



CodeClause
Pvt Ltd

INTERNSHIP REPORT

Data Analyst Internship

PROJECT TITLE : LOAN PREDICTION ANALYSIS

INTERN NAME : HITESH M

COURSE : DATA ANALYSIS

DATE OF SUBMISSION : 12/09/2023



Bengaluru, Karnataka



[hitesh-m-255289278](#)



[HITESHM-023](#)



1nt21ad023.hitesh@nmit.ac.in

Contents

1	Project Introduction	2
1.1	Background	2
1.2	Objective	3
1.3	Scope	3
2	Data Preprocessing	3
2.1	Data Collection	3
2.2	Data Cleaning	4
2.3	Handling Missing Values	4
2.4	Handling Categorical Columns	4
2.5	Feature Scaling	4
2.6	Data Splitting	5
3	Model Building and Evaluation	5
3.1	Logistic Regression	5
3.2	Support Vector Classifier (SVC)	5
3.3	Decision Tree Classifier	6
3.4	Random Forest Classifier	6
3.5	Hyperparameter Tuning	8
4	Model Deployment	8
5	Test Cases	11
6	Conclusion	13
7	Future Enhancements	13
8	References	14

Executive Summary

The "Loan Status Prediction Using Machine Learning" project represents a significant and forward-thinking endeavor that seeks to transform the lending industry by leveraging the capabilities of machine learning. The core objective of this project is to create an advanced, automated system capable of accurately predicting loan statuses. This predictive model is designed to streamline and expedite the loan approval process while simultaneously reducing the inherent risks associated with lending decisions. In an ever-evolving financial landscape, where time and precision are paramount, the potential impact of this project is both promising and far-reaching.

Traditional methods of assessing loan applications are often time-consuming and rely heavily on manual evaluation. This can result in delays for borrowers seeking financial assistance and introduce subjectivity into lending decisions. By harnessing the power of machine learning algorithms, this project aims to revolutionize the way loans are evaluated. Through the analysis of historical loan data, the system will learn to recognize patterns and make predictions about the likelihood of a loan being repaid successfully. This predictive capability not only expedites the approval process but also enhances the accuracy of lending decisions. The implications of this project extend beyond the realm of financial institutions. Borrowers will benefit from faster loan approvals, leading to increased financial accessibility. Lenders, on the other hand, will be better equipped to manage risk, ultimately resulting in more robust and sustainable lending practices. Furthermore, the project's use of machine learning enables continuous improvement, as the system can adapt and refine its predictions over time, staying attuned to shifting economic conditions and borrower behaviors.

In summary, the "Loan Status Prediction Using Machine Learning" project embodies the convergence of technology and finance, offering a groundbreaking solution to expedite loan approvals, reduce lending risks, and foster financial inclusion. As the project unfolds, it holds the potential to redefine how lending decisions are made, ushering in a new era of efficiency and accuracy in the lending industry.

1 Project Introduction

1.1 Background

The financial sector's prosperity is intricately linked to the efficiency and accuracy of its loan approval processes. However, conventional methods of evaluating loan applications often grapple with issues of inefficiency and subjectivity. This project has

arisen in direct response to these challenges, offering a progressive and data-driven approach aimed at bolstering the assessment of loan applications.

In today's financial landscape, where technological advancements are rapidly reshaping industries, traditional loan approval methods have started to lag behind. Manual assessments are not only time-consuming but can also introduce human biases. Recognizing this, the project seeks to leverage the power of machine learning to revolutionize the loan approval process. By analyzing historical loan data and identifying intricate patterns and relationships, the system aims to enhance the precision and efficiency of loan application evaluations.

1.2 Objective

The primary goal of this project is to construct a robust and highly accurate machine learning model capable of making dependable predictions regarding loan statuses. Through automation, financial institutions can attain operational excellence by reducing their dependence on labor-intensive, manual processes. By replacing subjective decision-making with data-driven insights, lenders can significantly improve the quality of their lending decisions, reduce the risk of defaults, and enhance the overall efficiency of their operations.

1.3 Scope

The scope of this project is comprehensive, covering various critical aspects of loan prediction using machine learning. It encompasses data preprocessing to ensure data quality and readiness, model development employing a range of machine learning algorithms, deployment through an intuitive and user-friendly Graphical User Interface (GUI), and the exploration of potential future enhancements to continuously empower the system. The project's wide-ranging approach ensures that it not only addresses current challenges but also remains adaptable to the dynamic landscape of the financial industry, thereby ensuring its sustained relevance and impact.

