SG - 7

Members:

Austin Steel

Sarah Starkey Lanham

Mahender Bandi

Sai Bhargav Reddy Kanuparthy

Executive Summary

After reviewing initial correlations and trends, we created a classification model to predict if students will pass or fail. This model gave us an 86% accuracy, 90% Precision and AUC of 0.927. From this model, we can see the most important factors in students passing or failing the test are ***Attendance*** and ***Hours Studied.***

What makes students pass or fail tests? Educators have wondered this for all of time, and considering there are so many different factors, it's hard to tell.

To answer this question, we looked at a dataset titled "Student Performance Factors". This dataset includes a wide variety of information about students, and their performance on a test.

The goal of this analysis will be to see if any specific factors or combinations of factors effect students passing or failing the test.

Background Info on Topic

- There are wide variety of factors which can affect student performance on tests both from school and their personal lives. Educators are constantly looking for how they can help students perform better by analyzing data and looking for correlations. While some factors Teachers can control, such as Teacher Quality, or Tutoring Hours, for many factors Teachers are unable to have any effect on. Many of these factors like Parental Involvement, Extracurricular Activities, Family Income, etc, can have a large impart on student performance as well. Given all the challenges that educators face today, modeling student data can help Teachers know which students may need more or extra support and predict from student data how they will perform.

