Data Science Project Report: Predicting Students' Final Grades

Problem Statement

The objective of this project is to develop a predictive model that estimates students' final grades based on various factors, including attendance, study hours, and socio-economic factors. Understanding the relationship between these variables and academic performance can help educators identify at-risk students and implement targeted interventions to improve educational outcomes.

Importance of the Study

1. Educational Insights: By analyzing how attendance, study habits, and socio-economic background influence academic performance, educators can gain valuable insights into student behavior and learning patterns.
2. Intervention Strategies: Identifying key predictors of student success allows schools to design effective intervention strategies aimed at improving attendance and study habits among students who may be struggling.

Dataset Features

The dataset used for this analysis will include the following features:

1. Attendance Rate: The percentage of classes attended by each student (e.g., 0% to 100%).
2. Study Hours: The average number of hours per week that a student dedicates to studying outside of class.
3. Socio-Economic Factors:
    - Parental Income Level: Categorical variable indicating income brackets (e.g., low, medium, high).
    - Parental Education Level: Categorical variable indicating the highest level of education attained by parents (e.g., high school, bachelor's degree, master's degree).
    - Access to Educational Resources: A variable indicating whether the student has access to resources such as internet.

Target Variable

The target variable for this project will be:

- Final Grade: The final grade achieved by each student in their course or program, typically represented as a percentage or letter grade (e.g., A-F)

We'll explore two different models — Linear Regression and Random Forest — to find the best approach for this task.

Exploratory Data Analysis

Before we dive into modeling, it's crucial to gain a comprehensive understanding of the data. Exploratory Data Analysis (EDA) serves as our compass, providing insights into the dataset's characteristics. This process includes:

- **data.info():** Provides a concise summary of the dataset's information, including the total number of non-null entries for each column, data types, and memory usage.
- **data.describe():** Generates summary statistics of the numerical columns, such as count, mean, standard deviation, and percentiles. These statistics offer insights into the distribution and variability of the data.

Performing EDA allows us to make informed decisions about feature selection and preprocessing, ensuring that our modeling approach is based on a solid understanding of the dataset.

Feature Selection

To build an effective predictive model, we need to choose the most relevant features. In this project, we select

'Attendance (%)', 'Midterm_Score', 'Final_Score',

'Assignments_Avg', 'Quizzes_Avg', 'Participation_Score',

'Projects_Score', 'Total_Score', , 'Study_Hours_per_Week',

'Internet_Access_at_Home',

'Parent_Education_Level', 'Family_Income_Level'.

These features are expected to have a significant impact on the final grade .

Model Building :Linear Regression

Our first approach is to employ a Linear Regression model. This model assumes a linear relationship between the selected features and the target variable (final grade). We train the model on a portion of the dataset and evaluate its performance using, Mean Absolute Error (MAE)

Linear regression operates on the assumption that the relationship between the features and target variable is approximately linear. It calculates a line of best fit that minimizes the difference between predicted and actual values. This line serves as our predictive model, enabling us to make accurate grade predictions.

2. Random Forest

For potential improvement, we introduce a Random Forest model. Unlike Linear Regression, Random Forest can capture non-linear relationships in the data and handle complex

interactions between features. We train the Random Forest model and evaluate its performance using the same set of metrics. This will allow us to compare the Random Forest model's predictive power with that of the Linear Regression model.

The Random Forest model is an ensemble of decision trees. It combines multiple decision trees to make more accurate predictions. Each tree considers a random subset of features and collectively contributes to the final prediction. This allows the model to capture non-linear relationships and interactions between features.

Validation Metrics

In both models, we use a set of validation metrics to assess their performance:The values of these metrices are shown in fig 2 below.

- **Mean Absolute Error (**MAE): Provides the average absolute difference between predicted and actual final grades.

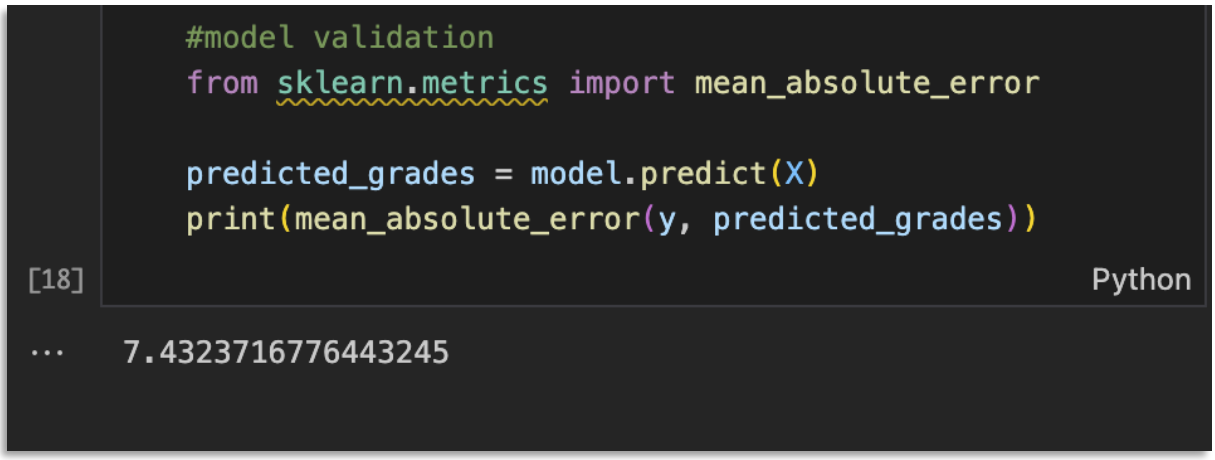

Fig1: Value of MAE

- A MAE of 7.432 suggests that, on average, the model's predictions deviate from the actual student performance scores by approximately 7.43 points.
- **Accuracy** is defined as the ratio of correctly predicted instances to the total instances in the dataset. It is calculated using the formula: An accuracy score indicates how often the model makes correct predictions. For example, an accuracy of 83 % means that 85 out of 100 predictions were correct.
- **Precision** measures the proportion of true positive predictions among all positive predictions made by the model: A precision score of 0.83 means that when the model predicts a student will perform well, it is correct 83% of the time.

- **Recall**, also known as sensitivity or true positive rate, measures the proportion of actual positives that were correctly identified by the model: A recall score of 1 indicates perfect recall; every student who actually performed well was correctly identified by the model.

- **Importance**: High recall is crucial in educational settings because failing to identify students who need help can have significant negative consequences on their academic success.

- The **F1 Score** combines both precision and recall into a single metric by calculating their harmonic mean: An F1 score of 0.91 indicates a strong balance between precision and recall; this suggests that while many students are accurately predicted as performing well (high precision), almost all students who actually performed well are also being identified (high recall).

```
Accuracy: 0.83
Precision: 0.83
Recall: 1.00
F1 Score: 0.91
```

Fig2:*Values for evaluation metrices.*

Conclusion

In this project, we've tackled the task of predicting student grades using two different models — Linear Regression and Random Forest. By selecting relevant features, training the models, and evaluating their performance, we gain valuable insights into their predictive power. This project exemplifies the power of predictive modeling in education, offering valuable insights for educators .