## A Project Report on

Investment-Recommendation

Submitted by

Girish Anil Amrutkar

**Abstract**

This project aims to develop an Investment Recommendation System using machine learning. Using a synthetic dataset simulating real-world financial scenarios, I created a system to provide personalized investment recommendations based on user-specific financial profiles. The project highlights the importance of user-specific details for relevant investment advice and addresses data quality and model interpretability challenges.

In the existing project , I have implemented the both type of the machine learning model (i.e. Regression and Classification) .I have implemented total eight model for the investment dataset are linear regression , Decision tree regressor ,random forest regressor ,Gradient boosting regressor ,Logistic regression ,Decision tree classifier , random forest Classifier ,Gradient Boosting Classifier .

**Table of Content**

1. Data Collection…………………………………………………………………………
2. Data Preprocessing……………………………………………………………………
3. Exploratory Data Analysis (EDA)…………………………………………………
4. Model Development …………………………………………………………………
5. Testing and Validation………………………………………………………………..
6. Prediction Insights ……………………………………………………………………
7. Prediction through user input …………………………………………………..
8. Conclusion ………………………………………………………………………………..

1.Data Collection

**1.1.Data Source**

The dataset was obtained from a synthetic data generator designed to mimic real- world investment-related data. The synthetic data contains user demographics and financial details.

**1.2.Data Type**

In the respective dataset ,it has the attributes Age ,Family Size , Income ,Monthly expenses ,Existing investments and assets, Debt obligations ,Financial goals, Risk tolerance level as independent Variable along with the Dependent variable Predicted Investment Return and Investment Category. It may have some missing values as it deals with the investment data and the future planes .So, it may have the missing values .For that purpose , it requires data preprocessing techniques .

**2.Data Preprocessing**

**2.1.Data Cleaning and Handling Missing Values**

As we have an idea about the missing values but ML model requires the clean dataset for the implementation of models ,there should not have a single empty value .So ,I have analyzed the missing value along with their count in each column and replace it with the mean of the respective column so that the dataset should not have the single missing value .

**2.2.Encoding Categorical Values**

Encoding is used to convert the categorical values in the numerical form for the better output through the ML model .Here in this case , we have converted the categorical values of ‘Financial Goal’ and ‘Risk Tolerance Level’ into the numerical values .

**2.3.Normalization**

Normalizing numerical features deals with the scaling the values of the numerical data so that they will fall under the specific range .In this case , I have set the range for that is 0 to 1 .In our dataset there are certain attributes like age , family size ,income ,monthly expenses and current investment and it is better to convert them in the specific range

**2.4.Handling the target Variables**

Handling target variables involves preparing the target variable (or variables) for machine learning modeling. This preparation can include encoding, transforming, or normalizing the target variable to ensure it is in a suitable format for the chosen algorithms. In this case , the dataset has gone through the Normalization for the predicted investment return and encoding for the investment category which is useful for the ML model .

In short , the data preprocessing technique involves the making the data suitable to implement the any ML model on that so that the accuracy and impact of the model along with the prediction or classification should be correct and accurate

**3.Explorative Data Analysis (EDA)**

**3.1.Distribution Of Numerical Features**

To understand the numerical features I have plotted the histogram with its KDE for numerical features to visualize their distributions. It Identifies the spread and central tendency of features, spotting any skewness or outliers.

**3.2.Relationship Between the Numerical features and investment Return.**

It’s better to know the relationship between the numerical feature and target variable for that purpose I have plotted the scatter plots to visualize relationships between numerical features and investment return. It reveals potential correlations and trends between features and the target variable.

**3.3. Count of Categorical Features**

To deal and know the categorical feature I used count plots for categorical features to show their frequency distributions. The main thing about it is that it highlights the distribution and balance of categorical variables.

