Credit Risk Modeling Using Machine Learning Classifiers

1. Introduction

Credit risk classification is a crucial task in the financial sector to assess the likelihood of a borrower defaulting on a loan. This project utilizes various machine learning models to predict credit risk based on key features such as a person's age, income, home ownership status, employment length, loan amount, and credit history. The objective is to develop a classification model that accurately predicts loan default and compares various models to find the most effective one.

2. Abstract

This project focuses on credit risk classification, an essential task in the financial sector aimed at predicting whether a borrower is likely to default on a loan. The project applies machine learning techniques using various models to predict loan default based on factors such as age, income, home ownership, employment length, loan amount, and credit history. The models are compared based on evaluation metrics, such as accuracy, precision, recall, F1 score, and ROC AUC, to identify the best performing model.

3. Project Objectives

- Develop a machine learning model to predict loan default based on a borrower’s characteristics.
- Compare multiple machine learning models to determine the most effective one for credit risk classification.
- Address challenges such as missing data and class imbalance.
- Improve model performance using hyperparameter tuning and feature engineering.

4. Problem Formulation

The problem is framed as a binary classification task where the goal is to predict whether a borrower will default (1) or not default (0) on a loan. The input features include various attributes such as age, income, home ownership status, employment length, loan amount, loan interest rate, and credit history. Given the imbalanced nature of the dataset (few defaults), the model must be designed to handle class imbalance effectively.

5. Dataset Overview

The dataset used for this project consists of 32,581 entries with the following columns:
- **person_age:** The age of the individual (int).
- **person_income:** The annual income of the person (int).
- **person_home_ownership:** Home ownership status (object).
- **person_emp_length:** Employment length in years (float).

- **loan_intent:** The purpose for which the loan was taken (object).
- **loan_grade:** The grade assigned to the loan (object).
- **loan_amnt:** The amount of the loan (int).
- **loan_int_rate:** The interest rate of the loan (float).
- **loan_status:** The default status of the loan (1: default, 0: no default) (int).
- **loan_percent_income:** Loan amount as a percentage of annual income (float).
- **cb_person_default_on_file:** Whether the person has a history of default (object).
- **cb_person_cred_hist_length:** Length of the credit history (int).

Missing values in the dataset were handled through imputation techniques for some columns (e.g., loan interest rate) while others like employment length had missing values filled with the median value.

6. Machine Learning Models

Various machine learning models were implemented to assess their performance for this classification problem:

1. Support Vector Machine (SVM)
2. K-Nearest Neighbors (KNN)
3. Decision Tree
4. Logistic Regression
5. Random Forest
6. AdaBoost
7. Bagging Classifier
8. Gradient Boosting
9. XGBoost
10. LightGBM

These models were evaluated based on metrics such as accuracy, precision, recall, F1 score, and ROC AUC. The results were compared to determine the most effective model for predicting credit risk.

