Predicting GVSU K-12 Students' Educational Success: Analyzing the Impact of Family and School Characteristics on Student Performance

Team Members:

Rohith Anugolu

Date: April 2, 2025

1. Summary of Insights

This study aims to predict student success in K-12 education at GVSU using key demographic and educational predictors. The response variable chosen for analysis is **SEGRADES (Student Grades)**, categorized into A, B, C, D, and Other. The predictors include **PARENT (Guardian), HAS_POST_SEC (Parent's Post-Secondary Education), PUBLIC_SCHOOL (Public vs. Private School), INTERNET_ACCESS (Home Internet Availability), NO_OF_ABSENT (Student Absenteeism), and INCOME_CAT (Household Income Category).** Two modeling techniques were employed: **Multinomial Logistic Regression and Linear Discriminant Analysis (LDA)**. Cross-validation was implemented using the **vfold_cv** method to ensure robust model evaluation. Performance metrics such as **accuracy, roc_auc, and confusion matrices** were used to compare model performance between the models as well as the train and test splits.

1.1 Key Insights:

- **Parental Structure**: Students from two-parent households had significantly better grade outcomes. Students with a single parent, either "Father" or "Mother" also had lower odds of lower grades.
- **Parental Education**: Students with post-secondary educated parents were less likely to receive lower grades.
- **Household Income**: Higher income was associated with higher grades. For students from "$200K or more" households, the odds of obtaining lower grades were very less.
- **Absenteeism**: More absences significantly predicted lower performance. Students absent "11-20 days" had increased odds of receiving a D or C.
- **School Type**: Students from public schools were more likely to receive lower grades when compared to non-public schools. Among public school students, **24.8% received a B and 57.3% received an A**. However, in non-public schools, **19.0% received a B and 63.6% received an A**, indicating stronger educational success.

1.2 Summary of findings:

- Overall**,** Household income and parental education showed significant influence on student grades**.**
- Students with higher absenteeism tended to have lower grades.
- Both models performed similarly in terms of accuracy and ROC-AUC.
- Test accuracy was slightly lower than training accuracy, suggesting a slight overfitting but within an acceptable range.

1.3 Recommendations for GVSU's K-12 Connect:

- Support programs that engage **both parents** in lower-income or single-parent households.
- Provide additional academic support in **public schools**, where students are more likely to score lower.
- Proactively address absenteeism, as it has a strong negative association with grades.
- Encourage parental education and awareness programs, emphasizing the long-term impact on children's academic outcomes.

2. Summary of Modeling Process

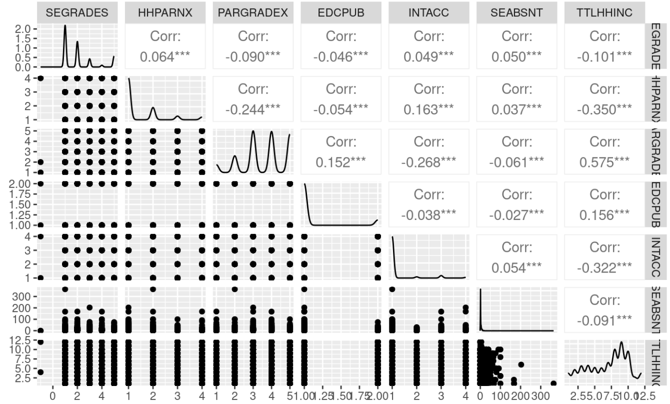
2.1 Data Preparation:

- The "pfi_2016" and "pfi_2019" datasets were imported and combined by selecting some common variables between them.
- Missing values were checked using skim() function, and no missing values were found.
- Relevant variables, including "SEGRADES," "HHPARN16X/HHPARN19X", "PARGRADEX," "EDCPUB," "INTACC," "SEABSNT," and "TTLHHINC," were selected for analysis.

