CPSC 392 Final Project

Ethan E. Lopez
Jane San

Introduction

Description of the data (e.g. What variables do you have? Where is this data from? How much data do you have? Anything notable about the data).

Sponsored by Kaggle, local universities aim to discover key patterns and trends in undergraduate outcomes using the past records of 10000 students. Rating academic achievements and skills in extra curricular activities, there are 9 variables geared towards evaluating student performance and traits in post graduation employment.

In these, 2 are categorical and 7 are continuous factors showcasing scholarly habits. *Internship_Experience* highlighting students that completed their first co-op's, *Placement* indicates the frequency of student successes securing a job in their specialized major field. *IQ*, *Prev_Sem_Result*, *CGPA*, *Academic_Performance*, *Extra_Curricular_Score*, *Communication_Skills*, and *Projects_Completed*, meanwhile, are the remaining continuous variables identifying class ratings at various levels of campus involvement.

Given the dataset is cleaned and preprocessed strategically, the records will open the gateway for further analysis in predictive modeling, allowing us to determine the underlying characteristics of students who excel vs. fall out of using their educational background for future opportunities. Most noteworthy about the data, on the other hand, is its inclusion of *Prev_Sem_Result* in conjunction with *CGPA*. This likely goes into affecting the performance of our models later on, referencing that the former is simply the *CGPA* from the semester before, introducing possibilities for high multi-collinearity.

Question #1 (Linear Regression): When predicting CGPA, how strong is a model using all other variables, and which predictor (among IQ, Projects_Completed, Internship_Experience, Communication_Skills, and Prev_Sem_Result) is most influential in improving R^2?

Methods

The goal of this question was to understand how well we could predict a student's CGPA using variables like IQ, number of projects completed, internship experience, communication skills, previous semester results and such.

For the initial step, the dataset was cleaned by removing any missing or invalid values. Additionally, Z-score normalization was performed since the variables had very different scales between variables (IQ scores vs Project counts). This step ensured that all variables were within similar ranges, treating them equally instead of giving more weights to larger numbers.

Next, we used a linear regression model. This defined the relationship between our independent variables (predictors) and the dependent variable, CGPA. We divided the data using an 80/20 train-test-split, where the first 80% was used to train the model on how to predict CGPA. The testing data was then used to check how accurate our predictions were.

To measure how strong our models performed, we used several evaluation metrics, including R^2 (R-squared), which communicated how much of the variation in CGPA was explained by the input variables. We also used RMSE (Root Mean Square Error), MAE (Mean Absolute Error), and MAPE (Mean Absolute Percentage Error) to check how close the predicted values were to the actual CGPA values.

After building the model using all predictors, we looked into which variable contributed the most to improving the R^2 of the model. For that, we used the Leave-One-Out method. In this approach, we removed one variable at a time and rebuilt the model to see how much R^2 changed (for example, removing IQ, then removing Internship_Experience, and so on). If removing one variable changed significantly, it meant that the variable was very important.

We visualized our results using a bar chart, comparing R^2 values for each model (with one variable left out), a scatter plot showing actual vs predicted CGPA values, and a bar chart of coefficients to see which features had the most influence on the prediction.

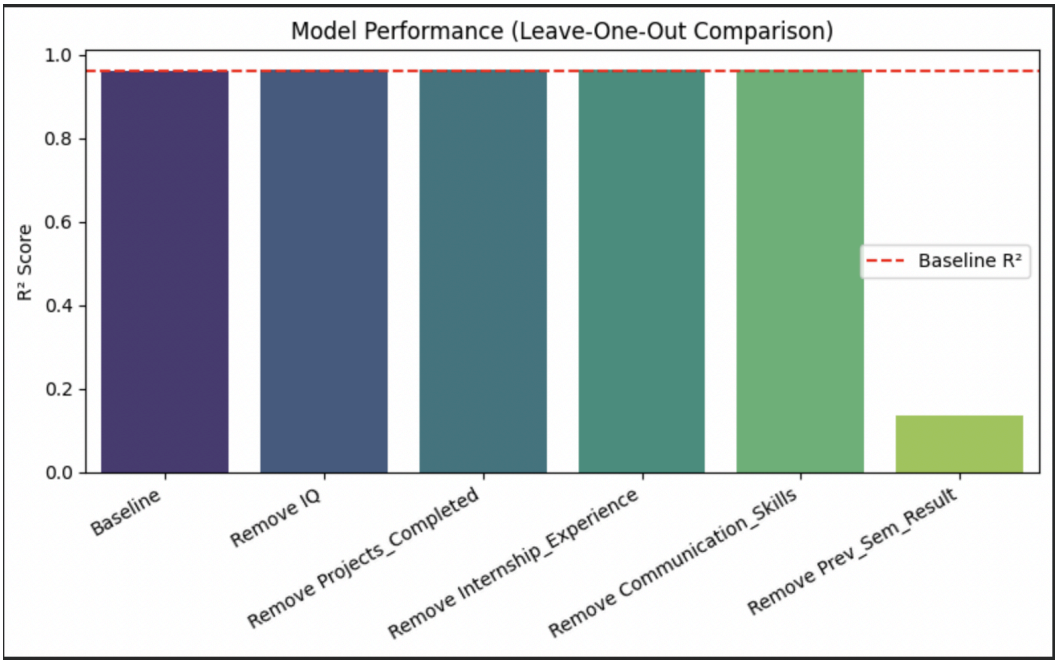

Figure 1 (Bar Chart): Model performance (Leave-One-Out R^2 Comparison)

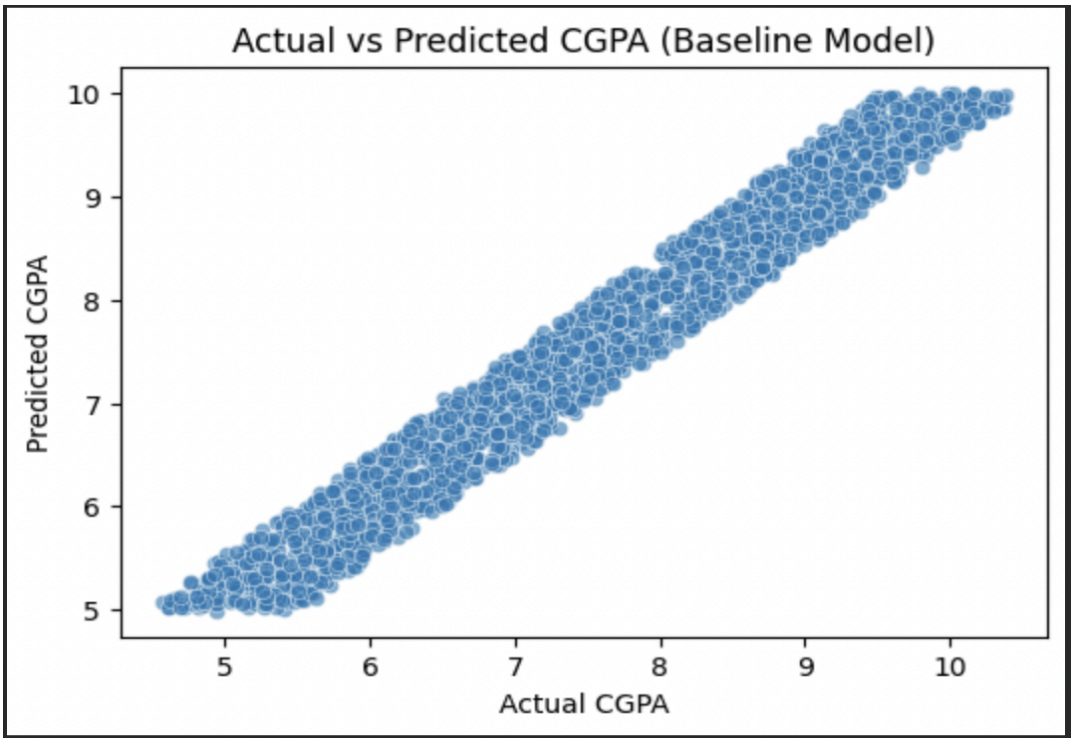

Figure 2 (Scatterplot): Actual vs Predicted CGPA showing strong linear alignment and high accuracy.

