University of Sunderland
Name: ROTIMI, Ayodeji Moses
Module Code: CETM46
CETM 46-DATA SCIENCE PRODUCT DEVELOPMENT
STUDENT ID : 229783515

Loan Default Web Application

Table of Contents

INTRODUCTION

The popularity of online personal loans has surged, with various platforms such as bank portals and Ant Credit Pays offering convenient access. Many of these platforms boast straightforward lending criteria, facilitating quick borrowing for users. (Zhu et al, 2023). Different artificial intelligence algorithms have been utilized for loan prediction purposes (Li et al., 2021).

This report aims at providing a comprehensive overview of the development process for a data science product: a web application designed to predict loan statuses. The core of this product is a machine learning model trained on relevant loan data. Leveraging this model, a Streamlit web application was built to provide users with the ability to input their information and receive a prediction regarding their loan approval status. The report will explore into the product's design, development, and project management aspects, highlighting critical decisions and methodologies employed throughout the process.

PRODUCT DESIGN SECTION

Data Source and Theme Selection and Specification

- Data Source

The effectiveness of machine learning models hinges on the quality of the data they are trained with. Greater volumes of processed datasets are crucial for AI projects, as they enable more effective model training and yield superior outcomes. (Robert Koch). The data used for the design of this product was sourced from Kaggle, an open-source platform for data science datasets. The dataset contained data of 28000 observations made across different features which loan status depends on. The table below describes the features available in the dataset as provided by the author. The dataset was gotten from https://www.kaggle.com/datasets/subhamjain/loan-prediction-based-on-customer-behavior

Table1: Showing the data description.

Column	Description
income	The income of the loan applicant
age	The age of the applicant
experience	Professional experience in years
Profession	User's Profession
Married	The Marital Status
House_ownership	Owned or rented or neither
Car_ownership	Owned a car or not
Current_job_years	Number of years in the current residence
City	The city of residence
State	State of residence
Risk_flag	Defaulted on loan or not

```
In [3]: #check the first five rows
        df.head()
```

Out[3]:

	Id	Income	Age	Experience	Married/Single	House_Ownership	Car_Ownership	Profession	CITY	STATE	CURRENT_JOB_YRS
0	1	1303834	23	3	single	rented	no	Mechanical_engineer	Rewa	Madhya_Pradesh	3
1	2	7574516	40	10	single	rented	no	Software_Developer	Parbhani	Maharashtra	9
2	3	3991815	66	4	married	rented	no	Technical_writer	Alappuzha	Kerala	4
3	4	6256451	41	2	single	rented	yes	Software_Developer	Bhubaneswar	Odisha	2
4	5	5768871	47	11	single	rented	no	Civil_servant	Tiruchirappalli[10]	Tamil_Nadu	3

Fig 1: First five rows from the data.

```
In [4]: df.tail()
```

Out[4]:

	Id	Income	Age	Experience	Married/Single	House_Ownership	Car_Ownership	Profession	CITY	STATE	CURRENT_JOB_YI
251995	251996	8154883	43	13	single	rented	no	Surgeon	Kolkata	West_Bengal	
251996	251997	2843572	26	10	single	rented	no	Army_officer	Rewa	Madhya_Pradesh	
251997	251998	4522448	46	7	single	rented	no	Design_Engineer	Kalyan-Dombivli	Maharashtra	
251998	251999	6507128	45	0	single	rented	no	Graphic_Designer	Pondicherry	Puducherry	
251999	252000	9070230	70	17	single	rented	no	Statistician	Avadi	Tamil_Nadu	

Fig 2: Screenshot of last five rows preview from the data.

- ## Theme Selection

The theme selection process involved the determination and setting out the overarching goal and target of the loan prediction web application. The nature of this project enhanced the theme to revolve around the development of a predictive model that evaluates and examines how likely a loan is to be approved given the applicants attributes. The chosen theme aims to tackle a prevalent challenge in the loan application journey: the uncertainty surrounding loan approval. Through the utilization of machine learning methods, the application endeavors to furnish users with insightful information regarding their loan approval status, enabling them to make educated decisions and enhance their financial circumstances.

- ## Specification

The specification was in two phases which included the training of model specification and the application specification. The first phase involved the process of preparing the data to satisfy all data specification and training the predictive model.

The Application features were defined based on user requirements and market research for loan defaults. Functional requirements, input forms, prediction output display, and result interpretation, were specified to ensure a better user experience. By defining

clear objectives and requirements, the project was effectively executed and deliver a high-quality product that meets user needs and expectations.

Application Domain/End User's Requirements Analysis.

The global loan market is segmented based on several factors such as type, provider type, interest rate, tenure period, region, and competitive landscape. Types of loans include Housing, Mortgage, Personal, Auto, Business, Home Improvement, and Others, which encompass Gold, Education, Agriculture, and Retail loans, among others. (KapooR, 2023).

Key features, such as a user-friendly interface, intuitive input forms, and easily interpretable prediction results, were identified through this process. An understanding of the loan approval domain assisted in defining the application's scope and prioritizing features according to user preferences.

Product Functional and Non-functional Requirements Specifications.

Requirements analysis involves gathering the requirements that specify the intended functionality and non-functionality of the software. (Dave, 2022). Based on the requirements analysis, both functional and non-functional requirements were specified for the application. Functional requirements outlined the features and capabilities of the web application, such as data input forms, and prediction output display. Non-functional requirements addressed aspects like performance, scalability, and speed, ensuring a flawless user experience.

Product Software Architecture Design

Selecting the appropriate web application architecture is a crucial decision that establishes the groundwork for the entire development process of the web application. (Andersson,2023). The web application architecture was designed to support the seamless integration of the machine learning model with the Streamlit web application. A layered architecture approach was adopted, with separate components for presentation, model inference, and data processing. This architecture facilitated modularity and extensibility, allowing for easy maintenance and future enhancements.

Product Use Cases Specifications

Use cases were established to showcase the interactions between users and the product. These encompassed scenarios such as the entry of applicant information, initiating model inference, and presenting prediction results. Specifications for use cases informed the development of user workflows and interface design enhancing great user experience.