- New variables were created and encoded as factors:
    - "GRADE" was derived from "SEGRADES" with levels "A," "B," "C," "D," and "Other".
    - "PARENT" was derived from "HHPARNX" with levels "Both," "Mother," "Father," and "None".
    - "HAS_POST_SEC" was derived from "PARGRADEX" with levels "Post Secondary" and "Not Post Secondary".
    - "PUBLIC_SCHOOL" was derived from "EDCPUB" with levels "Public" and "Non-Public".
    - "INTERNET_ACCESS" was derived from "INTACC" with levels "Yes" and "No".
    - "NO_OF_ABSENT" was derived from "SEABSNT" with levels "0-10 days," "11-20 days," and "Other".
    - "INCOME_CAT" was derived from "TTLHHINC" with levels "0-100K", "100-200K", and "200K or more".
- The combined dataset was divided into training and testing sets. Then we trained the model(s) on the training set using k-fold cross validation with 5 folds and tested on the testing set.

2.2 Exploratory Analysis:

We started with Exploratory Data Analysis to understand the relationship between potential predictor variables and student grades.



There's a small positive correlation between student grades and the number of parents in the household.

Internet access has a small positive correlation with student grades.

Student absences unexpectedly show a small positive correlation with student grades.

Parental education is strongly positively correlated with household income

3. Model Selection and Performance

To ensure robust model selection, **5-fold cross-validation** was applied to both **Multinomial Logistic Regression (MLR)** and **Linear Discriminant Analysis (LDA)** models. This method helps assess generalizability and prevent overfitting by evaluating performance across different data partitions.

Here's the regression equation of the log-odds of a student who got "Other" grade vs "A" grade:

$$
\begin{aligned}
\log\left(\frac{\widehat{p_{\mathrm{A}}}}{\widehat{p_{\mathrm{Other}}}}\right) = &-1.1039 + 0.46164431 PARENTFather + 0.16219859 PARENTMother \\
&+ 0.27154298 PARENTBoth + 0.10870736 HAS\_POST\_SECYes \\
&- 0.15197641 PUBLIC\_SCHOOLYes + 0.03235633 INTERNET\_ACCESSYes \\
&- 0.15881349 NO\_OF\_ABS\tilde{}ENT11 - 20days - 0.07506882 NO\_OF\_ABSENTOther \\
&+ 0.06715767 INCOME\_CAT100 - 200K + 0.08357023 INCOME\_CAT200Kormore
\end{aligned}
$$

Interpretation of coefficients:

Intercept (-1.1039): This is the baseline log-odds of getting an "A" grade when all predictor variables are zero.

PARENTFather (0.4616): Compared to the reference category (likely "No parent" or "Unknown"), having a father as the primary parent increases the log-odds of getting an "A" grade.
PARENTMother (0.1622): Having a mother as the primary parent also increases the log-odds of getting an "A" grade, but the effect is smaller than having a father.
PARENTBoth (0.2715): Having both parents increases the log-odds of getting an "A" grade, but less than having just a father.
HAS_POST_SECYes (0.1087): If a students' parent has post-secondary education, the log-odds of getting an "A" grade increase slightly.
PUBLIC_SCHOOLYes (-0.1520): Attending a public school reduces the log-odds of getting an "A" grade compared to the reference category (likely private school).
INTERNET_ACCESSYes (0.0324): Having internet access slightly increases the log-odds of getting an "A" grade.
NO_OF_ABSENT 11–20 days (-0.1588): Being absent for 11–20 days reduces the log-odds of getting an "A" grade compared to the reference category ("0–10 days" or "more than 20 days").
NO_OF_ABSENT Other (-0.0751): Other categories of absence (possibly irregular absences) also reduce the log-odds of getting an "A" grade.
INCOME_CAT100–200K (0.0672): Having a family income between $100K–$200K slightly increases the log-odds of getting an "A" grade.
INCOME_CAT200K or more (0.0836): Having a family income of $200K or further increases the log-odds of getting an "A" grade compared to lower-income categories**.**

4. Model Diagnostics

To evaluate the performance of our multinomial regression and linear discriminant analysis (LDA) models, we analyzed several key metrics, including accuracy, roc_auc, and confusion matrices for both training and test datasets.