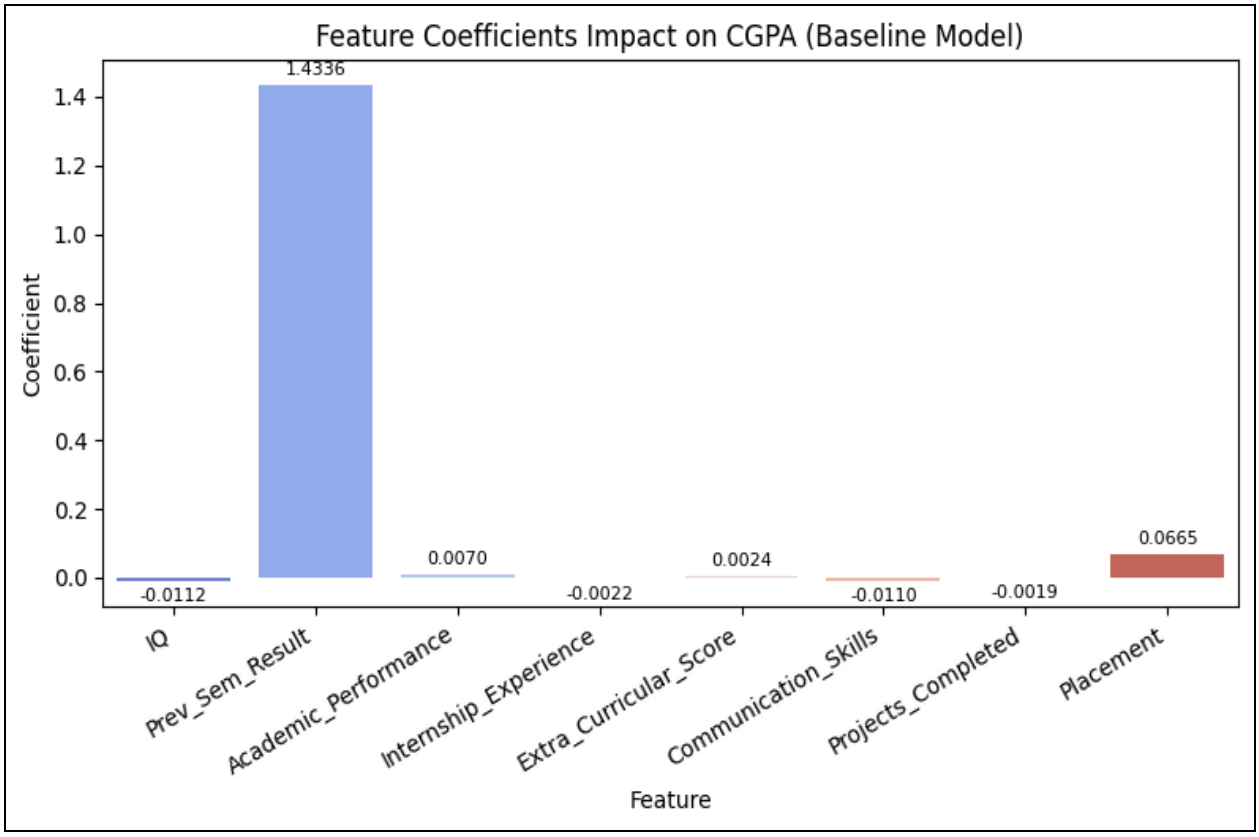

Figure 3 (Bar Chart): Feature Coefficients Impact on CGPA with coefficient measurements rounded to 4 decimal places

Regression Model Performance Metrics:

	Metric	Training Set	Testing Set
0	MSE	0.0821	0.0814
1	MAE	0.2479	0.2479
2	MAPE	3.4346	3.4329
3	RMSE	0.2866	0.2854
4	R^2	0.9621	0.9620

Figure 4 (Table): Exact performance metrics for the Linear Regression model

Results

The linear regression model performed extremely well, reaching an R^2 value of around 0.96, which meant the model explained around 96% of the variation in student CGPAs. This was a very strong result, showing that the chosen predictors were good indicators of academic performance.

On the other hand, when we looked at the Leave-One-Out comparison, we noticed that removing most features barely made any difference to the R^2 value. This meant the model was indifferent to the features of IQ, Projects_Completed, Internship_Experience, and Communication_Skills in determining CGPA. On the contrary, Prev_Sem_Result had the biggest impact on CGPA for when it was removed, R^2 dropped significantly more than the others. This told us that past academic results had the greatest impact on predicting future CGPA.

From the coefficient comparison chart, we also noticed that Prev_Sem_Result had the highest positive coefficient value. This meant that it contributed the most to increasing CGPA, while other variables such as Academic_Performance and Placement had smaller positive effects. Meanwhile, variables like IQ, Communication_Skills, and Projects_Completed showed very small coefficients, meaning they had little to no influence.

The Actual vs Predicted CGPA scatter plot showed almost a perfect straight diagonal line. This meant that the model's predictions were very close to the actual values, confirming our regression model to be highly accurate. Referring to the relationship between the input features, CGPA was well captured by the model, with Prev_Sem_Result being the most impactful variable in predicting CGPA.

Discussion

The findings clearly indicate that prior academic achievement (Prev_Sem_Result) serves as the indicator of a student's present CGPA. This is logical since students excelling in terms tend to sustain strong performance likely owing to steady study routines, self-discipline and grasp of their subjects. At the same time though, it's important to recognize that Prev_Sem_Result is also a student's previous CGPA, so the fact that CGPAs change little between semesters could've had something to do with the model being so accurate.

Although attributes such as IQ and Communication_Skills are commonly thought to significantly influence performance, their effect in this case was minimal, indicating that steady

academic effort and commitment serve much more reliable indicators of CGPA than overall intellect or interpersonal abilities. The Leave-One-Out analysis, however, demonstrated that the model's effectiveness was significantly influenced by Prev_Sem_Result alone. This was discouraging as it indicated the model was overly tailored to this variable and relied on it to generate precise predictions.

