CAPSTONE PROJECT

LOAN ELIGIBILITY PREDICTION USING MACHINE LEARNING

Presented By:
1. Vedant Roy-University School of Automation and Robotics-Artificial Intelligence and Machine Learning

OUTLINE

- Problem Statement
- Proposed System/Solution
- System Development Approach
- Data Visualization (EDA)
- Algorithm & Deployment
- Result
- Conclusion
- Future Scope
- References

PROBLEM STATEMENT

The process of loan approval in financial institutions involves significant risk assessment and decision-making. Manual evaluation of loan applications is time-consuming and prone to human error. The challenge is to develop an automated system that can predict the likelihood of loan approval based on historical data.

PROPOSED SOLUTION

- The proposed solution leverages machine learning techniques to predict the likelihood of loan approval. The solution will consist of:

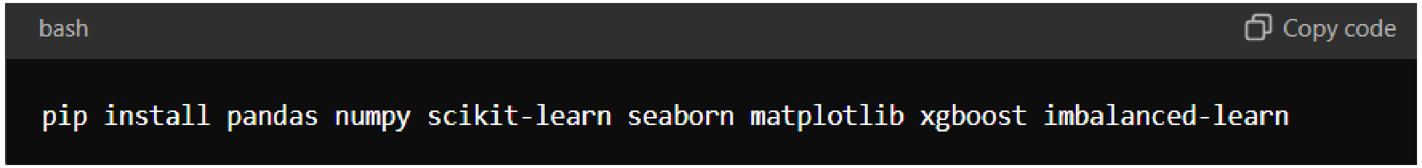
- **Data Collection:**
  - Historical loan application data including applicant details, loan amount, loan term, credit history, etc.
  - External data like economic indicators, regional data, etc. (if available).
- **Data Preprocessing:**
  - Cleaning the data to handle missing values and outliers.
  - Feature engineering to create relevant features that impact loan approval.
- **Machine Learning Algorithm:**
  - Implementation of algorithms such as Logistic Regression, Decision Trees, Random Forest, Gradient Boosting, and XGBoost.
  - Evaluation and comparison of these models.
- **Deployment:**
  - Development of an interface or application for loan approval prediction.
  - Deployment on a scalable platform.
- **Evaluation:**
  - Using metrics such as accuracy, precision, recall, F1-score, and AUC-ROC score to evaluate model performance.
  - Continuous monitoring and fine-tuning of the model.

SYSTEM DEVELOPMENT APPROACH

System Requirements

- Python environment set up
- Jupyter Notebook or any other Python IDE
- Required Python libraries (Pandas, NumPy, Scikit-learn, Seaborn, Matplotlib, XGBoost, Imbalanced-learn)

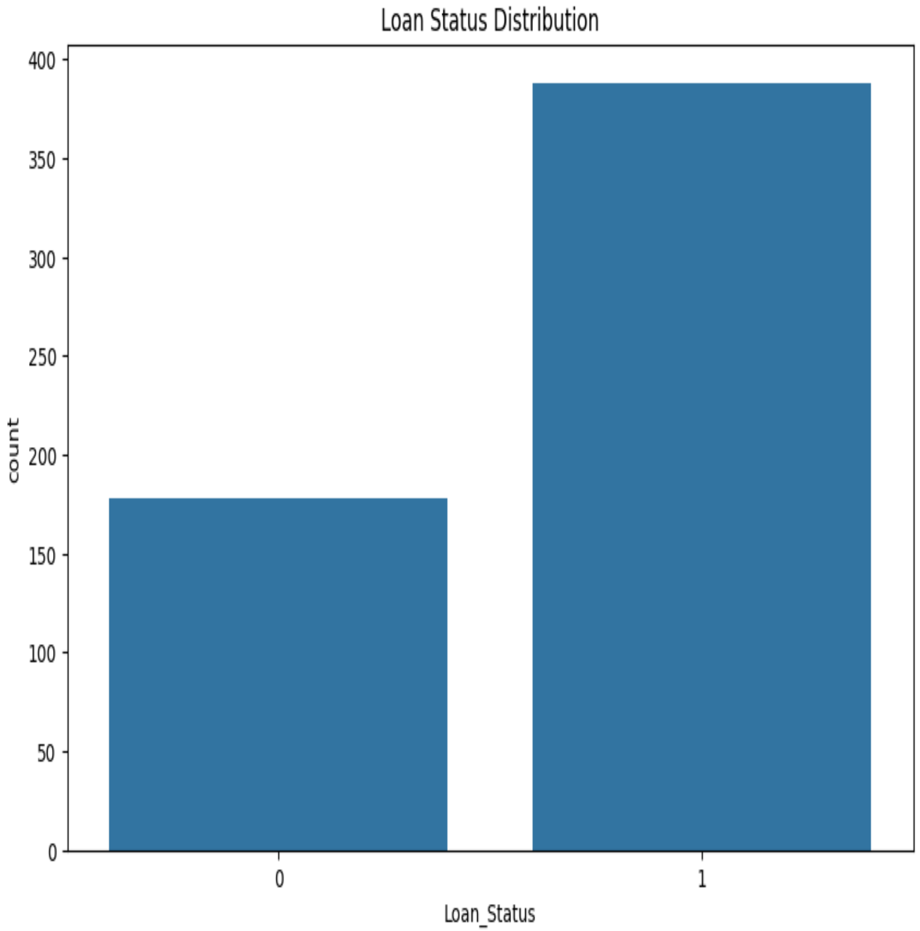
Library Requirements

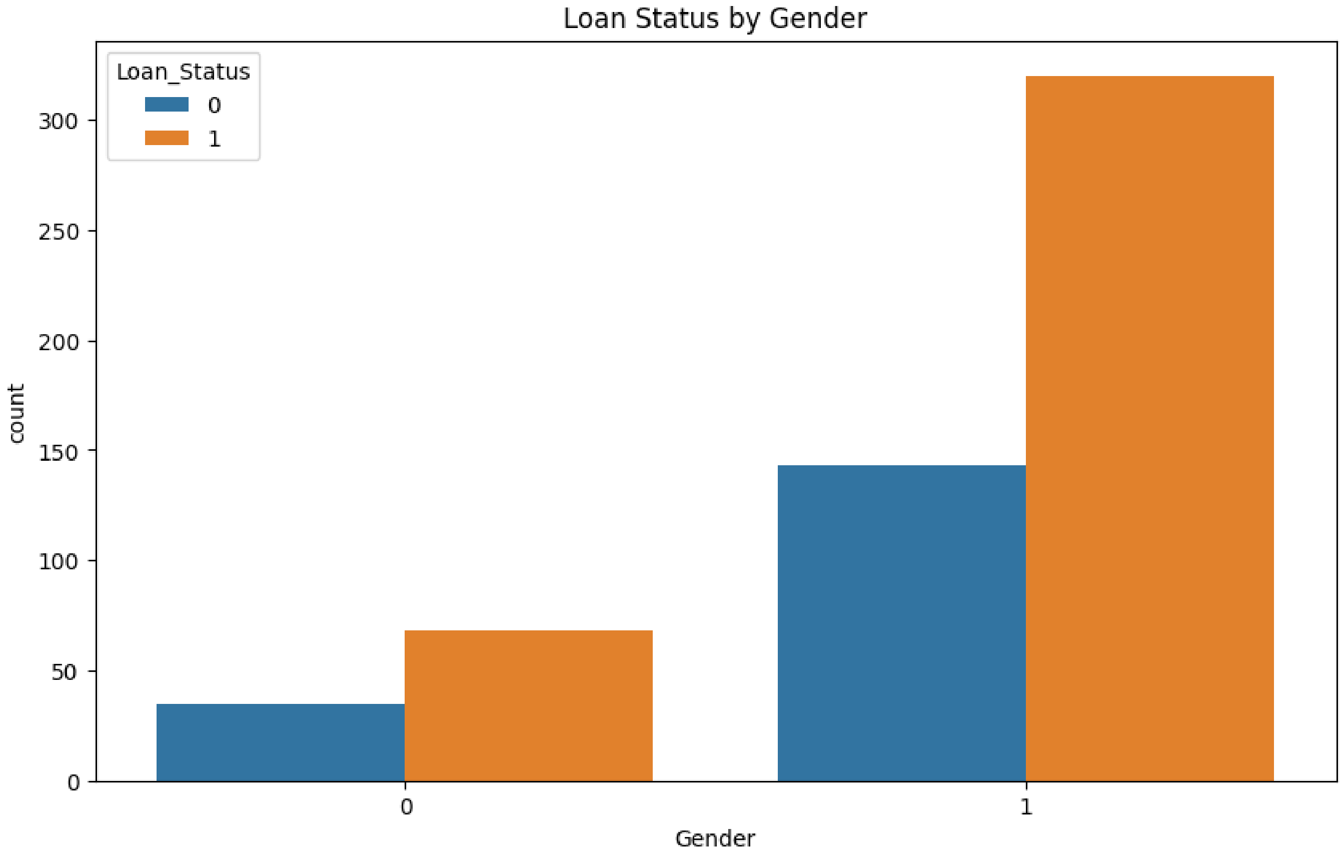


```bash
pip install pandas numpy scikit-learn seaborn matplotlib xgboost imbalanced-learn
```