4.1 Model Performance Metrics

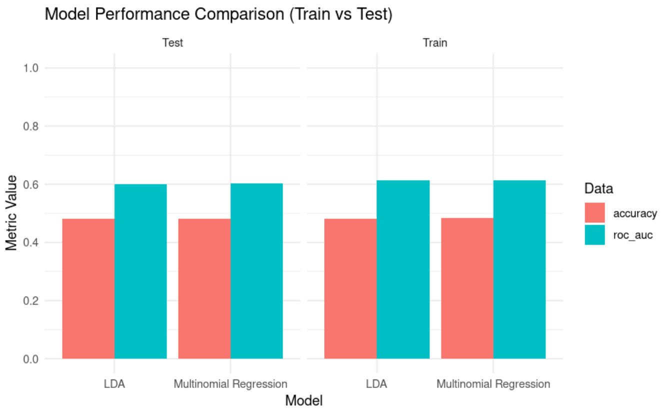
We assessed model quality using the following performance metrics:

- **Accuracy**: Measures the proportion of correct classifications among total predictions.
- **ROC-AUC**: Evaluates the model's ability to distinguish between classes. A higher AUC indicates better class separation.
- **Confusion Matrix**: Displays the breakdown of actual vs. predicted classifications, helping assess misclassification patterns.

4.2 Training vs. Testing Performance

We performed cross-validation on the training dataset and tested model generalizability on the test dataset.

Model	Dataset	Accuracy	ROC AUC
Multinomial Regression	Train	0.484	0.613
Multinomial Regression	Test	0.482	0.603
LDA	Train	0.482	0.612
LDA	Test	0.481	0.599



Overall Performance: Both models perform similarly, with both accuracy and ROC AUC scores around 0.48-0.61. This suggests that neither model has a clear advantage over the other in terms of overall performance.

Training vs. Test Performance:

- *Multinomial Regression*: The model performs slightly better on the training set (Accuracy: 0.484, ROC AUC: 0.613) compared to the test set (Accuracy: 0.482, ROC AUC: 0.603). This suggests a slight degree of overfitting.
- *LDA*: The model also performs slightly better on the training set (Accuracy: 0.482, ROC AUC: 0.612) compared to the test set (Accuracy: 0.481, ROC AUC: 0.599). This also suggests a slight degree of overfitting.

4.3 Confusion Matrices

Both models perform **well for grades A and B**, with most predictions aligning with actual values. **Misclassifications primarily occur in lower grade categories (Other, D, and C). However, there're extreme misclassification for lower grade.** Both models struggle with underrepresented classes (D, C, Other), incorrectly assigning most students to A or B. This suggests **a strong class imbalance issue** in the dataset. Also, there's **low sensitivity for lower grades.** Almost all "Other" students are classified as A or B, **causing an overestimation of high grades.**

Multinomial Regression					
GRADE	**Other**	**D**	**C**	**B**	**A**
Other	0	0	0	327	2361
D	0	0	0	165	254
C	0	0	0	589	1394
B	0	0	0	1173	5117
A	0	0	0	985	9357

Linear Discriminant Analysis					
GRADE	**Other**	**D**	**C**	**B**	**A**
Other	0	2	8	306	2372
D	0	1	12	144	262
C	0	7	32	535	1409
B	0	13	63	1044	5170
A	0	5	32	912	9393

5. Conclusion

The project aimed to predict student success at GVSU using **Multinomial Regression** and **Linear Discriminant Analysis (LDA)**. Both models demonstrated moderate performance, with accuracy around **48%** and ROC AUC scores near **0.60**, indicating limited predictive power. The confusion matrices revealed a bias toward correctly classifying higher grades (**A and B**) while struggling with lower ones (**C, D, and Other**), likely due to class imbalance. Although LDA showed a slight improvement in distinguishing lower grades, the overall difference was minimal. While the models provided some predictive capability, the relationships between the predictor variables and student grades require further exploration to fully understand the underlying dynamics. For GVSU's K-12 Connect, these findings highlight the importance of considering a holistic approach that addresses various aspects of a student's environment to foster academic success.