Overall, the regression analysis shows that CGPA can be predicted with very high accuracy using Prev_Sem_Result. The strong R^2 value of 0.96 proves that our model fits the data very well, though it may be learning too much from Prev_Sem_Result. Among all features, Prev_Sem_Result was the strongest driver of academic success, while other factors like IQ, internship experience, and projects provided smaller, less supportive contributions.

Question #2 (Logistic Regression): How well does a classification model predict whether a student gets placed after graduation, and which factor is most important in making that prediction?

Methods

The purpose of this question was to forecast whether a student will secure a job following graduation, considering their achievements and personal background traits. The primary variables utilized included: Placement (Y variable), CGPA, Prev_Sem_Result, IQ, Academic_Performance, Extra_Curricular_Score, Communication_Skills, Projects_Completed, and Internship_Experience.

Initially, we prepared the data by eliminating entries and transforming the categorical variable Internship_Experience into a numeric form where Yes was 1 and No was 0. Next, we applied Z-score normalization to all variables (including CGPA, IQ and Communication_Skills) to ensure each feature was on a consistent scale.

We then employed a Logistic Regression model because our outcome variable (Placement) was binary. A student being either placed or not placed, the data was divided into 80% for training and 20% for testing to assess the model's performance on the new unseen data.

To evaluate the model’s effectiveness, we computed these metrics:

- Accuracy: The frequency with which the model accurately forecasts placement.
- Precision: Among the students forecasted as “placed”, the count of those who were truly placed.
- Recall: Among all students who were truly placed how many were accurately detected by the model.
- F1 Score: An even metric that merges precision and recall.
- ROC AUC: A value ranging from 0 to 1 that indicates the model's effectiveness in distinguishing between placed and not placed students.

We depicted the outcomes with the help of two charts:

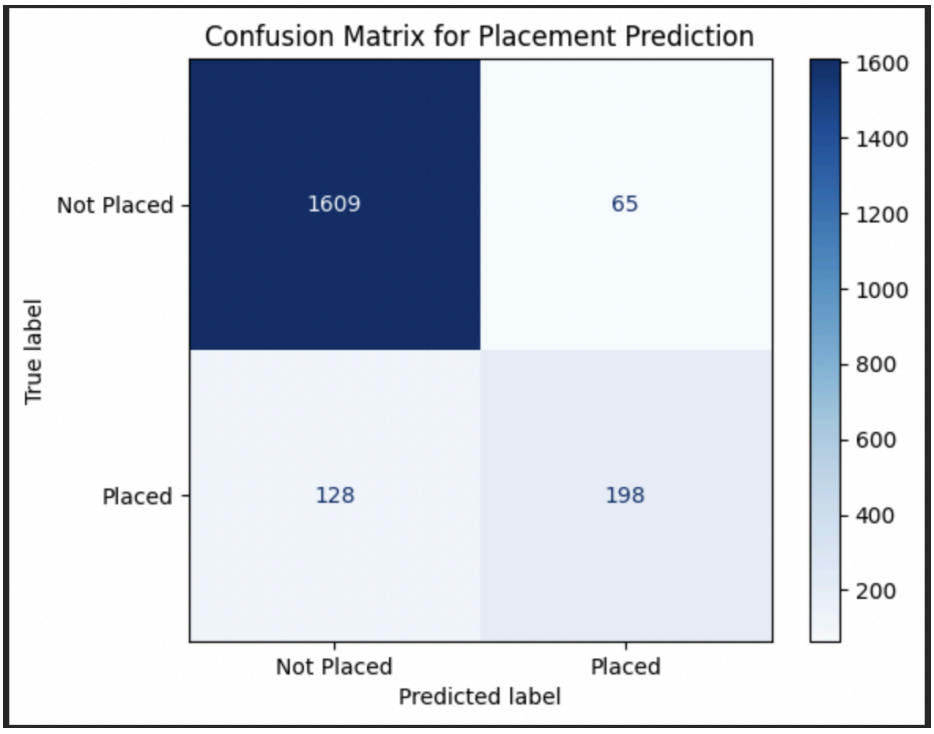

Figure 1 (Confusion Matrix): Feature Coefficients Impact on CGPA

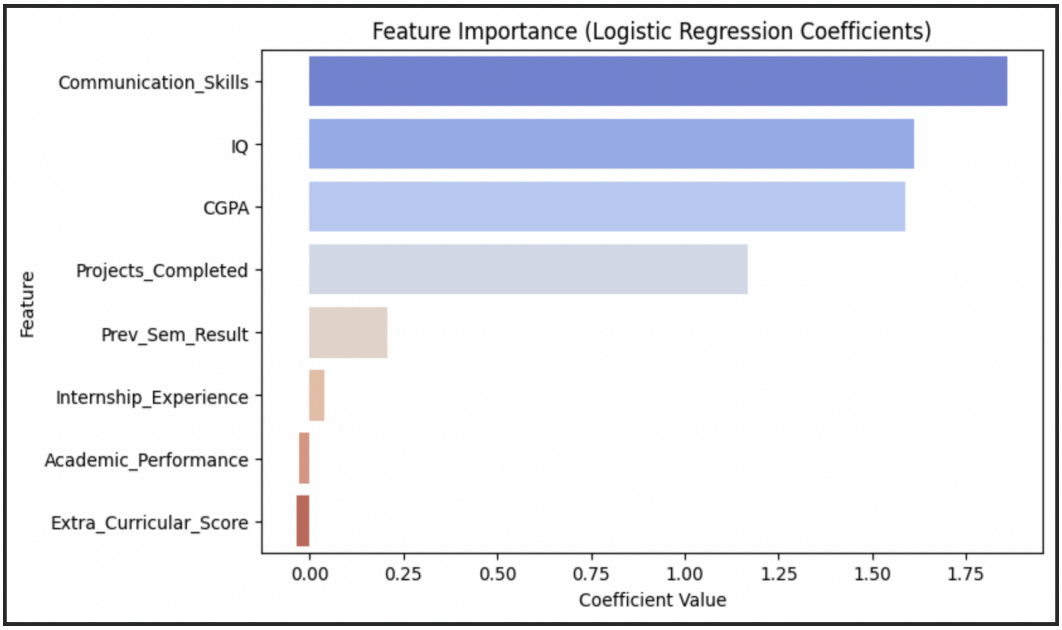

Figure 2 (Bar Chart): Feature Importance (Logistic Regression Coefficients)

Logistic Regression Model Performance Metrics:

	Metric	Training Set	Testing Set
0	Accuracy	0.9020	0.9035
1	Precision	0.7553	0.7529
2	Recall	0.6092	0.6074
3	F1 Score	0.6744	0.6723
4	ROC AUC	0.9405	0.9442

Figure 3: Exact performance metrics for the Logistic Regression model

Results

The Logistic Regression model attained a 90.35% accuracy on the testing dataset, indicating it was able to forecast student placement in most instances accurately. The confusion matrix indicated that the model accurately identified placed and non-placed students with just a small number of errors.

Based on the feature importance diagram, the four key factors influencing student placement were Communication_Skills, IQ, CGPA and Projects_Completed. Communication_Skills emerged as the dominant factor, demonstrating that students, with social and presentation abilities were more likely to achieve placement success. IQ also showed an effect suggesting that problem-solving and analytical abilities were elements in securing a position. CGPA remained the third most prominent factor, suggesting that consistent academic success and performance were still essential for employability. Projects_Completed also contributed by showing that hands-on experience helps students stand out when applying for jobs.