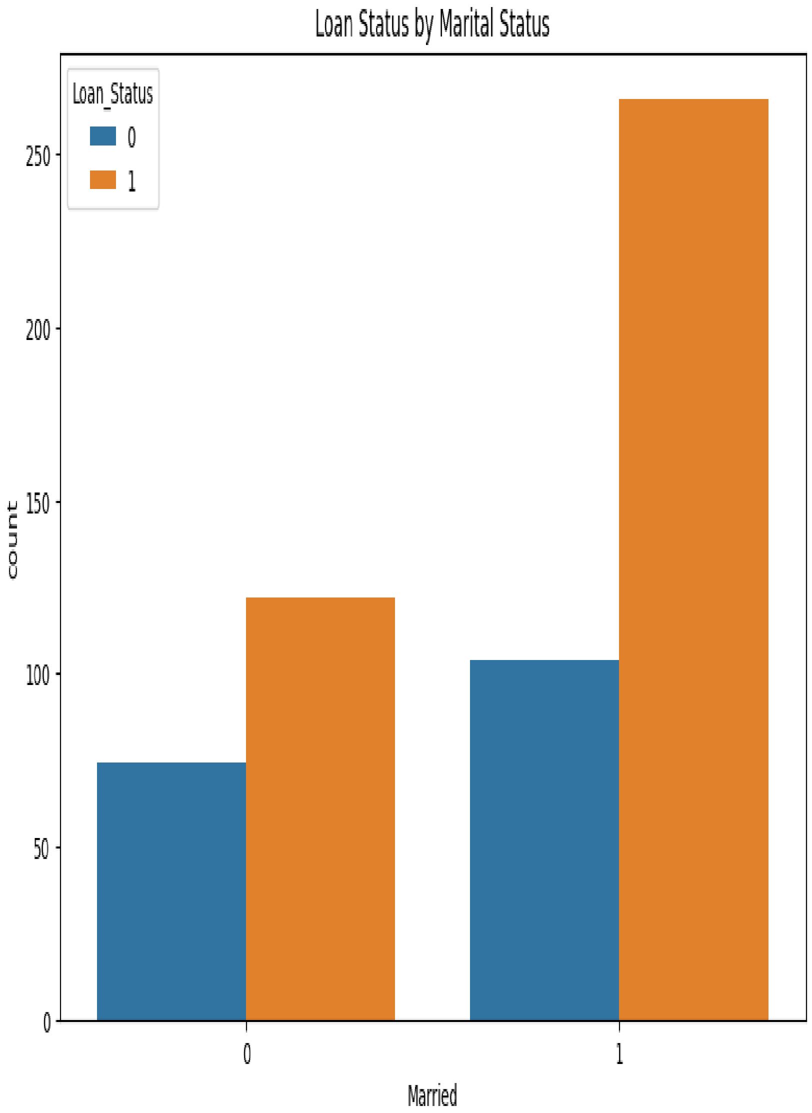
DATA VISUALIZATION (EDA)

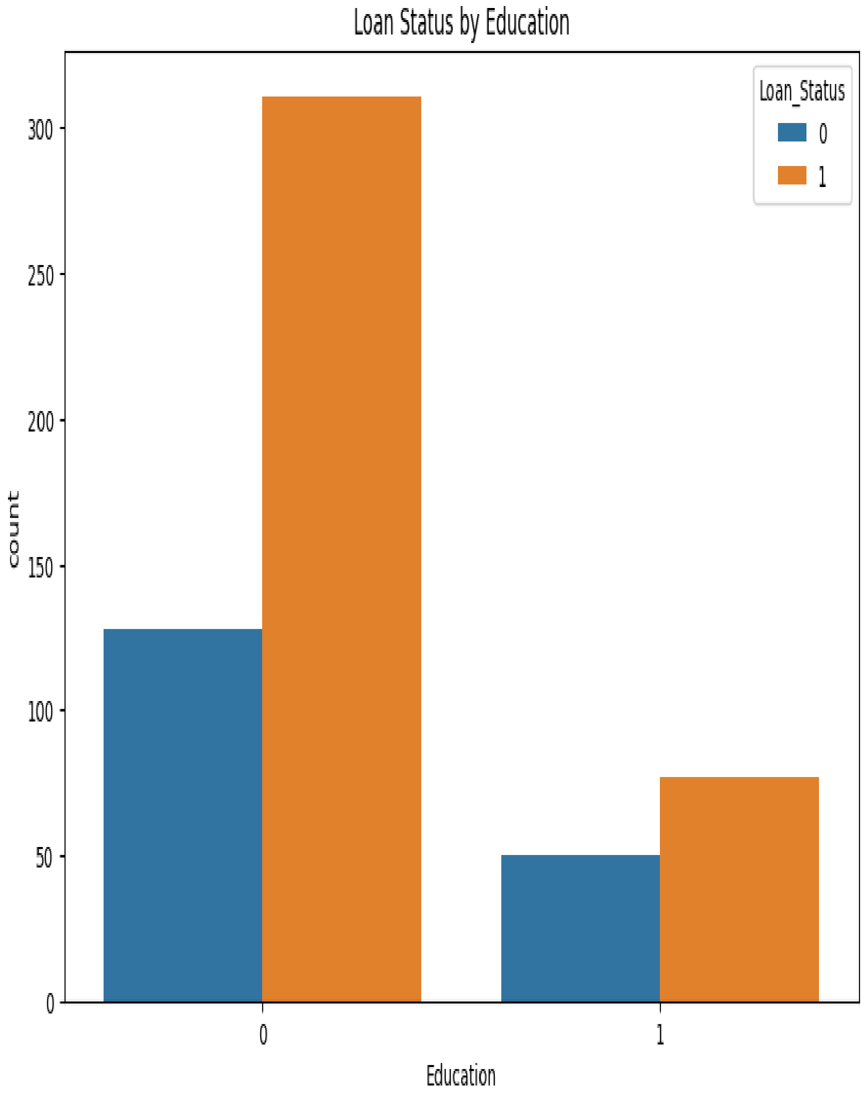
EDA was performed too understand the distribution and relationships of the data. The following visualizations were created to analyze the dataset:

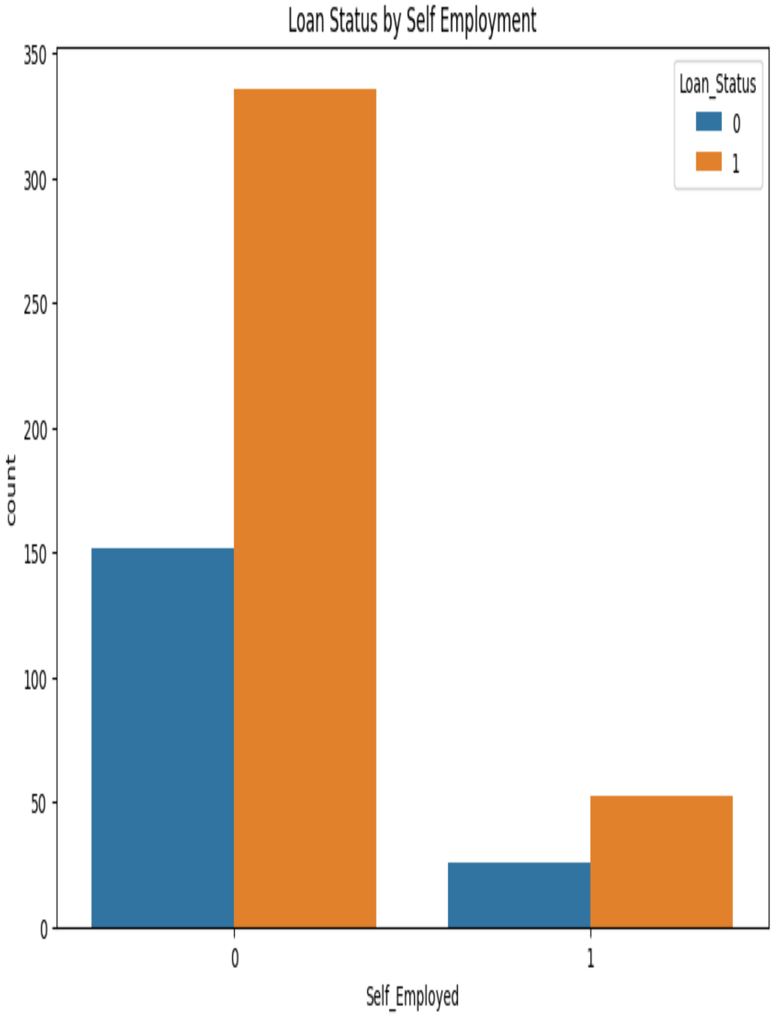




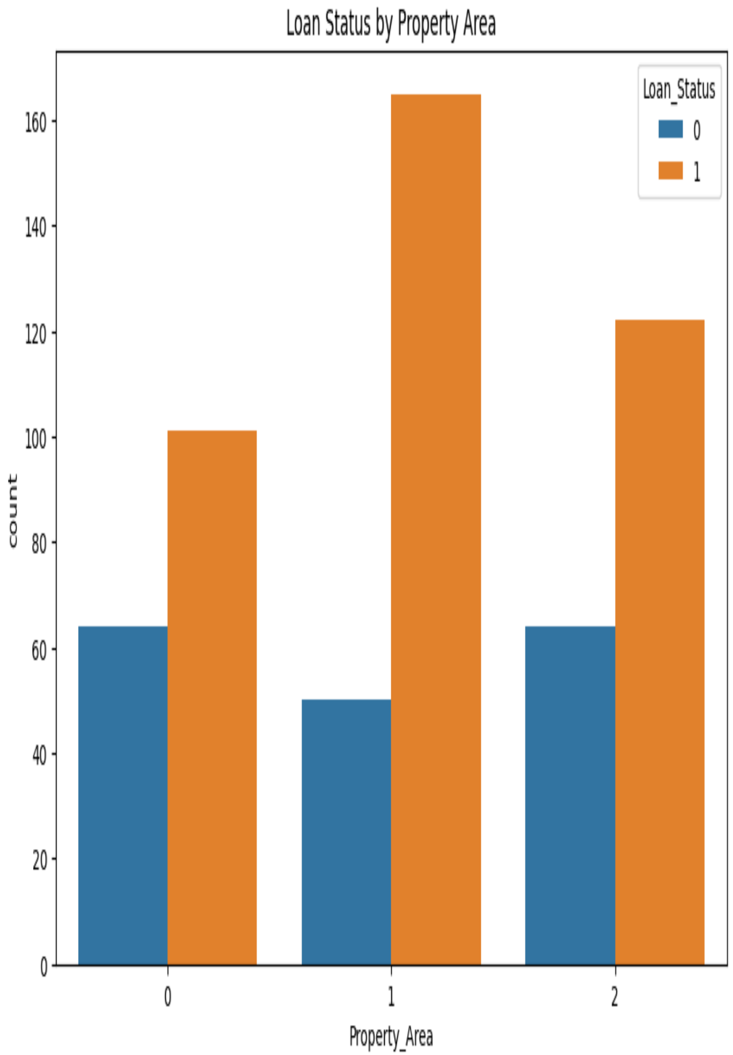
DATA VISUALIZATION (EDA)

