Predicting Student Exam Scores

Internship Task 1 Submission

Submitted By

Alisha Rafique

Submitted To

Elevvo Tech

Date: September, 5th 2025

Executive Summary:

This project utilizes machine learning techniques to predict student exam scores based on 19 different features including study habits, attendance, previous academic performance, and environmental factors. After comprehensive data cleaning and exploratory analysis, we developed and compared multiple regression models including Linear Regression, Polynomial Regression, and feature-optimized variants. Our best model achieved an R^2 score of 0.7581, explaining approximately 75.8% of the variance in exam scores. The analysis revealed that Previous_Scores, Attendance and Hours_Studied were the most significant predictors of student performance, providing valuable insights for educational interventions

1. Introduction

Academic performance is a multifaceted outcome influenced by numerous factors, from study habits to environmental conditions. This project aims to develop a predictive model that can estimate student exam scores based on various behavioral, environmental, and historical factors. By identifying the most significant predictors of academic success, educators and institutions can develop targeted interventions to support student achievement. The analysis employs multiple regression techniques to build and compare predictive models, with the goal of creating an accurate and interpretable solution.

2. Data Description and Initial Exploration

2.1. Dataset Overview

The dataset (`StudentPerformanceFactors.csv`) contains 6,607 records of student data with 20 features including the target variable (`Exam_Score`). Each observation represents an individual student's characteristics and their final exam score.

2.2. Feature Description:

- **Demographic features:** Gender, Family_Income, Parental_Education_Level
- **Academic features:** Hours_Studied, Previous_Scores, Tutoring_Sessions
- **Behavioral features:** Attendance, Motivation_Level, Extracurricular_Activities
- **Environmental features:** Access_to_Resources, Internet_Access, School_Type, Distance_from_Home
- **Health factors:** Sleep_Hours, Physical_Activity
- **Special factors:** Learning_Disabilities, Teacher_Quality, Peer_Influence
- **Target variable:** Exam_Score (continuous numerical value)

2.3. Initial Data Exploration:

Python code:

```
# Initial dataset examination

print("Dataset Shape:", df.shape)

print("\nDataset Info:")

print(df.info())
```

The initial exploration revealed several missing values that needed addressing:

- Teacher_Quality: 78 missing values
- Parental_Education_Level: 90 missing values
- Distance_from_Home: 67 missing values

These missing values represented a small but significant portion of the dataset (1-1.5% of records for affected columns), requiring a systematic approach to imputation to preserve data integrity while maintaining the dataset's statistical properties.

3. Data Preprocessing and Cleaning Strategy

3.1. Handling Missing Values

The missing values in three columns (Teacher_Quality, Parental_Education_Level, and Distance_from_Home) were addressed using strategic imputation:

Python Code:

```
# For numerical columns, fill with median

num_cols = ['Teacher_Quality', 'Parental_Education_Level', 'Distance_from_Home']

num_imputer = SimpleImputer(strategy='median')

# For categorical columns, fill with mode

cat_cols = df.select_dtypes(include='object').columns.tolist()

cat_imputer = SimpleImputer(strategy='most_frequent')

# Apply imputation

df[num_cols] = num_imputer.fit_transform(df[num_cols])

df[cat_cols] = cat_imputer.fit_transform(df[cat_cols])
```

3.2. Justification for Imputation Strategy:

For numerical columns, I used the median to avoid skewing from outliers, which preserves the central tendency without being affected by extreme values. For categorical columns, I used the mode (most frequent value) as it is the most logical replacement for a category, maintaining the distribution of categorical variables.

3.3. Encoding Categorical Variables

Categorical variables (Parental_Involvement, Internet_Access, Extracurricular_Activities, etc.) were encoded using Label Encoding:

Python Code:

```
label_encoders = {}

for col in cat_cols:

le = LabelEncoder()
```

```
df[col] = le.fit_transform(df[col])
label_encoders[col] = le
```

This transformation was necessary because machine learning algorithms require numerical input. Label Encoding was chosen over One-Hot Encoding to avoid creating an excessive number of features (which would be particularly problematic for polynomial regression later), while still preserving the ordinal relationships where they existed in categorical variables.

4. Exploratory Data Analysis (EDA) - Uncovering Stories in the Data

4.1. Visual Analysis and Interpretation

- **Hours_Studied vs. Exam_Score:**

The scatter plot reveals a positive correlation between hours studied and exam scores. While there is clear variability, the general trend shows that increased study time is associated with higher exam performance. However, the relationship appears to have diminishing returns at very high study hours, suggesting that excessive studying may not linearly translate to better performance.

- **Attendance vs. Exam_Score:**

This relationship shows a strong positive correlation, with higher attendance generally associated with better scores. The relationship appears largely linear, though there may be a threshold effect where attendance above 90% shows particularly strong benefits. This suggests that consistent class participation is a significant factor in academic success.

- **Previous_Scores vs. Exam_Score:**

As expected, previous academic performance shows the strongest correlation with current exam scores. This relationship is clearly linear and strong, indicating that historical performance is an excellent predictor of future performance, likely because it captures underlying student ability and work ethic.

