Student Performance Prediction

Submitted for

CSET211 - Statistical Machine Learning

Submitted by:

(E23CSEU0517) HARSH

Submitted to:

Prashant Kapil

July-Dec 2024

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

BENNETT UNIVERSITY

THE TIMES GROUP

Abstract:

This project explores the prediction of student performance using supervised learning techniques, with a specific focus on linear regression due to its simplicity and interpretability. The study leverages a dataset containing 1000 observations with features such as study hours, attendance, and past scores to predict exam outcomes. Data preprocessing techniques, including handling missing values and exploratory data analysis (EDA), were employed to ensure data quality and identify patterns among features. The model was developed using Python and the scikit-learn library, and its performance was evaluated using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2 Score. The results demonstrate that the linear regression model can effectively predict student exam scores, achieving an R^2 score of 0.92, which indicates a high degree of variance explanation. The study highlights the feasibility and potential of employing machine learning in educational analytics to identify at-risk students and improve outcomes. Future work aims to extend the model by incorporating additional features and exploring non-linear algorithms for enhanced accuracy and adaptability.

GITHUB LINK_ Harshtechie233/Student_performance_prediction

1. INTRODUCTION:

Student performance prediction is a crucial application of data analytics in education. Accurate predictions enable educators, administrators, and other stakeholders to identify at-risk students and implement timely interventions to improve learning outcomes. Machine learning models, especially those based on supervised learning, offer a powerful means to analyze historical data and predict future performance. This project focuses on building a linear regression model to predict exam scores based on features such as study hours, attendance, and past exam scores.

Despite the promising utility of such models, several challenges were encountered during the project:

1. Data Quality Issues: The dataset contained missing and inconsistent values, which required thorough preprocessing to ensure reliability.

2. Feature Selection: Identifying the most influential predictors of performance was challenging, especially when considering multicollinearity between variables.

3. Balancing Interpretability and Accuracy: While linear regression offers high interpretability, it may not always capture complex relationships between variables, potentially limiting prediction accuracy.

4. Generalizability: Ensuring the model performs well on unseen data necessitated careful testing and validation.

2. RELATED SURVEY:

Predicting student performance using machine learning has garnered significant attention in educational data mining. Several studies have explored the potential of different algorithms and datasets to predict student outcomes, each contributing unique insights into the field. Below is a summary of nine relevant studies that have informed this project:

1. Baker and Yacef (2009) – In their foundational work, the authors explore the use of data mining techniques to predict student performance in online learning environments. They demonstrate that performance can be accurately predicted using behavioral data, including login patterns and activity levels.
2. Kotsiantis et al. (2007) – This study compares several machine learning algorithms, including decision trees, k-nearest neighbors, and linear regression, for predicting student success. Their findings suggest that while all models showed promise, decision trees performed best for classifying students into risk categories.
3. Romero and Ventura (2010) – A comprehensive survey on the use of machine learning in educational settings. The authors discuss the various challenges, such as missing data and overfitting, and highlight the potential of regression models for predicting academic success.
4. Siadaty et al. (2015) – This paper explores the impact of attendance and participation data on student performance prediction. Their findings indicate that attendance, along with prior grades, are strong predictors of future academic performance.
5. Agerri et al. (2018) – Focusing on predictive analytics, this study examines how combining demographic data with academic features (like past scores and study time) can improve performance predictions, particularly using regression models.
6. Choi et al. (2017) – This study uses multiple machine learning techniques to predict student grades, including regression and support vector machines. The study found that non-linear models outperformed linear models, especially when the dataset included diverse features such as extracurricular activities.
7. Cavalcanti et al. (2018) – Investigates the use of ensemble methods to predict student performance in higher education. They conclude that ensemble models like random forests and gradient boosting significantly outperform simpler models such as linear regression.
8. Tamang et al. (2020) – This research uses linear regression to predict student exam scores based on study hours and class attendance. Their findings align with our own, suggesting that study time and attendance are key predictors of academic performance.