Other features like Prev_Sem_Result, Internship_Experience, Academic_Performance, and Extra_Curricular_Score, had smaller coefficient values, some of them negative, meaning they contributed less to the model's overall decision-making.

Discussion

The results suggest that both social skills and technical or academic proficiencies are essential in predicting a student's chances of obtaining post-graduation placement.

The significance of Communication_Skills as a key factor highlights that employers highly value communication, teamwork and presentation skills, rather than focusing exclusively on academic grades. Students who express their ideas clearly and cooperate effectively tend to perform well in interviews and professional environments.

Similarly IQ and Projects_Completed suggest that ability and practical experience influence hiring decisions. A high IQ supports thinking and technical assessments while project experience demonstrates real-world skills and problem-solving capabilities.

Although CGPA remains a marker, its reduced ranking compared to Communication_Skills and IQ suggests that depending solely on academic scores is inadequate for successful placement. Employers seem to desire a blend of academics, communication abilities and practical experience.

The model's precision and the clear patterns shown in the confusion matrix indicate that the logistic regression was effective for this classification task. Overall, the analysis shows that students most likely to get placed are those who combine strong communication skills, solid problem-solving ability, high academic performance, and hands-on project experience.

Question #3 (Clustering): What student groups can be formed based on IQ, Prev_Sem_Result, and CGPA, and how do these groups differ?

Methods

The goal of this question was to understand patterns in student groups, combining practical abilities with grades to view any associations between real-world skills and academic expectations.

After dropping nulls and resetting indices, the factors for IQ, Prev_Sem_Result, and CGPA were plotted on three separate graphs a couple of times to view their relationships. Comparing IQ vs. Prev_Sem_Result, Prev_Sem_Result vs. CGPA, and CGPA vs. Prev_Sem_Result, we attempted to distinguish obvious patterns by shifting point frequencies in the two sets.

The first set of the 3 scatterplots including all 10,000 original data points, we figured initially that it was going to be difficult to see shapes forming considering the thousands of data points overlapped between each other. Taking a more elementary approach, we thus created a second set of scatterplots sampling only the first 500 data points to see if we could spot clearer boundaries with more whitespace in the graphs. This approach worked marginally as we were able to distinguish multiple cluster groups, but we still didn't have enough information to determine the appropriate k value.

Considering the clusters we were able to see were relatively oblong shaped, however, we moved forward with deciding GMM (Gaussian Mixture Modeling) as our final clustering method before model fitting. Providing flexibility for choosing k in soft assignment without worrying about overlapping data points, circular boundaries, or data linkage, we saw it as the most effective technique for grouping the variables of IQ, Prev_Sem_Result, and CGPA. To determine k, we thus went ahead with z-scoring the continuous factors and using a unique method for GMM clustering, plotting the Bayesian Information Criterion (BIC) scores for different k values.

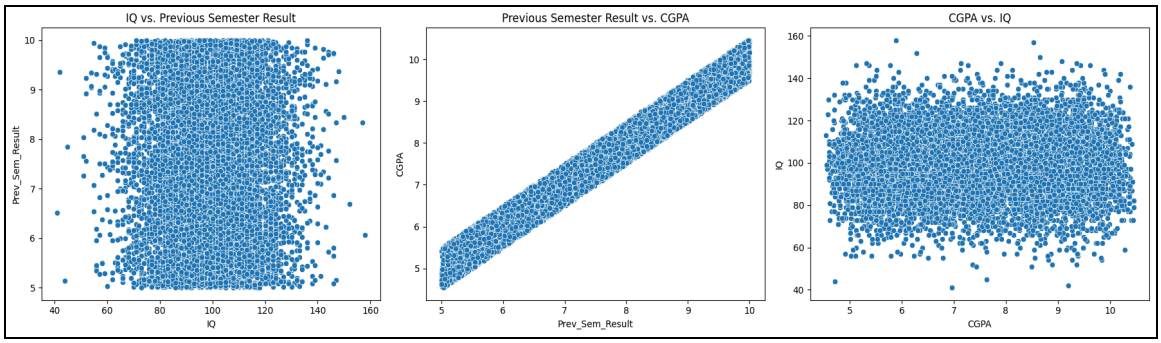

Figure 1: A series of scatterplots displaying the relationships between IQ, Prev_Sem_Result, and CGPA
(no obvious patterns showing)

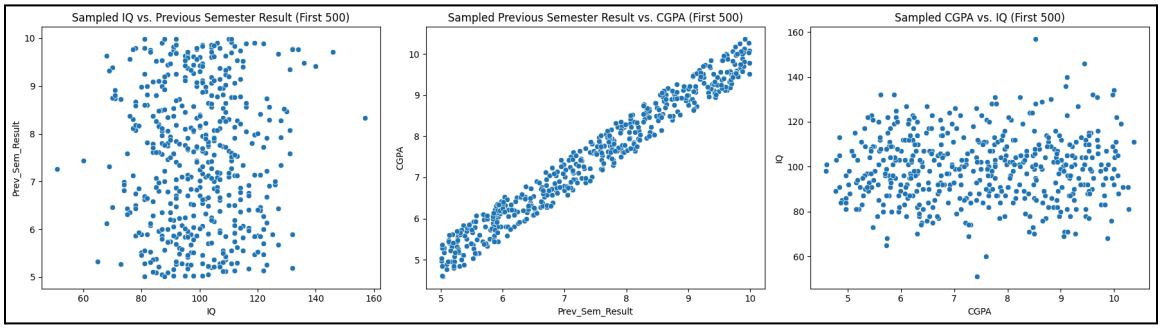

Figure 2: A series of scatterplots sampling the first 500 data points between IQ, Prev_Sem_Result, and CGPA
(some patterns showing but not enough to set an accurate k by ourselves)

Results

The results of GMM through the k-clustering BIC method ended up being moderately successful, both in unsupervised accuracy and conceptuality. In BIC plots, it's common for the minimum value with the lesser complexity to be chosen as the correct number of clusters in a dataset. According to the BIC plot for IQ, Prev_Sem_Result, and CGPA, this number was determined to be 5. We in turn fit a GMM model with 5 n_components to visualize final clusters and determine group characteristics.

Showcasing unique areas of student studies, the 5 clusters ended up being color coded in our original scatterplots to verify patterns and classify undergraduate group performances. All in all, clusters 0, 1, and 2 ended up dominating the upper halves of all three scatterplots, with clusters 3 and 4 presenting the lower halves, splitting student academic study habits into two main halves each with their own specialized subgroups explained in the discussion.