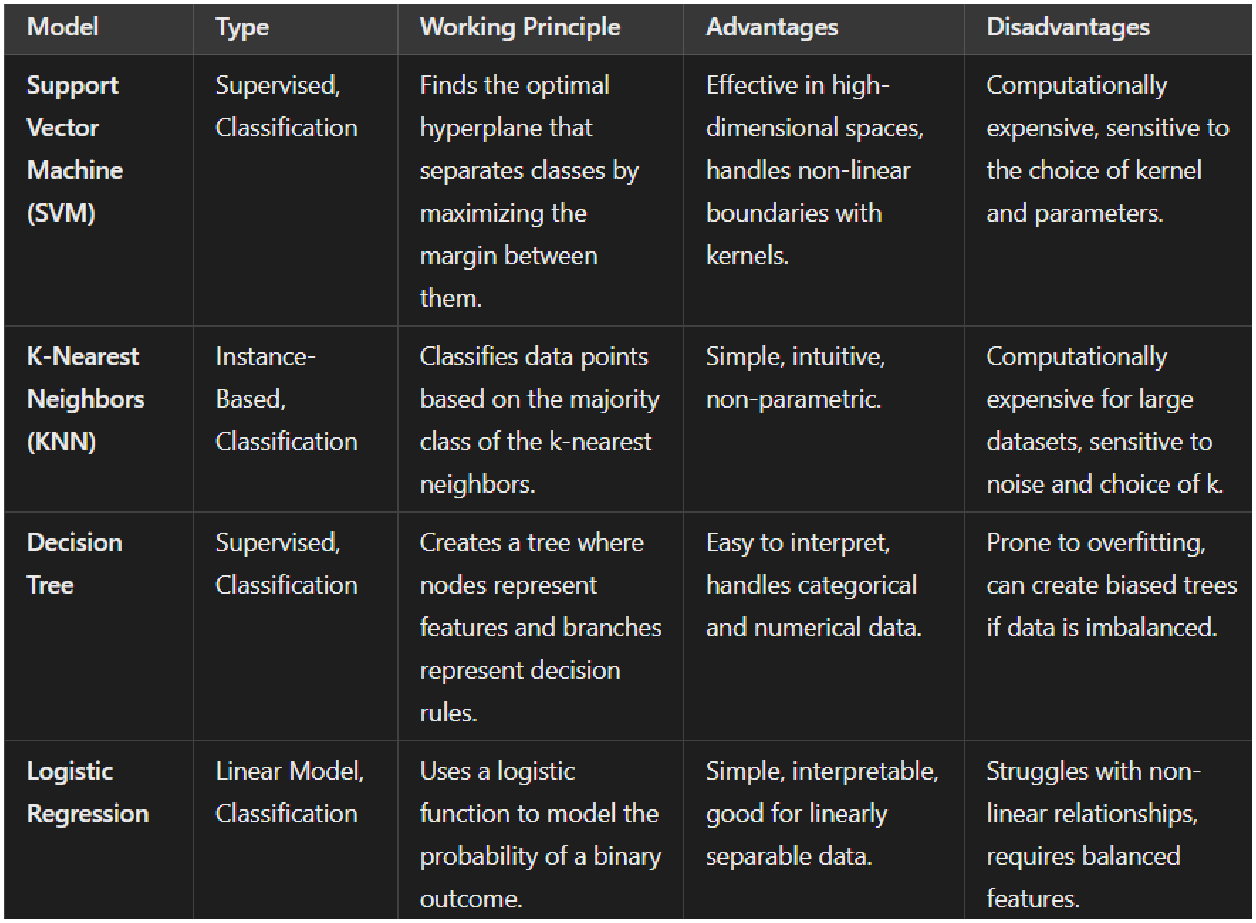
7. Model Performance Evaluation

Each model was trained using the training data (X_train, y_train) and tested on X_test, y_test. The performance metrics are outlined below.