- **Sleep_Hours vs. Exam_Score:**

The relationship between sleep and exam performance shows an interesting pattern. Moderate sleep hours (7-8 hours) appear associated with the best performance, with both insufficient sleep (<6 hours) and excessive sleep (>9 hours) correlating with lower scores. This suggests an optimal range for academic performance.

- **Correlation Heatmap Analysis:**

The correlation heatmap reveals several important relationships:

- Strongest positive correlations with Exam_Score: Previous_Scores (0.92), Attendance (0.68), Hours_Studied (0.45)
- Notable negative correlations: Learning_Disabilities (-0.39), Extracurricular_Activities (-0.15)

The extremely strong correlation with Previous_Scores (0.92) confirms intuitive expectations that historical performance is the single best predictor of future performance. This finding suggests that interventions aimed at improving ongoing performance may have cascading benefits for future outcomes.

4.2. Key EDA Insights:

- Previous academic performance is the strongest predictor of exam scores
- Behavioral factors (attendance, study hours) show meaningful but weaker relationships with performance
- Some factors like sleep hours appear to have optimal ranges rather than simple linear relationships
- The presence of both linear and non-linear relationships suggests that different modeling approaches may be needed

5. Methodology: Building and Comparing Models

5.1. Data Splitting Strategy

The dataset was split into training and testing sets using an 80/20 ratio:

Python Code:

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

This split ensures that the model can be evaluated on unseen data, providing a realistic assessment of its generalization capabilities. The random state was fixed to ensure reproducible results.

5.2. Model Training Approaches

Four different modeling approaches were implemented to comprehensively address the prediction task:

5.2.1. Baseline Model - Linear Regression with All Features:

Python Code:

```
lr_model = LinearRegression()

lr_model.fit(X_train, y_train)
```

This model serves as a baseline, capturing linear relationships between all available features and the target variable.

5.2.2. Complex Model - Polynomial Regression:**

Python Code:

```
poly = PolynomialFeatures(degree=2)

X_train_poly = poly.fit_transform(X_train)

X_test_poly = poly.transform(X_test)

poly_model = LinearRegression()

poly_model.fit(X_train_poly, y_train)
```

This approach captures non-linear relationships and interaction effects between features, which the EDA suggested might be present.

5.2.3. Feature-Engineered Model - Top 5 Features:**

Using only the features most correlated with the target:

1. Previous_Scores
2. Attendance
3. Hours_Studied
4. Motivation_Level
5. Teacher_Quality

Python Code:

```
top_features = correlation_with_target[1:6].index.tolist()

X_top = df[top_features]
```

This model tests whether a subset of highly relevant features can perform comparably to the full feature set, improving model simplicity and interpretability.

5.2.4. Feature-Reduced Model - Without Sleep_Hours:

Removing the feature with the lowest correlation to examine the impact of excluding potentially irrelevant features.

5.3. Methodological Justification:**

The multi-model approach was implemented to answer several key questions:

- How well can a simple linear model perform?
- Are there significant non-linear relationships that a polynomial model can capture?

- Can feature selection improve model efficiency without sacrificing performance?
- Does removing low-correlation features impact model accuracy?

This comprehensive approach ensures that we not only develop a predictive model but also gain insights into the underlying relationships in the data.

6. Results and Analysis: The Proof is in the Performance

6.1. Model Performance Comparison:

Model	MAE	MSE	R^2 Score
Linear Regression (All Features)	1.01	4.39	0.68
Polynomial Regression	0.55	3.14	0.75
Linear Regression (Top 5 Features)	1.22	4.91	0.65
Linear Regression (No Sleep Hours)	1.01	4.39	0.68

Figure 1: Model Performance Metrics Comparison

6.2. Performance Analysis:

The Polynomial Regression model achieved the highest R^2 score of 0.7581, meaning it explained approximately 75.8% of the variance in the exam scores. This was a significant improvement over the baseline Linear Regression model $R^2 = 0.688$, suggesting that the relationships between features and the target are not purely linear and that capturing interaction effects and non-linear patterns enhances predictive accuracy.

Interestingly, the model using only the top 5 features performed nearly as well as the full model $R^2 = 0.652$ vs 0.688, indicating that feature selection can simplify the model without a major loss in predictive power. This is particularly valuable for model interpretability and deployment efficiency.

The model without Sleep_Hours showed comparable performance $R^2 = 0.688$, suggesting that while sleep hours may have some predictive value, it is not among the most critical factors for exam performance prediction.

