Data Science (DS)

|  |
| --- |
| TECHNICAL REPORT |

|  |
| --- |
|  |

|  |  |
| --- | --- |
| FALL 23 |  |



Contents

Project Name ………………………………………….3

Technical Report………………………………………4

Highlights of Project…………………………………5

Abstract…………………………………………………..7

Data Collection………………………………………..8

Executive Summary………………………………….9

Methodology………………………………………….13

Results Section……………………………………….15

Conclusion………………………………………………21

|  |
| --- |
| Loan Approval Prediction System |

|  |  |  |
| --- | --- | --- |
| **TITLE: LOAN APPROVAL PREDICTION SYSTEM** | | |
|  | | |
| Cover Page:  Title: Loan Approval Prediction System.  Authors:  Syamala Durga: Data preprocessing [Team Lead]  (email: [dmar3@unh.newhaven.edu](mailto:dmar3@unh.newhaven.edu))  Asif Ali Shaik: Data Analysis  (email: [among4@unh.newhaven.edu](mailto:among4@unh.newhaven.edu))  Nithish Kumar Mayavan-Data Modeling  (email: [nmaya1@unh.newhaven.edu](mailto:nmaya1@unh.newhaven.edu))  Keerthi Kappera -Model deployment/Application Development  (email: [kkapp3@unh.newhaven.edu](mailto:kkapp3@unh.newhaven.edu)) |  |  |

|  |
| --- |
| Technical Report |

|  |  |
| --- | --- |
|  |  |
| Highlights of Project:The project aims to develop a predictive model that assesses the creditworthiness of loan applicants.The primary goal is to minimize the risk associated with loan approval by accurately predicting whether a borrower is likely to default on their loan.The model incorporates feature engineering and regularization techniques to improve accuracy and robustness.The use of Gaussian NB modeling is shown to be effective in capturing complex patterns in the data.The model provides reliable predictions with a high degree of confidence.The findings can provide valuable insight potential features influencing loan approval.The model incorporates feature engineering and hyperparameter tuning to improve performance.The use of a Gaussian NB is shown to be effective in capturingcomplex patterns in the data.The model provides reliable predictions with a high degree of accuracy. |
| Training and Testing data highlights:   * The training data is used to train the model and learn the patterns and relationships in the data * The testing data is used to evaluate the performance of the trained model on unseen data * The accuracy of the model is measured on the testing data to ensure it generalizes well to new data * The split of data into training and testing sets is typically done using a pre-defined ratio, such as 80/20 or 70/30 * It is important to ensure that the training and testing data are representative of the overall population and are not biased towards a specific subset of the data   Top of Form  Top of Form |  |

## Abstract

The loan approval prediction system streamlines the decision-making process by automating the assessment of loan applications. This efficiency leads to faster responses to applicants, reducing the time traditionally spent on manual reviews and enabling financial institutions to handle a larger volume of applications promptly.

In this project, we explore the use of advanced regression techniques for predicting loan approval. We first present a brief overview of the relevant literature on loan approval prediction, highlighting the limitations of traditional methods. We then describe our approach, which involves using a combination of advanced regression techniques to model the relationship between various factors.

We evaluate the performance of our approach using a real-world dataset and compare the results to those obtained using traditional methods. Our experiments show that our approach can achieve significantly higher accuracy in predicting loan approvals compared to traditional methods. We conclude by discussing the implications of our findings and potential directions for future work.