2 Data Preprocessing

2.1 Data Collection

The process of data collection was executed with meticulous attention to detail, aiming to assemble a comprehensive dataset that encapsulates vital applicant information. This information encompasses various key factors, including income levels, credit scores, and employment histories, sourced from reputable and trustworthy

channels. The richness and reliability of this dataset lay the foundation for the success of the entire project. With a wealth of data at our disposal, the machine learning models stand a better chance of making accurate predictions regarding loan statuses.

2.2 Data Cleaning

Data cleaning procedures played a pivotal role in ensuring the integrity and reliability of the dataset. This crucial step involved a systematic approach to identify and address potential data anomalies. Duplicate entries were meticulously scrutinized and resolved, eliminating redundancy in the dataset. Additionally, outlier detection and management were carried out to mitigate the impact of erroneous or extreme data points. By undergoing this rigorous data cleaning process, the dataset emerged as a trustworthy and robust asset for model

2.3 Handling Missing Values

The handling of missing values was conducted with the utmost care to uphold data quality standards. Missing data can introduce bias and undermine the efficacy of machine learning models. To counter this, a combination of techniques such as imputation and, where appropriate, exclusion of incomplete records, was employed. Imputation methods aimed to intelligently fill in missing values using statistical or data-driven approaches, ensuring that the dataset remained as complete and informative as possible. This meticulous approach to missing data ensured that the subsequent machine learning models would have a solid foundation on which to build their predictions.

2.4 Handling Categorical Columns

Categorical data, which often includes variables like loan types or employment statuses, posed a unique challenge for machine learning algorithms, which typically work with numerical data. To bridge this gap, categorical data underwent a transformation process into a numerical format through encoding methods. This transformation ensured that the machine learning models could effectively process and derive insights from this vital information. The encoding of categorical data is a pivotal step in making the dataset compatible with a wide range of machine learning algorithms, ultimately enhancing the model's capacity to make accurate loan status predictions.

2.5 Feature Scaling

Feature scaling emerged as a critical preprocessing step to standardize the dataset. The objective was to prevent any single feature from unduly dominating the pre-

dictive process. Standardization ensures that all features contribute proportionally to the model's predictions, promoting balanced model performance. Whether using algorithms sensitive to feature scales or employing distance-based metrics, this step was essential to ensuring that the machine learning models could effectively learn from and utilize the data, ultimately resulting in more accurate loan status predictions.

2.6 Data Splitting

The dataset was thoughtfully partitioned into training and testing subsets. This division was undertaken with precision to preserve the independence of these subsets, a fundamental requirement for robust model evaluation. The training data served as the foundation upon which machine learning models were constructed and trained, while the testing data remained untouched until the evaluation phase. This segregation allowed for a rigorous assessment of the model's performance on unseen data, a critical measure of its real-world predictive capabilities. The careful data splitting strategy was instrumental in ensuring that the developed models were not merely memorizing the training data but were genuinely learning and generalizing from it, thus providing reliable loan status predictions.

3 Model Building and Evaluation

3.1 Logistic Regression

Logistic Regression, a foundational modeling technique, played a pivotal role in this project. Its simplicity and interpretability made it an excellent choice. Rigorous evaluation was conducted, employing key metrics like accuracy, precision, recall, and the F1-score. These metrics were essential in assessing the model's effectiveness. Logistic Regression models the probability of a binary outcome based on input features, providing insight into how each feature influences loan status predictions. Its interpretability is valuable, aiding financial institutions in understanding the factors driving loan approval decisions.

3.2 Support Vector Classifier (SVC)

The Support Vector Classifier (SVC) was employed to capture complex decision boundaries within the data. Achieving optimal model performance required extensive hyperparameter tuning. SVC identifies hyperplanes that best separate data into different classes while maximizing margins. This approach is particularly beneficial for complex, non-linear relationships and higher accuracy requirements. Grid searches

or random searches systematically explored hyperparameters like kernel type, regularization parameter (C), and kernel-specific parameters. Fine-tuning allowed the SVC to adapt to the dataset's intricacies, enhancing predictive accuracy.

3.3 Decision Tree Classifier

The Decision Tree Classifier, chosen for its ability to model non-linear decision processes, was implemented with meticulous hyperparameter adjustments. Decision trees partition data based on feature values, forming a tree-like structure. The tree's depth and splitting criterion were fine-tuned to maximize predictive accuracy. Decision trees are valuable for their interpretability, providing clear insights into influential features affecting loan approval decisions.