PRODUCT DEVELOPMENT SECTION

Selection of Appropriate Software Tools

Python was chosen as the primary programming language for its extensive libraries and frameworks suitable for data science tasks. The selection of appropriate software tools, platforms, and hardware methodologies for the project, centered on developing a Streamlit application for loan prediction, required meticulous evaluation to ensure compatibility, efficiency, and scalability.

Streamlit prioritizes speed and interactivity. It serves as a web application framework designed to facilitate the creation and development of Python-based web applications. It includes built-in functionalities for tasks like capturing user inputs and displaying interactives outputs (Richards, 2023). The Streamlit framework was selected for building the web application due to its simplicity and ease of use. Hosting was done on the streamlit cloud platform which, provide utmost scalability and reliability for the web app.

Product Development Software Engineering Methodology.

The product development software engineering methodology for this project involved a sequential approach, starting with the preparation of the acquired loan dataset and subsequently transitioning to the development of the Streamlit application.

- ### Data Preprocessing

  The first step in the product development was the preparation of the acquired dataset. By meticulously addressing data preprocessing tasks including carrying out Exploratory Data Analysis, data cleaning, data transformation and feature engineering, the raw loan dataset is transformed into a clean, structured format suitable for model training. This sets the stage for subsequent steps in the development process.

- ## Model Training

  The next step of development focused on training the predictive model using machine learning algorithms. This involved the splitting of the preprocessed loan dataset into training and testing sets, selecting appropriate features, and applying machine learning techniques to build the predictive model. Various machine learning classification algorithms were used and evaluated which included the RandomForest, Gradient Boosting and the Decision Tree. Random Forest produces a collection of decision trees known as an ensemble. To ensure diversity among these individual decision trees, Breiman employed a randomization technique that synergizes effectively with bagging or random subspace methods. (Kulkarni & Sinba). Decision Trees are flexible Machine Learning algorithms capable of handling various tasks, including classification, regression, and multioutput tasks. They are powerful tools capable of effectively modeling complex datasets. (Geron, 2019). Gradient Boosting operates by incrementally adding predictors to an ensemble, with each subsequent predictor correcting the errors of its predecessor. Unlike AdaBoost, which adjusts instance weights at each iteration, this method focuses on fitting the new predictor to the residual errors generated by the previous predictor. (Geron, 2019). These classification algorithms were fitted with the preprocessed training data.

- ## Model Evaluation

  After training, the models were all evaluated to assess its performance and accuracy. Performance metrics such as accuracy, precision, recall, AUC score, and the confusion matrix which compared the predicted labels, and the true labels were calculated to measure the model's effectiveness in predicting loan statuses.

  Model evaluation involves both quantitative and qualitative assessment of machine learning models' performance. This process aids in gauging how effectively a model can generalize to unfamiliar data. (Singh, 2023). Results from the evaluation metrics can be displayed below.

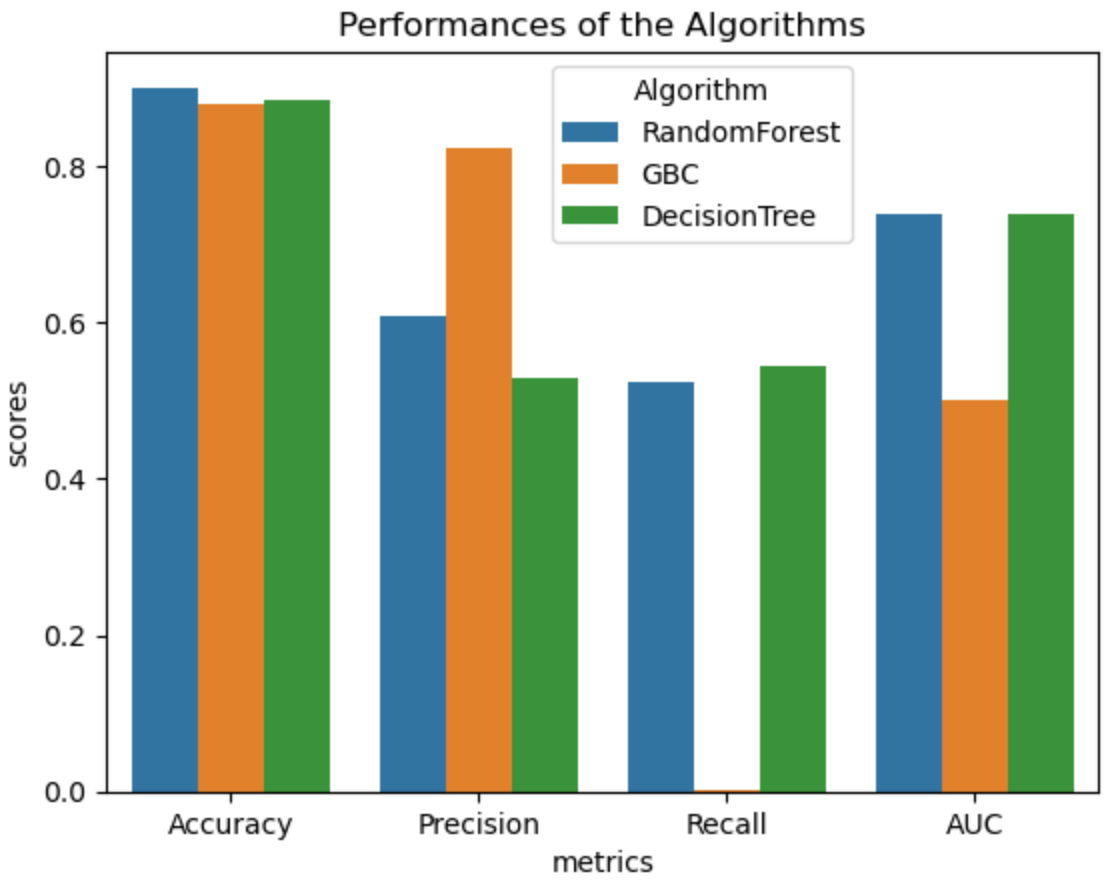


Fig 3: Performances of the Algorithms.