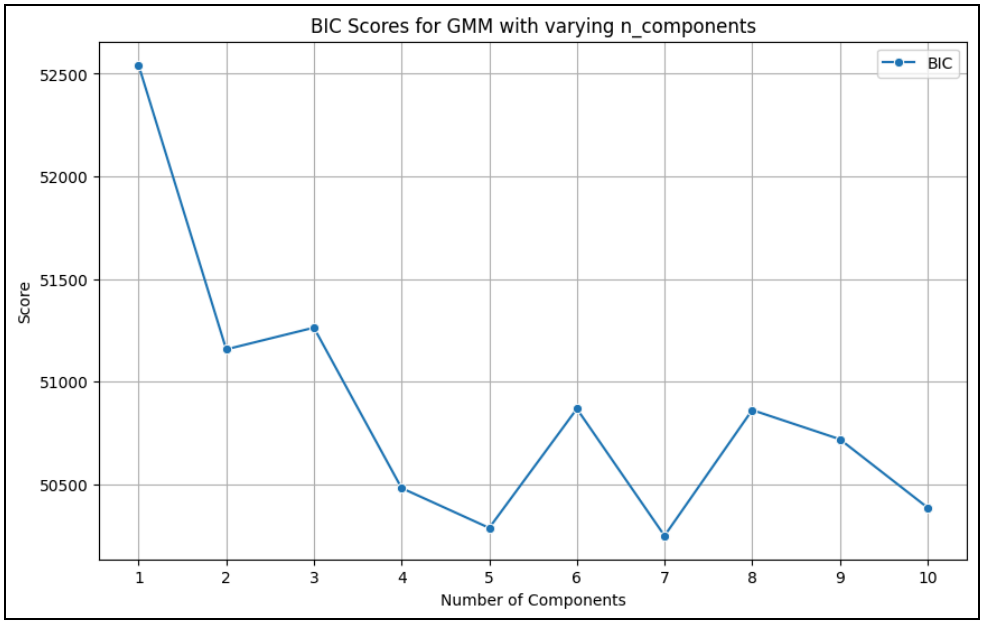

Figure 3: The BIC (Bayesian Information Criterion) Plot comparing scores vs. the number of k clusters (7 was the minimum but 5 was chosen for better simplicity)

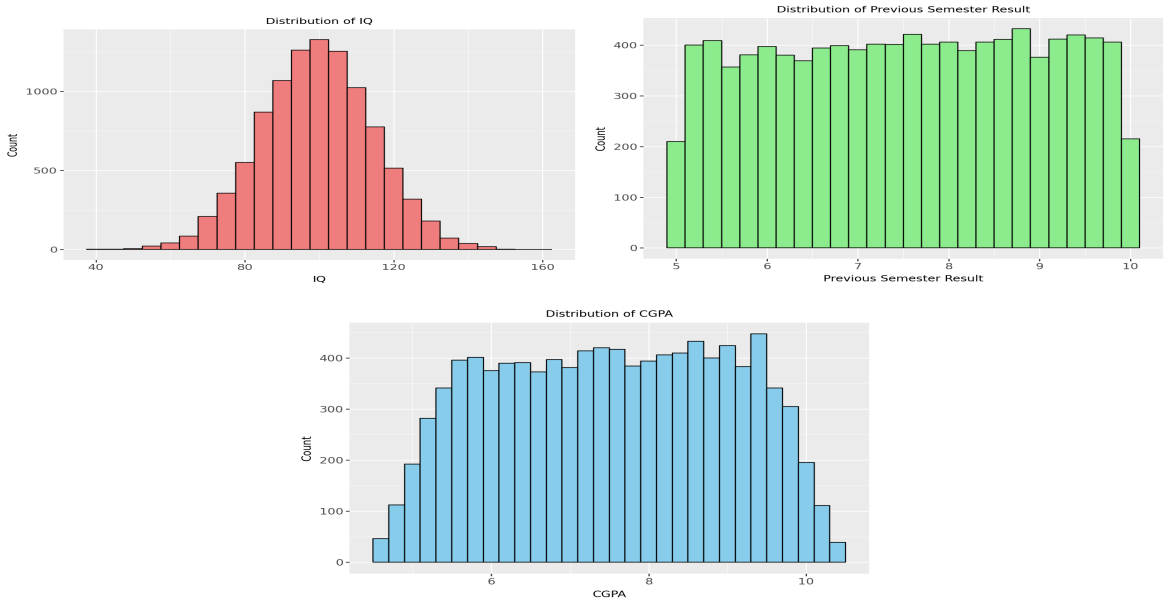

Figure 4: Ggplot histogram distributions of IQ, Prev_Sem_Result, and CGPA (all variables are normal, which validates using GMM for clustering)

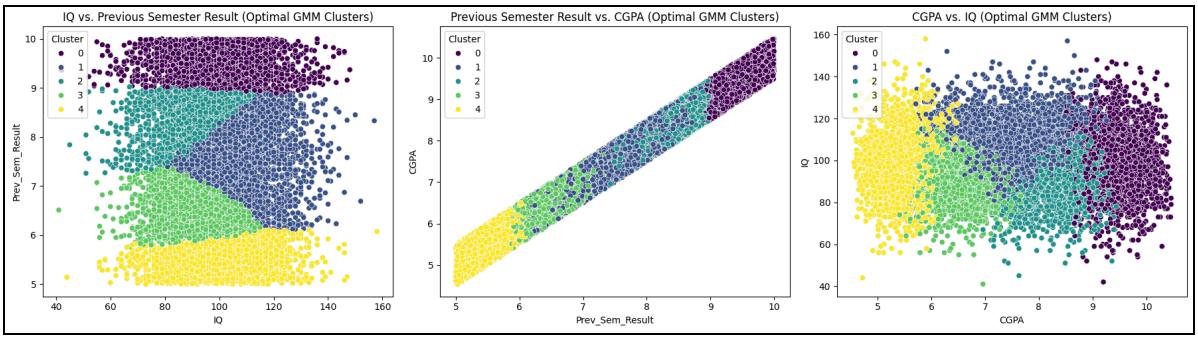

Figure 5: The results of the 5 GMM clusters (0-4) on the original scatterplots

Discussion

The clusters calculated from GMM (labeled as clusters 0-4 in the scatterplots) show five distinct groups students may be classified under based on their IQ, Prev_Sem_Result, and CGPA scores.

The first cluster (Cluster 0) shows the most advanced performing students overall. Having the highest recorded Prev_Sem_Result and CGPA scores (9.49 and 9.48 averages), they are consistently topping the results as the purple group in the scatterplots for IQ vs. Prev_Sem_Result and Prev_Sem_Result vs. CGPA. Despite having mixed IQ scores (99.58 avg.), their class determination is likely the strongest through mass dedication to school and coursework, labeling them as the *High Achievers*.

The second cluster (Cluster 1) also having good Prev_Sem_Result and CGPA scores, their best strengths are found in their IQs (109.54 avg.). Seen as the blue group in the scatterplots, though they are moderate in class performance, they are seen to have the most advanced cognitive abilities nonetheless. Reflecting this group to have superior problem-solving and reasoning skills from their intelligence quotients, they are viewed as the *Gifted Performers*.

Cluster 2 as the sky or light blue in the scatterplots also shows promise referencing both above average Prev_Sem_Result and CGPA scores (8.40). While struggling a little in the IQ (88.52 avg.) section, this shows that their abilities lie in solid academic standing from dedicated efforts to effective learning strategies and natural fits to their class studies. Viewed as the second group next to the *High Achievers* (Cluster 0) to earn high grades in the semester, these students are classified as the *Exceptional Achievers*.