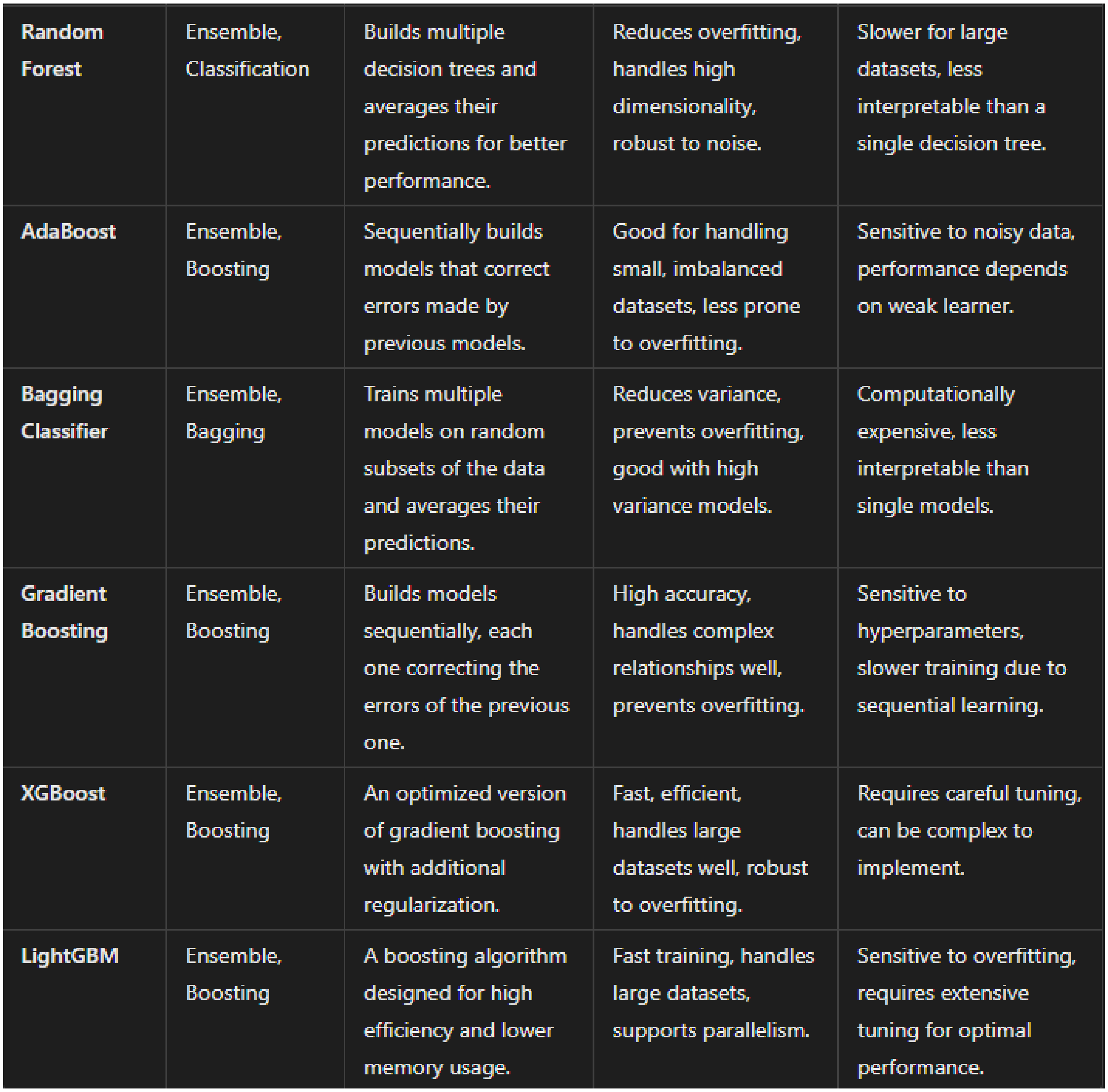
Evaluation Metrics:

- **Accuracy:** The percentage of correctly predicted instances.
- **Precision:** The proportion of positive predictions that were correct.
- **Recall:** The proportion of actual positives that were identified correctly.
- **F1 Score:** The harmonic mean of precision and recall.
- **ROC AUC:** The area under the receiver operating characteristic curve.