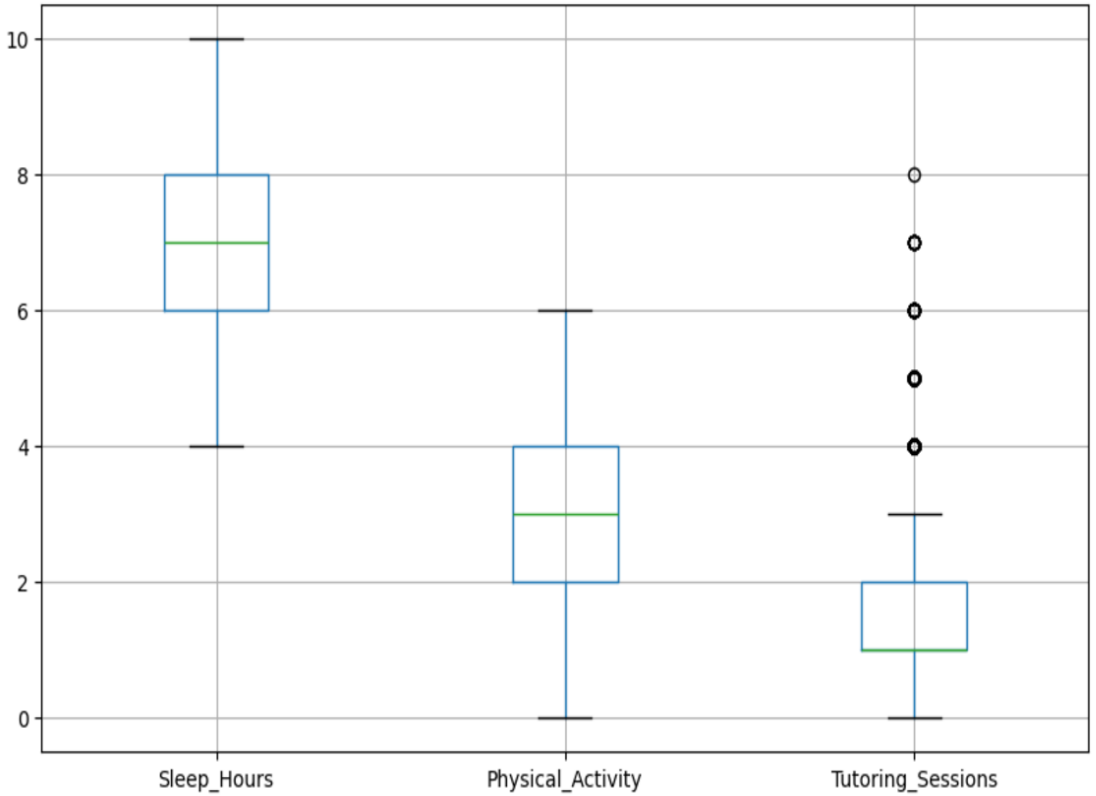
Data Set Description

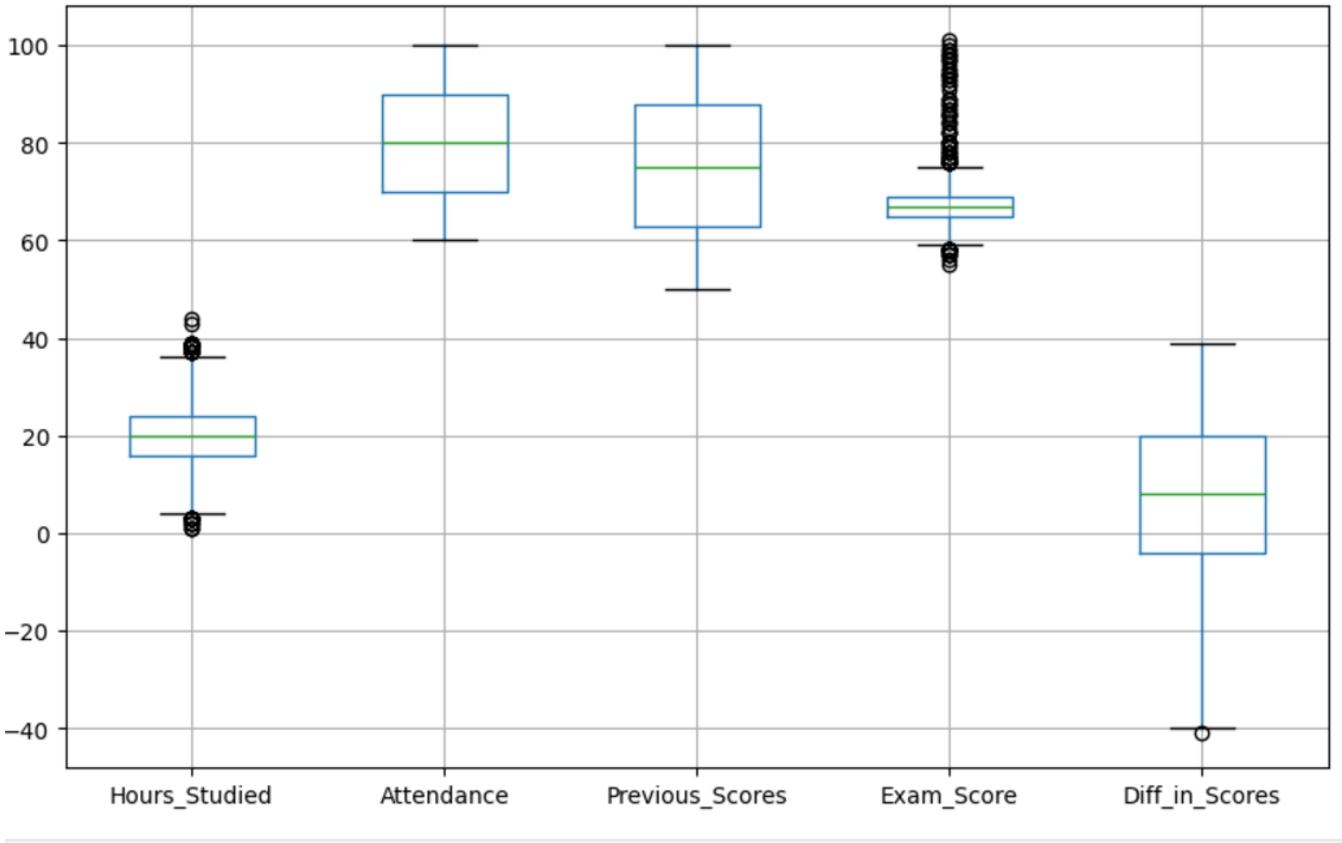
- This data set provides an overview of various factors effecting students' performance on exam scores. It considers variables such as parental involvement, tutoring hours, family income, teacher quality, and other aspects that may affect students' performance on exams.
- In total there are 20 columns (variables) and 6607 row entries. Most of the variables are categorical rankings such as low, medium, high. The rest are numerical variables.
- Less than 1% of the entries had any missing values.
- Target variable: Exam Scores**

