that learn from outcomes and decide which action to take next. After each action, the algorithm receives feedback that helps it determine whether the choice it made was correct, neutral or incorrect. It is a good technique to use for automated systems that have to make a lot of small decisions without human guidance.
   Example- In reinforcement learning, a robot can be trained to perform various tasks through trial and error while receiving feedback in the form of rewards. If the robot does anything wrong the system will warn the robot and deduct some points but when the robot completes the given task it will receive a reward, by this the robot will automatically learn what is wrong and what is right.

Types of Machine Learning, How ML model is trained:

1. **Batch/Offline Machine Learning-**

   It is a conventional way to train a machine learning model, in which you can use whole data to train the model. The training process occurs in distinct stages.

   First of all you will save your data offline on your personal server to train the model and then you deploy the data on an online server.

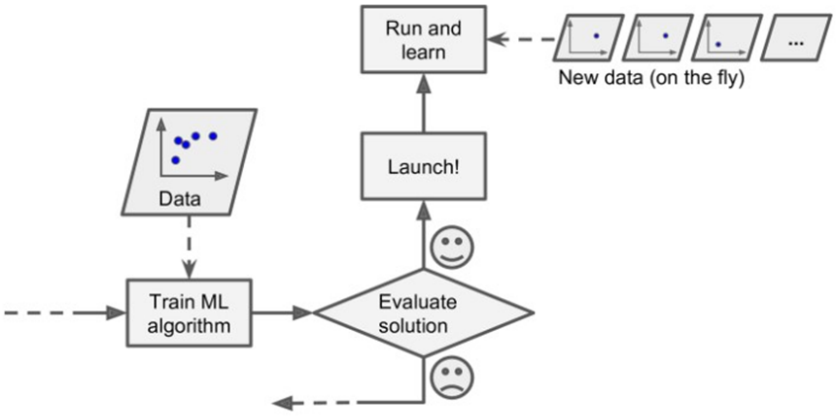


(Fig. 1.2)

2. Online Machine Learning-

It is quite similar to batch learning, which is done incrementally. In this, we provide data to our model in small batches sequentially. These are also called mini-batches, and after each batch, our model improves.

Online ML is a type of machine learning where data is acquired sequentially and is utilized to update the best predictor for future data at each step.



(Fig. 1.3)

Types of Machine Learning, How ML model learns:

1. **Instance Based Learning:**

   In this are model does not train or learn anything but stores the data and when a user gives an new input, on the basis of the stored data our model predicts and it is also known as lazy learner. Each time whenever a new query is encountered, its previously stores data is examined. And assign a target function value for the new instance.

2. **Model Based Learning:**

   In this model trains and learns from the data and they find mathematical relationships between input and output columns. The model draws a curve by learning the data points and by that when a user gives a new input, if the data point lies under the curve answer is no and if data point lies outside the curve answer is yes.

Difference between Supervised Learning and Unsupervised Learning

Supervised Learning	Unsupervised Learning
Supervised learning algorithms are trained using labeled data.	Unsupervised learning algorithms are trained using unlabeled data.
The goal of supervised learning is to train the model so that it can predict the output when it is given new data.	The goal of unsupervised learning is to find the hidden patterns and useful insights from the unknown dataset.
Supervised learning model takes direct feedback to check if it is predicting correct output or not.	Unsupervised learning model does not take any feedback.
In supervised learning, input data is provided to the model along with the output.	In unsupervised learning, only input data is provided to the model.

(Table no.2)

Applications of Machine Learning:

1. **Retail Sector:**

   Demand Forecasting: ML is used to predict customer demand for products, optimizing inventory management, and reducing costs.

   Association Rule Learning: This technique helps discover patterns and associations among products in the data, enabling recommendations and cross-selling.

   Example - **Amazon** employs ML for personalized product recommendations and demand forecasting to stock products efficiently.

2. **Banking and Finance Sector:**

   Fraud Detection: ML models are crucial in identifying fraudulent activities by analyzing transaction patterns and behavior.

   Anomaly Detection: They can detect unusual activities or deviations from standard behavior that may indicate fraud or other security threats.

   Example - **Loan Eligibility Prediction Model**: Banks use ML to assess a customer's creditworthiness and predict their eligibility for loans, improving decision-making.

3. **Transport Sector:**

   Route Optimization: ML models analyze real-time traffic data and historical patterns to suggest the shortest and fastest routes for vehicles.

   Traffic Prediction: They can predict the traffic conditions, helping drivers avoid delays.

   Example - **Ola**, a ride-hailing service, employs ML for route optimization and dynamic pricing based on demand and traffic conditions.

4. **Automobile Sector:**

   Self-Driving Cars: ML, especially deep learning and computer vision, is used to build autonomous vehicles that can perceive their environment and make driving decisions.

   Manufacturing Automation: Automation in car manufacturing processes, including quality control, predictive maintenance, and robotics, improves efficiency and product quality.

   Example - **Tesla**, utilizes ML to power its Autopilot feature, enabling self-driving capabilities in their electric vehicles.

Introduction to Support Vector Machine (SVMs)

SVMs is a type of supervised learning algorithm that can be used for regression or classification tasks. The main idea behind SVMs is to find a hyperplane that maximally separates the different classes in the training data. This is done by finding the hyperplane that has the largest margin, which is defined as the distance between the hyperplane and the closest data points from each class.

Once the hyperplane is determined, new data can be classified by determining on which side of the hyperplane it falls. SVMs are particularly useful when the data has many features, and/or when there is a clear margin of separation in the data.

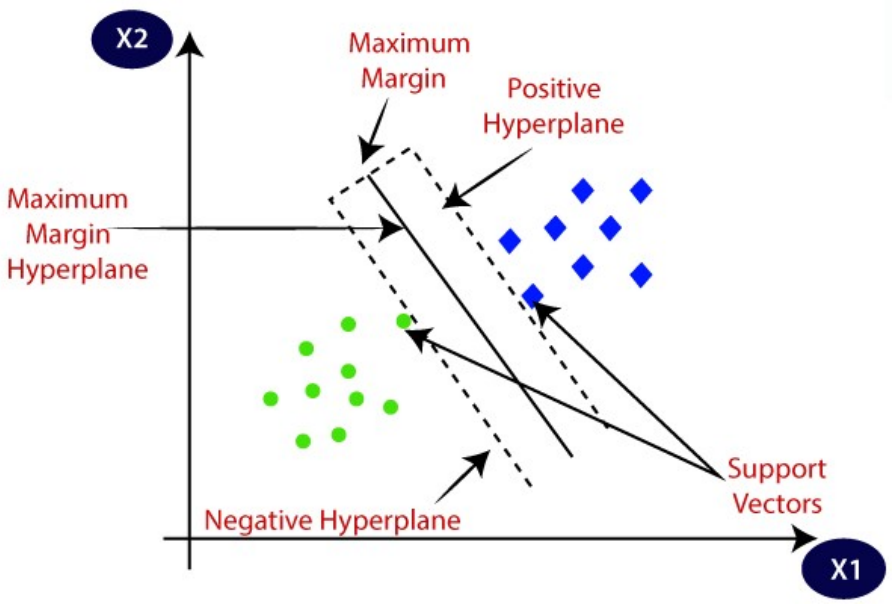
They are known for their ability to handle high-dimensional data, and resist overfitting, making them a valuable tool in ML.

What is a Support Vector Machines

SVMs is a relatively simple Supervised ML Algorithm used for classification or regression. It is more preferred for classification but is sometimes very useful for regression as well. Its primary purpose is to find a hyperplane that best separates data points into different classes while maximizing the margin between the classes. In 2-dimensional space, this hyper-plane is nothing but a line. In SVMs, we plot each

data item in the dataset in an N-dimensional space, where N is the number of features/attributes in the data.

- Hyperplane: In the context of SVMs, a hyperplane is a high-dimensional linear decision boundary that separates data points into different classes.
- Support Vectors: Support vectors are data points that are closest to the decision boundary (the hyperplane) and have the smallest margin.
- Margin: The margin is the distance between the support vectors and the decision boundary (hyperplane). The objective of SVMs is to maximize this margin while still correctly classifying the data points.
- Training an SVMs: To train an SVMs, you provide it with a labeled dataset, and it learns to find the optimal hyperplane that separates the data points.
- Prediction: Once trained, the SVMs can classify new data points by evaluating which side of the hyperplane they fall on.