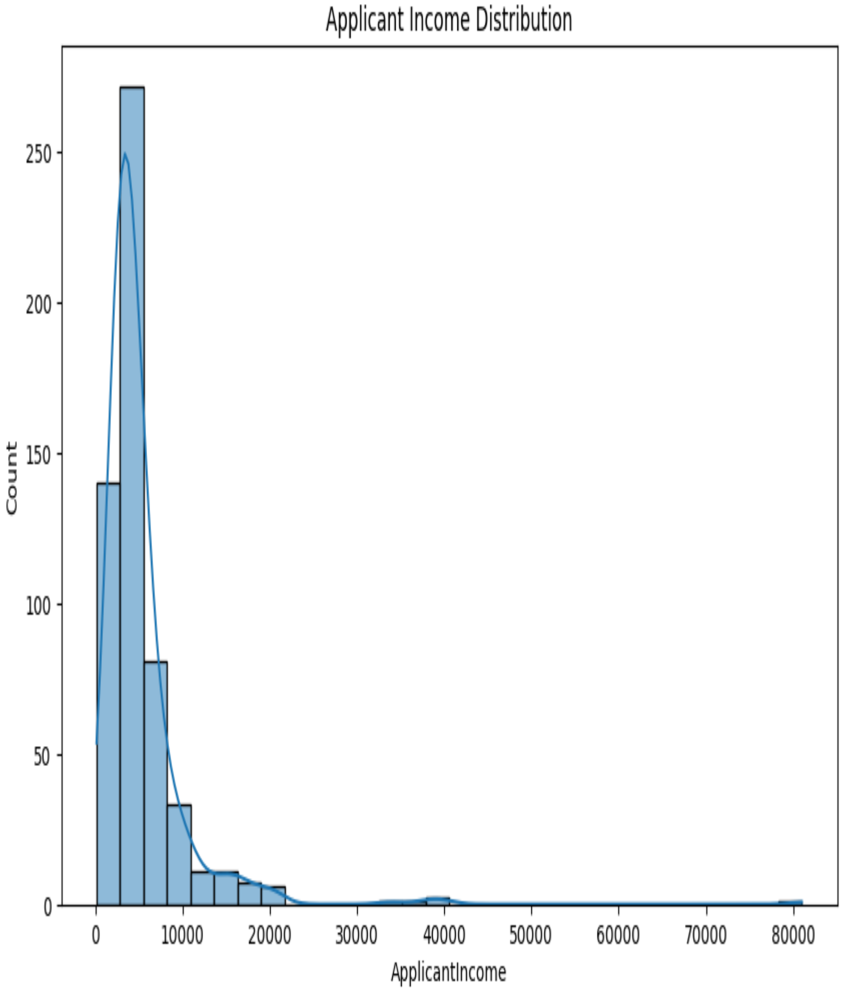


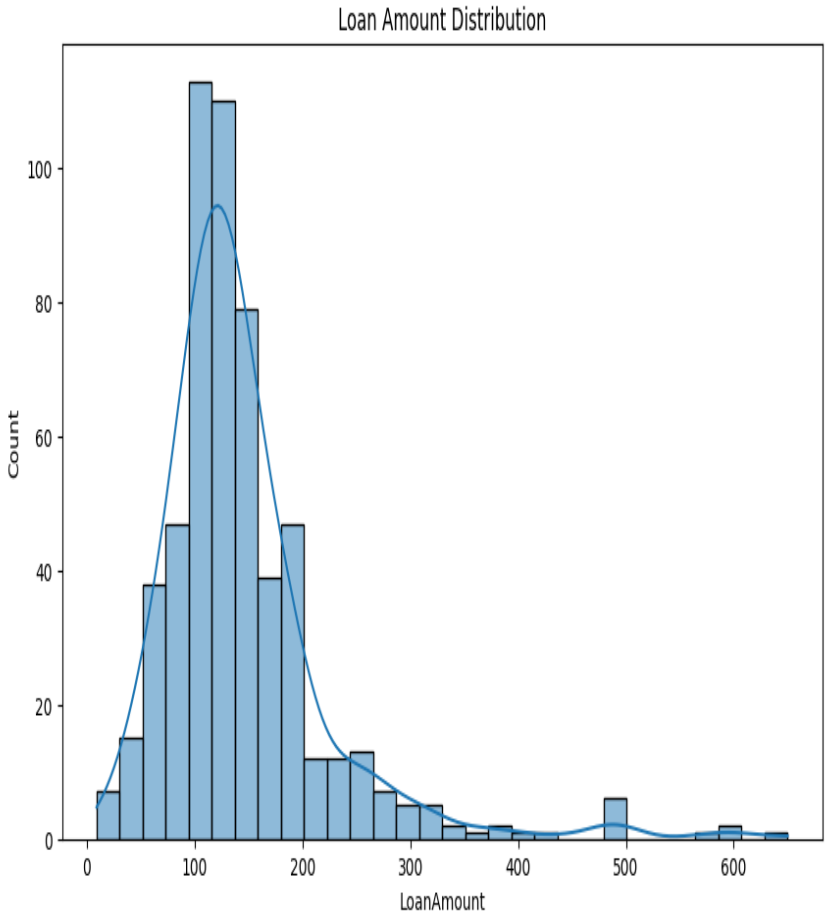




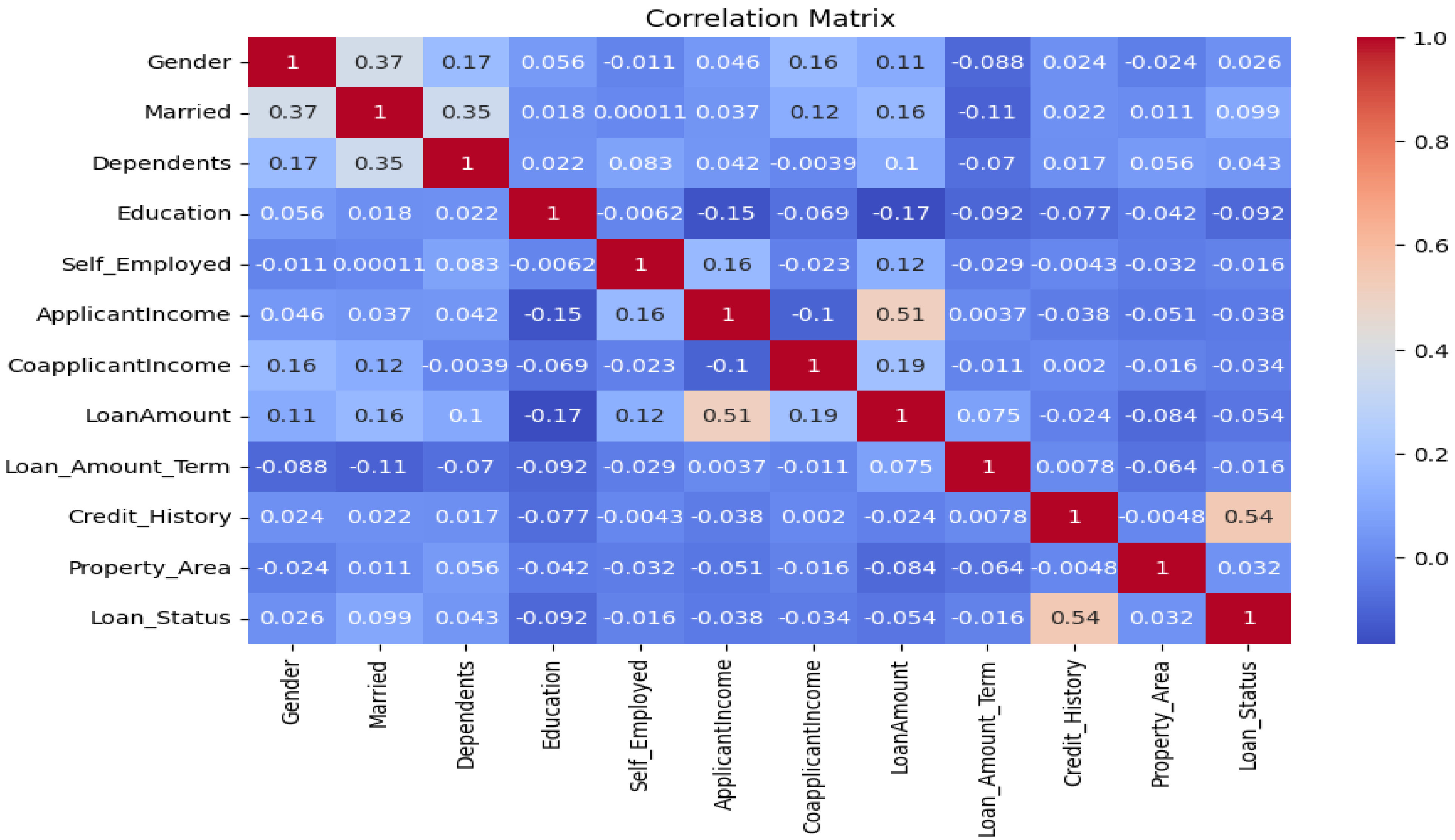
DATA VISUALIZATION (EDA)







DATA VISUALIZATION (EDA)



Correlation Matrix

	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
Gender	1	0.37	0.17	0.056	-0.011	0.046	0.16	0.11	-0.088	0.024	-0.024	0.026
Married	0.37	1	0.35	0.018	0.00011	0.037	0.12	0.16	-0.11	0.022	0.011	0.099
Dependents	0.17	0.35	1	0.022	0.083	0.042	-0.0039	0.1	-0.07	0.017	0.056	0.043
Education	0.056	0.018	0.022	1	-0.0062	-0.15	-0.069	-0.17	-0.092	-0.077	-0.042	-0.092
Self_Employed	-0.011	0.00011	0.083	-0.0062	1	0.16	-0.023	0.12	-0.029	-0.0043	-0.032	-0.016
ApplicantIncome	0.046	0.037	0.042	-0.15	0.16	1	-0.1	0.51	0.0037	-0.038	-0.051	-0.038
CoapplicantIncome	0.16	0.12	-0.0039	-0.069	-0.023	-0.1	1	0.19	-0.011	0.002	-0.016	-0.034
LoanAmount	0.11	0.16	0.1	-0.17	0.12	0.51	0.19	1	0.075	-0.024	-0.084	-0.054
Loan_Amount_Term	-0.088	-0.11	-0.07	-0.092	-0.029	0.0037	-0.011	0.075	1	0.0078	-0.064	-0.016
Credit_History	0.024	0.022	0.017	-0.077	-0.0043	-0.038	0.002	-0.024	0.0078	1	-0.0048	0.54
Property_Area	-0.024	0.011	0.056	-0.042	-0.032	-0.051	-0.016	-0.084	-0.064	-0.0048	1	0.032
Loan_Status	0.026	0.099	0.043	-0.092	-0.016	-0.038	-0.034	-0.054	-0.016	0.54	0.032	1

ALGORITHM & DEPLOYMENT

Algorithm and Deployment

Algorithm Selection

- **Logistic Regression:** Chosen for its simplicity and interpretability.
- **Decision Trees:** For capturing non-linear relationships.
- **Random Forest:** For better accuracy by reducing overfitting.
- **Gradient Boosting:** For improving model performance through boosting.
- **XGBoost:** For efficient and scalable implementation of gradient boosting.

Data Input

- Historical loan data with features like applicant income, co-applicant income, loan amount, loan term, credit history, etc.

Training Process

- Splitting the data into training and testing sets.
- Training each algorithm on the training set.
- Performing cross-validation and hyperparameter tuning.

Prediction Process

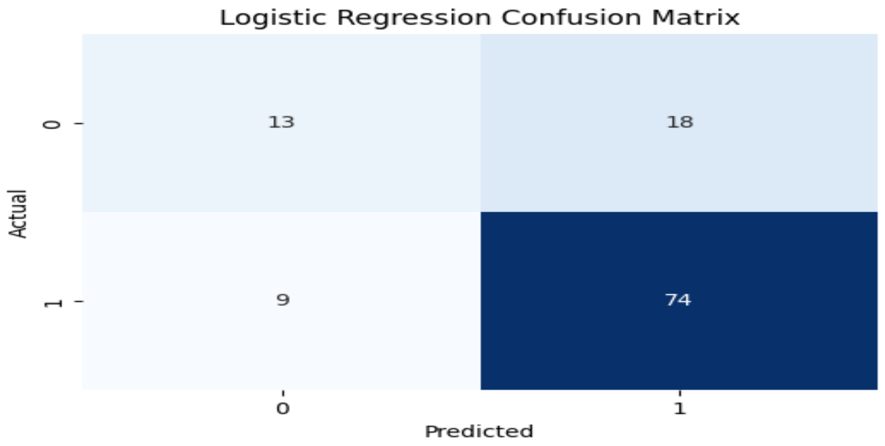
- Making predictions on the test set.
- Evaluating model performance using the selected metrics.

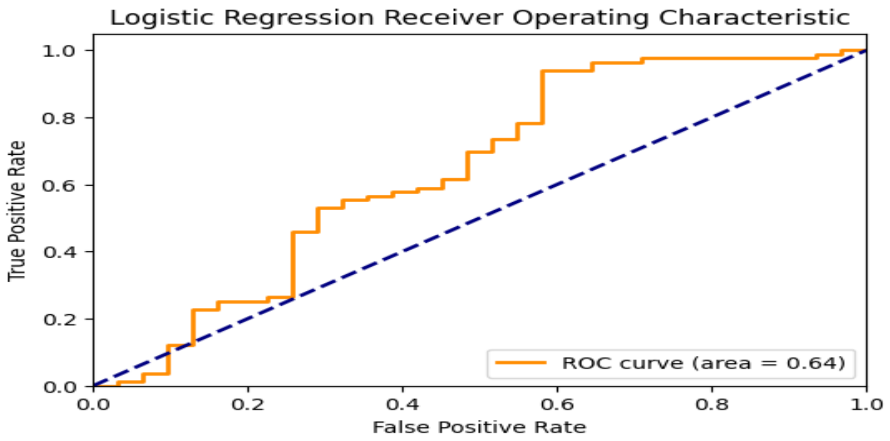
RESULT

The machine learning models were trained and evaluated to predict the likelihood of loan approval. Below are the detailed results for each model:

1. Logistic Regression

- **Classification Report:**
  - Precision: 0.59 for class 0, 0.80 for class 1
  - Recall: 0.42 for class 0, 0.89 for class 1
  - F1-score: 0.49 for class 0, 0.85 for class 1
  - Accuracy: 76.32%
  - AUC-ROC Score: 0.6555
- **Confusion Matrix:**
  - True Positives: 74
  - True Negatives: 13
  - False Positives: 18
  - False Negatives: 9