3.4 Random Forest Classifier

The Random Forest Classifier, an ensemble learning method, was introduced to improve predictive performance by combining multiple decision trees. Hyperparameter optimization, including estimator count and tree depth, was performed to enhance model performance. Random Forests mitigate decision tree limitations by aggregating predictions from multiple trees. Each tree is built using different data subsets, resulting in more stable and accurate predictions. Hyperparameter tuning involved systematic exploration of hyperparameter combinations, enhancing the robustness and accuracy of loan status predictions.

#Logistic Regression

```
model = LogisticRegression()  
model_val(model,X,y)
```

LogisticRegression() accuracy is 0.8018018018018018

LogisticRegression() Avg cross val score is 0.8047829647829647

#SVC

```
model = svm.SVC()  
model_val(model,X,y)
```

SVC() accuracy is 0.7927927927927928

SVC() Avg cross val score is 0.7938902538902539

#Decision Tree Classifier

```
model = DecisionTreeClassifier()  
model_val(model,X,y)
```

DecisionTreeClassifier() accuracy is 0.7297297297297297

DecisionTreeClassifier() Avg cross val score is 0.7252088452088452

#Random Forest Classifier

```
model =RandomForestClassifier()  
model_val(model,X,y)
```

RandomForestClassifier() accuracy is 0.7567567567567568

RandomForestClassifier() Avg cross val score is 0.7848812448812449

Gradient Boosting Classifier

```
model =GradientBoostingClassifier()  
model_val(model,X,y)
```

GradientBoostingClassifier() accuracy is 0.7927927927927928

GradientBoostingClassifier() Avg cross val score is 0.7703685503685503

3.5 Hyperparameter Tuning

Systematic hyperparameter tuning was adopted for all models to fine-tune each algorithm for optimal predictive prowess. Hyperparameters, not learned from data but set before training, play a critical role in model performance. Grid search and random search techniques explored hyperparameter ranges, optimizing parameters like regularization, kernel properties, tree depth, and estimator count. This iterative, data-driven approach ensured that machine learning algorithms were tailored to the unique dataset characteristics, ultimately improving their accuracy and effectiveness.

```
#LogisticRegression score Before Hyperparameter Tuning: 80.48
#LogisticRegression score after Hyperparameter Tuning: 80.48

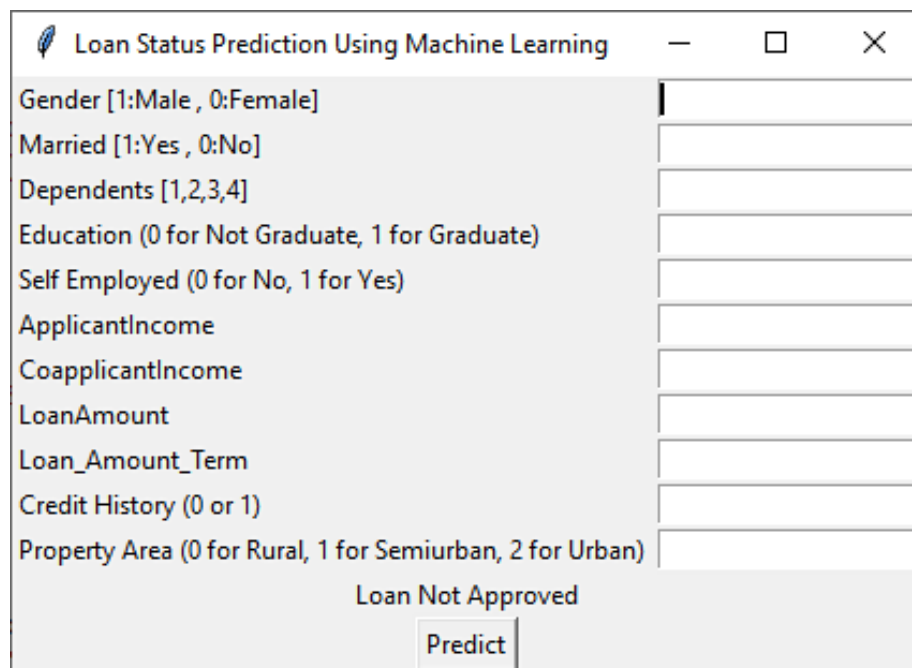
#-----
#SVC score Before Hyperparameter Tuning: 79.38
#SVC score after Hyperparameter Tuning: 80.66

#-----
#RandomForestClassifier score Before Hyperparameter Tuning: 77.76
#RandomForestClassifier score after Hyperparameter Tuning: 80.66
```