8. Comparison of Machine Learning Models



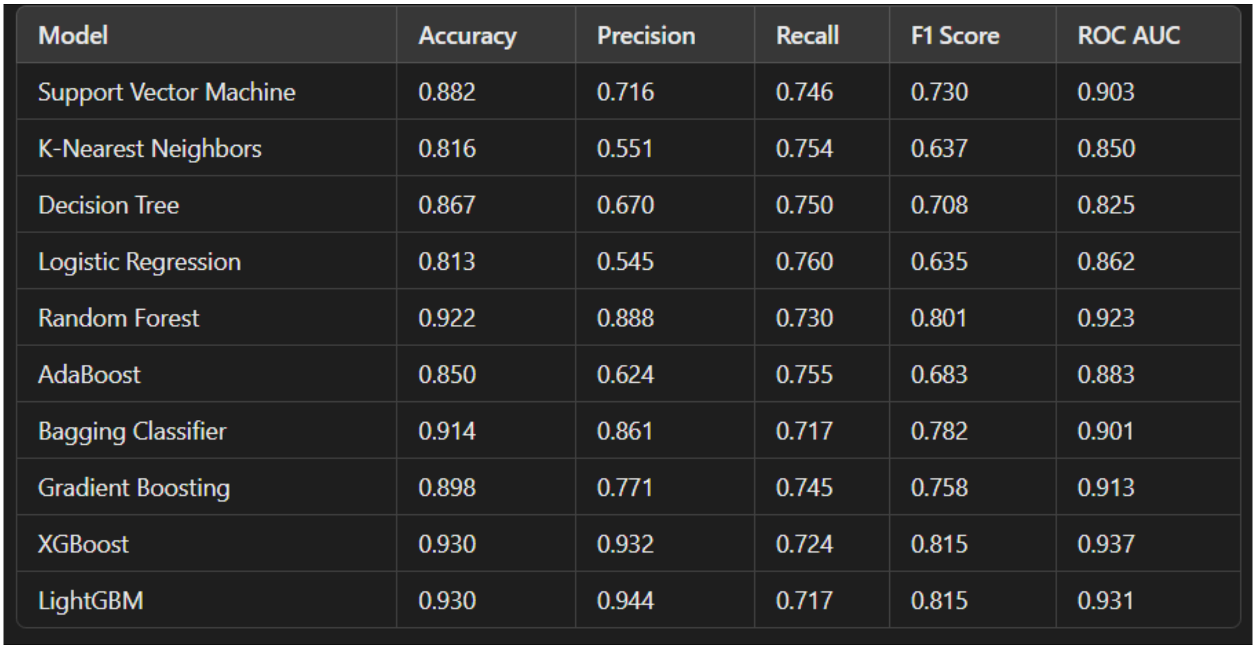
Model	Type	Working Principle	Advantages	Disadvantages
Support Vector Machine (SVM)	Supervised, Classification	Finds the optimal hyperplane that separates classes by maximizing the margin between them.	Effective in high-dimensional spaces, handles non-linear boundaries with kernels.	Computationally expensive, sensitive to the choice of kernel and parameters.
K-Nearest Neighbors (KNN)	Instance-Based, Classification	Classifies data points based on the majority class of the k-nearest neighbors.	Simple, intuitive, non-parametric.	Computationally expensive for large datasets, sensitive to noise and choice of k.
Decision Tree	Supervised, Classification	Creates a tree where nodes represent features and branches represent decision rules.	Easy to interpret, handles categorical and numerical data.	Prone to overfitting, can create biased trees if data is imbalanced.
Logistic Regression	Linear Model, Classification	Uses a logistic function to model the probability of a binary outcome.	Simple, interpretable, good for linearly separable data.	Struggles with non-linear relationships, requires balanced features.



Random Forest	Ensemble, Classification	Builds multiple decision trees and averages their predictions for better performance.	Reduces overfitting, handles high dimensionality, robust to noise.	Slower for large datasets, less interpretable than a single decision tree.
AdaBoost	Ensemble, Boosting	Sequentially builds models that correct errors made by previous models.	Good for handling small, imbalanced datasets, less prone to overfitting.	Sensitive to noisy data, performance depends on weak learner.
Bagging Classifier	Ensemble, Bagging	Trains multiple models on random subsets of the data and averages their predictions.	Reduces variance, prevents overfitting, good with high variance models.	Computationally expensive, less interpretable than single models.
Gradient Boosting	Ensemble, Boosting	Builds models sequentially, each one correcting the errors of the previous one.	High accuracy, handles complex relationships well, prevents overfitting.	Sensitive to hyperparameters, slower training due to sequential learning.
XGBoost	Ensemble, Boosting	An optimized version of gradient boosting with additional regularization.	Fast, efficient, handles large datasets well, robust to overfitting.	Requires careful tuning, can be complex to implement.
LightGBM	Ensemble, Boosting	A boosting algorithm designed for high efficiency and lower memory usage.	Fast training, handles large datasets, supports parallelism.	Sensitive to overfitting, requires extensive tuning for optimal performance.