While the first 3 clusters identify outstanding performers, the other 2 clusters work to summarize the results of the remaining students. Cluster 3, the green areas in the scatterplots,

are seen as learners that are staying within the average bounds of Prev_Sem_Result and CGPA (6.49 avg.). And while not exceeding expectations in their IQ (91.05 avg.) and report cards, they are still viewed with potential to contribute to their studies through proficient scores in their classes. So all in all, by not excelling but still meeting college requirements, Cluster 3 students can overall be viewed as the *Moderate Performers*.

Cluster 4 students (the yellow group), however, are observed to run more of a risk in their education with below average Prev_Sem_Result and CGPAs (5.47 avg.), indicating lower performance in their classes. Interestingly enough, these students are also observed to achieve exceptional IQ scores (100.79 avg.), labeling them second to the *Gifted Performers* (Cluster 1) in terms of practical skills and real world application. Their grades not reflecting their capabilities, this indicates the presence of external factors affecting their learning. Events like demotivation, poor study habits, weak time management, family responsibilities, jobs, and other challenges could be influencing their poor performance, and it's important these students get the resources they need to aid in difficult times and personal dilemmas so they can focus more on their studies. For now, we'll refer to them as the *Straggling Achievers*.

Colleges with these groups in mind may be able to take several actions to help each student class. Monitoring the actions of *High Achievers*, *Gifted Performers*, and *Exceptional Achievers*, universities may target these learners with further opportunities to showcase their skills in professional environments. Inviting and encouraging these students to take part in student jobs, graduate programs, and internships, these groups are likely to benefit the most from expanding their resume portfolio and preparing for the job market post-graduation.

For *Moderate Performers*, while they may also participate in these events, it's best to first focus on their individual strengths in specialized hands-on projects, clubs, research, and networking opportunities. Building their experiences in extra curricular activities and connections, it's important for this group to lift their professional abilities so they may compete successfully with those of higher CGPAs when applying to future positions.

Finally, for the *Straggling Achievers*, though it's important they also apply to jobs, the main priority of these students is getting their academic studies up to speed. Universities offering these students more connections to departments like counseling, tutoring sessions, studying workshops, and special accommodations, this group may benefit from reaching out to school programs and other services. Discussing their personal delays, university staff will ensure these students get the support and materials they need to succeed in class.

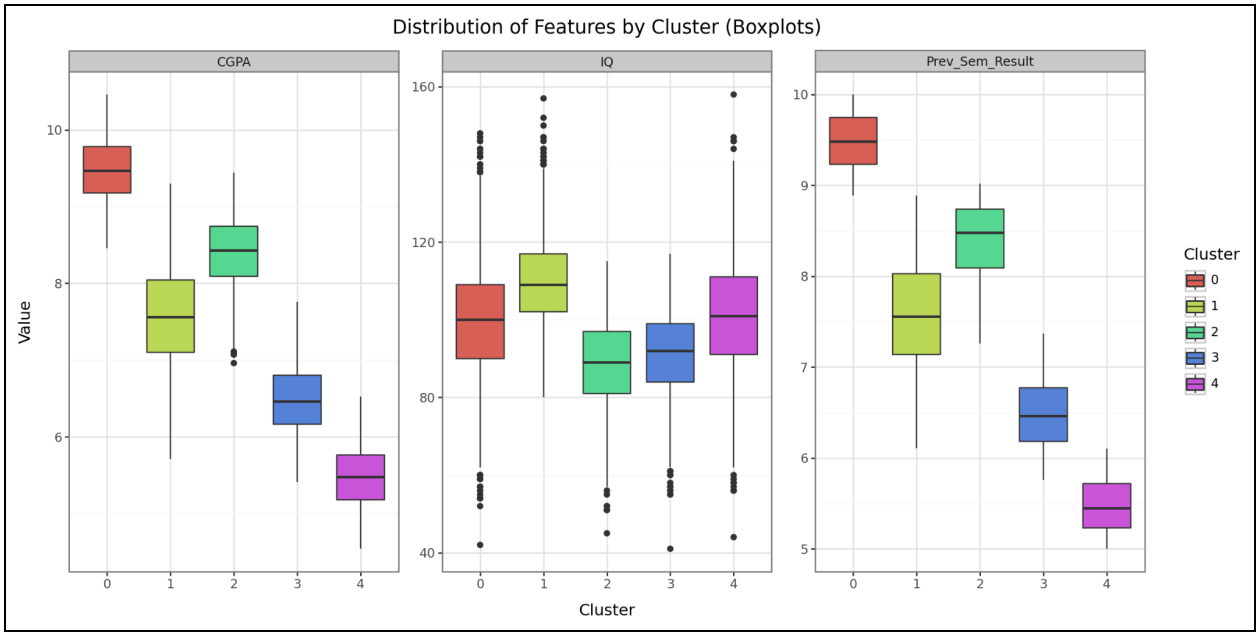

Figure 6: A series of ggplot boxplots showing average distributions between each cluster

Mean Values For Cluster	IQ	Prev_Sem_Result	CGPA
0	99.584651	9.486386	9.475209
1	109.535242	7.571821	7.561574
2	88.515894	8.396881	8.402477
3	91.050177	6.485525	6.491222
4	100.789843	5.473793	5.474673

Figure 7: A table displaying exact mean values for variables in clusters 0-4

Question #4 (Dimensionality Reduction): Can we use PCA to simplify the features when predicting Internship_Experience, and how many components do we need to keep the model's accuracy similar to using all features?

Methods

The main purpose of this question was to see if dimensionality reduction could offer unique insights to a classification model without overcomplicating it.

Dropping nulls, categorical variables for Internship_Experience and Placement were one-hot encoded into binary 1/0 values for their respective “yes/no” responses. Z-scoring the continuous intervals, we decided from here to create two models geared toward classifying student Internship_Experience using all other variables.

The first model fit was a Logistic Regression using an 80/20 train-test-split, evaluating each variable’s coefficients for the log odds probability of predicting whether a student had an internship or not. We felt it necessary to create this baseline model to serve as a comparison for features and accuracy in the PCA model, though to build PCA the process was a little more complex than originally anticipated.

An essential beginning step in principal component analysis (PCA) being to build the linear combinations, we initialized this step by streamlining our scaled training data into sklearn’s PCA() function. From here, to determine how many components we needed for the model, we used scree and cumulative variance plots to evaluate the number of PCs at which both the explained variance dropped significantly (elbow point) and where cumulative variance explained at least 95% of the data.

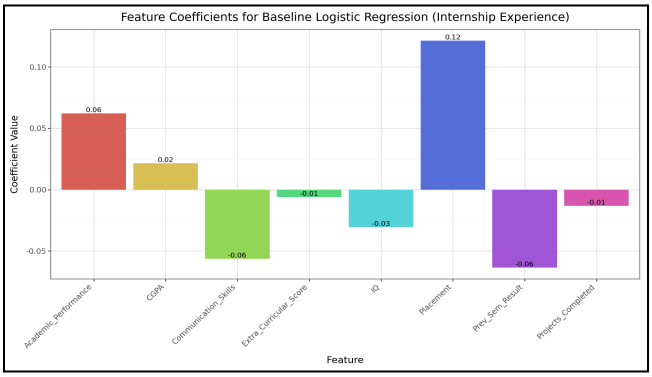

Figure 1: A bar chart comparison for logistic regression coefficients
(log odds for each individual variable appear initially simplistic to understand)

Results

Observing the results of the scree and cumulative variance plots, the optimal number of principal components was determined to be 6. Every other PC after PC6 explaining very little

variance, we knew at this point in acknowledging the problem that only the first six PCs were going to be used to answer the question. Though because the PCs themselves were linear combinations of all other variables, it was difficult to gauge originally how this helped classify Internship_Experience.