Based on the above displayed figure, it is evident that all models demonstrated strong accuracy performance, with Random Forest emerging as the top performer. Gradient Boosting Classifier (GBC) excelled in precision, although it exhibited a lower recall score compared to the other models. Additionally, its AUC score was also lower than the rest. The other two models showed similar AUC scores, indicating a tie in that aspect.

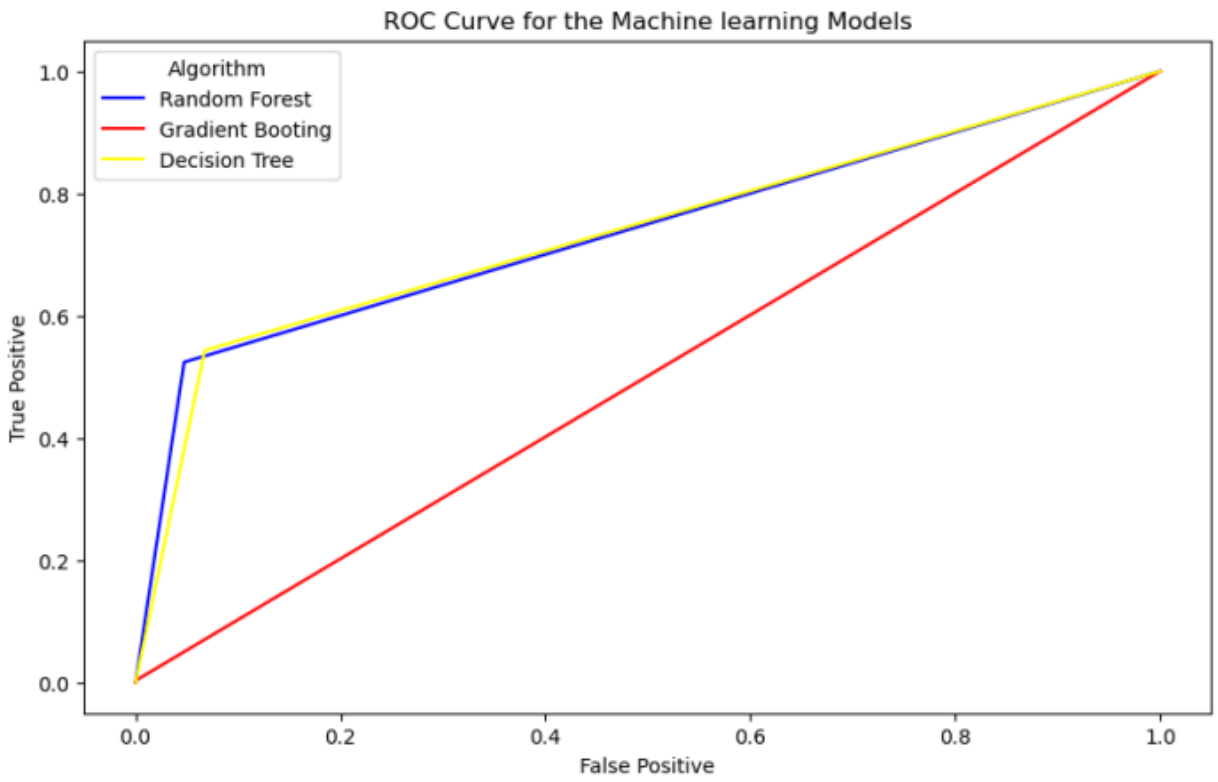


Fig 4: ROC Curve for the classifiers.

Random Forest exhibits a superior ROC curve compared to the Algorithm, implying better classification accuracy overall. Gradient Boosting achieves the highest ROC curve among the models, signifying its superior performance with the highest true positive rate for a given false positive rate. Decision Tree's ROC curve falls between Random Forest and Gradient Boosting, indicating performance superior to the Algorithm but not as strong as Gradient Boosting. The Confusion Matrix for the models can be displayed below:

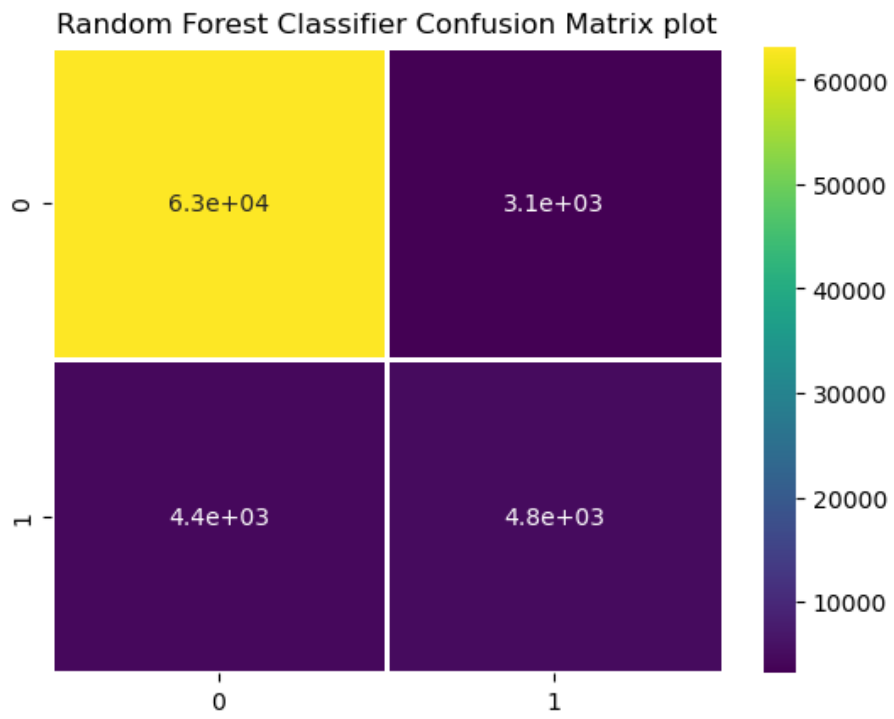


Fig 5: RandomForest Classifier Confusion Matrix

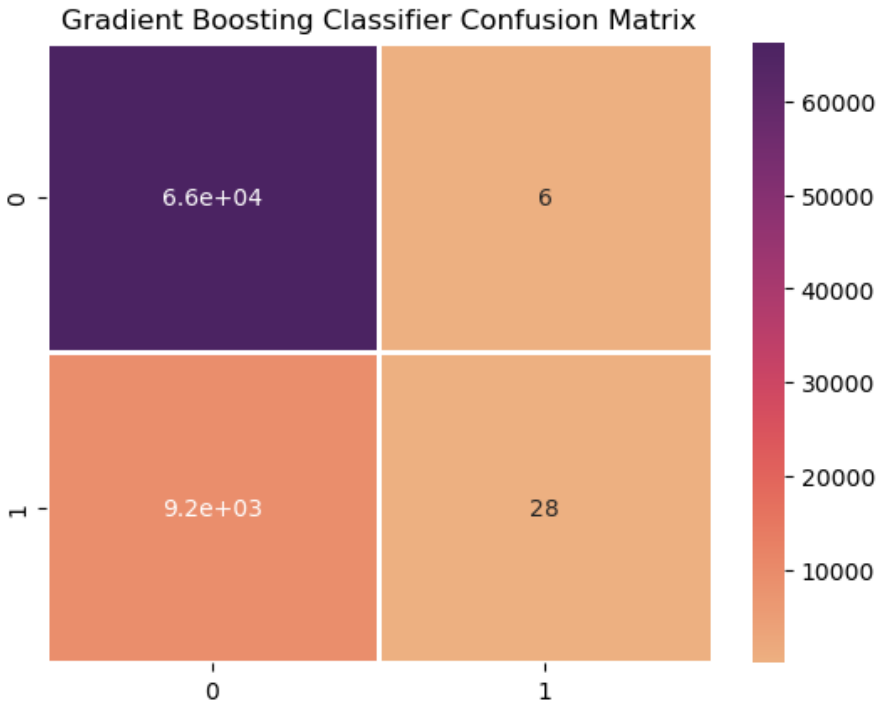


Fig 6: Gradient Boosting Classifier Confusion Matrix

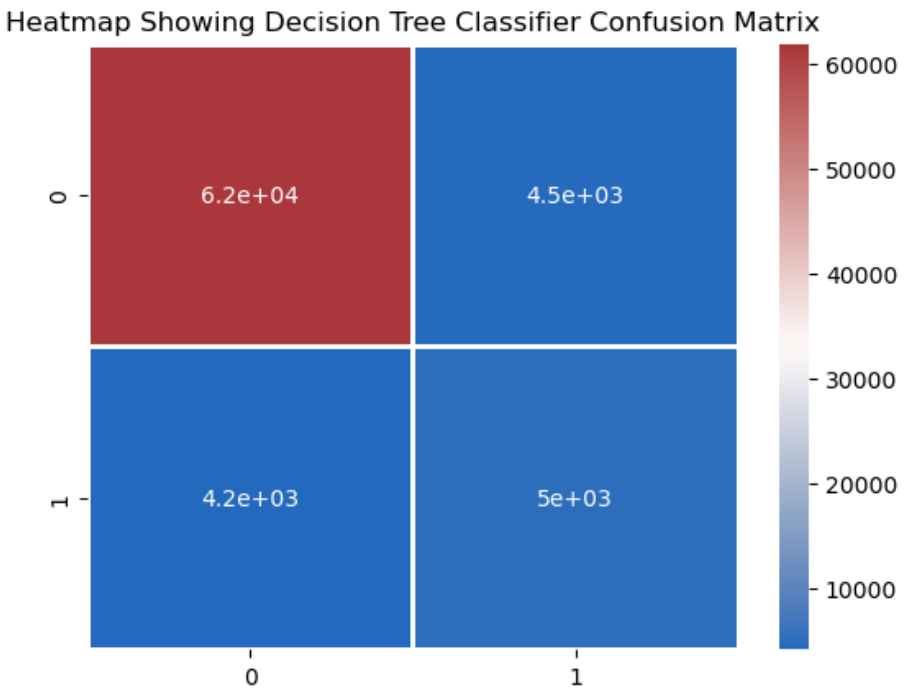


Fig 7: Decision Tree Classifier Confusion Matrix

- Streamlit Application Development

  With a trained and validated predictive model in place, the focus shifted to developing the Streamlit application. The development process involved designing the user interface, integrating the predictive model, implementing data input forms, and creating output displays for prediction results. Streamlit's intuitive API and rapid development capabilities facilitated the creation of an interactive and user-friendly application.

- Integration and Testing

  As development progressed, the predictive model was integrated into the Streamlit application, ensuring seamless interaction between the frontend interface and the backend model. Comprehensive testing was conducted to verify the functionality, performance, and reliability of the application.

- Application Deployment

  Once the development phase was complete and the application met the desired quality standards, it was deployed to streamlit cloud platform which provide us with a great production environment.

System Testing Method

Testing constitutes a crucial aspect of every software development cycle, demanding significant time and resources from companies. (Lakshmi & Mallika, 2017). Comprehensive testing was conducted to ensure the reliability and accuracy of the web application. End-to-end testing was performed to validate the entire workflow, from user input to prediction output. Automated testing tools were utilized to identify and address issues promptly.

User Evaluation Plan and Methods

The plan was to conduct a beta testing with a sample target applicant. To invite users to interact with the application by inputting their information to view their predicted loan application status. Feedback will be collected from each user after the interaction with the web app.

PROJECT MANAGEMENT SECTION

Timeline Gannt Chart

The product development stages timeline gannt chart can be found in Appendix 1.