**3.4.Corelation Heatmap**

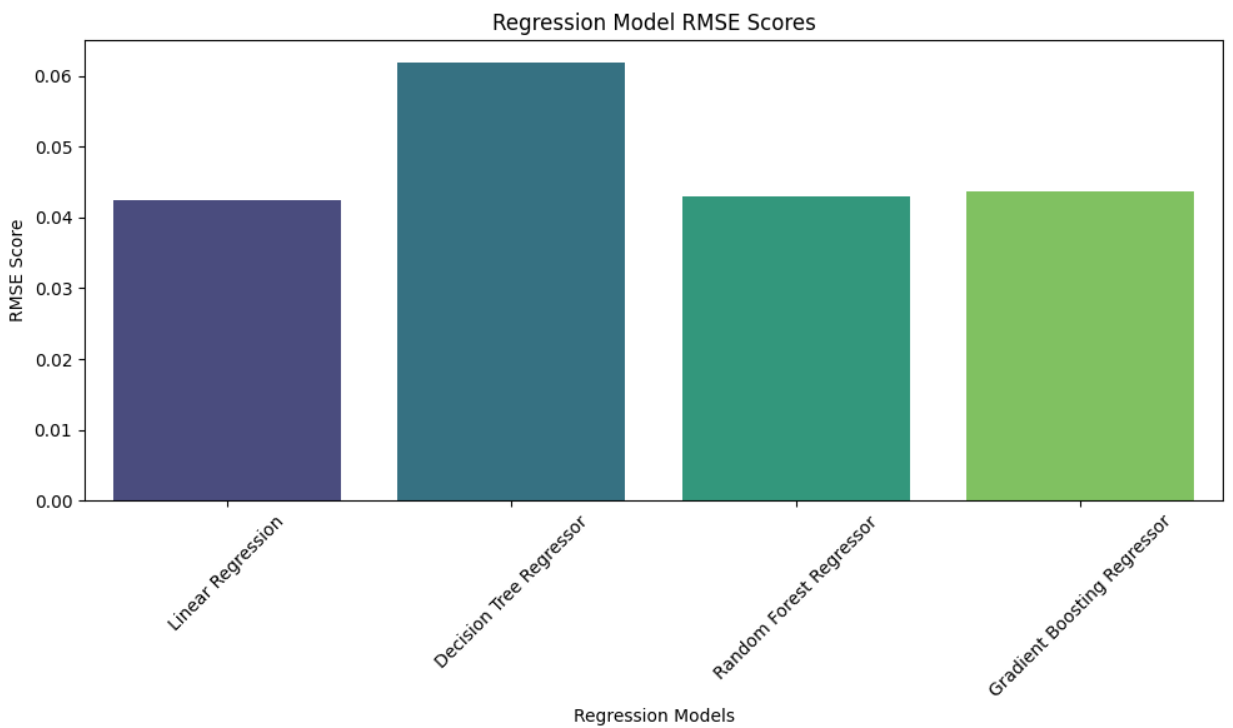
It displays a heatmap of correlations between numerical features. It also Identifies strong and weak correlations among features, which is crucial for feature selection and understanding relationships. In short it helps to find the relationship between the attributes .

**4.Model development**

The model development is the most important part of the developing the recommendation system. For that purpose I have divided the dataset in the two sets one is training set and other is testing dataset. Here in this case I have set the test size as 0.2 means 20% of the dataset is used for the testing dataset and remaining is as training dataset ( 80%) .In the dataset I have 2000 rows among that 1600 rows will be used for the training dataset and remaining 400 rows will be for testing .

After the splitting of the dataset , I have implemented the both types of the machine learning models ( regression and Classification model) with their accuracy parameter like RMSE ,accuracy factor .The regression Ml model which I have implemented are listed below with their RMSE value

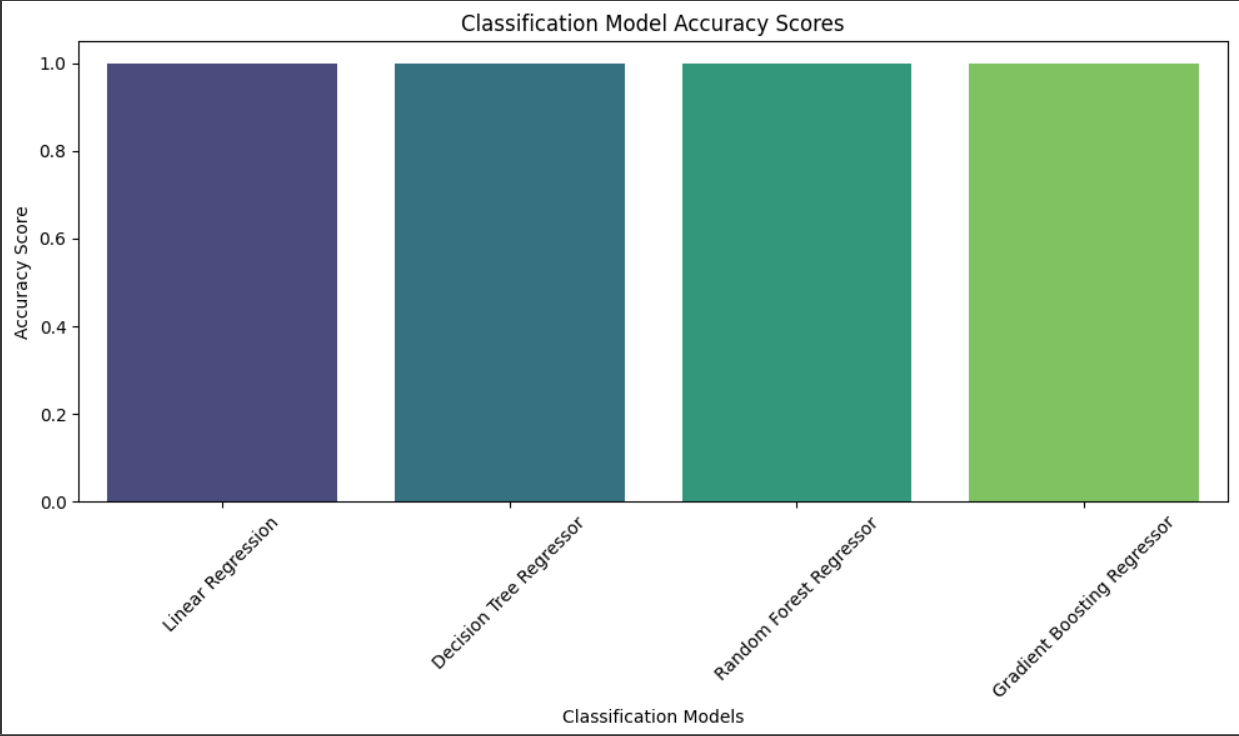
1. Linear Regression : 0.0424911306929846
2. Decision Tree Regressor: 0.06193813816138068
3. Random Forest Regressor: 0.042905866605952685
4. Gradient Boosting Regressor: 0.0437670025602668