(Fig. 1.4)

Introduction to Logistic Regression(LR)

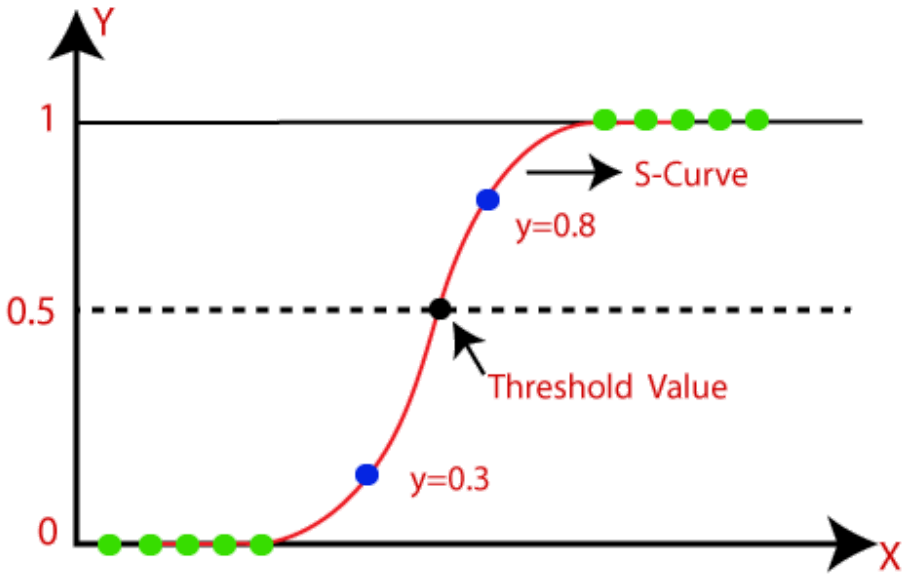
LR is one of the most popular ML algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using

a given set of independent variables. LR predicts the output of a categorical dependent variable.

What is Logistics Regression

LR is a supervised learning technique, where the outcome must be a categorical value. It can be either Yes or No, 0 or 1, True or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1. LR is much similar to Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas LR is used for solving the classification problems.

In LR, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1). LR is a significant ML algorithm because it has the ability to provide probabilities and classify new data using continuous datasets. LR can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification. The below image is showing the logistic function:



(Fig. 1.5)

Sigmoid function in LR

The sigmoid function is a mathematical function used to map the predicted values to probabilities. It ranges all the value from 0 to 1

It maps any real value into another value within a range of 0 and 1. The value of the LR must be between 0 and 1, which cannot go beyond this limit, so it forms a curve like the “S” form.

The S-form curve is called the Sigmoid function or the logistic function.

In LR, we use the concept of the threshold value, which defines the probability of either 0 or 1. Such as values above the threshold value tends to 1, and a value below the threshold values tends to 0.

Loan Eligibility Prediction

- Loan prediction is a process in which a lender assesses an applicant's eligibility for a loan by examining various background factors. These factors typically include the applicant's credit score, income, prevalent interest rate, tentative loan tenure, nature of employment of the applicant, market value of the securities, desired loan amount, any previous loans(if any then the time of repaying the loan), employment status, and assets etc.
- The key principle is to determine whether the applicant's profile aligns with the attributes of past borrowers who have successfully repaid their loans in a timely manner. ML algorithms leverage this historical data and comparison with other applicants to create a data science problem aimed at predicting the loan status of a new applicant based on similar criteria.
- Through a series of steps, including data preprocessing, feature selection, and model training, ML algorithms are trained to recognize patterns and relationships between these factors and loan approval decisions. Once trained, these models can evaluate new loan applications, providing a probability estimate of loan approval or denial based on the applicant's information.

Lenders can use these predictions to make well-informed and objective decisions, setting threshold probabilities for approval or denial.

- This method enhances the efficiency and accuracy of the loan approval process, as it can handle large datasets, reduce human bias, and make real-time, data-driven lending decisions. By making the process more objective and equitable, loan prediction using ML benefits both financial institutions and loan applicants, ultimately leading to more informed and fair loan approval outcomes.
- The process of loan prediction typically involves the following steps:

1. Data Collection: Gathering historical data on loan applicants, which includes various attributes such as credit score, income, employment history, loan amount, and loan outcome (approved or denied).
2. Data Preprocessing: Cleaning and preparing the data by handling missing values, outliers, and ensuring data consistency. Categorical variables may be encoded.
3. Feature Selection: Identifying and selecting the most relevant features that strongly influence loan approval or denial. This step helps reduce noise and improve model performance.
4. Model Selection: Choosing an appropriate ML algorithm for the task. The choices are logistic regression and support vector machines.
5. Model Training: Using the historical data, the selected ML model is trained to learn patterns and relationships between the features and loan approval outcomes. This step involves adjusting the model's parameters to optimize its performance.
6. Model Evaluation: Assessing the trained model's predictive performance using evaluation metrics like accuracy score. This helps determine how well the model generalizes to new data and whether it's suitable for the task.
7. Prediction: When a new loan application is submitted, the trained model is utilized to predict the probability of loan approval or denial based on the applicant's information.

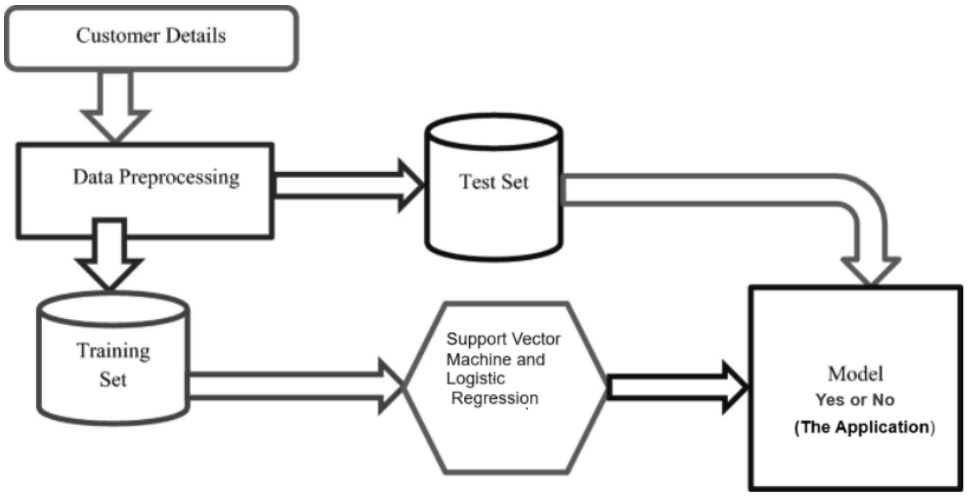
8. Decision Making: Lenders use the predicted probability, typically applying a predefined threshold, to make informed decisions about whether to approve or deny the loan application. The threshold can be adjusted to balance risk and the number of approved loans.

These steps are fundamental to the loan prediction process using ML. By following this structured approach, financial institutions can improve the efficiency and accuracy of their lending decisions, ultimately benefiting both lenders and loan applicants.

Factors that are used to predict the Loan Eligibility:

1. **Credit Score:** The most important factor considered in loan eligibility criteria is credit score. A positive credit score is vital for loan eligibility, as it demonstrates a borrower's past repayment reliability and trustworthiness.
2. **Applicant Income:** The applicant's income is a crucial factor, as it directly influences the loan amount they can borrow and their repayment capacity.
3. **Loan Amount:** The requested loan amount is essential, as it should align with the applicant's financial situation and repayment capability.
4. **Co-applicant Income:** Co-applicant income can boost the combined household income, potentially increasing the loan eligibility amount.
5. **Married:** Marital status can affect eligibility, as married individuals may have combined incomes or shared financial responsibilities that impact their loan application.
6. **Dependents:** The number of dependents a person has can influence their ability to repay a loan, as more dependents may require higher income to cover expenses.
7. **Education:** Educational qualifications can be a factor, as higher education might lead to better job prospects and higher income, positively impacting loan eligibility.