So to level the playing field, we fit another Logistic Regression model using the same 80/20 train-test-split with the same standardized variables, except this time, we used the six principal components to classify the value of Intership_Experience. It was from here that we built metrics tables, ROC AUC curves, and charts for coefficient and accuracy analysis to compare the performance of both models and determine PCA effectiveness.

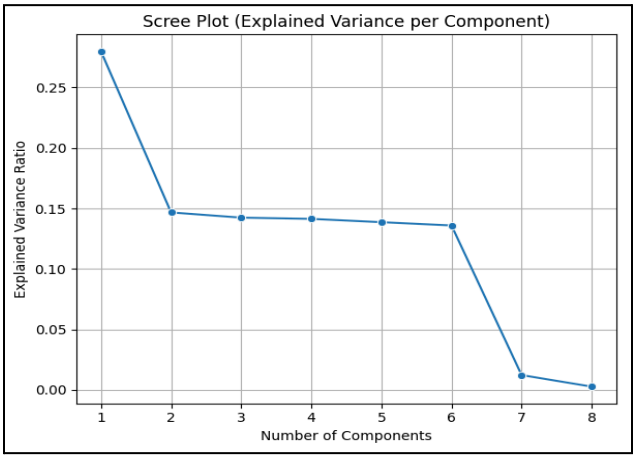

Figure 2: A scree plot showing each principal component's explained variance in the data (after six PCs, the seventh PC explains very little, so 6 is the elbow)

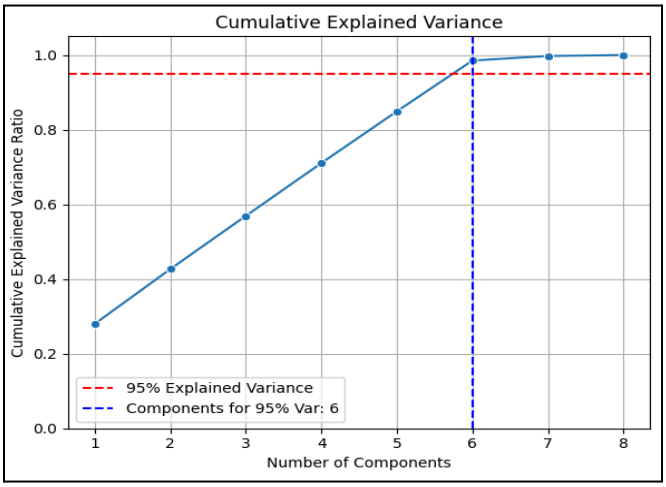

Figure 3: A cumulative variance plot showing the sums of explained variance in the data

(6 PCs explain at least 95% of the data)

Discussion

Answering the question, the number of principal components we needed to keep the PCA's model accuracy similar to using all features in the original Logistic Regression was 6. With 6 PCs in mind, the testing accuracies of both models ended up matching at 60.35%, though achieving testing ROC AUCs of 0.50 and 0.51 for baseline and PCA randomness respectively, these metrics showed weak overall predicting power for determining Internship_Experience.

The baseline log and PCA models referencing the confusion matrix were strongest in classifying true negative cases for those who didn't have internships (1207), whereas for those with internships, they ended up classifying 0. This resulted in zeros for precision and recall in both training and testing metrics in the sense that no true positives were found. F1 scores being a combination of the two previous metrics, this also ended up being 0.

As the PCA model demonstrated weak predicting power, its principal components also inherently did nothing to simplify features when fitted into a Logistic Regression. PC1 dominated by positive loadings in CGPA and Prev_Sem_Result, this clashed with PC3's positive loadings in Communication_Skills and Extra_Curricular_Score. PC2 combining Academic_Performance with positive influences in Communication_Skills, PCs 4 and 6 contrarily showed strong negative influences in Communication_Skills. Features like IQ, Extra_Curricular_Score, and Projects_Completed that are normally considered core skills in earning an internship were also seen to have their fair shares of negative loadings in a majority of PCs, showing limited consistency in what factors went into being a qualified intern.

In sum, with the PCA model showing poor accuracies in zero recall and precision scores, this method was proven ineffective at simplifying features when classifying Internship_Experience. The ROC AUCs being along the boundary presented a loss for colleges implementing the model to classify students who had prior internship experiences. A score of 0.5 indicates the model at this point is random guessing, and simply flipping a coin between internship and no internship, this is resulting in no one having prior internship experience. Conceptually, this is highly inaccurate, showing the model fails entirely in understanding real life scenarios.