9. Results and Insights

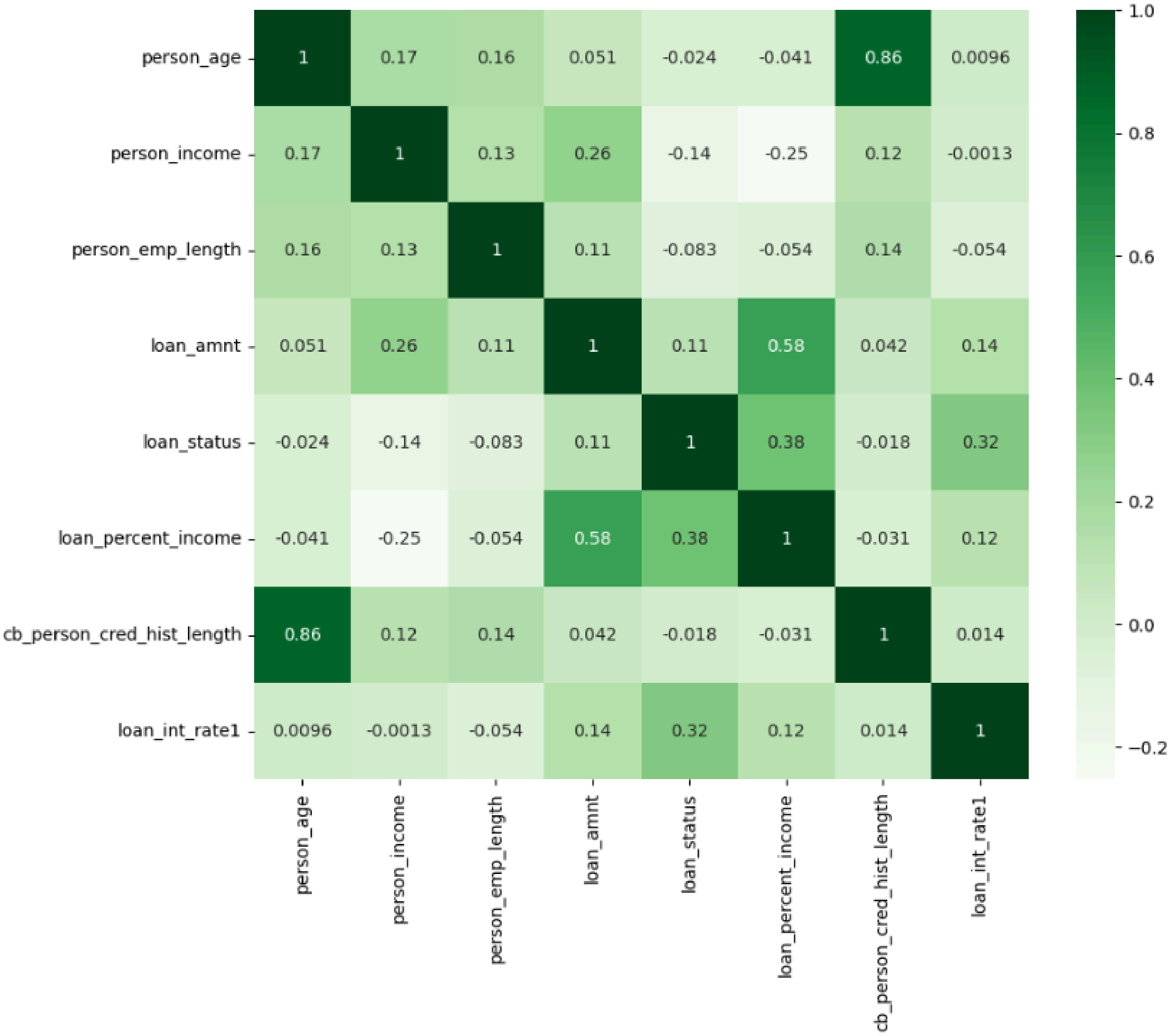
The models showed varying performances with regard to different evaluation metrics. Overall, the XGBoost and LightGBM models provided the highest accuracy, with values of 93.0%, and the ROC AUC scores of 0.937 and 0.931 respectively, indicating a strong ability to distinguish between defaulters and non-defaulters.



Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
Support Vector Machine	0.882	0.716	0.746	0.730	0.903
K-Nearest Neighbors	0.816	0.551	0.754	0.637	0.850
Decision Tree	0.867	0.670	0.750	0.708	0.825
Logistic Regression	0.813	0.545	0.760	0.635	0.862
Random Forest	0.922	0.888	0.730	0.801	0.923
AdaBoost	0.850	0.624	0.755	0.683	0.883
Bagging Classifier	0.914	0.861	0.717	0.782	0.901
Gradient Boosting	0.898	0.771	0.745	0.758	0.913
XGBoost	0.930	0.932	0.724	0.815	0.937
LightGBM	0.930	0.944	0.717	0.815	0.931

- Support Vector Machine performed well in ROC AUC but had lower precision compared to ensemble methods.
- Random Forest and Gradient Boosting both offered strong accuracy and balanced metrics, making them suitable for robust classification tasks.
- Logistic Regression and K-Nearest Neighbors provided lower precision and F1 scores, showing that they were less effective in this scenario.

Correlation Matrix



10.Learning

Through this credit risk classification project, I have gained several important insights related to machine learning algorithms, data preprocessing, and model evaluation:

- **Understanding of Classification Algorithms:**
  We explored a variety of classification algorithms, from simple models like Logistic Regression to complex ensemble methods like Random Forest and Gradient Boosting. Each algorithm provided a unique approach to handling the credit risk data, offering varying degrees of accuracy, interpretability, and computational efficiency. The comparison of these models taught us that there is no one-size-fits-all solution in machine learning; model selection depends on the trade-offs between accuracy, performance, and complexity.

- **Handling Missing Data:**
  Our dataset contained missing values, particularly in features like `person_emp_length` and `loan_int_rate`. We learned how to handle missing data through imputation techniques to ensure that the models could be trained without losing valuable information. This step was critical in maintaining the integrity of the dataset and ensuring that the results were reliable.
- **Feature Selection and Engineering:**
  We learned the importance of selecting the right features and creating new ones to improve model performance. For instance, features like `loan_percent_income` provided valuable insight into a person's financial risk relative to their income. Additionally, categorical features such as `person_home_ownership` and `loan_intent` were encoded properly to be utilized by machine learning models.
- **Performance Metrics and Model Evaluation:**
  We evaluated models using various performance metrics, including Accuracy, Precision, Recall, F1 Score, and ROC-AUC. This allowed us to assess each model's strengths and weaknesses. For example, while XGBoost and LightGBM provided the highest accuracy, Logistic Regression and Decision Trees were more interpretable. This comparison emphasized that different models excel in different contexts, and evaluation should be based on the project's specific objectives.
- **Imbalanced Data Consideration:**
  Credit risk classification often involves imbalanced datasets where defaults are relatively rare. By focusing on metrics like Precision, Recall, and ROC-AUC rather than just Accuracy, we learned to handle class imbalance more effectively. This approach ensured that our models did not disproportionately favor the majority class at the expense of correctly identifying risky customers.
- **Practical Application in Credit Risk:**
  The project provided us with a practical understanding of how machine learning can be applied to real-world problems like credit risk assessment. We learned how predictive models can help financial institutions make better lending decisions, reduce risk, and ensure the stability of their loan portfolios.

11. Conclusion

In this project, multiple machine learning models were compared to classify credit risk. The ensemble-based methods, specifically XGBoost and LightGBM, demonstrated superior performance in both accuracy and ROC AUC, making them the recommended models for this task. Future work can include hyperparameter tuning and feature engineering to further improve the performance of these models.

12.Github link

https://github.com/sridhar98765/ml-project