8. **Self Employed:** Self-employed individuals may face different loan eligibility criteria compared to salaried individuals, considering variations in income stability and documentation
9. **Loan Amount Term:** The loan term affects the monthly repayment amount, and a longer term might increase eligibility, but it can also result in higher interest payments.
10. **Property Area:** The location of the property can be a factor, as it might impact the property's value and, consequently, the loan amount that can be granted.

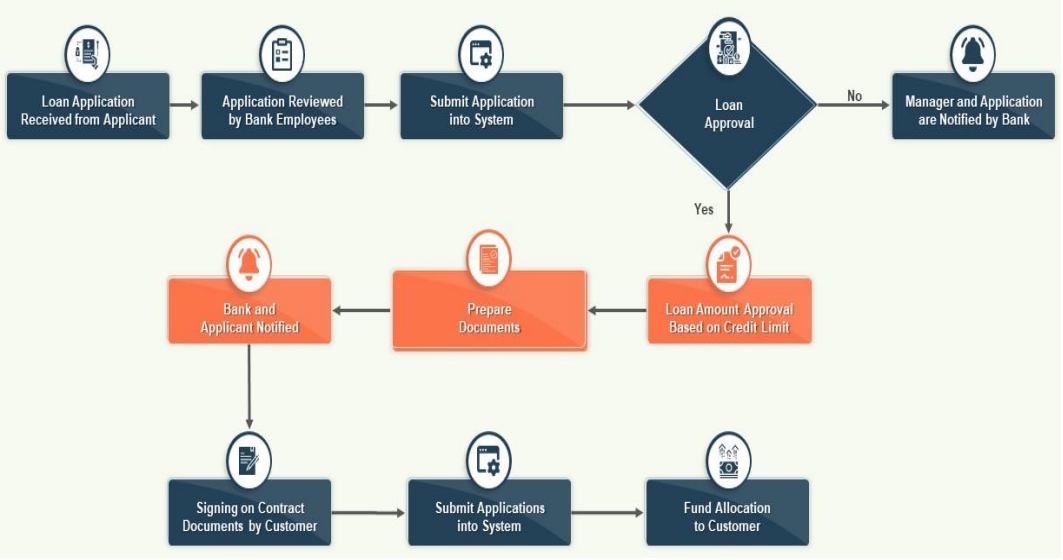


(Fig. 1.6)

Traditional Method

- Traditional processes determine the risk by manually looking at the applicant's income, credit history, and several other dynamic parameters and creating a data-driven risk model.
- As evaluating loan eligibility, manually going through heaps of paperwork just to check if an applicant qualifies for a particular loan scheme or not can invariably elongate the lending lifecycle.

- The traditional method of assessing Loan Eligibility has been deeply entrenched in the financial industry for many years. In this age-old approach, the evaluation of an individual's eligibility for a loan primarily revolves around a manual and somewhat rigid process.
- Key determinants in this method include factors like income, credit history, employment status. Each of these elements is individually scrutinized by loan officers, who exercise their judgment to make lending decisions.
- While this approach has been in use for decades, it comes with certain inherent limitations. It is often time-consuming, prone to errors, and can be influenced by biases. Moreover, it may not fully capture an applicant's true financial capability and potential, as it relies on historical data and simple rule-based assessments.
- In an era where technology and data analytics have revolutionized the financial sector, the traditional method of loan eligibility assessment is increasingly being replaced by more sophisticated and data-driven approaches, offering greater efficiency and objectivity in decision-making processes.



(Fig. 1.7)

Disadvantages of Traditional Loan Eligibility Method:

1. **Bias and Subjectivity:** Traditional methods can introduce bias, as they rely on human judgment. Loan officers may unconsciously favor or discriminate against applicants based on subjective factors such as race, gender, or personal biases.
2. **High Operational Costs:** Manual processes is expensive for financial institutions due to the need for a substantial workforce to assess loan applications.
3. **Risk of Error:** Human error in data entry and calculations can lead to inaccuracies in loan decisions, potentially approving loans to individuals who should have been denied or vice versa.
4. **Inefficiency:** Manual evaluation of loan applications is time-consuming and labor-intensive. It can result in delays in loan processing, causing frustration for applicants.
5. **Limited Data:** Traditional methods often use a limited set of historical data and rely heavily on credit scores and a few financial factors. This can lead to an incomplete picture of an applicant's creditworthiness.
6. **Lack of Flexibility:** Traditional systems may not easily adapt to changing economic conditions or evolving lending practices, making them less responsive to market dynamics.

Why is Loan Prediction Needed?

Loan prediction is a vital component of the modern lending landscape, addressing the shortcomings of manual loan application processing. The traditional approach to evaluating loan applications is not only time-consuming but also prone to errors, inefficiencies, and, at times, subjectivity that can introduce biases.

By contrast, loan prediction, with the aid of ML tools and techniques, offers a transformative solution. It enables financial organizations to swiftly assess loan applications, categorize high-risk applicants for rejection, identify creditworthy customers for approval, and designate certain cases for manual review.

This streamlined process, when incorporating ML, has the potential to significantly reduce loan processing times, making the lending process more efficient and responsive. **In fact, it is estimated that loan processing times can be cut by almost 40% when such predictive models are put into action**. This not only benefits financial institutions by minimizing risks but also ensures that qualified borrowers can access the loans they need in a more timely and equitable manner.

Advantages of Loan Eligibility Prediction:

1. **Lower Operational Costs:** Automation reduces the need for manual data entry and assessment, which can significantly lower operational costs for lending institutions. This cost reduction can be passed on to borrowers in the form of lower interest rates or fees.
2. **Efficiency:** Loan eligibility prediction automates the assessment process, which significantly reduces the time and effort required for manual application reviews. This leads to faster decision-making and quicker disbursement of funds.
3. **Risk Management:** By using predictive models, lenders can better assess the risk associated with a loan application. This results in more accurate risk profiles for borrowers, helping lenders make informed decisions about interest rates, loan terms, and whether to approve or deny an application.
4. **Quick Feedback:** Borrowers can receive immediate feedback on the likelihood of loan approval. If an application is denied, they can explore other options rather than waiting for a lengthy manual review process.
5. **Fraud Detection:** Predictive models can help identify fraudulent loan applications by analyzing patterns and anomalies in the data. This safeguards the institution from potential losses due to fraudulent activities.
6. **Reduced Human Bias:** ML models base their decisions on data and algorithms, minimizing the impact of human bias in the decision-making process. This can lead to fairer lending practices and help avoid discrimination.

Chapter - III: Practical Implementation of the Project

Introduction

The loan providing companies find it hard to use traditional methods of evaluating loan eligibility often rely heavily on manual assessment, leading to potential biases and delays. Additionally, the loan applications can overcome manual review processes, making it difficult for timely decision-making. This inefficiency may result in missed opportunities for the institution and frustration for loan applicants. To address this challenge, the business seeks to implement a data-driven solution using machine learning algorithms to automate and optimize the loan prediction process. The goal is to improve efficiency, reduce bias, and enable faster, more objective decision-making, ultimately enhancing the overall effectiveness of the loan approval system.