Data Preprocessing and Basic Analysis

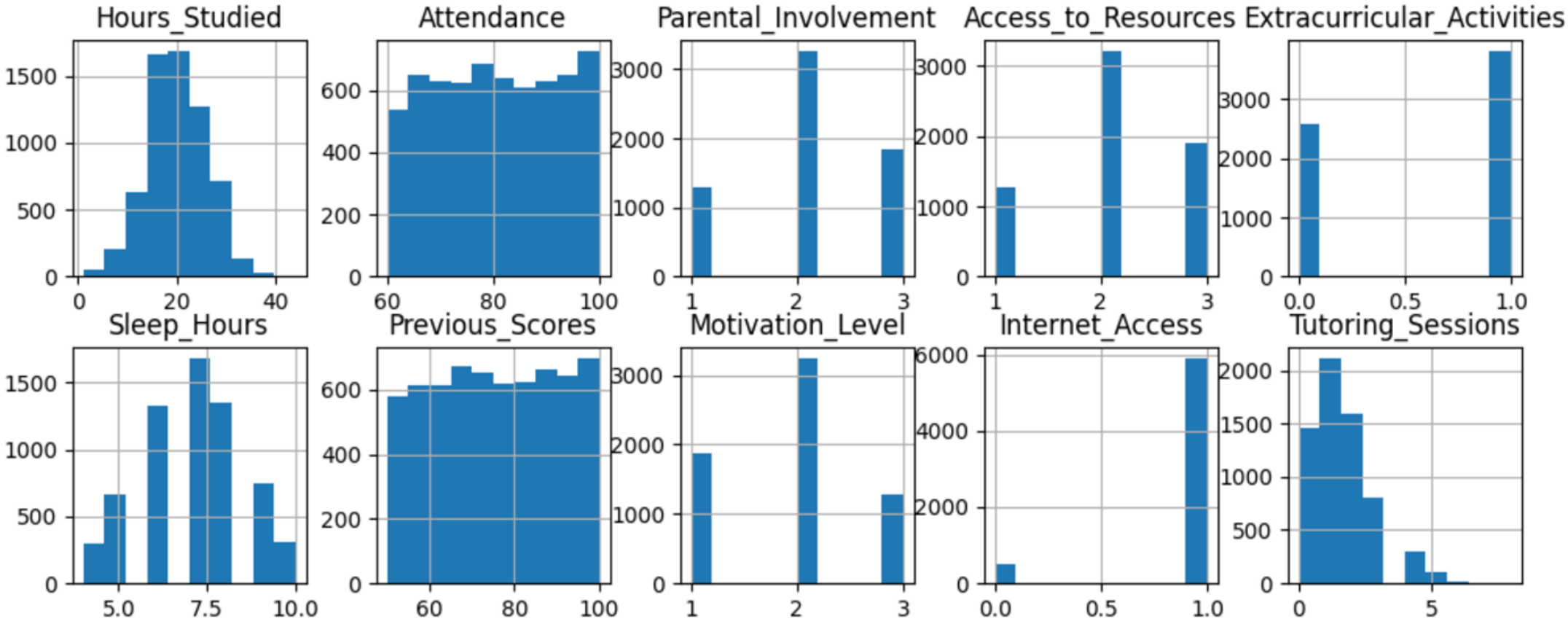
- Check for Null Values.
- Deleted missing values, leaving us with 6378 row entries.
- Replaced categorical variables with integers (0, 1, or 1, 2, 3).
- Added Difference in Scores column.
- Added columns for Exam Score Rating and Previous Score Rating for Pass or Fail as 0, or 1.