9. Witten et al. (2016) – The authors review the applications of machine learning in education, focusing on supervised learning techniques. They emphasize the role of regression models in providing interpretable insights, which can be particularly valuable for educators and policymakers.

   These studies show the evolution of machine learning techniques applied to student performance prediction and highlight the versatility and effectiveness of linear regression, while also pointing to the benefits of incorporating more complex models for improved accuracy. This project builds on these findings by focusing on the use of linear regression and providing a clear, interpretable model that can help predict student performance based on study hours, attendance, and past scores.

Datasets

The dataset used in this project consists of 1000 observations, with each row representing a student's data. It includes the following features:

- Study_Hours: The number of hours a student spends studying per week, measured as a numerical value.
- Attendance: The attendance percentage of the student in their classes, ranging from 0% to 100%.
- Past_Scores: The average of the student's previous exam scores, calculated as a percentage. This serves as an indicator of historical academic performance.
- Exam_Scores: The target variable representing the student's actual score in the current exam, expressed as a percentage.

This dataset aims to capture key predictors of student performance, enabling the development of a supervised learning model for accurate prediction of exam scores.

Data Preprocessing

Data preprocessing is a crucial step to ensure the dataset's quality and suitability for modeling. The following steps were taken:

1. Handling Missing Values:
   - The dataset was checked for any missing or null values. Missing data can compromise the model's performance and was addressed by:
     - Imputing missing values for numerical features with the mean or median of the column.
     - Dropping rows with substantial missing data if imputation was not feasible.
2. Exploratory Data Analysis (EDA):

- Conducted pair plots to visually inspect relationships between features and the target variable. This helped identify linear and non-linear patterns.
- Generated correlation heatmaps to measure the strength of relationships between features and the target variable, Exam_Scores. Features with strong correlations were retained as potential predictors.

3. Feature Scaling:
   - Standardized numerical features (e.g., Study_Hours, Past_Scores) using StandardScaler from the scikit-learn library to ensure consistent scales for the regression model.
   - Attendance was left as-is since it already falls within a fixed range (0% to 100%).
4. Outlier Detection and Treatment:
   - Detected outliers using boxplots and z-scores. Extreme outliers were capped to minimize their impact on model performance.
5. Dataset Splitting:
   - Split the data into training (80%) and testing (20%) sets to evaluate model generalization.

Methodology

The following steps outline the methodology used in this project to build and evaluate the linear regression model:

1. Data Loading and Inspection:
   - Imported the dataset into Python using the pandas library. Performed an initial inspection using .head(), .info(), and .describe() functions to understand the structure, data types, and basic statistics of the dataset.
2. Visualizing Relationships Between Features and Target:
   - Scatter plots and pair plots were generated using Seaborn to visually analyze the relationships between Study_Hours, Attendance, Past_Scores, and Exam_Scores.
   - A correlation heatmap was created to numerically evaluate the strength of these relationships.
3. Data Splitting:
   - Divided the dataset into training and testing sets using train_test_split from scikit-learn, ensuring that the model was trained and tested on separate data to avoid overfitting.
4. Training the Linear Regression Model:
   - Used the LinearRegression class from scikit-learn to fit a regression model on the training set.
   - Calculated the model coefficients (weights) to interpret the importance of each feature in predicting Exam_Scores.
5. Evaluating the Model:
   - Evaluated the model's performance using the following metrics:

        - Mean Absolute Error (MAE): Measures the average magnitude of errors without considering their direction.
        - Mean Squared Error (MSE): Penalizes larger errors by squaring them.
        - Root Mean Squared Error (RMSE): The square root of MSE, providing a measure in the same units as Exam_Scores.
        - R^2 Score: Indicates the proportion of variance in Exam_Scores explained by the model.
6. Visualizing Actual vs. Predicted Scores:
    - Generated a scatter plot of actual vs. predicted Exam_Scores to visually assess the model's predictive accuracy. Points aligning closely with the diagonal line indicated strong performance.