Risk Assessment on Personal Information Protection and Data

Security/Governance

In the information era, the issue of protecting personal information has garnered unparalleled focus (Wang et al, 2022). Risk assessment was conducted to identify potential threats to personal information protection and data security/governance. Regular security audits and penetration testing were performed to ensure the integrity and confidentiality of user data.

Quality Control on Software Development

Quality control procedures were implemented to enhance great standards across the development lifecycle which include code reviews, automated testing, and continuous integration methodologies. Routine quality assurance assessments were carried out to verify compliance with coding norms and industry best practices.

Basic Customer/User Relationship Management

Effective customer relationship management was essential for engaging with users and gathering feedback. This includes the establishment of regular communication channels to keep the users informed and involved in the development process.

Basic Product Marketing Strategy

This entails the devising of a basic product marketing strategy to promote the loan prediction web application. This includes the creation of a website for the application, running advertisements on social media channels. The advent of the internet revolutionized the business landscape by introducing numerous digital marketing strategies. (Olson et al, 2021). The goal was to increase awareness and drive user adoption of the application.

CONCLUSION

The development of the loan prediction web application involved a structured approach encompassing product design, development, and project management. By critically analyzing each aspect of the process, from data source selection to marketing strategy, the project aimed to deliver an effective solution. Furthermore, continuous iteration and improvement will be key to ensuring the success and sustainability of the application in meeting the needs of its users.

REFERENCES

Richards, T. (2023). Streamlit for Data Science: Create Interactive Data Apps in Python. United Kingdom: Packt Publishing.

Xu Zhu, Qingyong Chu, Xinchang Song, Ping Hu, & Lu Peng. (2023). Explainable prediction of loan default based on machine learning models. Data Science and Management, 6(3), 123-133. https://doi.org/10.1016/j.dsm.2023.04.003.

Koch, R. (2023.). An Introduction to Machine Learning Datasets and Resources. Retrieved from https://www.clickworker.com/customer-blog/machine-learning-datasets/#:~:text=Machine%20learning%20models%20are%20only,and%20achieve%20the%20best%20results.

Wang, C., Guo, F., & Ji, M. (2022). Analysis of Legal Issues of Personal Information Protection in the Field of Big Data. Journal of environmental and public health, 2022, 1678360. https://doi.org/10.1155/2022/1678360 (Retraction published J Environ Public Health. 2023 Jun 28; 2023:9852305)

Li, M., Yan, C., & Liu, W. (2021, December). The network loan risk prediction model based on Convolutional neural network and Stacking fusion model. Applied Soft Computing, 113, 107961. https://scholar.google.com/scholar_lookup?title=The%20network%20loan%20risk%20prediction%20model%20based%20on%20Convolutional%20neural%20network%20and%20Stacking%20fusion%20model&publication_year=2021&author=M.%20Li&author=C.%20Yan&author=W.%20Liu

Kapoor, V. (2023, February 27). Loan Market Analysis, Share, Trends, Demand, Size, Opportunity & Forecast. TechSci Research. Business Specialist. https://www.linkedin.com/pulse/loan-market-analysis-share-trends-demand-size-opportunity-varun-kp

Dave, D. J. (2022). Identifying Functional and Non-functional Software Requirements from User App Reviews and Requirements Artifacts. Montclair State University. https://digitalcommons.montclair.edu/cgi/viewcontent.cgi?article=2014&context=etd

Andersson, M. (2023). The Importance of Web Application Architecture: Understanding Factors Influencing Decision-Making and Identifying Potential Improvements for Effective Web Development. Handelshögskolan Karlstad Business School. https://www.diva-portal.org/smash/get/diva2:1772594/ATTACHMENT01.pdf

Géron, A. (2019). Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems (5th ed.). O'Reilly.

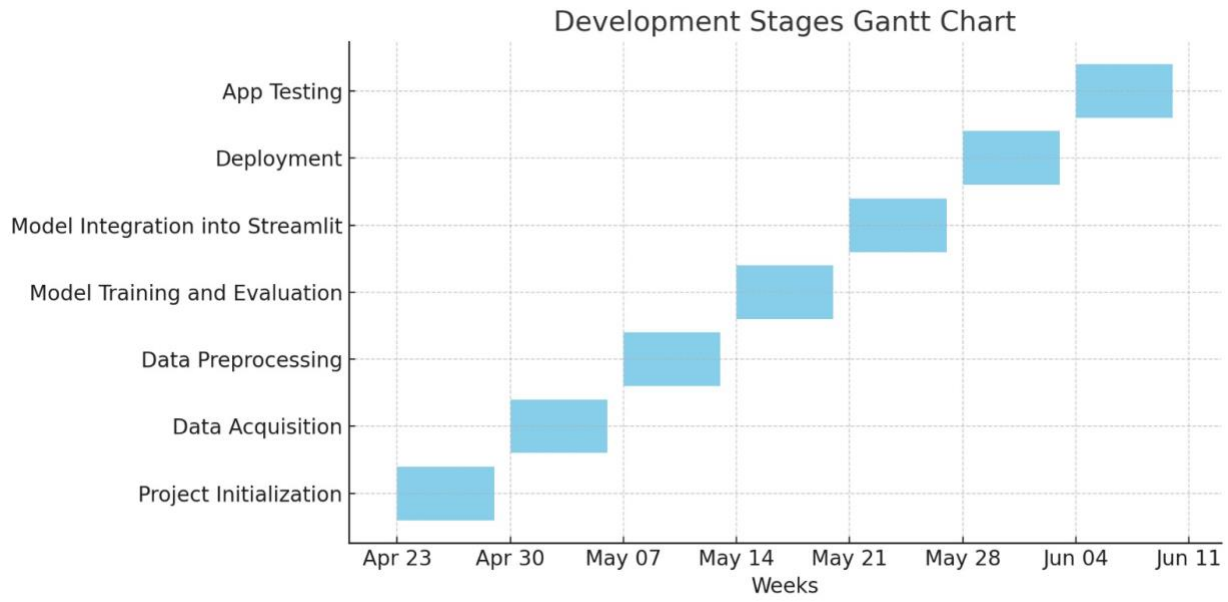
Kulkarni, V. Y., & Sinha, P. K. (2013). Random Forest Classifiers: A Survey and Future Research Directions. International Journal of Advanced Computing, 36(1), 1144. ISSN: 2051-0845. https://adiwijaya.staff.telkomuniversity.ac.id/files/2014/02/Random-Forest-Classifiers_A-Survey-and-Future.pdf

Singh, J. (2023, June 13). Model Evaluation in Machine Learning: A Comprehensive Guide. https://medium.com/@jasmeet.dotnettricks/model-evaluation-in-machine-learning-a-comprehensive-guide-e4f199fc2c9d

Lakshmi, D. & Mallika, S.. (2017). A Review on Web Application Testing and its Current Research Directions. International Journal of Electrical and Computer Engineering (IJECE). 7. 2132. 10.11591/ijece.v7i4.pp2132-2141.