Boxplots of Numerical Variables

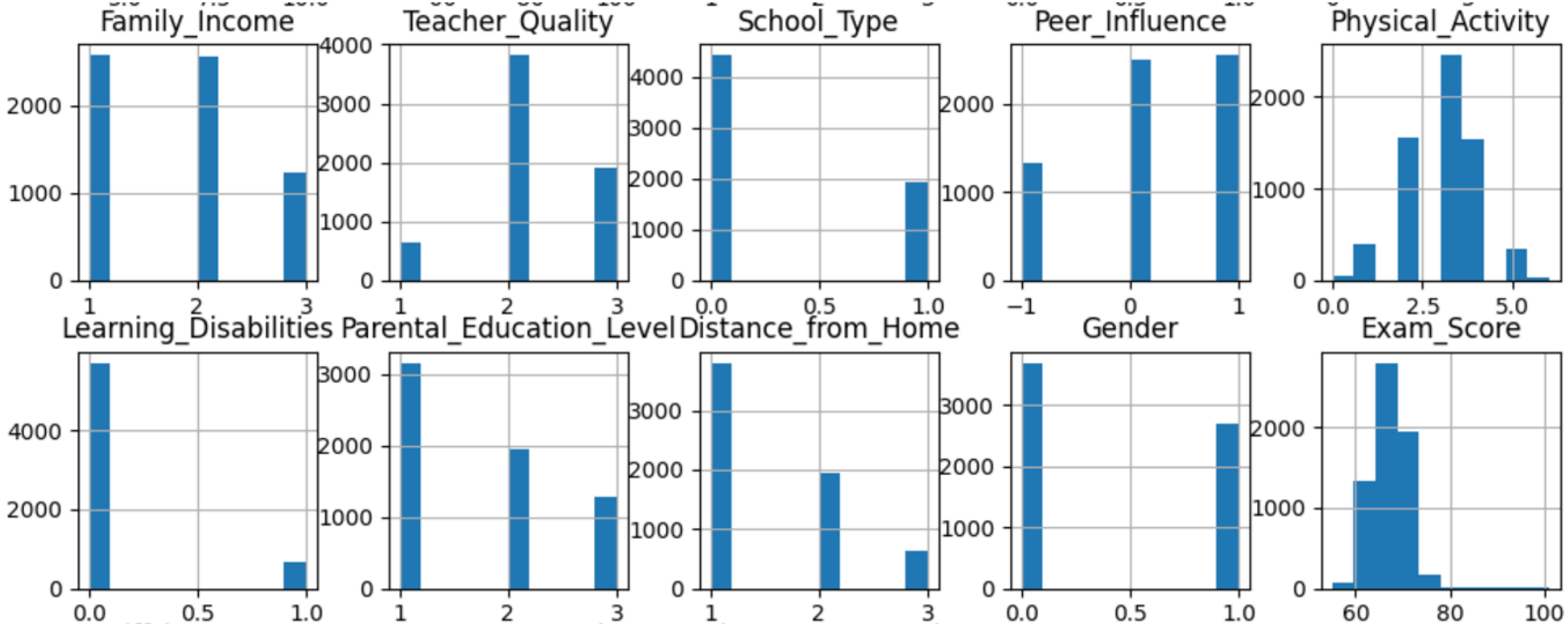




Distribution of Variables



Distribution of Variables



Preprocessing More...

- Data Imbalance!!
- Used SMOTE to transform data.

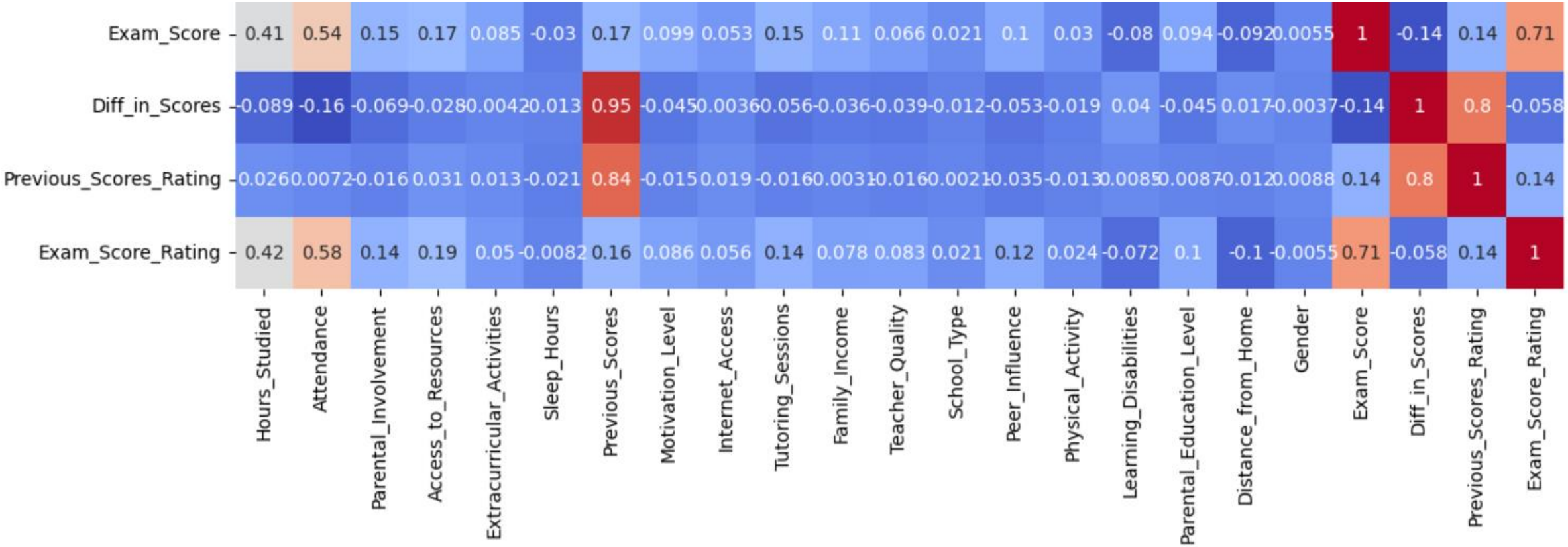
Initial Data:

```
Count of 0: 4797
Count of 1: 1581
```