The Classification model that I have implemented are listed belowas follows:

1. Logistic Regression: 1.0
2. Decision Tree Classifier : 1.0
3. Random Forest Classifier: 1.0
4. Gradient Boosting Classifier : 1.0

In case of the classification model I have got the perfect accuracy for all four model means , the prediction of all four models are true about the investment type .So , we can only focus on the prediction of the investment return as above we discussed about the regression models .



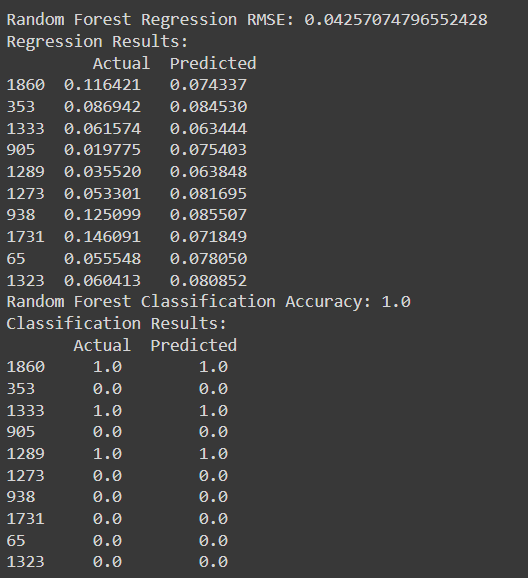
**5.Testing and Validation**

In this section again I gone through the training of the ML models on the training dataset only as the accuracy of the regression models are not as expected .for that purpose I have used the grid search CV to fine tune the model in which I got the Cross validation score as ‘1’ which actually increases the accuracy of the of the regression models.

But , the accuracy of the regression models as compared to the classification models are less but the predicted values are some how are near .So ,we can say the accuracy is near about 83% and for the classification it is 100% .

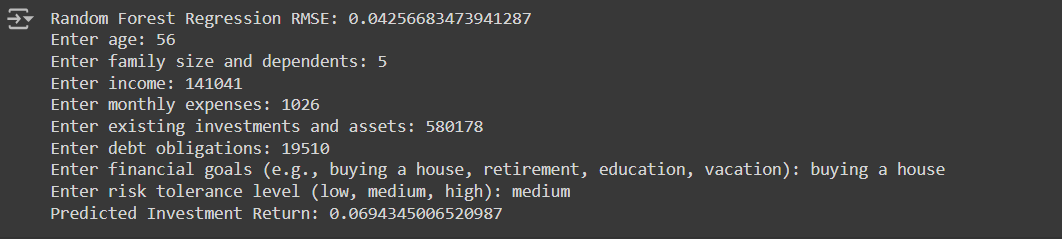
1. **Prediction Insights**

The prediction insights includes the Accuracy ,RMES score of the Regression and classification models along with the actual value and predicted value .It also includes the comparison between the actual value and predicted value as follows :-



**7.Prediction through user Input**

Apart from the terms like accuracy , RMSE score and predicted value ,the model also works on the user input also .To predict the investment returns we take the input from the users like age , family size and dependents, income ,monthly expenses ,existing investments and assets ,debt obligations , financial goals (e.g., buying a house, retirement, education, vacation), risk tolerance level (low, medium, high).On that basis ,the prediction of the investment return can be done .



**8.Conclusion**

Here We have implemented the Investment prediction model for investment returns on that basis a user can decide whether to invest in a stock/bond/mutual funds/real estate/stocks by analysing the return value. Higher the return value ,it is good to invest in that and if it has a lower value the user should think twice before investing in that .

Here ,I have used both the regression type of the model for the investment return prediction and classification type of the model to analyse the investment category .The models which I have implemented are listed below :

1)Linear Regression

2)Decision Tree Regressor

3)Random Forest Regressor

4)Gradient Boosting Regressor

5)Logistic Regression

6)Decision Tree Classifier

7)Random Forest Classifier

8)Gradient Boosting Classifier