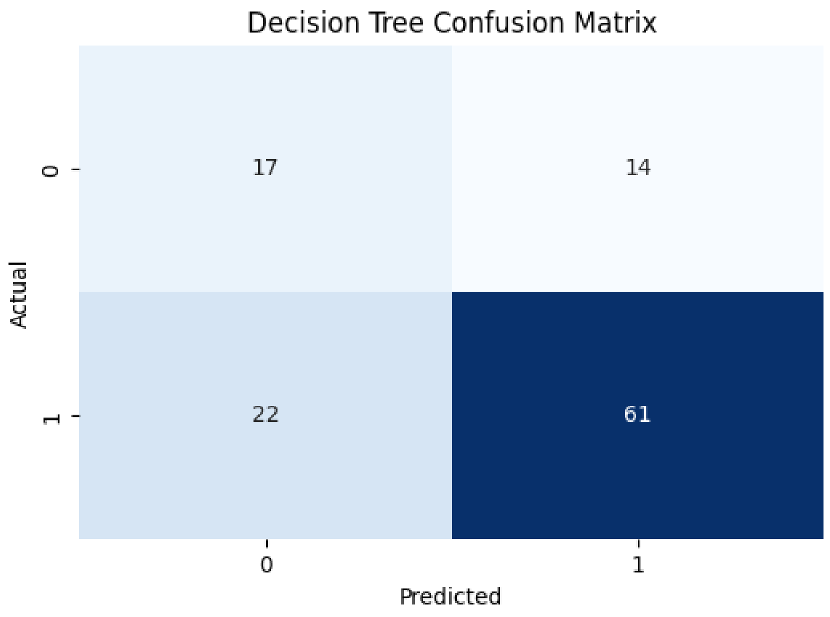


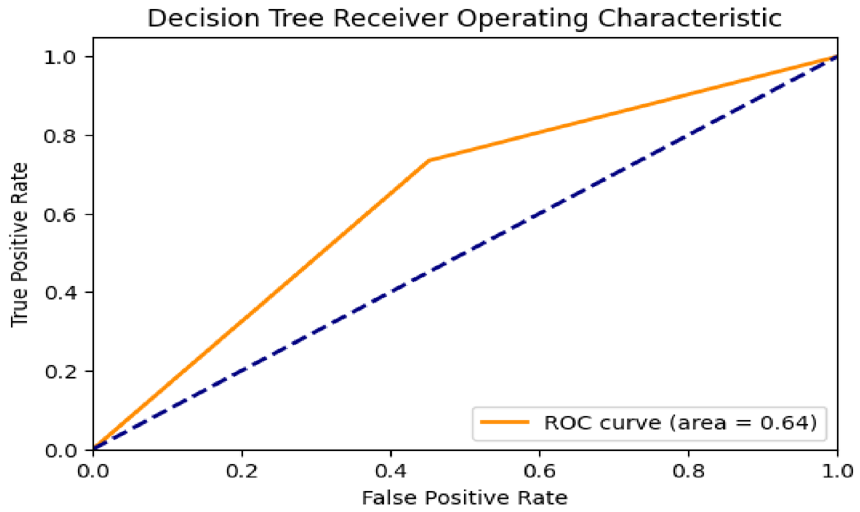


RESULT

2. Decision Tree

- **Classification Report:**
  - Precision: 0.44 for class 0, 0.81 for class 1
  - Recall: 0.55 for class 0, 0.73 for class 1
  - F1-score: 0.49 for class 0, 0.77 for class 1
  - Accuracy: 68.42%
  - AUC-ROC Score: 0.6417
- **Confusion Matrix:**
  - True Positives: 61
  - True Negatives: 17
  - False Positives: 14
  - False Negatives: 22

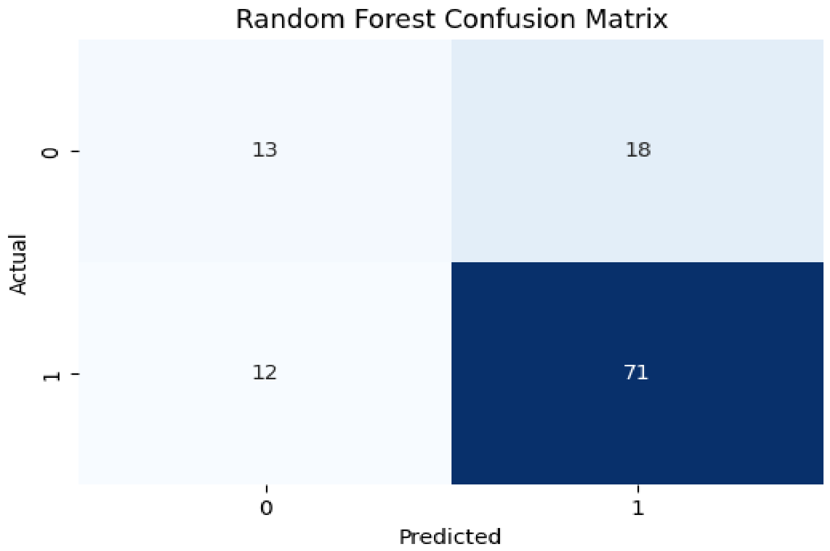


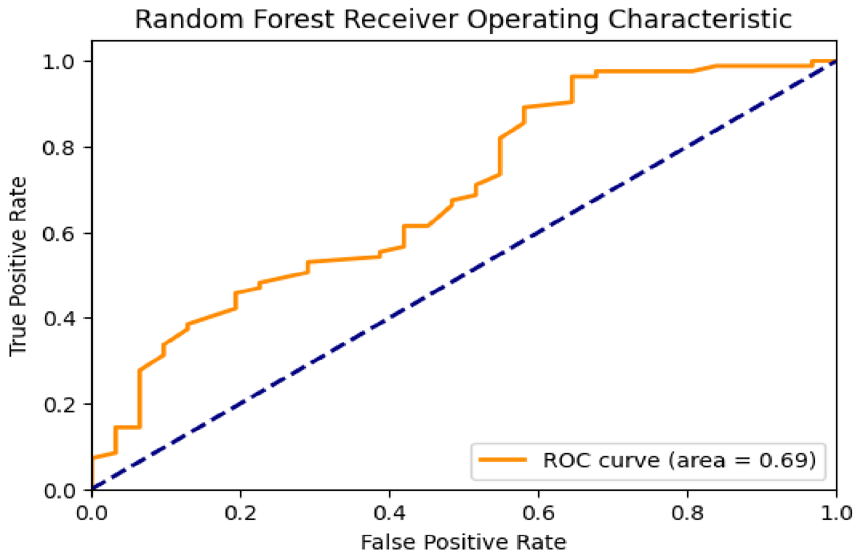


RESULT

3. Random Forest

- **Classification Report:**
  - Precision: 0.52 for class 0, 0.80 for class 1
  - Recall: 0.42 for class 0, 0.86 for class 1
  - F1-score: 0.46 for class 0, 0.83 for class 1
  - Accuracy: 73.68%
  - AUC-ROC Score: 0.6374
- **Confusion Matrix:**
  - True Positives: 71
  - True Negatives: 13
  - False Positives: 18
  - False Negatives: 12

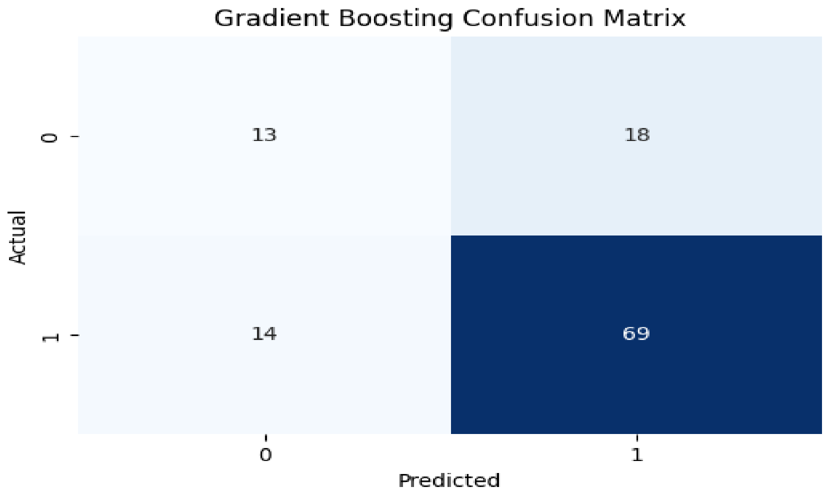




RESULT

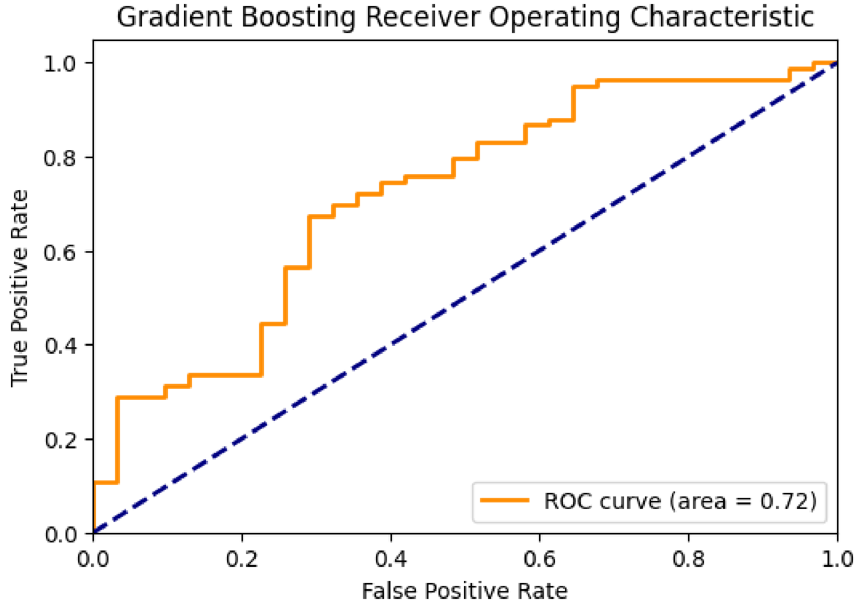
4. Gradient Boosting

- **Classification Report:**
  - Precision: 0.48 for class 0, 0.79 for class 1
  - Recall: 0.42 for class 0, 0.83 for class 1
  - F1-score: 0.45 for class 0, 0.81 for class 1
  - Accuracy: 71.93%
  - AUC-ROC Score: 0.6253
- **Confusion Matrix:**
  - True Positives: 69
  - True Negatives: 13
  - False Positives: 18
  - False Negatives: 14