Data Collection:

The dataset contains 13 features:

|  |  |  |
| --- | --- | --- |
| 1 | Loan | A unique id |
| 2 | Gender | Gender of the applicant Male/female |
| 3 | Married | Marital Status of the applicant, values will be Yes/ No |
| 4 | Dependents | It tells whether the applicant has any dependents or not. |
| 5 | Education | It will tell us whether the applicant is Graduated or not. |
| 6 | Self Employed | This defines that the applicant is self-employed i.e. Yes/ No |
| 7 | Applicant Income | Applicant income |
| 8 | Co applicant Income | Co-applicant income |
| 9 | Loan Amount | Loan amount (in thousands) |
| 10 | Loan Amount Term | Terms of loan (in months) |
| 11 | Credit History | Credit history of individual’s repayment of their debts |
| 12 | Property Area | Area of property i.e. Rural/Urban/Semi-urban |
| 13 | Loan Status | Status of Loan Approved or not i.e. Y- Yes, N-No |

Executive Summary:

Introductory Section:

As the data are increasing daily due to digitization in the banking sector, people want to apply for loans through the internet. Artificial intelligence (AI), as a typical method for information investigation, has gotten more consideration increasingly. Individuals of various businesses are utilizing AI calculations to take care of the issues dependent on their industry information. Banks are facing a significant problem in the approval of the loan. Daily there are so many applications that are challenging to manage by the bank employees, and also the chances of some mistakes are high. Most banks earn profit from the loan, but it is risky to choose deserving customers from the number of applications. One mistake can make a massive loss to a bank. Loan distribution is the primary business of almost every bank. This project aims to provide a loan to a deserving applicant out of all applicants. An efficient and non-biased system that reduces the bank’s time employs checking every applicant on a priority basis. The bank authorities complete all other customer’s other formalities on time, which positively impacts the customers. The best part is that it is efficient for both banks and applicants. This system allows jumping on particular applications that deserve to be approved on a priority basis. There are some features for the prediction like- ‘Gender’, ‘Married’, ‘Dependents’, ‘Education’, ‘Self\_ Employed’, ‘Applicant Income’, Co applicant Income, ‘Loan Amount’, ‘Loan Amount Term’, ‘Credit History’, ‘Property Area’, ‘Loan Status’.

Loan Approval Prediction based on Machine Learning Approach:

The main objective of this project is to predict whether assigning the loan to particular person will be safe or not. This project is divided into four sections (i)Data Collection (ii) Comparison of machine learning models on collected data (iii) Training of system on most promising model (iv) Testing.

Exploring the Machine Learning Algorithm for Prediction the Loan Sanctioning Process:

Extending credits to corporates and individuals for the smooth functioning of growing economies like India is inevitable. As increasing number of customers apply for loans in the banks and non- banking financial companies (NBFC), it is really challenging for banks and NBFCs with limited capital to device a standard resolution and safe procedure to lend money to its borrowers for their financial needs. In addition, in recent times NBFC inventories have suffered a significant downfall in terms of the stock price. It has contributed to a contagion that has also spread to other financial stocks, adversely affecting the benchmark in recent times. In this project, an attempt is made to condense the risk involved in selecting the suitable person who could repay the loan on time thereby keeping the bank’s nonperforming assets (NPA) on the hold. This is achieved by feeding the past records of the customer who acquired loans from the bank into a trained machine learning model which could yield an accurate result. The prime focus of the project is to determine whether or not it will be safe to allocate the loan to a particular person. This project has the following sections (i) Collection of Data, (ii) Data Cleaning and (iii) Performance Evaluation. Experimental tests found that the Naïve Bayes model has better performance Evaluation. Experimental tests found that the Naïve Bayes model has better performance than other models in terms of loan forecasting. With the enhancement in the banking sector lots of people are applying for bank loans but the bank has its limited assets which it has to grant to limited people only, so finding out to whom the loan can be granted which will be a safer option for the bank is a typical process. So, in this project we try to reduce this risk factor behind selecting the safe person to save lots of bank efforts and assets. This is done by mining the Big Data of the previous records of the people to whom the loan was granted before and based on these records/experiences the machine was trained using the machine learning model which give the most accurate result The main objective of this project is to predict whether assigning the loan to particular person will be safe or not. This paper is divided into four sections (i)Data Collection (ii) Comparison of machine learning models on collected data (iii) Training of system on most promising model (iv) Testing. In this project we predict the loan data by using some machine learning algorithms they are classification, logic regression, Decision Tree, and gradient boosting.