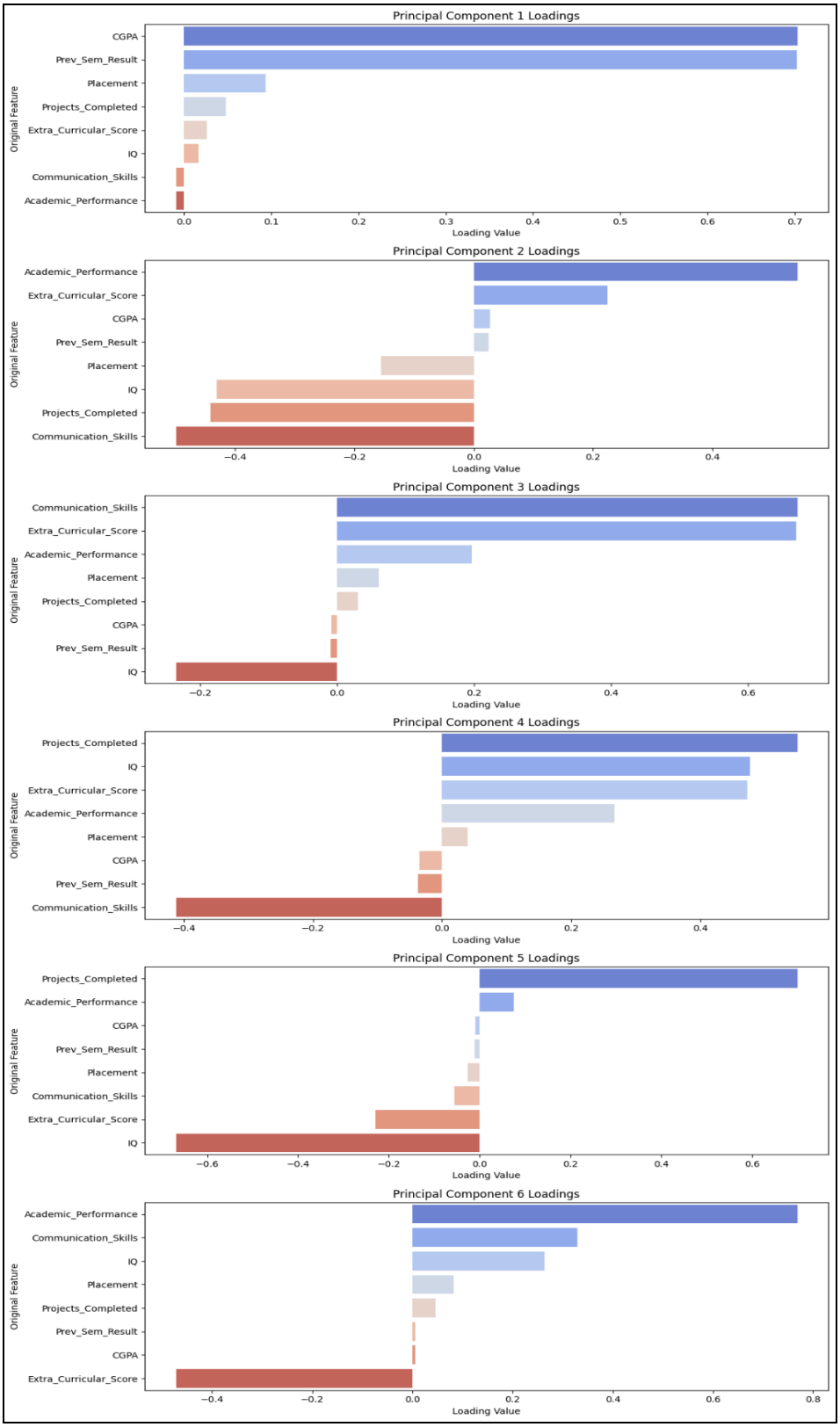

Figure 4: A series of bar charts showing loadings for the 6 principal components chosen in predicting Internship_Experience

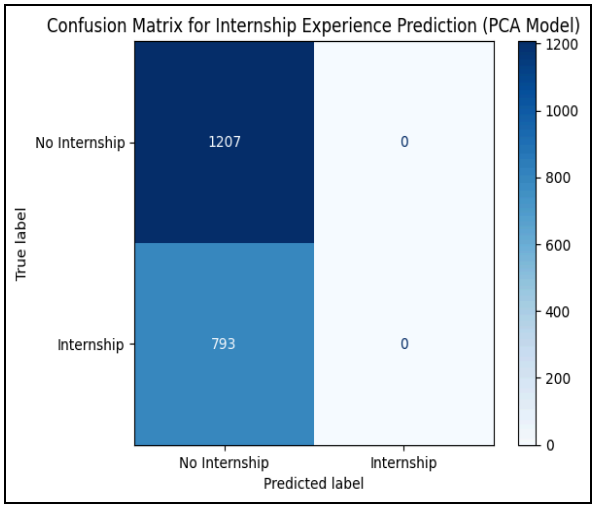

Figure 5: A confusion matrix displaying the testing results of the PCA model
(PCA is strongest at predicting true negatives (Dark Blue), but predicts 0 true positives for Internship_Experience (bottom right corner))

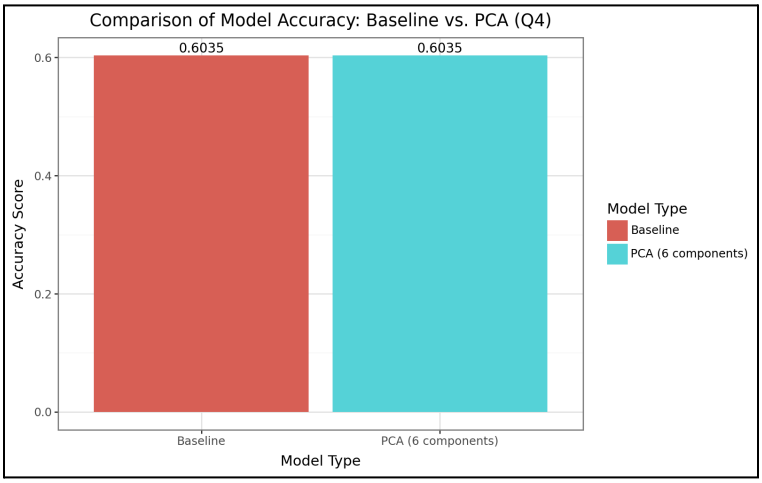

Figure 6: A ggplot bar chart comparison of Logistic Regression (Baseline) vs. PCA testing accuracies
(accuracies are similar, but they are very poor)

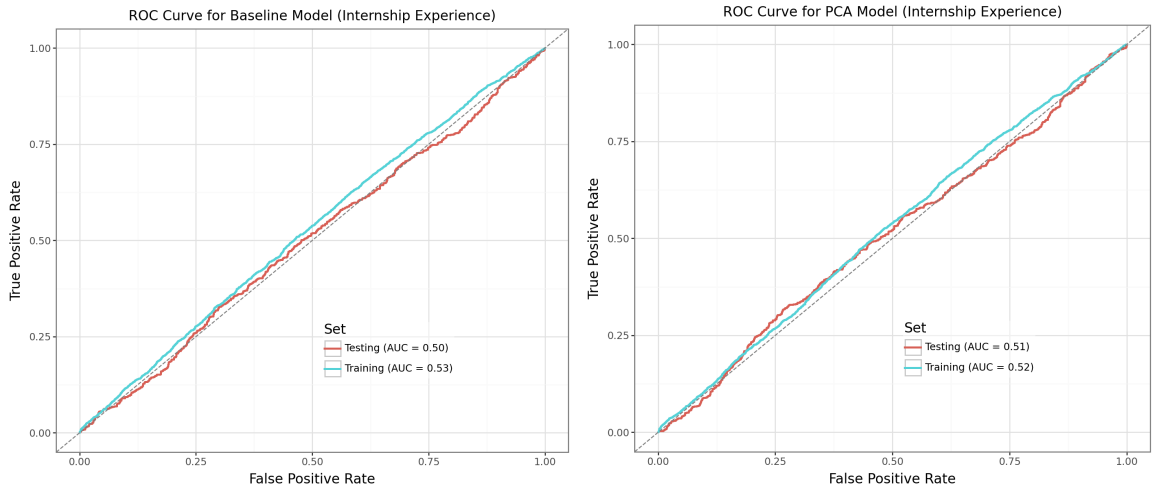

Figure 7: The ROC AUC training and testing curves for Logistic Regression (Baseline) vs. PCA (all curves are ~0.5, meaning they have weak predictive power and are random guessing for Internship_Experience)

	Metric	Training Set	Testing Set
0	Accuracy	0.6036	0.6035
1	Precision	0.0000	0.0000
2	Recall	0.0000	0.0000
3	F1 Score	0.0000	0.0000
4	ROC AUC	0.5254	0.5013

Figure 8: A table displaying exact performance metrics for the Baseline Logistic Reg model

	Metric	Training Set	Testing Set
0	Accuracy	0.6036	0.6035
1	Precision	0.0000	0.0000
2	Recall	0.0000	0.0000
3	F1 Score	0.0000	0.0000
4	ROC AUC	0.5231	0.5104

Figure 9: A table displaying performance metrics for the PCA model