Gradient Boosting Confusion Matrix

Actual \ Predicted	0	1
0	13	18
1	14	69

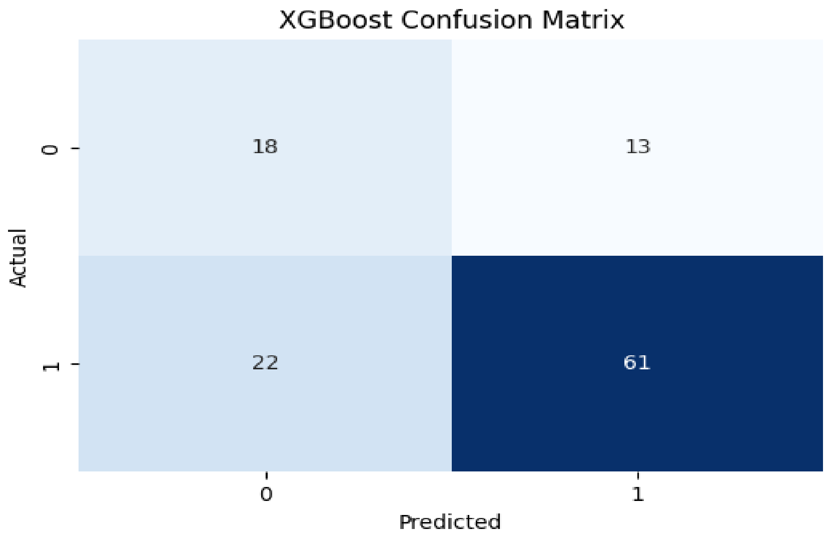


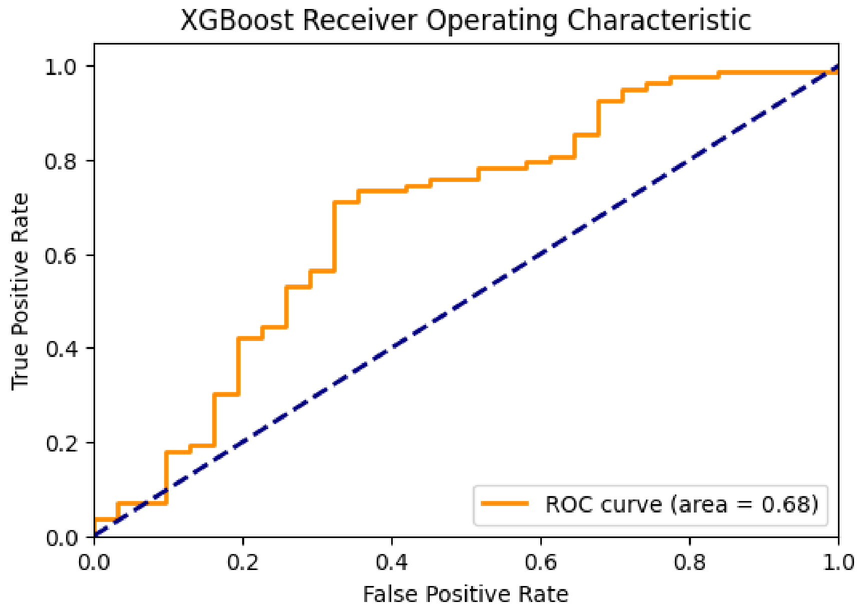
Gradient Boosting Receiver Operating Characteristic

RESULT

5. XGBoost

- **Classification Report:**
  - Precision: 0.45 for class 0, 0.82 for class 1
  - Recall: 0.58 for class 0, 0.73 for class 1
  - F1-score: 0.51 for class 0, 0.78 for class 1
  - Accuracy: 69.30%
  - AUC-ROC Score: 0.6578
- **Confusion Matrix:**
  - True Positives: 61
  - True Negatives: 18
  - False Positives: 13
  - False Negatives: 22





RESULT

Model Comparison

The results indicate that the Logistic Regression model achieved the highest accuracy of 76.32%, followed by Random Forest at 73.68%. However, in terms of the AUC-ROC score, XGBoost slightly outperformed Logistic Regression, indicating better performance in distinguishing between the classes.

CONCLUSION

- The project successfully developed and evaluated several machine learning models to predict loan approval. Logistic Regression demonstrated the highest accuracy (0.76), indicating its reliability for this task. The Random Forest model also performed well, with an accuracy of 0.74, suggesting it as a strong candidate for deployment. Although the XGBoost model had a slightly lower accuracy (0.69), it provided the best AUC-ROC score (0.658), suggesting a robust classification capability.
- During the implementation, some challenges were encountered, such as handling missing values and class imbalances. These were addressed using data preprocessing techniques like filling missing values and employing the SMOTE algorithm for oversampling.
- Potential improvements for this project include further hyperparameter tuning of the models and exploring additional features that could improve prediction accuracy. Additionally, combining multiple models through ensemble methods might yield better results.
- Accurate loan approval predictions are crucial for financial institutions to minimize risks and streamline the approval process. The models developed in this project can significantly contribute to achieving these goals, ensuring efficient and reliable loan processing.

FUTURE SCOPE

There are several potential enhancements and expansions for this loan approval prediction system, which can further improve its accuracy, scalability, and usability. These future improvements could include:

1. Incorporating Additional Features:

- **Credit Score:** Integrating credit scores can significantly enhance the model's predictive power, providing a more comprehensive view of an applicant's creditworthiness.
- **Employment History:** Including details about an applicant's employment history, such as job stability and length of employment, can add valuable context to the prediction.
- **Financial Indicators:** Other financial indicators, such as debt-to-income ratio, savings, and investment details, can provide a more holistic assessment of the applicant's financial health.

REFERENCES

2. Implementing Real-Time Data Processing:

- **Real-Time Data Integration:** Enhancing the model to process real-time data can improve prediction accuracy and applicability. This would involve streaming data pipelines that continuously update the model with the latest information.
- **Dynamic Model Updating:** Implementing mechanisms for dynamically updating the model as new data comes in, ensuring the model remains relevant and accurate over time.

3. Deploying the Model as a Web Service:

- **API Integration:** Deploying the model as a web service with API endpoints allows for seamless integration with financial institution systems. This can enable automated, real-time loan approval decisions.
- **Scalable Infrastructure:** Ensuring the deployment infrastructure is scalable to handle large volumes of requests efficiently, possibly leveraging cloud services for flexibility and reliability.
- **User-Friendly Interface:** Developing a user-friendly interface for stakeholders to interact with the model, visualize results, and gain insights without needing in-depth technical knowledge.

REFERENCES

- 1. **Kaggle Dataset:** Loan data was obtained from Kaggle. The dataset can be accessed at [Kaggle: Loan Data Set] (https://www.kaggle.com/datasets/burak3ergun/loan-data-set).

- 2. **Source Dataset Information:** The original dataset was sourced from Analytics Vidhya's practice problem, "Loan Prediction III". More details can be found at [Analytics Vidhya: Loan Prediction III] (https://datahack.analyticsvidhya.com/contest/practice-problem-loan-prediction-iii/).

- 3. **Logistic Regression:** Hosmer, D. W., & Lemeshow, S. (2000). "Applied Logistic Regression". John Wiley & Sons.

- 4. **Decision Trees:** Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). "Classification and Regression Trees". Wadsworth & Brooks/Cole Advanced Books & Software.

- 5. **Random Forests:** Breiman, L. (2001). "Random Forests". Machine Learning, 45(1), 5-32. DOI: [10.1023/A:1010933404324] (https://doi.org/10.1023/A:1010933404324).

REFERENCES

- 6. **Gradient Boosting:** Friedman, J. H. (2001). "Greedy Function Approximation: A Gradient Boosting Machine". Annals of Statistics, 29(5), 1189-1232. DOI: [10.1214/aos/1013203451](https://doi.org/10.1214/aos/1013203451).

- 7. XGBoost: Chen, T., & Guestrin, C. (2016). "XGBoost: A Scalable Tree Boosting System". Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794. DOI: [10.1145/2939672.2939785](https://doi.org/10.1145/2939672.2939785).