Proposed System :

• This proposed model will characterize the behavior of customers based on their record. These records are taken from the customers and create a data set. With the help of these data sets and training machine learning model, we predict that the customer’s loan will pass or not.

• The aim of this Paper is to provide a quick, immediate, and easy way to choose the deserving applicants. It can provide special advantages to the bank. The Loan Prediction System can automatically calculate the weight of each feature taking part in loan processing and on new test data same features are processed with respect to their associated weight. A time limit can be set for the applicant to check whether his/her loan can be sanctioned or not.

• Loan Prediction System allows jumping to specific application so that it can be check on priority basis. This Project is exclusively for the managing authority of Bank/finance company, whole process of prediction is done privately no stakeholders would be able to alter the processing.

Advantages Of Proposed System :

• The advantage of this system is that we provided some conditions by setting the algorithms and just by evaluating the details, we get to know eligibility criteria that client is eligible or not.

• The proposed system also scales relatively well to high dimensional data.

• The proposed system is relatively memory efficient.

• The risk of over-fitting is less in our proposed system.

• A small change to the data does not greatly affect the hyper plane

## 

## Methodology :

We have collected the data from Kaggle.

To collect data for the loan approval prediction system project, one can search for relevant datasets on Kaggle using keywords such as "loan approval predictor” and filter the results by relevant criteria such as the data type, size, and source.

In this project we have used Gaussian NB model (short for Gaussian Naïve Bayes) Gaussian Naive Bayes is a probabilistic classification algorithm that is part of the broader Naive Bayes family. It is particularly useful for classification tasks, where the goal is to predict the class of an observation based on its features. The "Naive" in Naive Bayes refers to the assumption of independence between features, which simplifies the underlying probability calculations.

The methodology of Gaussian NB involves several steps:

1) For Gaussian Naive Bayes, estimate the mean and standard deviation of each feature for each class in the training set. These parameters will be used to model the Gaussian distribution.

2) Gaussian NB uses several techniques to improve efficiency and reduce overfitting, such as classification-based algorithms for faster training. This makes it faster and more accurate than other classification methods.

Overall, Gaussian NB is a powerful and efficient machine learning algorithm that can be used for regression and classification tasks. Its methodology of training data and combining their predictions can be effective for many different types of data.

Discussion:

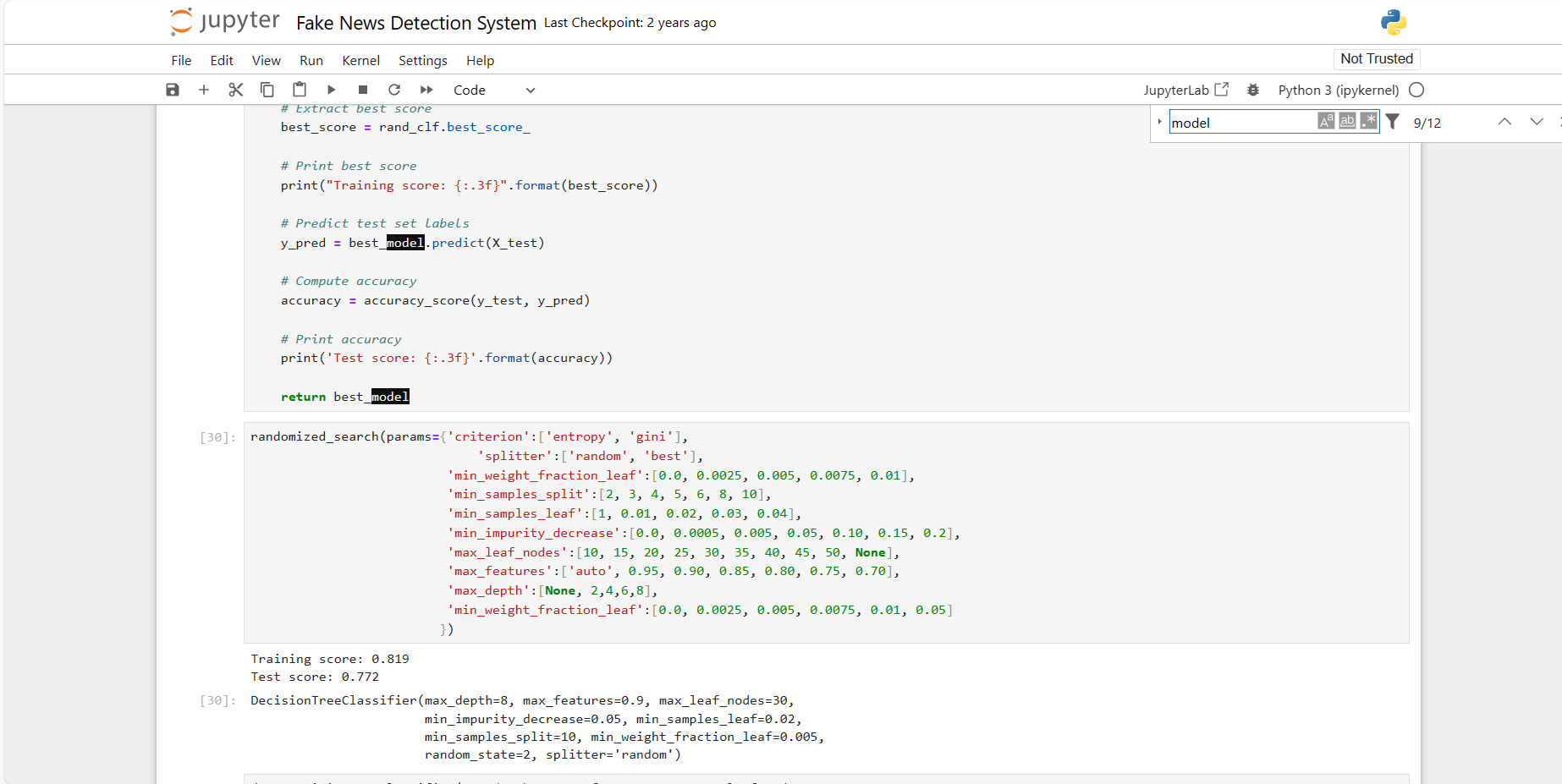
Predicting house prices is a common problem in real estate and finance. There are many factors that can affect house prices, such as location, size, number of bedrooms and bathrooms, and the overall condition of the property. Advanced regression techniques can be used to model the relationship between these factors and the price of a house, and to make predictions about future prices.

One approach for loan approval prediction system is to use Gaussian NB, It is a gradient boosting framework that uses decision trees as the base learner. It is designed to be efficient and to reduce overfitting. Using Gaussian NB for predicting house prices can be an effective approach. The methodology of Gaussian NB involves training decision trees on the data and then combining their predictions to make a final prediction. One advantage of using Gaussian NB for predicting house prices is its efficiency. It uses histogram-based algorithms and leaf-wise growth for decision trees, which can make training faster and more accurate than other gradient boosting methods. This can be especially useful when working with large datasets or when training the model on multiple cores.