Data Set Used

Loan_ID	Gender	Married	Dependen	Education	Self_Empl	ApplicantI	Coapplicar	LoanAmou	Loan_Amc	Credit_His	Property_	Loan_Stat
LP001002	Male	No	0	Graduate	No	5849	0		360	1	Urban	Y
LP001003	Male	Yes	1	Graduate	No	4583	1508	128	360	1	Rural	N
LP001005	Male	Yes	0	Graduate	Yes	3000	0	66	360	1	Urban	Y
LP001006	Male	Yes	0	Not Gradu	No	2583	2358	120	360	1	Urban	Y
LP001008	Male	No	0	Graduate	No	6000	0	141	360	1	Urban	Y
LP001011	Male	Yes	2	Graduate	Yes	5417	4196	267	360	1	Urban	Y
LP001013	Male	Yes	0	Not Gradu	No	2333	1516	95	360	1	Urban	Y
LP001014	Male	Yes	3+	Graduate	No	3036	2504	158	360	0	Semiurbar	N

Rows and Columns (614, 13)

(Fig. 2)

Project Work

Importing the libraries

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn import svm
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
```

(Fig. 2.1)

Loading the dataset and processing

```
loan_dataset = pd.read_csv("loan dataset.csv")
```

```
# Taking the overview of the dataframe by looking at first 5 rows
loan_dataset.head()
```

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History
0	LP001002	Male	No	0	Graduate	No	5849	0.0	NaN	360.0	1.0
1	LP001003	Male	Yes	1	Graduate	No	4583	1508.0	128.0	360.0	1.0
2	LP001005	Male	Yes	0	Graduate	Yes	3000	0.0	66.0	360.0	1.0
3	LP001006	Male	Yes	0	Not Graduate	No	2583	2358.0	120.0	360.0	1.0
4	LP001008	Male	No	0	Graduate	No	6000	0.0	141.0	360.0	1.0

(Fig. 2.2)

```
# Statistical analysis of numerical columns only
loan_dataset.describe()
```

	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History
count	614.000000	614.000000	592.000000	600.00000	564.000000
mean	5403.459283	1621.245798	146.412162	342.00000	0.842199
std	6109.041673	2926.248369	85.587325	65.12041	0.364878
min	150.000000	0.000000	9.000000	12.00000	0.000000
25%	2877.500000	0.000000	100.000000	360.00000	1.000000
50%	3812.500000	1188.500000	128.000000	360.00000	1.000000
75%	5795.000000	2297.250000	168.000000	360.00000	1.000000
max	81000.000000	41667.000000	700.000000	480.00000	1.000000

(Fig. 2.3)

```
# Calucalting the number of missing values in each column
loan_dataset.isnull().sum()
```

```
Loan_ID              0
Gender              13
Married              3
Dependents          15
Education            0
Self_Employed       32
ApplicantIncome      0
CoapplicantIncome    0
LoanAmount          22
Loan_Amount_Term    14
Credit_History      50
Property_Area        0
Loan_Status          0
dtype: int64
```

(Fig. 2.4)

```
# Dropping the missing values
loan_dataset = loan_dataset.dropna()
```

```
# checking whether the missing values are dropped
loan_dataset.isnull().sum()
```

```
Loan_ID              0
Gender               0
Married              0
Dependents           0
Education            0
Self_Employed        0
ApplicantIncome      0
CoapplicantIncome    0
LoanAmount           0
Loan_Amount_Term     0
Credit_History       0
Property_Area        0
Loan_Status          0
dtype: int64
```

(Fig. 2.5)

```
# label encoding
loan_dataset.replace({'Loan_Status':{'N': 0,'Y': 1}},inplace = True)
# inplace = True it modifies the original dataset
```

```
# checking whether the values of Loan_Status as converted
loan_dataset.head()
```

Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
1	Graduate	No	4583	1508.0	128.0	360.0	1.0	Rural	0
0	Graduate	Yes	3000	0.0	66.0	360.0	1.0	Urban	1
0	Not Graduate	No	2583	2358.0	120.0	360.0	1.0	Urban	1
0	Graduate	No	6000	0.0	141.0	360.0	1.0	Urban	1
2	Graduate	Yes	5417	4196.0	267.0	360.0	1.0	Urban	1

(Fig. 2.6)

Converting all the categorical features into numerical features

```
loan_dataset = loan_dataset.replace({'Gender':{'Male':1,'Female':0},'Married':{'No':0,'Yes':1},
                                     'Education':{'Graduate':1,'Not Graduate':0},'Self_Employed':{'No':0, 'Yes':1},
                                    'Property_Area':{'Rural':0,'Semiurban':1,'Urban':2}})
```

```
# Checking whether all the categorical features are converted into numerical features
loan_dataset.head()
```

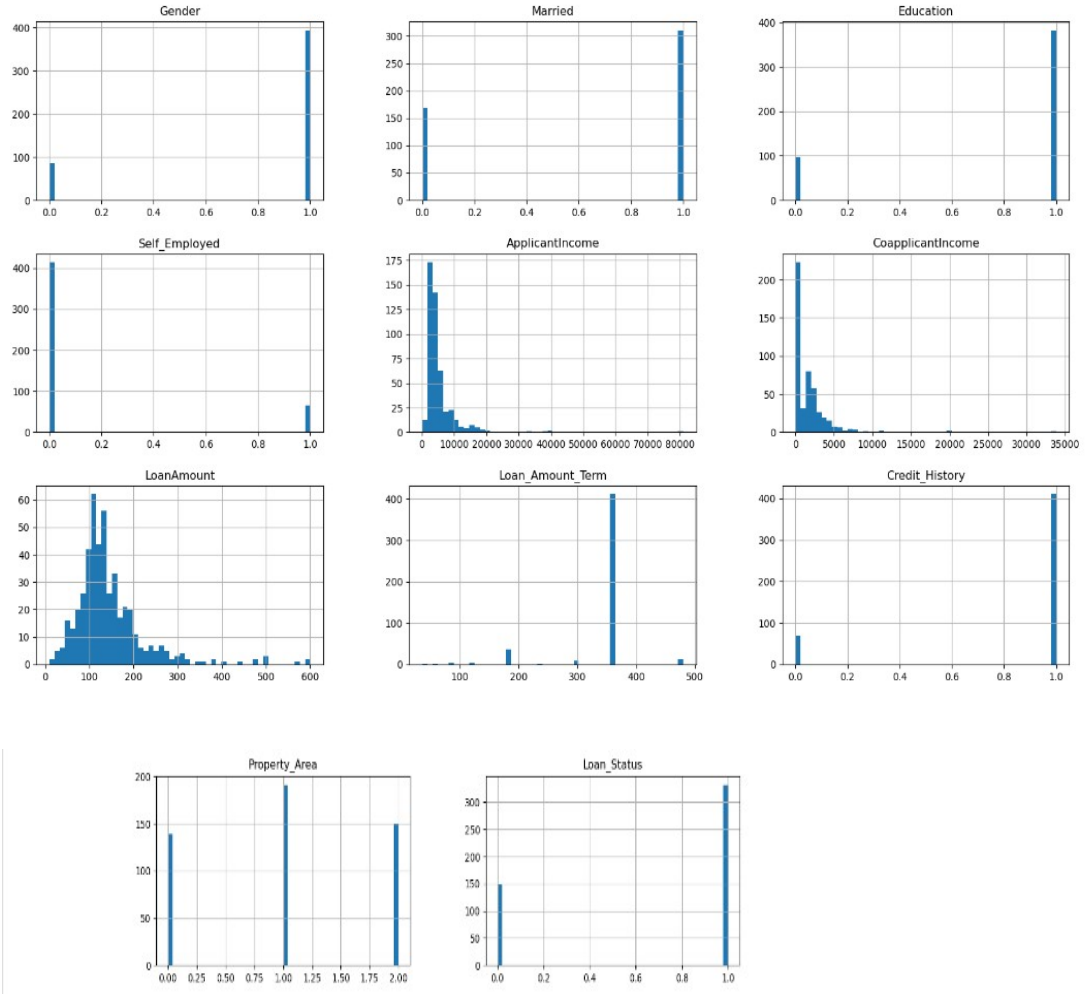
	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History
1	LP001003	1	1	1	1	0	4583	1508.0	128.0	360.0	1.0
2	LP001005	1	1	0	1	1	3000	0.0	66.0	360.0	1.0
3	LP001006	1	1	0	0	0	2583	2358.0	120.0	360.0	1.0
4	LP001008	1	0	0	1	0	6000	0.0	141.0	360.0	1.0
5	LP001011	1	1	2	1	1	5417	4196.0	267.0	360.0	1.0

(Fig. 2.7)

Data Visualization

To find the relationship between various features or columns in the dataset

```
# Plotting histrogram for each numerical features
loan_dataset.hist(bins=50,figsize=(20,15))
plt.show()
```



(Fig. 2.8)

Train Test Split

```
x_train, x_test, y_train, y_test = train_test_split(x,y,test_size = 0.2,random_state = 42, stratify = y)
```