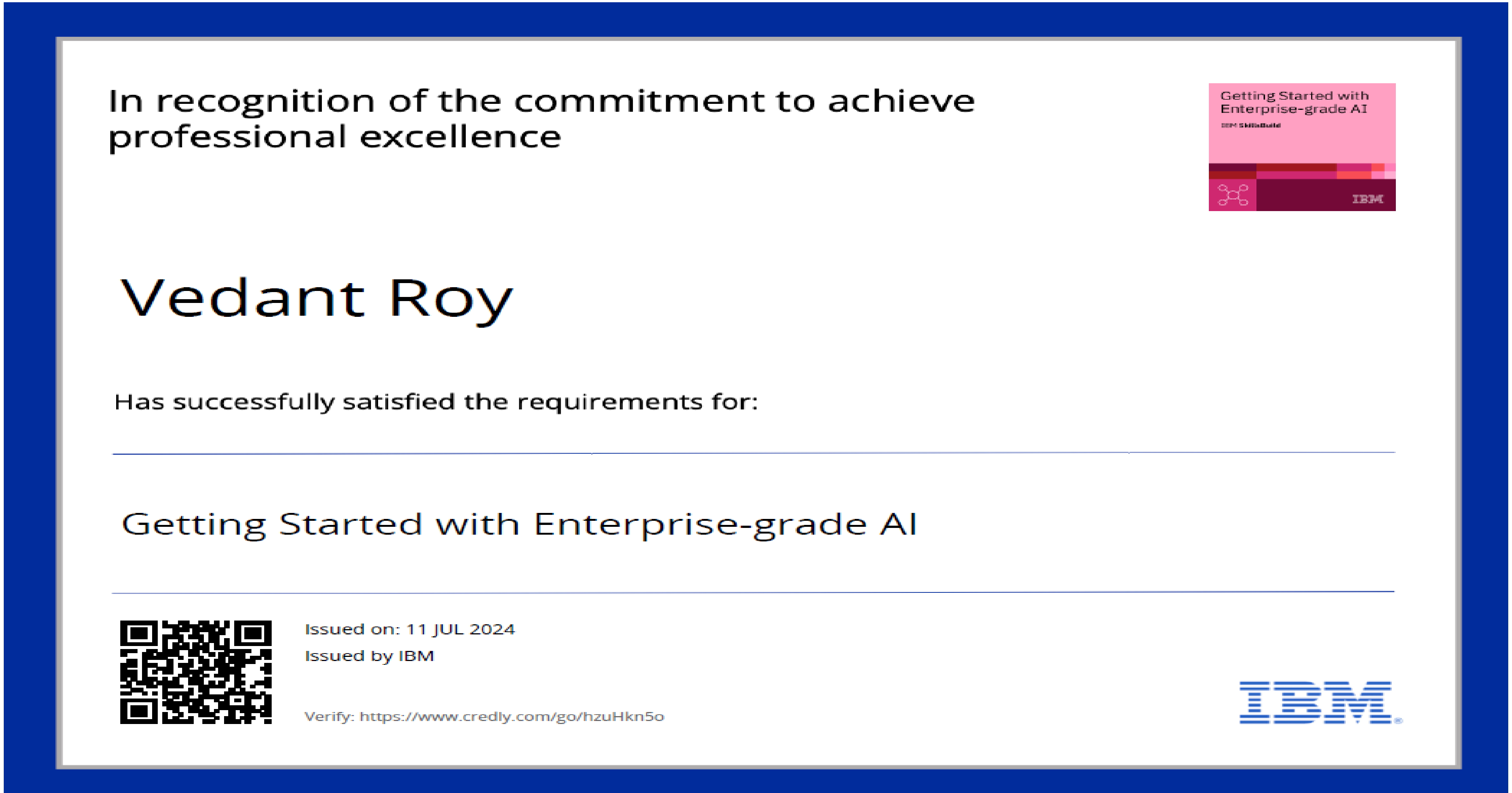
- 8. **SMOTE:** Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). "SMOTE: Synthetic Minority Over-sampling Technique". Journal of Artificial Intelligence Research, 16, 321-357. DOI: [10.1613/jair.953](https://doi.org/10.1613/jair.953).

- 9. **Model Evaluation Metrics:** Fawcett, T. (2006). "An Introduction to ROC Analysis". Pattern Recognition Letters, 27(8), 861-874. DOI: [10.1016/j.patrec.2005.10.010](https://doi.org/10.1016/j.patrec.2005.10.010).

- 10. **Data Preprocessing Techniques:** Han, J., Kamber, M., & Pei, J. (2011). "Data Mining: Concepts and Techniques". Elsevier.

- 11. **Project Code and Outputs**: The complete code and outputs for this project can be accessed in the Google Colab notebook available at: https://colab.research.google.com/drive/18VV0DSwpSBA2SrK7peF-Z9Sdt-PrOGDX?usp=sharing

COURSE CERTIFICATE (GETTING STARTED WITH ENTERPRISE- GRADE AI)



Getting Started with Enterprise-grade AI
IBM SkillsBuild
IBM

In recognition of the commitment to achieve professional excellence

Vedant Roy

Has successfully satisfied the requirements for:

Getting Started with Enterprise-grade AI

Issued on: 11 JUL 2024
Issued by IBM

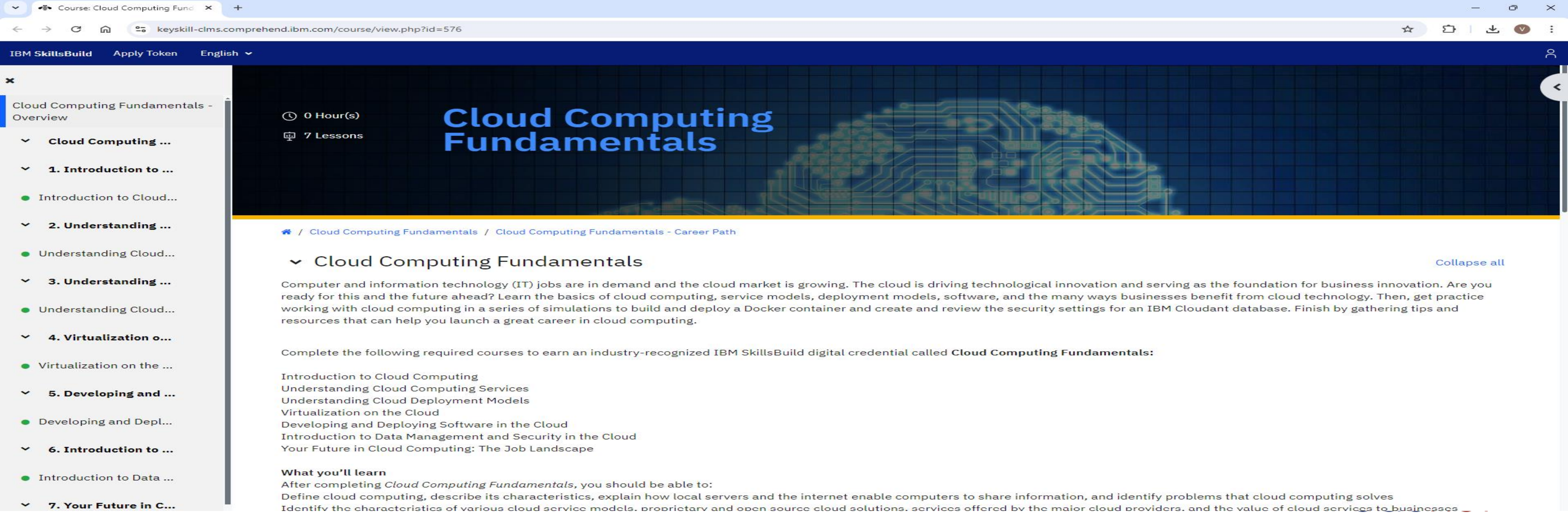
Verify: https://www.credly.com/go/hzuHkn5o

IBM

COURSE CERTIFICATE 2 (CLOUD COMPUTING FUNDAMENTALS)

Note- Didn’t got certificate from credly website despite completing the course 100% so sharing the screenshots as proof of completion

Part- 1



Cloud Computing Fundamentals

Computer and information technology (IT) jobs are in demand and the cloud market is growing. The cloud is driving technological innovation and serving as the foundation for business innovation. Are you ready for this and the future ahead? Learn the basics of cloud computing, service models, deployment models, software, and the many ways businesses benefit from cloud technology. Then, get practice working with cloud computing in a series of simulations to build and deploy a Docker container and create and review the security settings for an IBM Cloudant database. Finish by gathering tips and resources that can help you launch a great career in cloud computing.