## Results Section

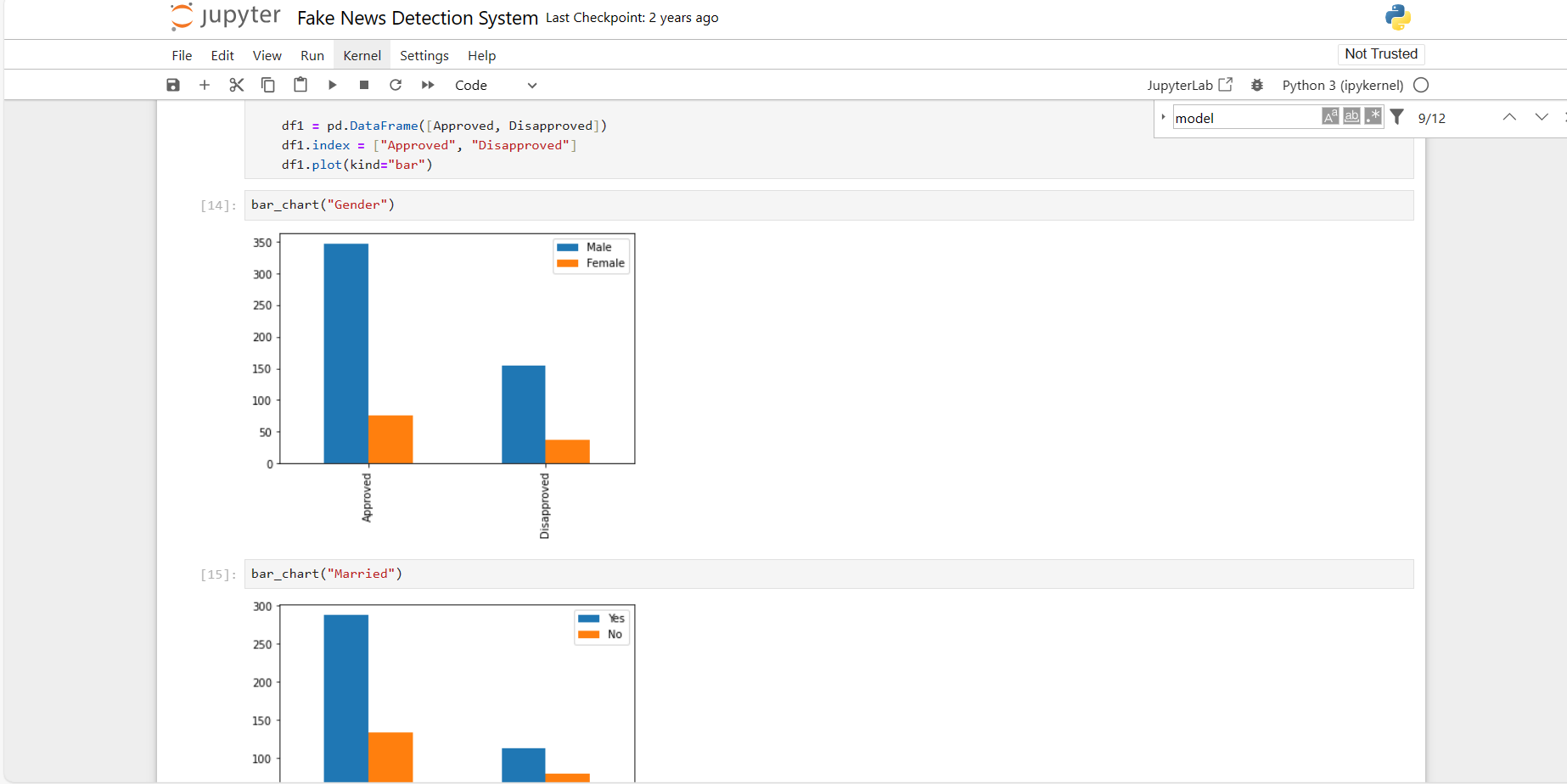
The results of the loan approval prediction system project can be presented in a variety of ways, depending on the specific metrics and evaluation methods used. Some common ways to present the results include:

Model Performance:



**Visualizations:**

The below graph showcases the trend between the columns that has been taken from the data set.