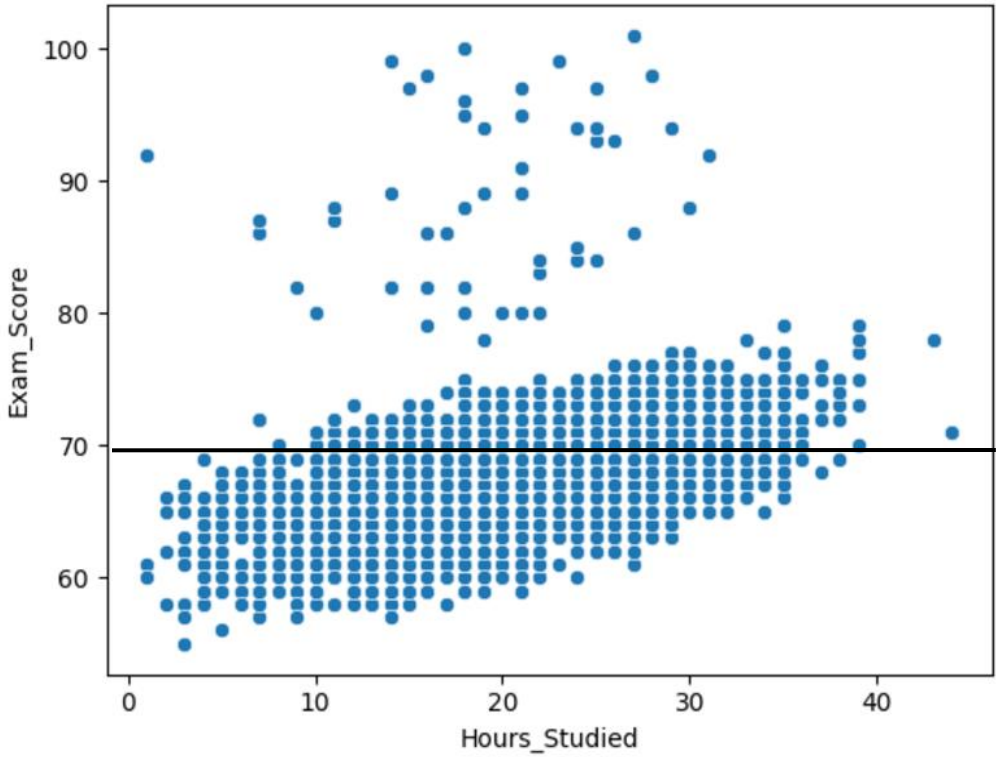
Data after balancing:

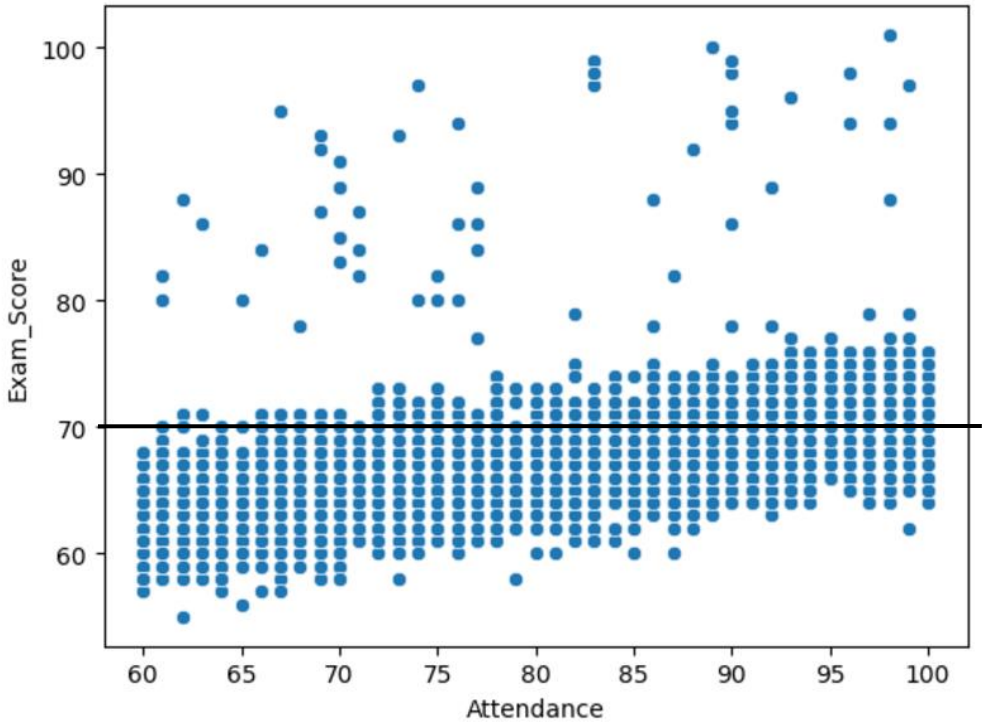
```
Count of 0: 4797
Count of 1: 4797
```