Olson, E. M., Olson, K. M., Czaplewski, A. J., & Key, T. M. (2021). Business strategy and the management of digital marketing. Business Horizons, 64(2), 285-293. https://doi.org/10.1016/j.bushor.2020.12.004

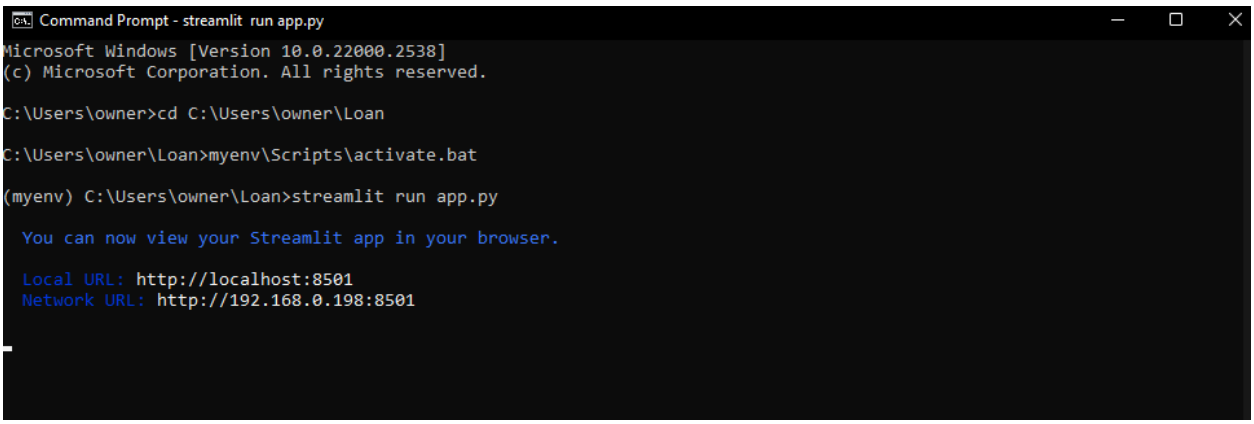
APPENDIX 1



Timeline Gantt Chart

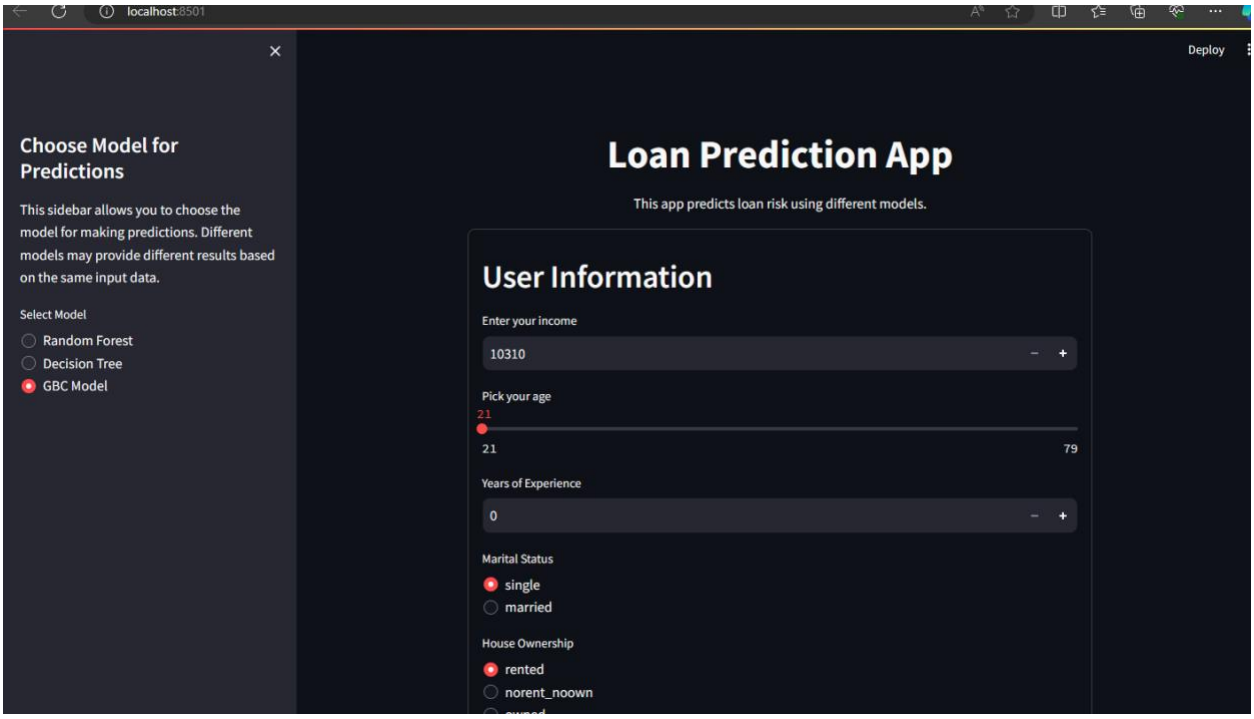
APPENDIX 2

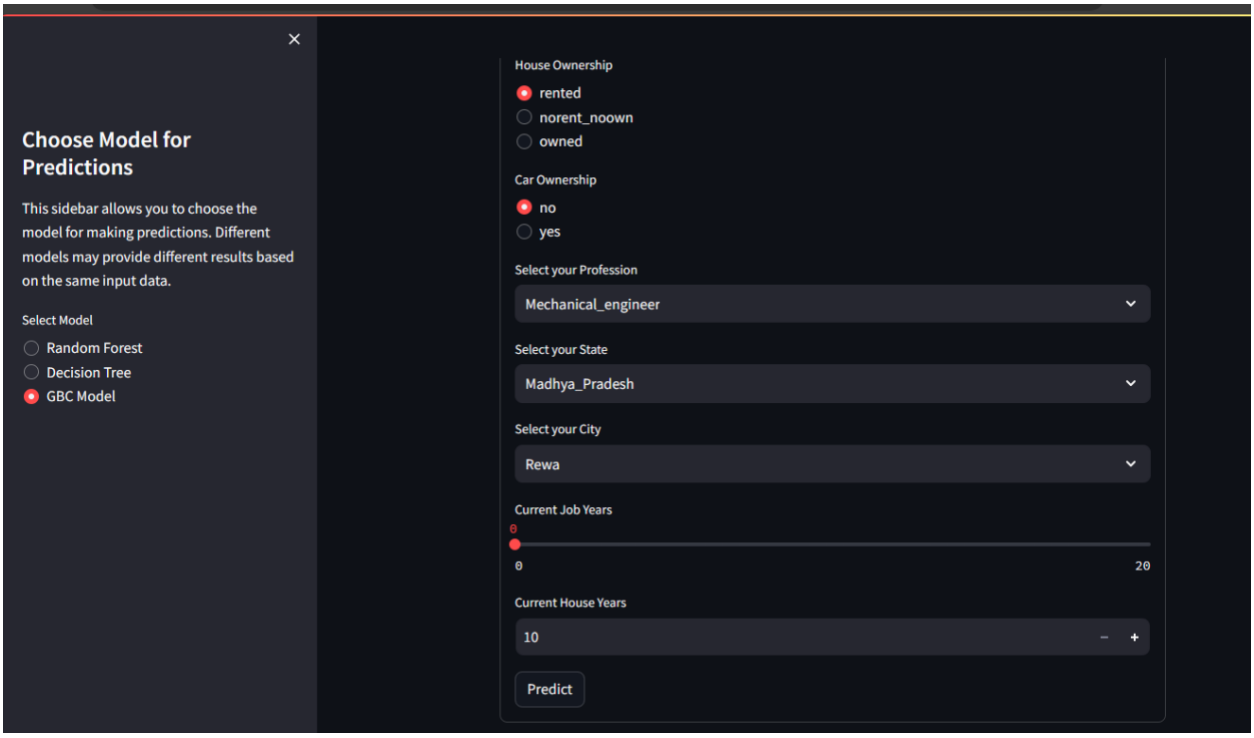
Running the application

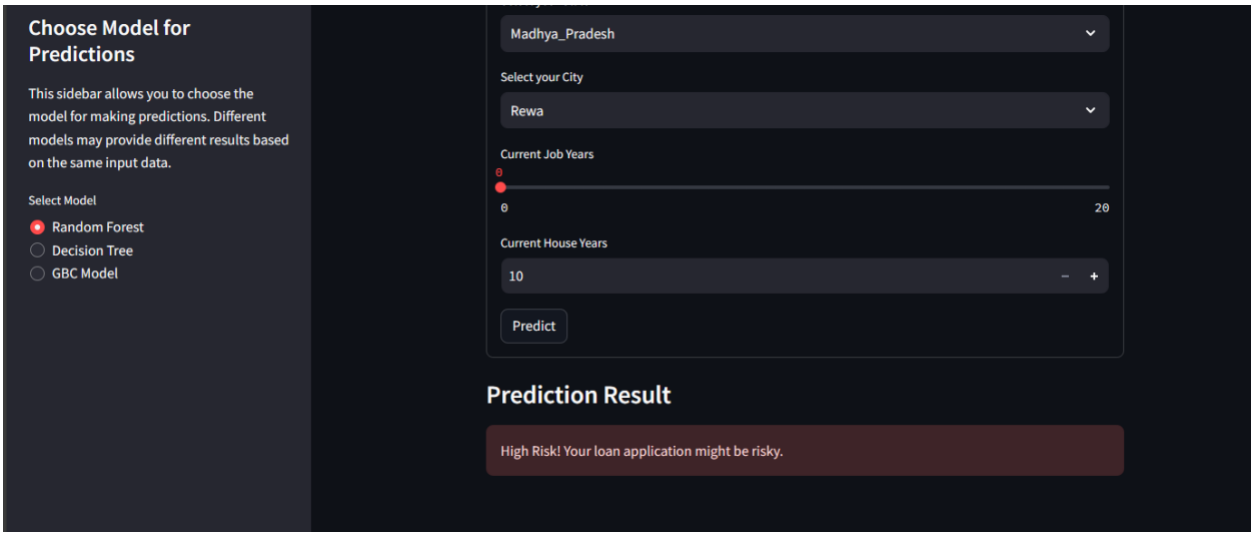


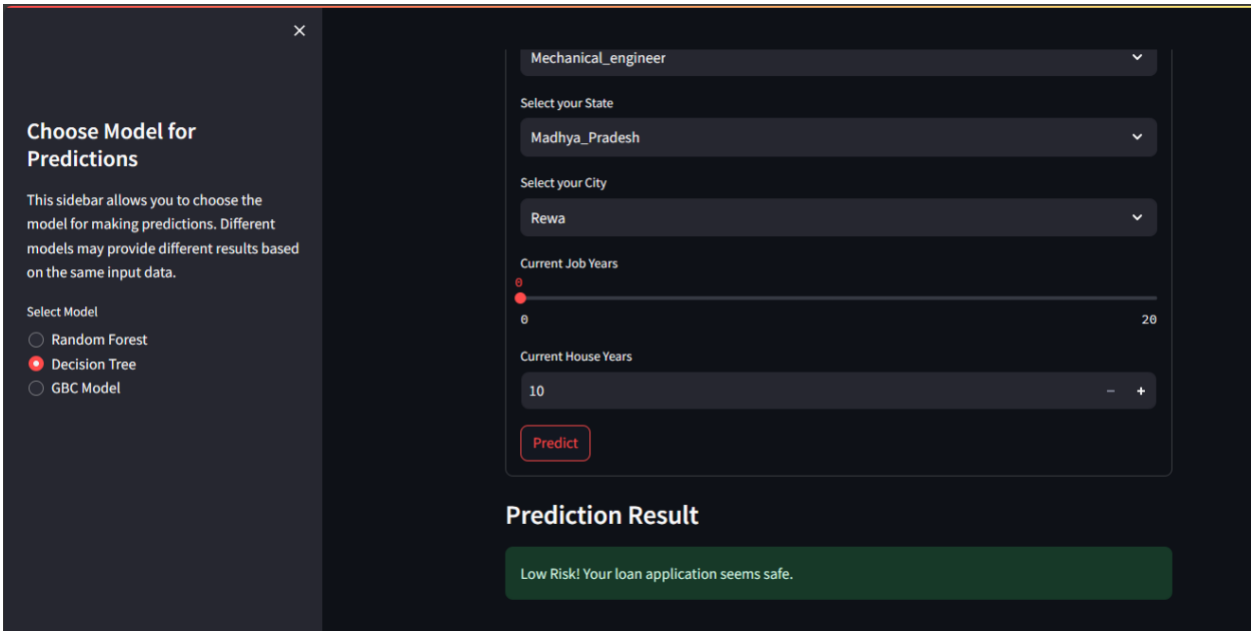
APPENDIX 3

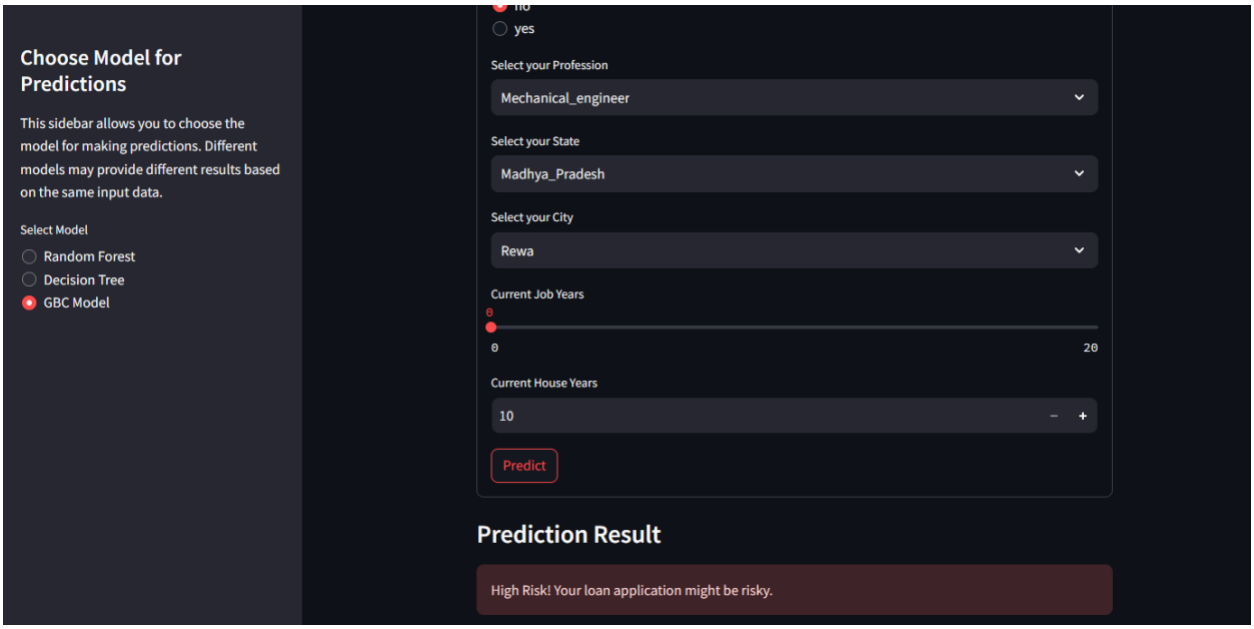
App Screenshots




Choose Model for Predictions
This sidebar allows you to choose the model for making predictions. Different models may provide different results based on the same input data.
Select Model