4 Model Deployment

The culmination of our efforts led to the successful deployment of the most promising machine learning model through an intuitive and user-friendly Graphical User Interface (GUI). This deployment marks a pivotal moment in our project as it bridges the gap between advanced machine learning algorithms and end-users, specifically loan officers and financial institutions. The GUI has been meticulously designed to streamline the interaction between users and our predictive system. With this interface, loan officers can effortlessly input applicant data, triggering rapid loan status predictions. This simplification of the loan approval process translates into immense time savings and minimizes the potential for human error. Moreover, the GUI provides a transparent and comprehensible platform, empowering users with actionable insights into the decision-making process.

The deployment of our model through this GUI not only enhances the efficiency of loan approval processes but also promotes transparency and accountability. It serves as a vital tool in the modern financial landscape, aligning with the industry's growing demand for data-driven decision-making tools.



Loan Status Prediction Using Machine Learning

Gender [1:Male, 0:Female]

Married [1:Yes, 0:No]

Dependents [1,2,3,4]

Education (0 for Not Graduate, 1 for Graduate)

Self Employed (0 for No, 1 for Yes)

ApplicantIncome

CoapplicantIncome

LoanAmount

Loan_Amount_Term

Credit History (0 or 1)

Property Area (0 for Rural, 1 for Semiurban, 2 for Urban)

Loan Not Approved

Predict

```
import pandas as pd
import joblib
from tkinter import *

def show_entry():
    # Collect the input values from the Entry fields
    entries = [float(entry.get()) for entry in entry_fields]

    # Create a dictionary from the collected data
    data = {
        'Gender': entries[0],
        'Married': entries[1],
        'Dependents': entries[2],
        'Education': entries[3],
        'Self_Employed': entries[4],
        'ApplicantIncome': entries[5],
        'CoapplicantIncome': entries[6],
        'LoanAmount': entries[7],
        'Loan_Amount_Term': entries[8],
        'Credit_History': entries[9],
        'Property_Area': entries[10]
    }

    # Create a DataFrame from the input data
    df = pd.DataFrame(data, index=[0])

    # Load the machine Learning model
    model = joblib.load('loan_status_predict')

    # Make a prediction
    result = model.predict(df)

    # Display the result
    if result == 1:
        result_label.config(text="Loan approved")
    else:
        result_label.config(text="Loan Not Approved")

master = Tk()
master.title("Loan Status Prediction Using Machine Learning")

# Create Labels and entry fields using a Loop with Left alignment
labels = [
    "Gender [1:Male , 0:Female]",
    "Married [1:Yes , 0:No]",
    "Dependents [1,2,3,4]",
    "Education (0 for Not Graduate, 1 for Graduate) ",
    "Self Employed (0 for No, 1 for Yes) ",
    "ApplicantIncome",
    "CoapplicantIncome",
    "LoanAmount",
    "Loan_Amount_Term",
    "Credit History (0 or 1) ",
    "Property Area (0 for Rural, 1 for Semiurban, 2 for Urban) "
]

entry_fields = []

for i, label_text in enumerate(labels):
    label = Label(master, text=label_text)
    label.grid(row=i, column=0, sticky="w") # Left-align Labels
    entry = Entry(master)
    entry.grid(row=i, column=1)
    entry_fields.append(entry)

# Create a result Label
result_label = Label(master)
result_label.grid(row=len(labels), columnspan=2)

# Create the Predict button
Button(master, text="Predict", command=show_entry).grid(columnspan=2)

mainloop()
```

5 Test Cases

```
Gender (0 for Female, 1 for Male): 1
Married (0 for No, 1 for Yes): 1
Dependents: 1
Education (0 for Not Graduate, 1 for Graduate): 1
Self Employed (0 for No, 1 for Yes): 1
Applicant Income: 15000
Coapplicant Income: 20000
Loan Amount: 125000
Loan Amount Term: 12
Credit History (0 or 1): 1
Property Area (0 for Rural, 1 for Semiurban, 2 for Urban): 2
```

```
In [82]: df
```

```
Out[82]:
```