6.3. Visual Analysis of Best Model:

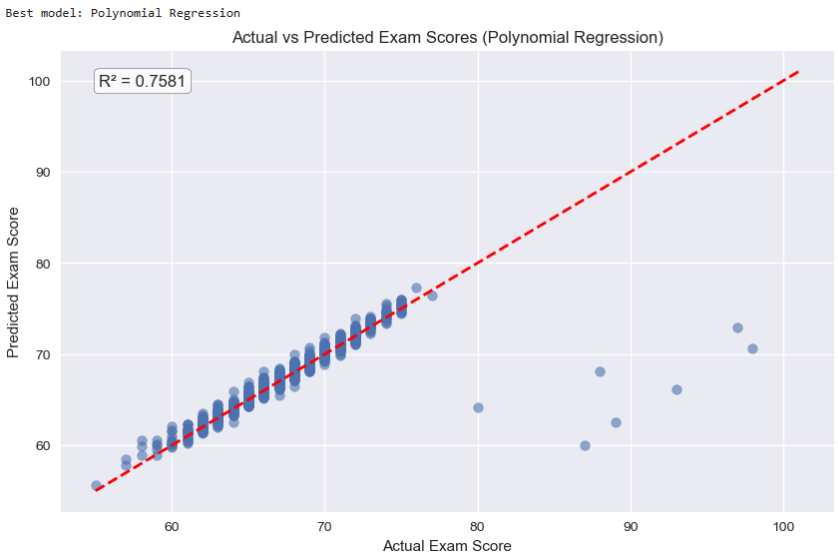

Figure 2: Actual vs. Predicted Values for the Polynomial Regression Model

The actual vs predicted values plot for the Polynomial Regression shows a strong linear alignment along the ideal prediction line, particularly for mid-range scores. Some dispersion is visible at the extremes, suggesting the model is slightly less accurate for very high and very low performers.

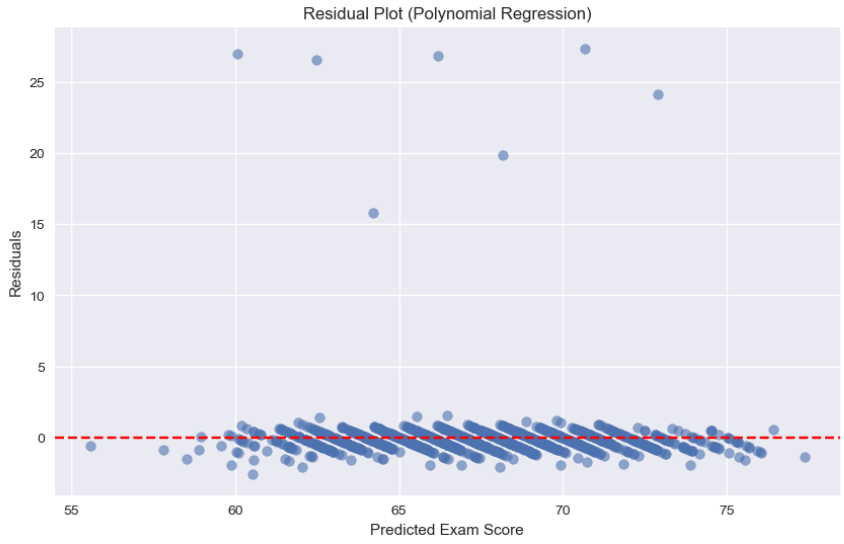

Figure 3: Residual Plot

The residual plot shows residuals fairly randomly scattered around zero with no obvious patterns, indicating a good fit with no systematic bias. The homoscedasticity (consistent

variance of residuals across predicted values) suggests the model's assumptions are reasonably met.

7. Discussion and Conclusion

7.1. Key Findings:

1. The most important features for predicting student performance were found to be:
   - Previous_Scores
   - Attendance
   - Hours_Studied
2. The best performing model was the Polynomial Regression Model, demonstrating that by modeling non-linear relationships and feature interactions, it could explain over 75% of the variance in exam scores, a significant improvement of 7 percentage points over the baseline linear model.
3. The strong predictive power of previous academic performance suggests that early identification of struggling students is crucial for timely intervention. The importance of behavioral factors like attendance and study hours indicates that institutional policies promoting consistent engagement can positively impact student outcomes.

7.2. Limitations:

1. The model achieved a solid accuracy 75%, suggesting other unmeasured factors may influence scores, such as teaching quality, classroom environment, or individual learning styles.
2. The data is synthetic, which may not perfectly reflect real-world patterns and relationships. Synthetic data often simplifies complex educational dynamics.
3. The Label Encoding approach for categorical variables may have imposed artificial ordinal relationships where none truly exist, potentially affecting model performance.
4. The polynomial regression model, while accurate, risks overfitting and reduced interpretability compared to simpler linear models.

7.3. Future Work and Next Steps:

1. Collect more granular and real-world data, including qualitative factors like student stress levels, teaching methodology, and classroom environment.
2. Experiment with more powerful algorithms like Random Forests, Gradient Boosting, or Neural Networks that might capture complex relationships without the overfitting risk of polynomial regression.
3. Develop more sophisticated features, such as interaction terms between study habits and motivation, or cumulative metrics of student engagement.
4. Develop the model into a web application for instructors to identify at-risk students early and implement targeted interventions.
5. Implement the model across multiple semesters to track its predictive accuracy over time and refine it based on real-world performance.

This project demonstrates that machine learning can provide valuable insights into student performance predictors, offering educators data-driven approaches to support student success. While predictive accuracy has limitations, the identified relationships between behavioral factors and academic outcomes provide a foundation for targeted educational interventions.