Random Forest
Decision Tree
GBC Model
House Ownership
rented
norent_noown
owned
Car Ownership
no
yes
Select your Profession
Mechanical_engineer
Select your State
Madhya_Pradesh
Select your City
Rewa
Current Job Years
0
0
20
Current House Years
10
Predict


Choose Model for Predictions
This sidebar allows you to choose the model for making predictions. Different models may provide different results based on the same input data.
Select Model
Random Forest
Decision Tree
GBC Model
Madhya_Pradesh
Select your City
Rewa
Current Job Years
0
0
20
Current House Years
10
Predict
Prediction Result
High Risk! Your loan application might be risky.


Choose Model for Predictions
This sidebar allows you to choose the model for making predictions. Different models may provide different results based on the same input data.
Select Model
Random Forest
Decision Tree
GBC Model
Mechanical_engineer
Select your State
Madhya_Pradesh
Select your City
Rewa
Current Job Years
0
0
20
Current House Years
10
Predict
Prediction Result
Low Risk! Your loan application seems safe.


Choose Model for Predictions
This sidebar allows you to choose the model for making predictions. Different models may provide different results based on the same input data.
Select Model
Random Forest
Decision Tree
GBC Model
yes
Select your Profession
Mechanical_engineer
Select your State
Madhya_Pradesh
Select your City
Rewa
Current Job Years
0
0
20
Current House Years
10
Predict
Prediction Result
High Risk! Your loan application might be risky.