```
print(x.shape,x_train.shape,x_test.shape)
```

```
(480, 11) (384, 11) (96, 11)
```

```
print(y.shape,y_train.shape,y_test.shape)
```

```
(480,) (384,) (96,)
```

(Fig. 2.9)

Support Vector Machine Model

```
1 svm = svm.SVC(kernel = 'linear')
```

```
1 svm.fit(x_train,y_train)
```

```
SVC(kernel='linear')
```

```
1 y_pred = svm.predict(x_test)
```

```
1 y_pred
```

```
array([1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 0, 1, 0, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1,
       1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1,
       1, 1, 0, 1, 1, 1, 1, 0, 1, 0, 0, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1,
       1, 1, 1, 1, 0, 0, 1, 1], dtype=int64)
```

```
1 a = round(accuracy_score(y_test,y_pred)*100,2)
```

```
1 a
```

```
83.33
```

(Fig. 2.10)

Logistic Regression Model

```
1 log_reg = LogisticRegression()
```

```
1 log_reg.fit(x_train,y_train)
```

```
C:\Users\DELL\anaconda3\lib\site-packages\sklearn\linear_model\_logistic.py:814: ConvergenceWarning: lbfgs failed to converge
(status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
  n_iter_i = _check_optimize_result(
```

```
LogisticRegression()
```

```
1 y_pred1 = log_reg.predict(x_test)
```

```
1 y_pred1
```

```
array([1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 0, 1, 0, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1,
       1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1,
       1, 1, 0, 1, 1, 1, 1, 0, 1, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 0, 1,
       1, 1, 1, 1, 0, 0, 1, 1], dtype=int64)
```

```
1 b = round(accuracy_score(y_test,y_pred)*100,2)
```

```
1 b
```

```
83.33
```

(Fig. 2.11)

Models Evaluation

```
models = pd.DataFrame({
    "Model": ["Support Vector Machine Model","Logistic Regression Model"],
    "Score": [a,b]
})
```

```
models
```

	Model	Score
0	Support Vector Machine Model	83.33
1	Logistic Regression Model	83.33

(Fig. 2.12)

```
# Predicting when new user come to apply for loan
y_pred2 = svm.predict([[1,1,0,1,1,5000,1000,100,360,1,1]])
```

```
C:\Users\DELL\anaconda3\lib\site-packages\sklearn\base.py:450: UserWarning: X does not have valid feature names, but SVC was fitted with feature names
  warnings.warn(
```