Initial Findings: Heatmap



Scatterplots of Highest Correlations





Pivot Tables

Key: 0=Fail
1=Pass

Distance_from_Home	Far	Moderate	Near
Exam_Score_Rating			
0	526	1515	2756
1	110	426	1045

Learning_Disabilities	No	Yes
Exam_Score_Rating		
0	4246	551
1	1464	117

Family_Income	High	Low	Medium
Exam_Score_Rating			
0	870	2032	1895
1	360	550	671

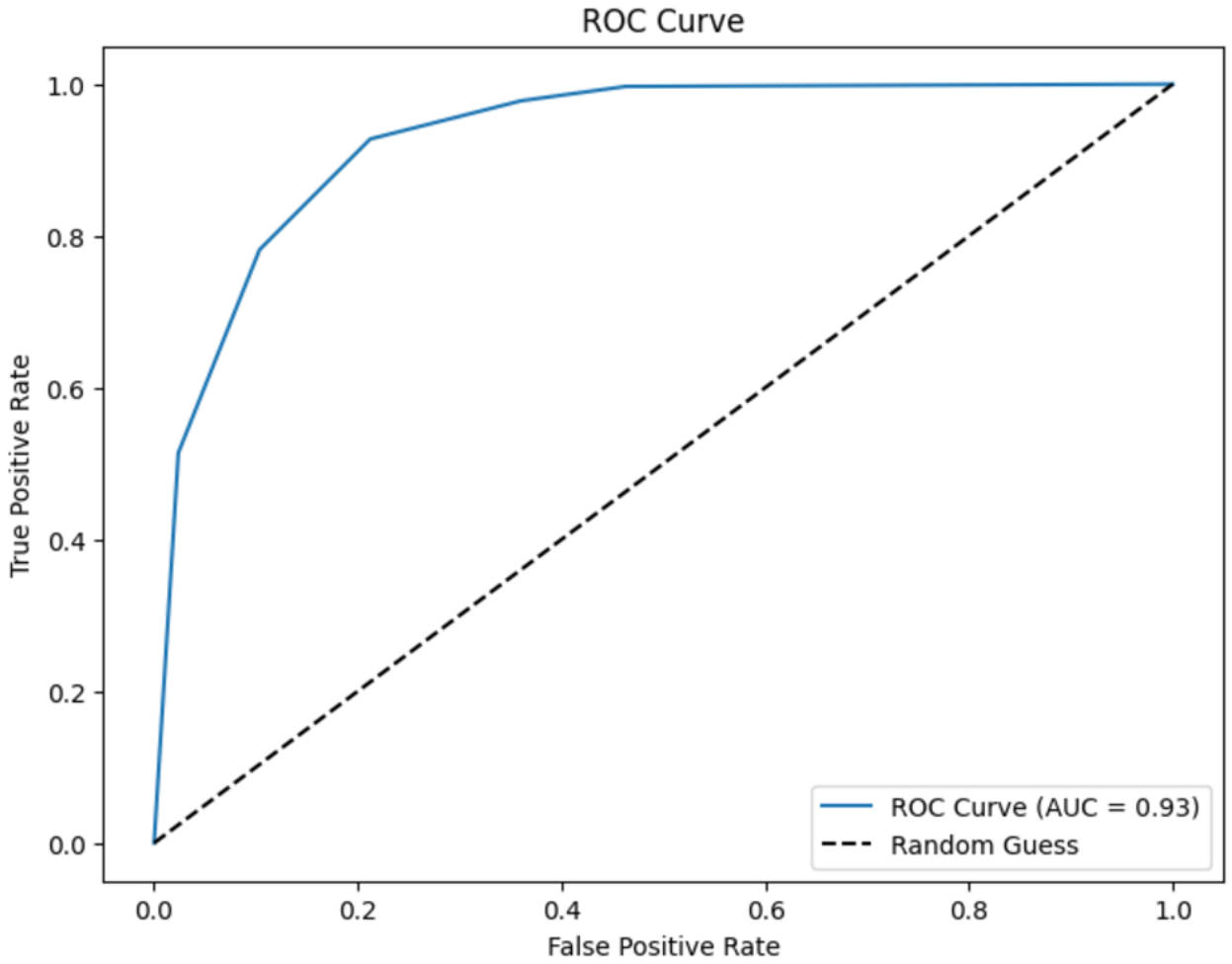
School_Type	Private	Public
Exam_Score_Rating		
0	1450	3347
1	494	1087