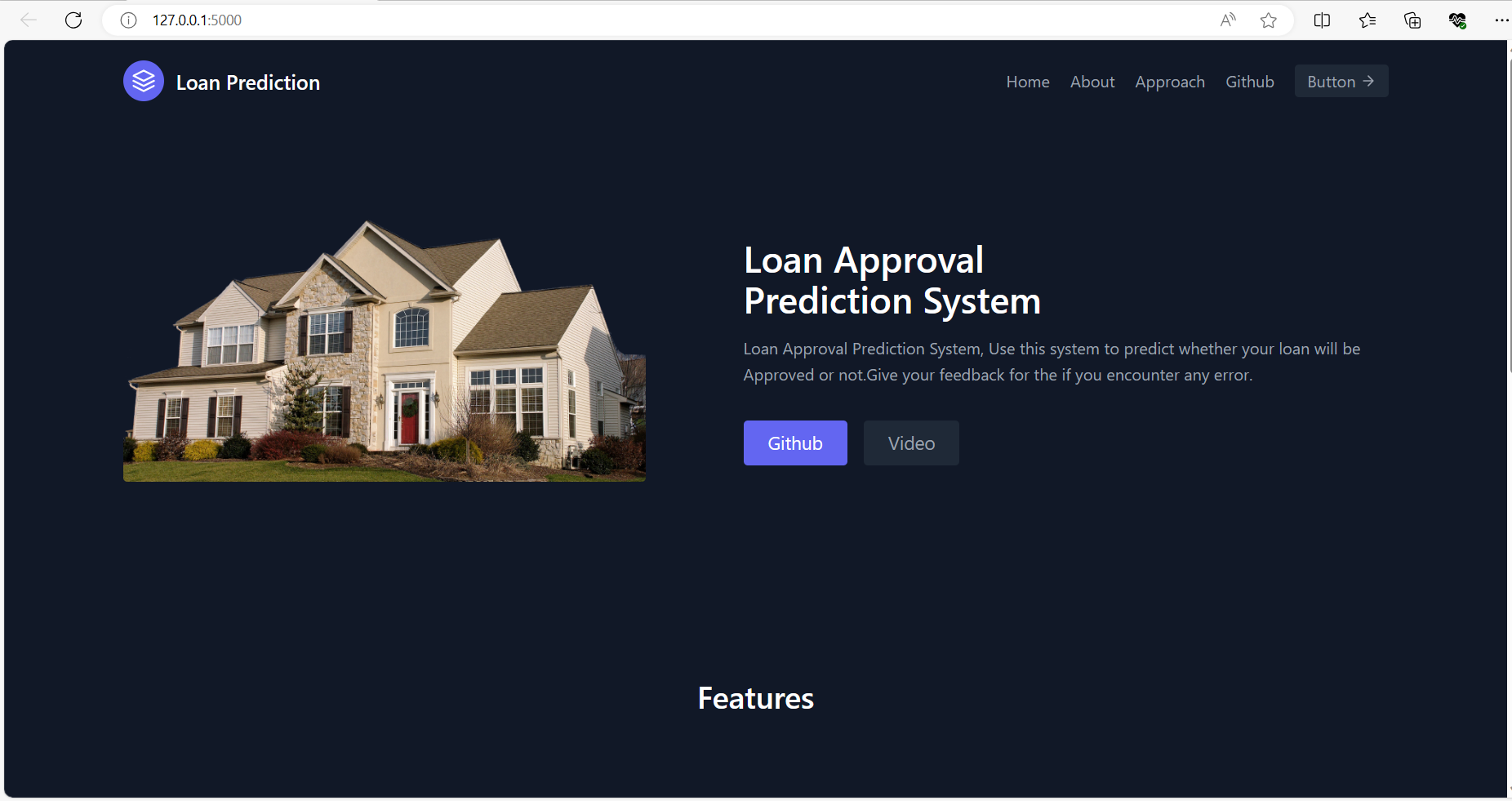


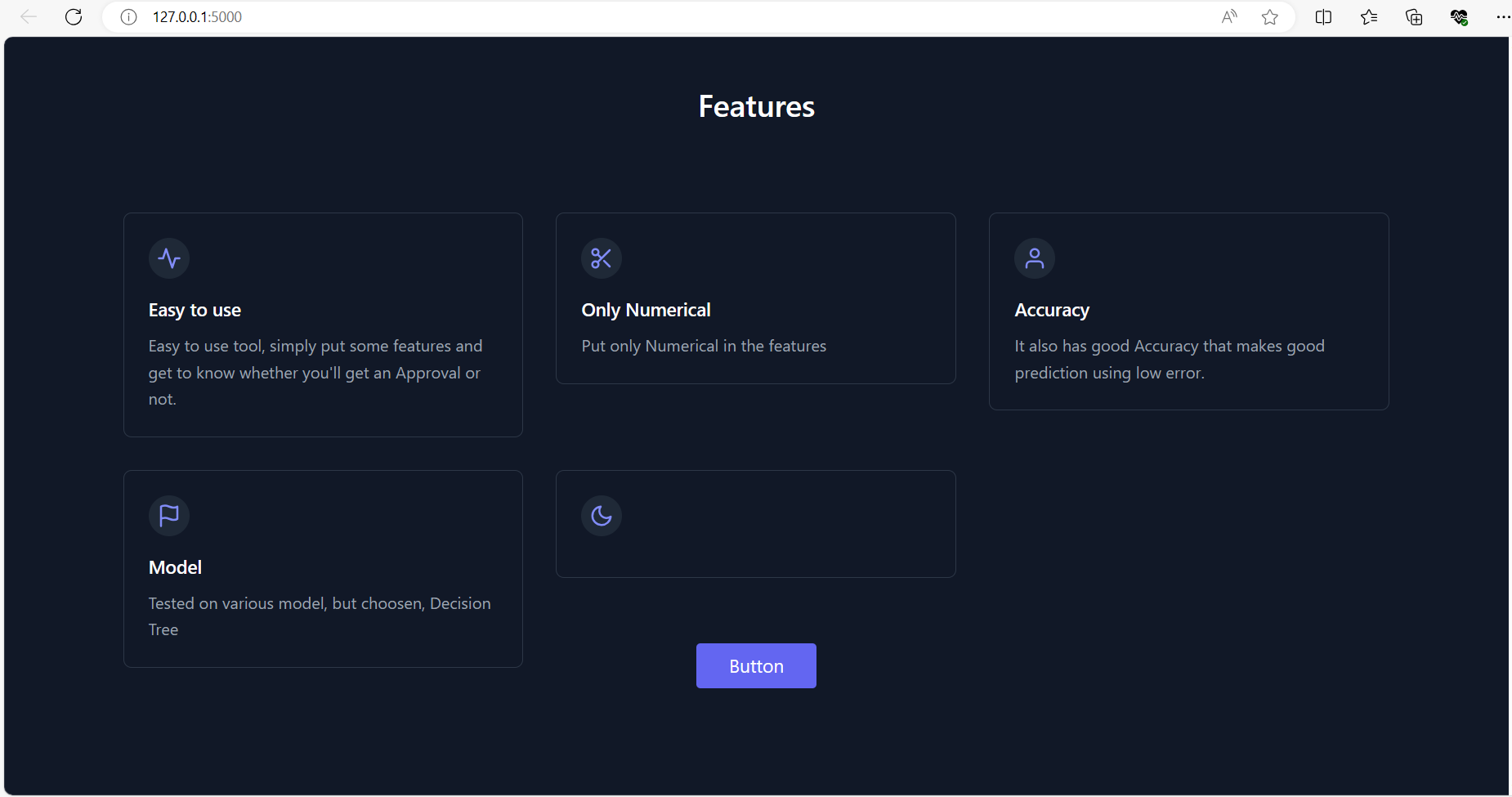
A chart of different colored squares

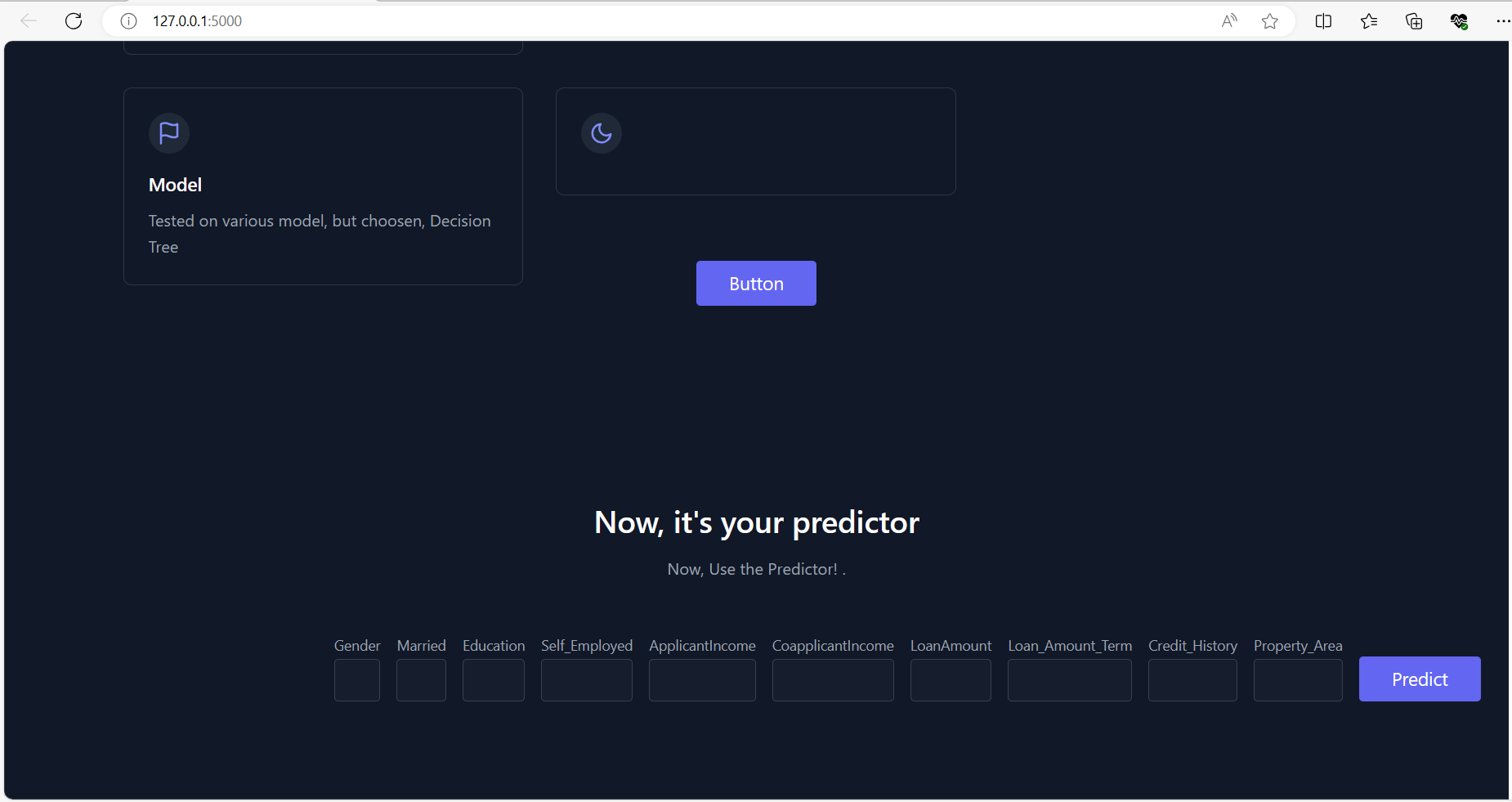
Description automatically generated with medium confidence

The above heatmap shows the correlation between Loan Amount and Applicant Income. It also shows that Credit History has a high impact on Loan Status.

USER INTERFACE/APPLICATION







A house with a garage and a tree

Description automatically generatedA yellow face with black eyes and a sad expression

Description automatically generated

## 

## Conclusion

## In conclusion, the Gaussian NB developed in this study can accurately predict loan approvals using a range of factors. The model is able to accurately capture complex patterns in the data and provide reliable predictions. The model's ability to effectively incorporate complex relationships and regularize the data allows for improved accuracy and robustness in the predictions. This can provide valuable insight for borrowers, helping them make informed decisions and optimize their investment potential.

## Contributions/References

[Loan-Approval-Prediction-Dataset (kaggle.com)](https://www.kaggle.com/datasets/architsharma01/loan-approval-prediction-dataset)