```
print(y_pred2)
```

```
[1]
```

(Fig. 2.13)

Chapter - IV: Opportunities

During my tenure as a data intern at the company, I had the opportunity to explore a wide range of roles and responsibilities that contributed to my professional growth. I was exposed to the intricacies of data analysis and interpretation, which opened doors to various career paths in data analytics and business intelligence. These skills could lead to roles such as data analyst, business analyst, or data scientist within the organization.

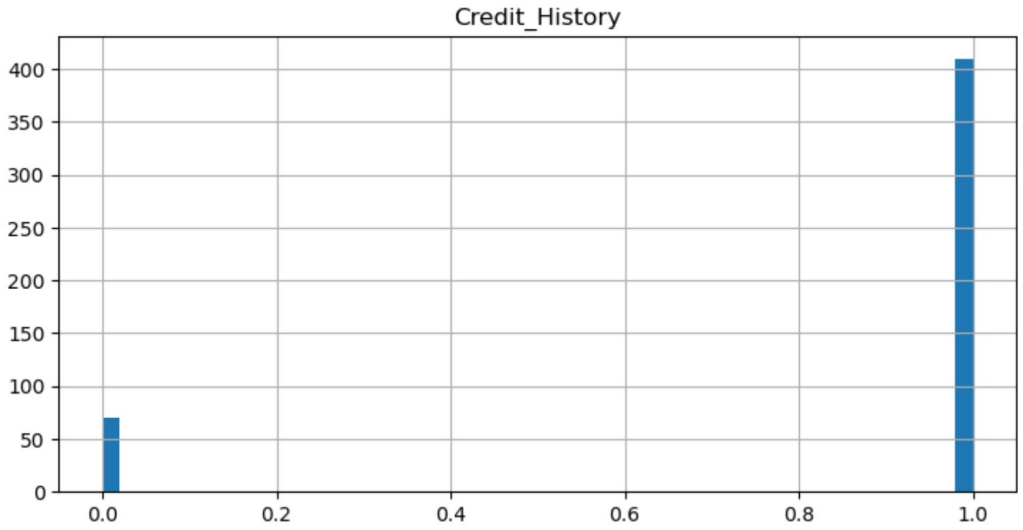
I had the chance to work with various teams, allowing me to build a strong network and gain insights into different aspects of the business. This experience could potentially lead to roles in various domains, where I could use my analytical skills to streamline workflows and enhance overall organizational performance.

My exposure to data-driven decision-making has equipped me with valuable skills that can be applied in various domains, helping the company leverage data to target customers effectively.

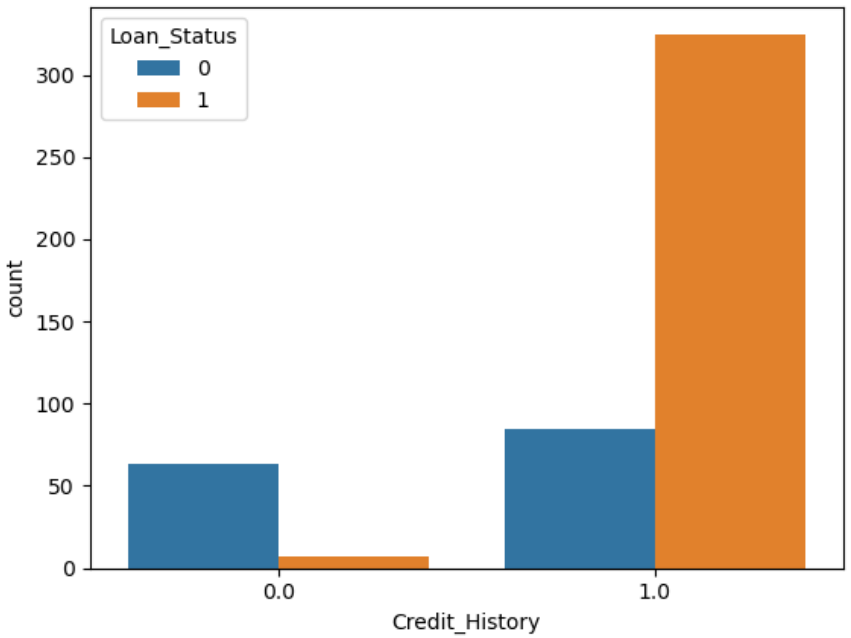
In conclusion, (COMPANY NAME) provided a diverse array of opportunities within the organization, including roles in data analysis, business intelligence, operations, and marketing. The experience has equipped me with a versatile skill set that can be applied to various career paths and have broadened my horizons in terms of future professional prospects within the company.

Chapter - V: Data Analysis & Interpretation

For the 1st visualization that is- The relationship between Credit History and Loan Status:

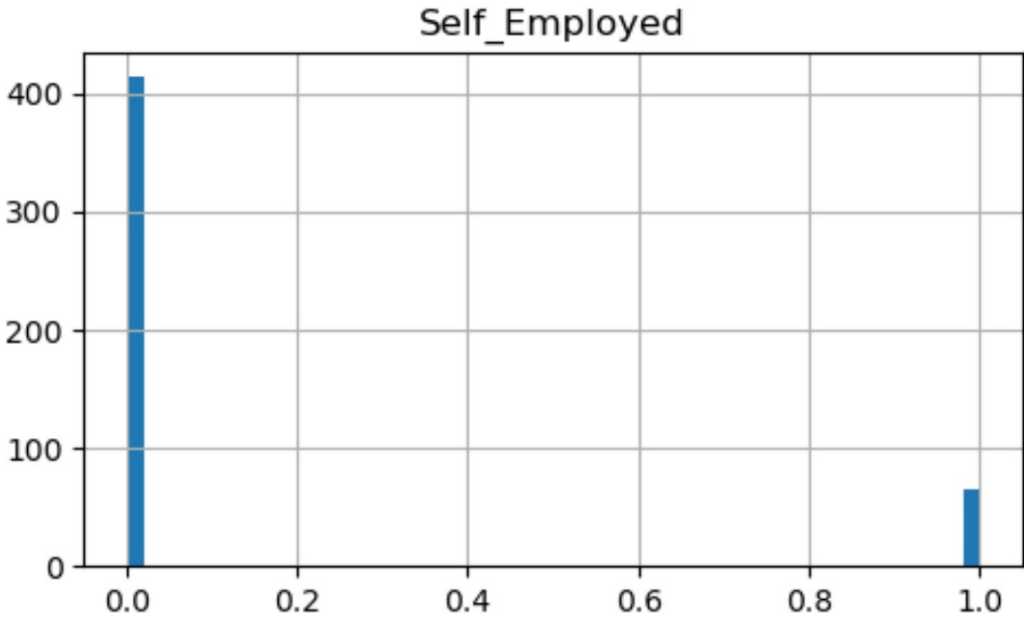


(Fig. 2)

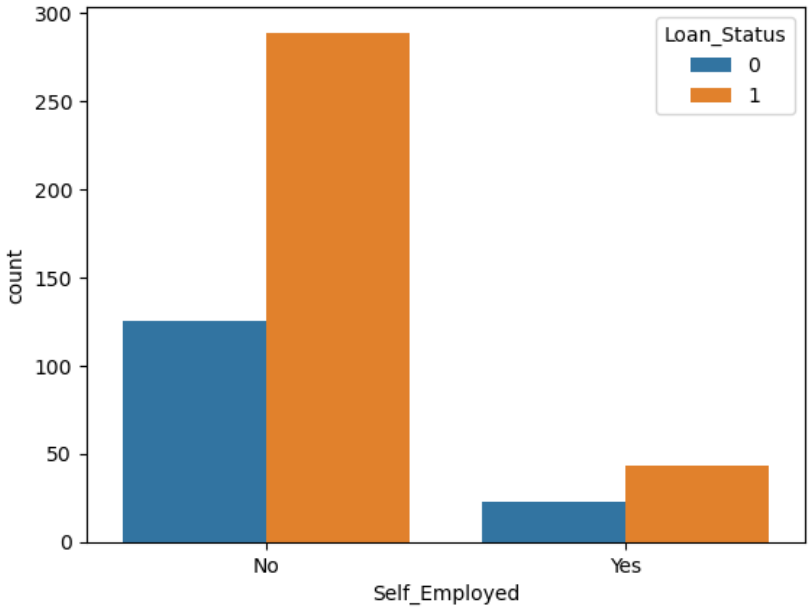


(Fig. 2.1)

For the 2nd visualization that is-What insights can we get by analyzing the relationship between Self Employed and Loan Status:

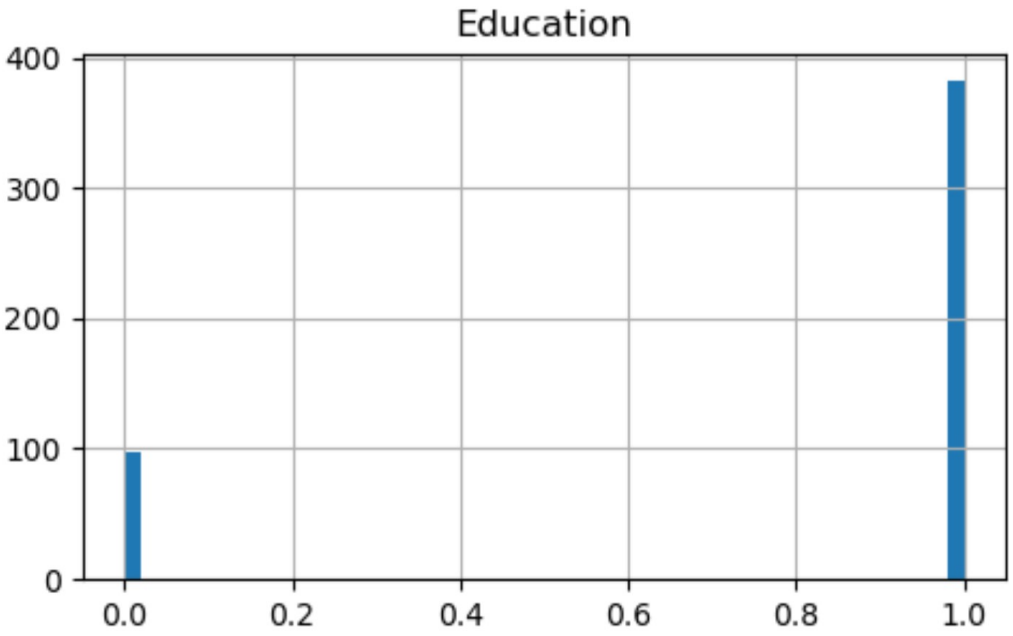


(Fig. 3)

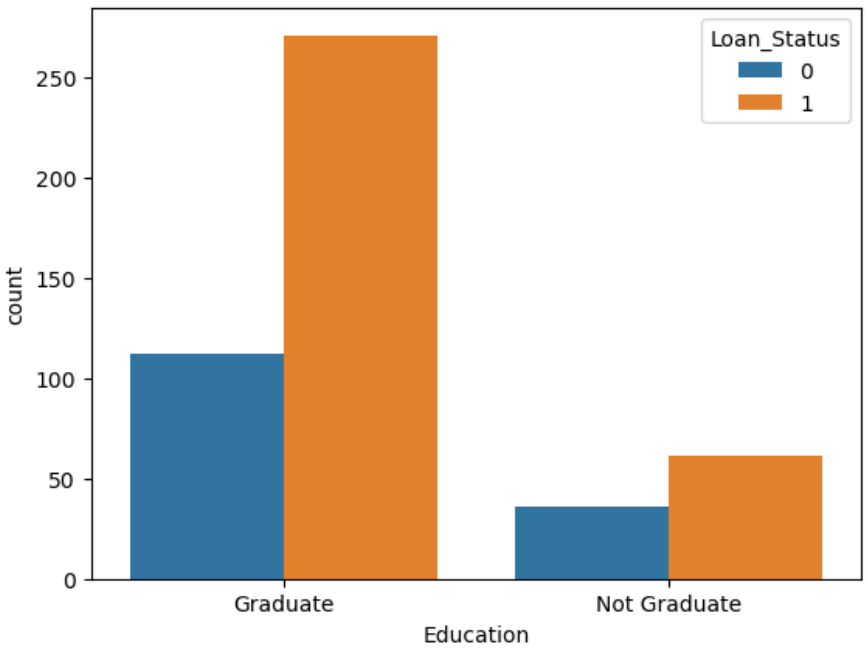


(Fig. 3.1)

For the 3rd visualization that is- What insights can be gained from examining the correlation between Education and the Loan Status:

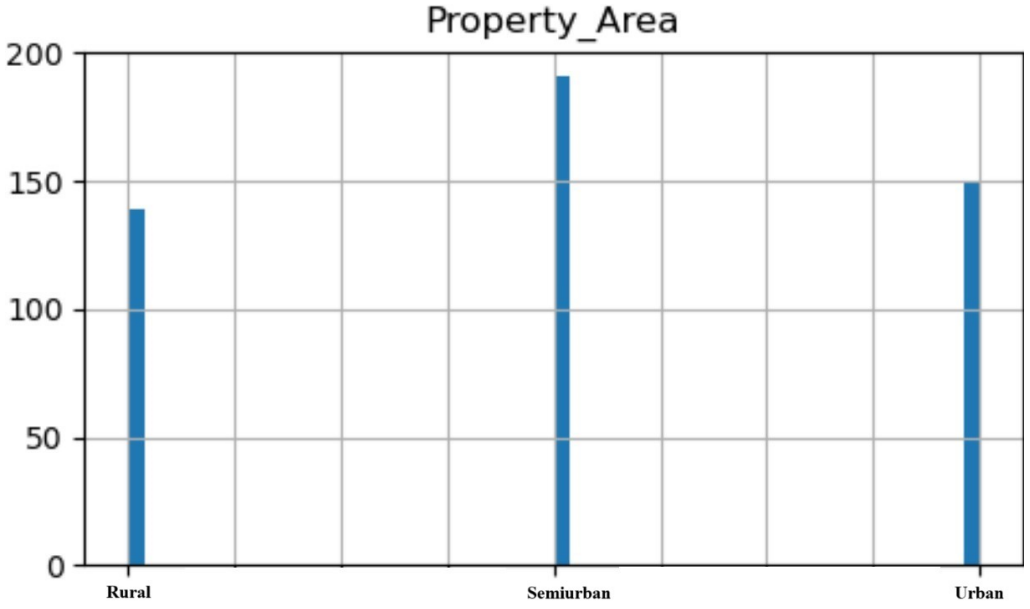


(Fig. 4)

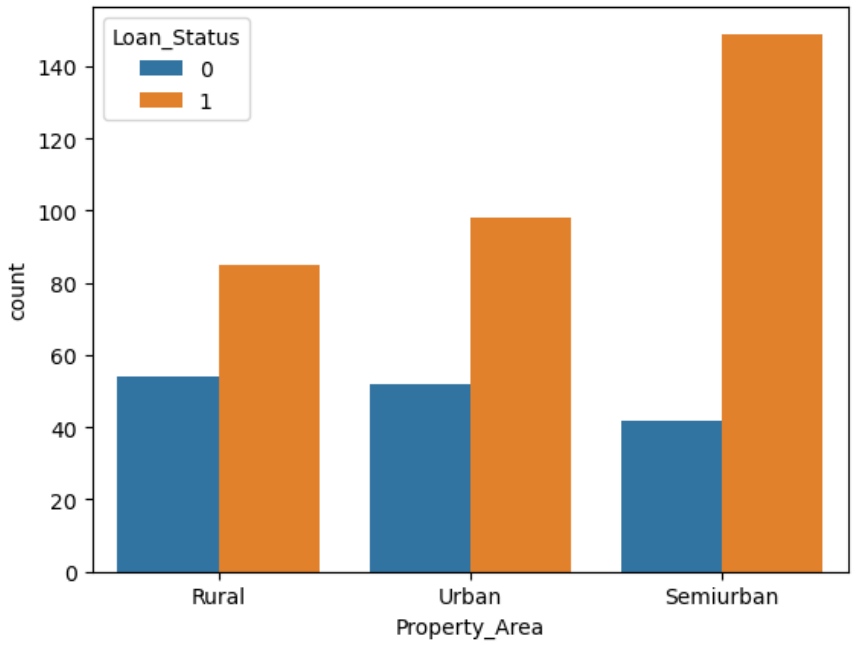


(Fig. 4.1)

For the 4th visualization that is- Which property area people are more likely to apply for the loans and what insights can be derived from exploring the relationship between different Property Area and Loan Status:

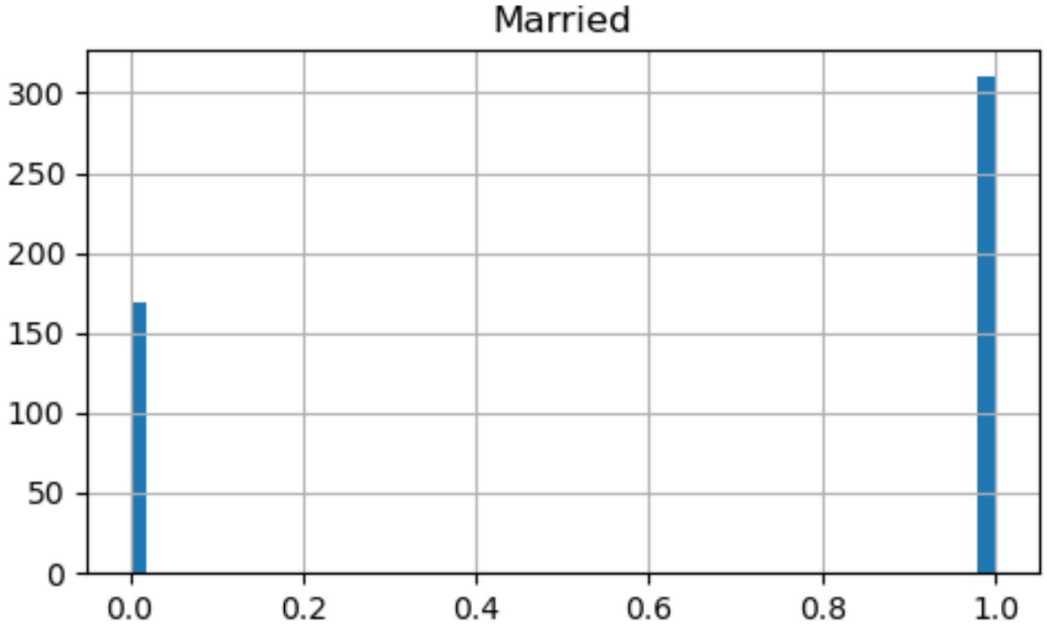


(Fig. 5)

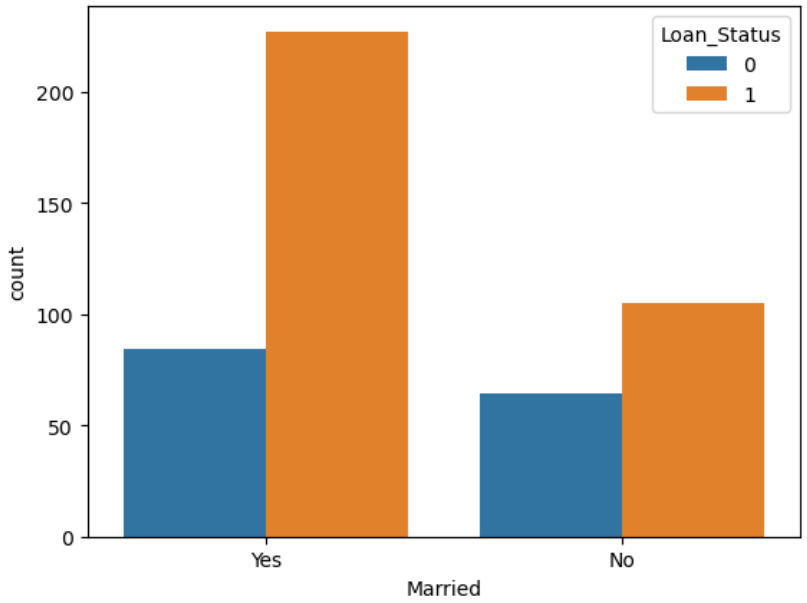


(Fig. 5.1)

For the 5th visualization that is- What information can we gain by analyzing the features of Married and Loan Status:

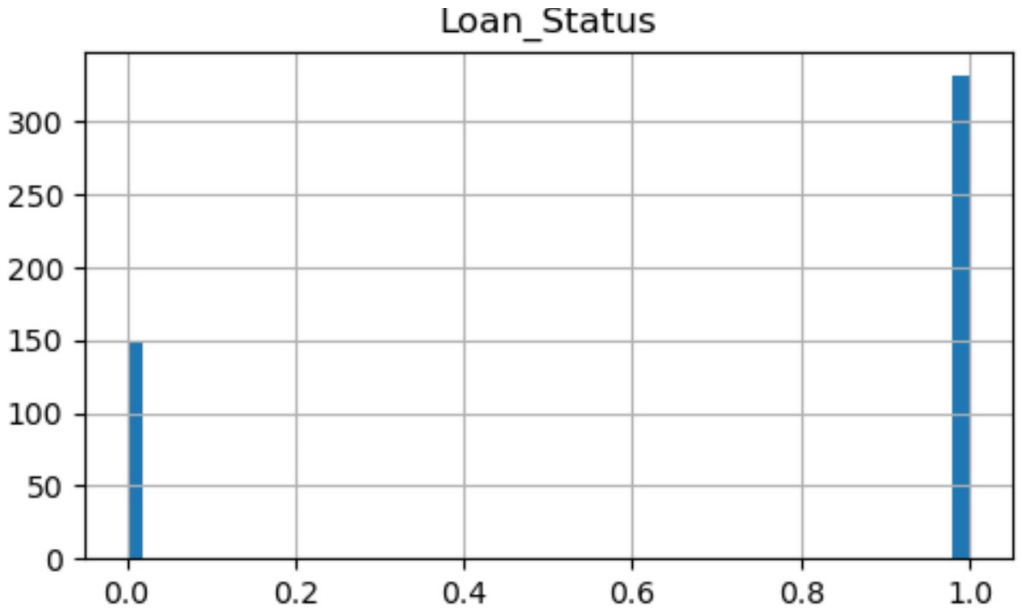


(Fig. 6)

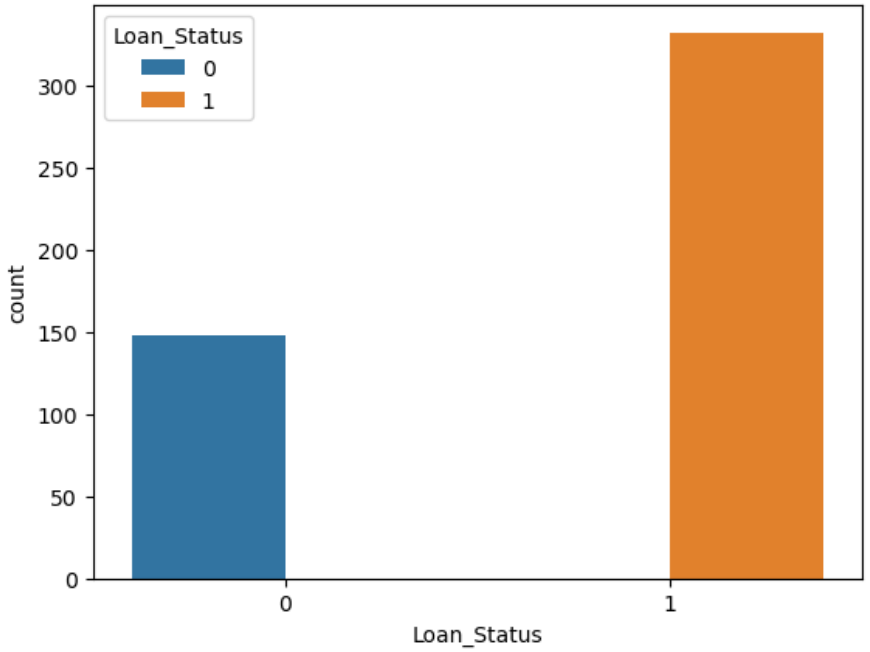


(Fig. 6.1)

For the 6th visualization that is- What is the number of applicants approved for a loan, and how many applicants are rejected among the total applicants:



(Fig. 7)



(Fig. 7.1)

Conclusion(On the basis of data analysis)

By visualizing and analyzing the relationship between various factors and loan status, it becomes apparent that credit history or credit score emerges as the most influential factor affecting loan status.

By identifying the relationship between credit history and loan status, it is observed that a good credit history(which is represented by 1) increases the likelihood of loan approval compared to a poor credit score(which is represented by 0).