Teacher_Quality	High	Low	Medium
Exam_Score_Rating			
0	1359	521	2917
1	546	126	909

Motivation_Level	High	Low	Medium
Exam_Score_Rating			
0	897	1472	2428
1	380	392	809

Test

- Classification Model
- Target variable: Exam_Score_Rating
- Split data: 80% Train / 20% Test
- Results:

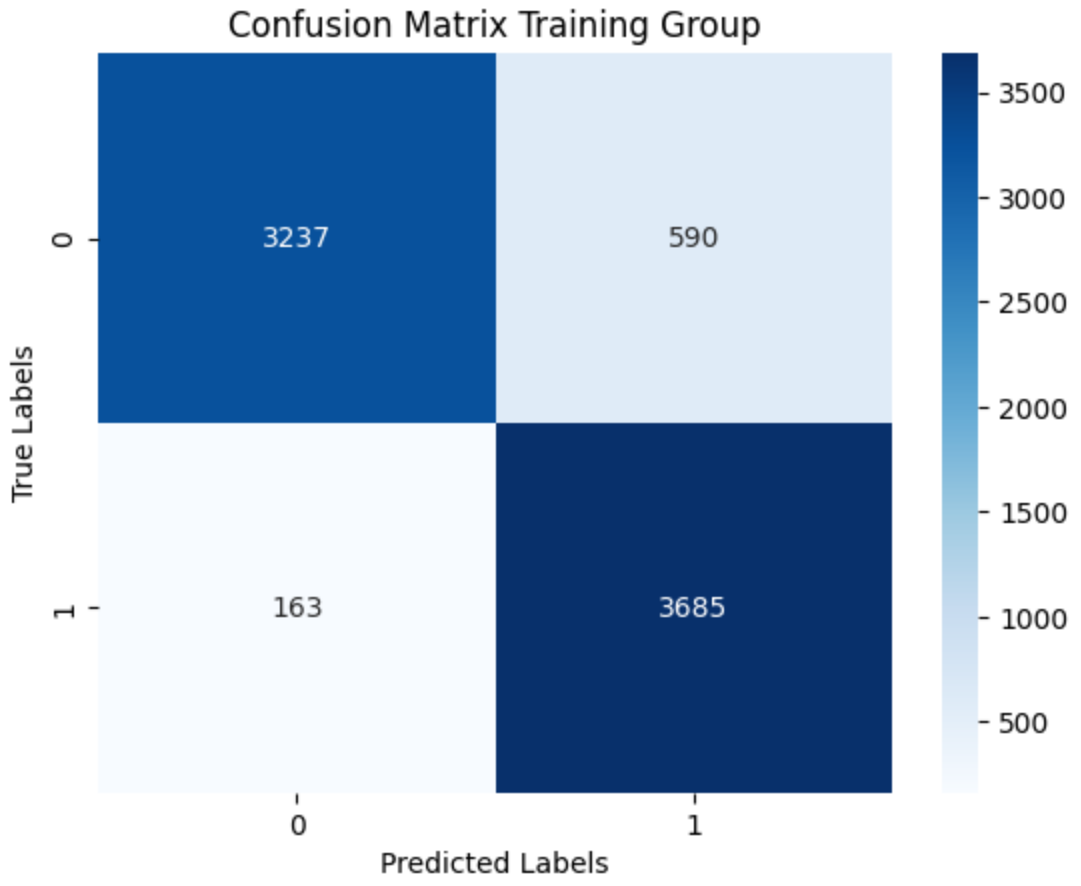


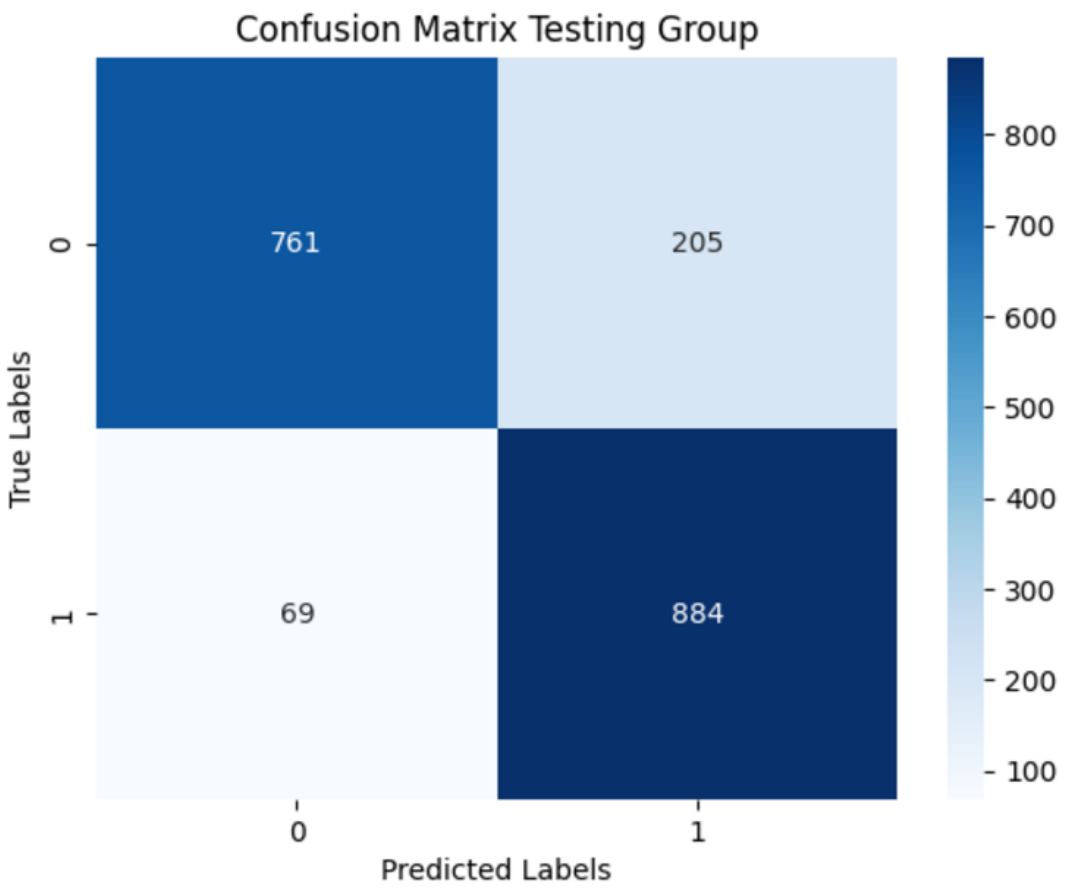
Test Results

Training Accuracy: 0.9017589576547231
Test Accuracy: 0.8572173006774362

	precision	recall	f1-score	support
0	0.90	0.79	0.84	970
1	0.81	0.91	0.86	949
accuracy			0.85	1919
macro avg	0.86	0.85	0.85	1919
weighted avg	0.86	0.85	0.85	1919

AUC: 0.9271718466535583





Conclusions / Recommendations

Passing students go to class and study and then perform well on tests. Students who don't attend class or put in study hours do poorly on testing.

Model for whether students will pass or fail built is very accurate with small overfitting

Most important passing factors:

Attendance and Hours Studied

Factors against passing:

Learning Disabilities and Distance From Home.

Reference Material and Additional Reading

- Data set: https://www.kaggle.com/datasets/lainguyn123/student-performance-factors
- https://link.springer.com/article/10.1007/BF01537904
- https://www.tandfonline.com/doi/pdf/10.1080/00098655.1994.9956043
- https://journals.sagepub.com/doi/epdf/10.1177/0741932508327460