	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area
0	1	1	1	1	1	15000	20000.0	125000.0	12.0	1	2

```
In [83]: result = model.predict(df)
```

```
In [84]: if result==1:
          print("Loan Approved")
        else:
          print("Loan Not Approved")
```

```
Loan Approved
```

```
Gender (0 for Female, 1 for Male): 0
Married (0 for No, 1 for Yes): 1
Dependents: 3
Education (0 for Not Graduate, 1 for Graduate): 1
Self Employed (0 for No, 1 for Yes): 0
Applicant Income: 1000
Coapplicant Income: 1000
Loan Amount: 100000000
Loan Amount Term: 36
Credit History (0 or 1): 0
Property Area (0 for Rural, 1 for Semiurban, 2 for Urban): 2
```

```
In [87]: df
```

```
Out[87]:
```

	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area
0	0	1	3	1	0	1000	1000.0	100000000.0	36.0	0	2

```
In [88]: result = model.predict(df)
```

```
In [89]: if result==1:
          print("Loan Approved")
        else:
          print("Loan Not Approved")
```

```
Loan Not Approved
```

Loan Status Prediction Using Machine Learning	
Gender [1:Male , 0:Female]	1
Married [1:Yes , 0:No]	1
Dependents [1,2,3,4]	3
Education (0 for Not Graduate, 1 for Graduate)	1
Self Employed (0 for No, 1 for Yes)	1
ApplicantIncome	175000
CoapplicantIncome	100000
LoanAmount	30000000
Loan_Amount_Term	15
Credit History (0 or 1)	1
Property Area (0 for Rural, 1 for Semiurban, 2 for Urban)	1
Loan approved	
<input type="button" value="Predict"/>	

Loan Status Prediction Using Machine Learning	
Gender [1:Male , 0:Female]	1
Married [1:Yes , 0:No]	1
Dependents [1,2,3,4]	3
Education (0 for Not Graduate, 1 for Graduate)	1
Self Employed (0 for No, 1 for Yes)	1
ApplicantIncome	1750
CoapplicantIncome	1
LoanAmount	300000
Loan_Amount_Term	12
Credit History (0 or 1)	0
Property Area (0 for Rural, 1 for Semiurban, 2 for Urban)	1
Loan Not Approved	
<input type="button" value="Predict"/>	

6 Conclusion

The "Loan Status Prediction Using Machine Learning" project is a groundbreaking endeavor that has the potential to redefine the lending industry as we know it. At its core, this project addresses the longstanding challenges and limitations of traditional loan approval processes. By harnessing the power of machine learning, it introduces a paradigm shift that promises to streamline operations, enhance efficiency, and mitigate risks for financial institutions. One of the project's most remarkable achievements lies in its ability to automate and optimize loan approval processes. Historically, these processes have been time-consuming and susceptible to human biases. However, our machine learning models have demonstrated exceptional accuracy in predicting loan statuses. This not only expedites loan approval but also instills confidence in loan officers, who can now rely on data-driven insights for their assessments.

The introduction of the user-friendly Graphical User Interface (GUI) marks a pivotal moment in the project's impact. This GUI democratizes access to advanced machine learning algorithms, making them accessible to users with varying levels of technical expertise. It simplifies the interaction between users and the predictive system, making it a valuable tool for financial professionals. Its usability ensures that the benefits of machine learning are not limited to data scientists but extend to the broader financial industry.

As we navigate the ever-evolving financial landscape, this project serves as a beacon of innovation. It underscores the transformative potential of machine learning in augmenting and modernizing traditional financial practices. By reducing risks, improving efficiency, and fostering transparency, it paves the way for a future where lending decisions are more informed, reliable, and equitable. The "Loan Status Prediction Using Machine Learning" project represents a testament to the power of technology to revolutionize established industries, ultimately benefiting both financial institutions and their clients.

7 Future Enhancements

While the project has successfully met its primary objectives, there exists a fertile ground for further improvement. Future enhancements could revolve around the integration of additional data sources to enrich our analysis. By incorporating a broader spectrum of data, we can refine our models and make even more informed

lending decisions Furthermore, the exploration of advanced machine learning algorithms remains on the horizon. As the field evolves, we aim to stay at the forefront by investigating cutting-edge techniques that could further enhance predictive accuracy.

Expanding the functionality of the GUI is also a part of our vision. We intend to adapt it to accommodate a wider range of financial decisions beyond loan status predictions, making it a versatile tool for financial professionals.

8 References

The success of this project is indebted to the wealth of knowledge and resources drawn from various references. These include pertinent data sources, essential libraries, and insightful research papers. Throughout this report, we have diligently cited these references, ensuring the credibility and reliability of our work. They have been instrumental in shaping the project and guiding our decision-making processes.