After identifying the correlation between self-employment and loan status, it is evident that a higher proportion of non-self-employed individuals are applying for loans compared to self-employed individuals based on my data.

After analyzing the relationship between education and loan status, it is observed that a significant number of individuals with higher education are more likely to apply for loans, according to my data.

Based on my evaluation of the relationship between property area (such as rural, semi-urban, and urban) and loan status, it is evident that individuals from semi-urban areas are more likely to apply for loans and less likely to face rejection compared to those from rural and urban areas. People from rural areas are less likely to apply for loans and more likely to face rejection.

After interpreting the relationship between married and loan status, it is evident that a significant number of loan applicants are married individuals, and they have a higher likelihood of approval, as per my data.

In our dataset, we initially had data for 614 individuals. However, after excluding missing values, we are left with 480 people. Among them, 332 individuals were approved for the loan due to their good credit history, while 148 individuals were denied the loan based on their credit score.

Chapter - VI: Learning & Observations

Learning :

My internship has been a remarkable journey of learning and growth. It provided me with hands-on experience in the field. I learned the intricacies of real-world applications of data analytics and machine learning, from data cleaning and feature engineering to model building and interpretation. The challenges I encountered during my internship, such as managing time and workload, adapting to a new work culture, and ensuring fairness in predictive models, equipped me with essential life skills. This experience reinforced the significance of adaptability, and perseverance in achieving professional success.

During my internship, I learned more than just technical skills. It has taught me the importance of teamwork and collaboration in achieving common goals, being flexible when things get tough. I also realized that it's essential to find a balance between work and personal life for overall well-being. Overall, my internship was not just a valuable educational experience, but also a transformative journey that shaped me into a more capable and well-rounded professional.

Observations :

My observations during the internship have provided valuable insights into the inner workings of the organization and the industry as a whole. I've observed the significance of continuous and transparent communication, both within teams and with clients, as a driving force behind successful projects. Additionally, I've noticed how adaptability and the ability to embrace change are vital qualities in a rapidly evolving business environment. Overall, my observations have reinforced the notion that continuous learning and a proactive approach are key to thriving in the professional world.

Chapter - VII: Challenges faced

During my internship on (COMPANY NAME), I encountered various challenges that significantly impacted the project's progress and outcomes. One of the primary challenges I faced was data quality and availability. Ensuring the data used for training and testing the prediction model was accurate, complete, and up-to-date proved to be a demanding task. Often, there were many missing data that had to be addressed through data cleaning and imputation, which consumed a considerable amount of time and resources.