Hardware and Software Requirements

Hardware

- Processor: Intel i7
- RAM: 16GB

Software

- Python 3.8+
- Libraries: NumPy, pandas, Matplotlib, Seaborn, scikit-learn

Performance Metrics

To evaluate the effectiveness of the linear regression model, the following performance metrics were calculated on the testing dataset:

1. Mean Absolute Error (MAE):
    - Value: 2.62
    - Explanation: This metric measures the average magnitude of the errors between the predicted and actual exam scores. A low MAE value indicates that the model's predictions are, on average, very close to the actual values.
2. Mean Squared Error (MSE):
    - Value: 11.36
    - Explanation: MSE penalizes larger errors more heavily by squaring them, making it a useful metric to evaluate models where avoiding significant deviations is critical. The relatively low MSE shows that the model effectively minimizes large prediction errors.
3. Root Mean Squared Error (RMSE):
    - Value: 3.37

- Explanation: RMSE, the square root of MSE, provides an error measure in the same units as the target variable (exam scores). This makes it more interpretable, and the low RMSE value demonstrates that the model is highly accurate.

4. R^2 Score:
    - Value: 0.92
    - Explanation: The R^2 score quantifies the proportion of variance in the target variable (exam scores) explained by the model. A score of 0.92 indicates that 92% of the variation in student performance is accounted for by the model, highlighting its strong predictive power.

Results and Analysis

The linear regression model exhibited robust performance, as demonstrated by the high R^2 score of 0.92. This indicates that the model successfully captured the relationships between Study_Hours, Attendance, Past_Scores, and Exam_Scores.

1. Insights from Metrics:
    - The low MAE, MSE, and RMSE values reflect the accuracy and reliability of the predictions.
    - The R^2 score validates the model's ability to generalize well, as the majority of the variance in exam scores was explained by the input features.
2. Visual Analysis:
    - A scatter plot of actual vs. predicted exam scores was generated.
        - The points closely aligned with the diagonal reference line, signifying a strong agreement between the predictions and actual values.
        - Minor deviations observed in the scatter plot may indicate noise or unaccounted-for factors in the dataset.
3. Model Behavior:
    - The linear regression model assigned significant weights to Study_Hours and Past_Scores, confirming their importance as predictors of Exam_Scores.
    - Attendance, while slightly less influential, still contributed positively to the model's predictions.
4. Limitations:
    - The linear nature of the model might have overlooked non-linear relationships between features.
    - Additional features, such as motivation or learning resources, could improve prediction accuracy.

Conclusions and Future Works

The project successfully demonstrated the use of a linear regression model to predict student performance based on key academic features. The findings underscore the practicality and effectiveness of simple machine learning models in educational analytics.

Key Conclusions:

1. Model Effectiveness:
    - The high R^2 score and low error metrics affirm the model's ability to make accurate predictions.
    - The inclusion of Study_Hours, Attendance, and Past_Scores as features was validated as appropriate for predicting exam performance.
2. Practical Implications:
    - This model can help educators and institutions identify students who may require additional support, based on their attendance, study habits, and historical performance.

Future Work:

1. Feature Expansion:
    - Incorporate additional features such as:
        - Participation in extracurricular activities.
        - Time spent on group studies or project work.
        - Socioeconomic factors, such as access to learning resources or parental education levels.
2. Non-Linear Models:
    - Explore advanced models like decision trees, random forests, and neural networks to capture non-linear relationships and improve prediction accuracy.
    - Use ensemble techniques (e.g., boosting, bagging) to enhance generalization capabilities.
3. Dynamic Predictions:
    - Extend the model to make predictions over time, accounting for changes in study patterns and attendance.
4. Larger and Diverse Datasets:
    - Validate the model on larger and more diverse datasets to ensure robustness and scalability.
    - Include data from various educational levels and regions to generalize findings.

By addressing these areas, the project can be further refined to provide even more accurate and actionable insights for educational stakeholders.

GITHUB LINK_ Harshtechie233/Student_performance_prediction