Complete the following required courses to earn an industry-recognized IBM SkillsBuild digital credential called **Cloud Computing Fundamentals:**

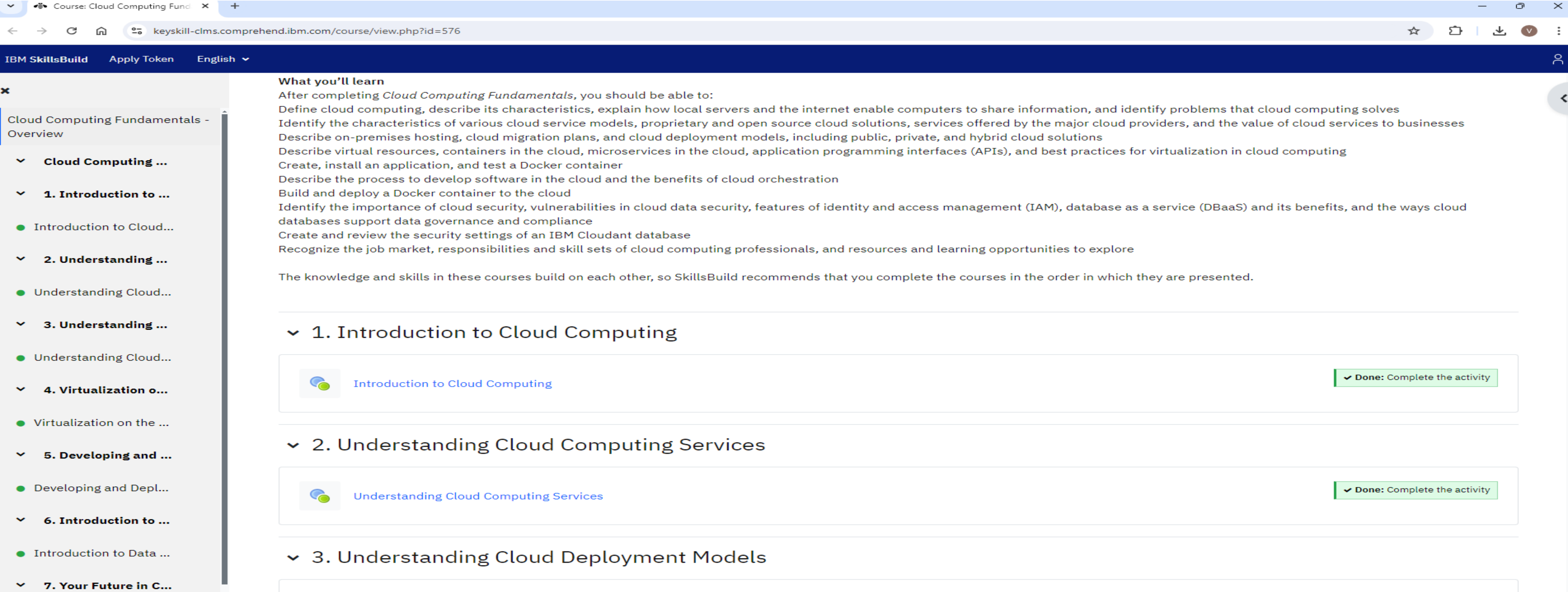
Introduction to Cloud Computing
Understanding Cloud Computing Services
Understanding Cloud Deployment Models
Virtualization on the Cloud
Developing and Deploying Software in the Cloud
Introduction to Data Management and Security in the Cloud
Your Future in Cloud Computing: The Job Landscape

What you’ll learn
After completing *Cloud Computing Fundamentals*, you should be able to:
Define cloud computing, describe its characteristics, explain how local servers and the internet enable computers to share information, and identify problems that cloud computing solves
Identify the characteristics of various cloud service models, proprietary and open source cloud solutions, services offered by the major cloud providers, and the value of cloud services to businesses

foundation

COURSE CERTIFICATE 2 (CLOUD COMPUTING FUNDAMENTALS)

Part- 2



What you'll learn

After completing *Cloud Computing Fundamentals*, you should be able to:

Define cloud computing, describe its characteristics, explain how local servers and the internet enable computers to share information, and identify problems that cloud computing solves
Identify the characteristics of various cloud service models, proprietary and open source cloud solutions, services offered by the major cloud providers, and the value of cloud services to businesses
Describe on-premises hosting, cloud migration plans, and cloud deployment models, including public, private, and hybrid cloud solutions
Describe virtual resources, containers in the cloud, microservices in the cloud, application programming interfaces (APIs), and best practices for virtualization in cloud computing
Create, install an application, and test a Docker container
Describe the process to develop software in the cloud and the benefits of cloud orchestration
Build and deploy a Docker container to the cloud
Identify the importance of cloud security, vulnerabilities in cloud data security, features of identity and access management (IAM), database as a service (DBaaS) and its benefits, and the ways cloud databases support data governance and compliance
Create and review the security settings of an IBM Cloudant database
Recognize the job market, responsibilities and skill sets of cloud computing professionals, and resources and learning opportunities to explore

The knowledge and skills in these courses build on each other, so SkillsBuild recommends that you complete the courses in the order in which they are presented.

1. Introduction to Cloud Computing

Introduction to Cloud Computing — ✓ **Done:** Complete the activity

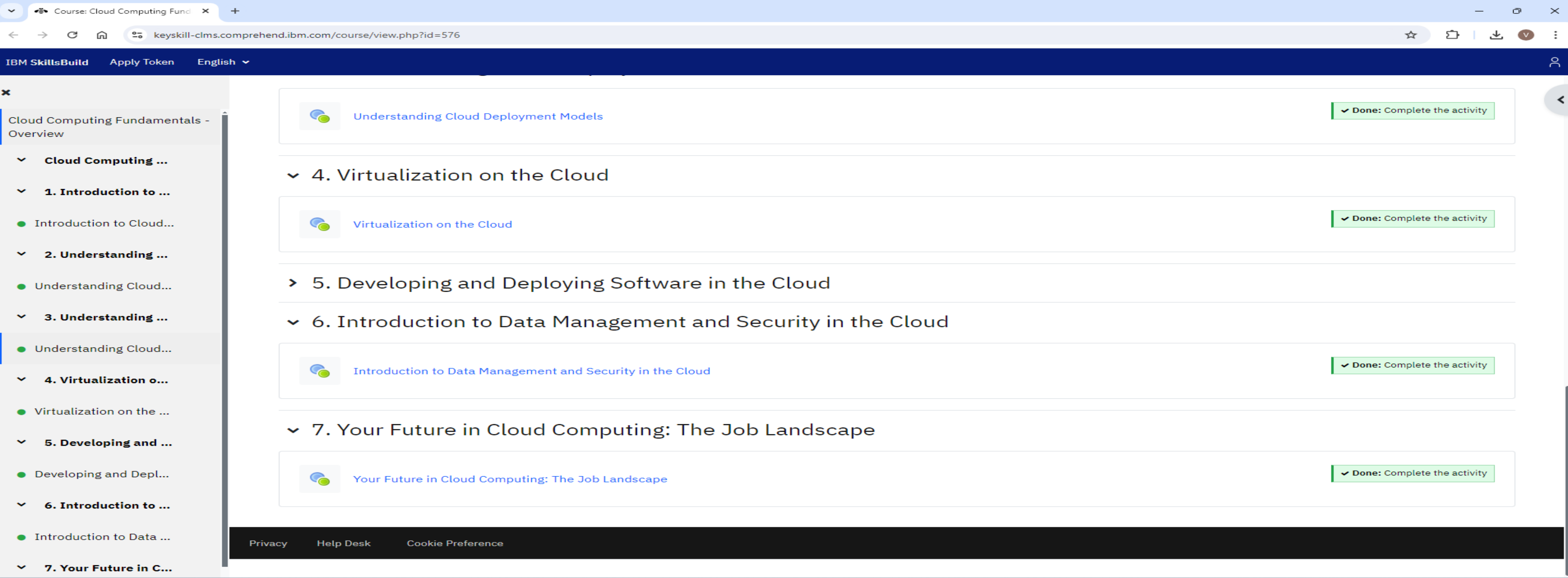
2. Understanding Cloud Computing Services

Understanding Cloud Computing Services — ✓ **Done:** Complete the activity

3. Understanding Cloud Deployment Models

COURSE CERTIFICATE 2 (CLOUD COMPUTING FUNDAMENTALS)

Part- 3



Course: Cloud Computing Fund
keyskill-clms.comprehend.ibm.com/course/view.php?id=576
IBM SkillsBuild
Apply Token
English

Cloud Computing Fundamentals - Overview

- Cloud Computing ...
- 1. Introduction to ...
- Introduction to Cloud...
- 2. Understanding ...
- Understanding Cloud...
- 3. Understanding ...
- Understanding Cloud...
- 4. Virtualization o...
- Virtualization on the ...
- 5. Developing and ...
- Developing and Depl...
- 6. Introduction to ...
- Introduction to Data ...
- 7. Your Future in C...

Understanding Cloud Deployment Models — ✓ Done: Complete the activity

4. Virtualization on the Cloud

Virtualization on the Cloud — ✓ Done: Complete the activity

5. Developing and Deploying Software in the Cloud

6. Introduction to Data Management and Security in the Cloud

Introduction to Data Management and Security in the Cloud — ✓ Done: Complete the activity

7. Your Future in Cloud Computing: The Job Landscape

Your Future in Cloud Computing: The Job Landscape — ✓ Done: Complete the activity

THANK YOU