Significant challenge was the selection of relevant features for the prediction model. Deciding which factors, such as credit score, income, or employment type, would most accurately predict loan eligibility required a deep understanding of the lending industry.

The most crucial challenge was to ensure that the loan eligibility prediction model was fair and unbiased, free from discrimination based on gender, race, or other sensitive attributes. This necessitated the use of fairness-aware algorithms and careful examination of the model's outputs for potential biases.

I also encountered challenges related to time management and workload within the work culture. Balancing the demands of my internship project alongside other responsibilities such as attending meetings, training, and handling administrative tasks often felt like managing conflicting priorities. It was vital for me to learn effective prioritization and time management to meet deadlines while keeping a healthy work-life balance intact.

In conclusion, my internship on (COMPANY NAME) was a valuable learning experience marred by various challenges, ranging from data quality and feature selection to model tuning, fairness considerations, and project report writing. Despite these challenges, I viewed them as opportunities for personal and professional growth, helping me develop essential skills such as adaptability, time management, communication, and resilience.

Chapter - VIII: Suggestions & Recommendations

After a period of 60 days I felt like the organization in which I was working had a really amazing work culture but there was a scope for improvement in the way they used to familiarize their interns with the projects and the problem with which we were dealing, and providing end-to-end guidance to the interns as per my experience.

The organization should consider investing in regular training and skill development programs for its interns and employees. Continuous learning and upskilling are essential in keeping up with rapidly evolving industries and technologies. Establish a culture of continuous improvement in data analytics processes.

This can be achieved by regularly reviewing and optimizing data collection, analysis, and reporting methods. In addition to this, implementing a more structured onboarding program i.e. mentorship initiatives, and a feedback mechanism will contribute to a more enriching experience for interns.

Annexure - I: Bibliography

https://youtube.com/@campusx-official?si=lU8Eh00xKgvfSSII

https://www.geeksforgeeks.org/instance-based-learning/

https://www.geeksforgeeks.org/introduction-to-support-vector-machines-svm/

https://www.projectpro.io/article/loan-prediction-using-machine-learning-project-source-code/632 https://youtube.com/@campusx-official?si=lU8Eh00xKgvfSSII

https://www.researchgate.net/figure/Comparison-of-Machine-Learning-ML-and-Deep-Learning-DL_tbl1_336000398

https://www.javatpoint.com/supervised-machine-learning

https://towardsdatascience.com/unsupervised-learning-and-data-clustering-eeecb78b422a

https://www.javatpoint.com/machine-learning-with-anomaly-detection

https://www.zendesk.com/in/blog/machine-learning-and-deep-learning/

https://www.techtarget.com/searchenterpriseai/definition/machine-learning-ML