UNCOVERING THE KEY DRIVERS OF CAMPUS PLACEMENT SUCCESS

Introduction

Campus Placement involves participating, identifying, and hiring young talented people for entry level positions in the job environment. Kesavaraj, G., & Pattnaik (2012) define it as a process whereby organizations meet and select intelligent and committed youth from various colleges and institutes who have the enthusiasm to prove and better themselves. Over the years, the need for talented and self-motivated people who can work tirelessly and with commitment has grown Bhargavi, S., & Yaseen, A. (2016). According to the National Association of Colleges and Employers (2023), close to two thirds of employers now use skill-based hiring practices for new entry-level hires. Other selection methods used by employees include interviews, aptitude tests, presentations, written assessments and many more. The higher demand being placed on graduates possessing the attributes render them to be work ready and have significant implications for graduate and selection practices. (Cabalero, C, & Walker, A, 2010). With this knowledge, it is important that institutions and students understand what places them in a better position for campus placement.

Project Motivation

Choosing the 'Factors Affecting Campus Placement' dataset highlights the roles played by higher institutions in preparing students for the job market. Knowing what influences Campus placement will help institutions to improve both their academic and on-academic to help students transition from school to work.

The dataset sourced from Kaggle offers comprehensive data on academic, demographic, and professional factors influencing campus placements. Key variables such as secondary school and higher secondary school percentages, degree specialization, and salary provide an opportunity to analyze both academic and non-academic factors enabling a holistic understanding of placement success. Beyond academics, skills such as communication, teamwork, and adaptability contribute significantly to campus placement outcomes as they foster collaborative and dynamic work environments (Andrews & Higson, 2008).

Since there is an organizational demand to hire the right kind of talent, analyzing this dataset will provide insights into what increases the chances of successful campus placements. According to a report from the National Association of Colleges and Employers (NACE, 2020), around 55-60% of students who had completed their internships were offered full-time positions due to their development of industry-relevant skills, making them more attractive to recruiters. By analyzing the factors influencing campus placement, this study aims to bridge the gap between students' preparation for jobs and industrial requirements. Educational institutions can better prepare students for the job market, thus benefiting both the employers and the fresh graduates.

Objectives

Understanding how academic performance impacts campus placement. This seeks to explore how high or low the academic performance on secondary, higher secondary and degree correlates with campus placement.

The influence Internship has on Campus Placement. Does prior experience in internship increase, decrease or have no impact on whether a student is placed.

To develop a classification model predicting placement success based on academic and non-academic factors.

Literature Review

Role of Academic Performance in Campus Placement Success According to York, Gibson, and Rankin (2015), student success encompasses 'academic achievement, satisfaction, acquisition of skills and competences, persistence, attainment of learning objectives, and career success'. Continuous excellent academic performance reflects a student's intellectual capacity, work ethic, and discipline. McCann and Hewitt (2023) found that higher academic performance significantly influences a student's ability to secure work placements. Specifically, students excelling in their first year in Business and Economics had higher chances of campus placement.

Furthermore, students who utilized school resources, such as libraries and tutoring services, performed better academically. de Araujo and Murray (2010) demonstrated that access to university-provided resources positively affected academic performance, which in turn improved placement outcomes. This highlights the importance of institutional support in enhancing students' academic and career success.

Impact of Internship experience on Campus Placement Internships play a critical role in determining career success as they bridge the gap between academic and demands of the workplace by offering students exposure to the real-world challenges in the employment sector. According to Galbraith, D., & Mondal, S. (2020), there is a high correlation between internships and job placement. Students who complete internships are more likely to get job offers according to the NACE 2020. Over 81% of the new graduates' state that their experiences from their internship helped them adjust their career plans.

Research done by Hart (2008) with 301 organization showed that despite employers being relatively satisfied with the entry level skills of the graduates, they were not confident that the graduate had what it took for advancement and promotion in the organization. This shows that having an internship would give the graduates a competitive edge as they start applying theoretical knowledge to the real world, gaining soft skills and discovering skills that you need to develop Ellis, C. et al. (2017).

According to Manjunath, D. R. (2020), acquiring the first-class and having the knowledge for the relevant career in the job specification is not enough, you will be competing among others who have similar qualifications. Therefore, the unique mix of interpersonal skills and personal qualities are what makes you stand out in a crowd of all academics. Therefore, achieving this milestone is not only about getting a job but also obtaining skills and attributes that will enable you to be successful throughout your working life.

Academic performance and internships are key to campus placement success as they showcase discipline and give students discipline and adaptability. Real-time exposure will prepare graduates to meet the employers' expectations. Future studies should explore additional factors like mentorship programs, co-curricular activities, and non-academic skills.

Data Description

A dataset named “Factors Affecting Campus Placement,” from famous database website Kaggle is used in this project which contains interesting categorical and numerical attributes related to the placement of students in the workplaces. The target variable is the status which is a categorical variables representing placed or not placed, other key variables are mentioned in the table in the appendix.

Methodology

The analysis starts with data collection from the Campus Recruitment dataset published on Kaggle which is loaded to RStudio for preprocessing and exploration.

During the data preprocessing, the data was checked and being handled its duplicates and missing values, and irrelevant columns are dropped to ensure data quality.

This is followed by Exploratory Data Analysis (EDA) to discover the trends and relationships using statistical summaries and visualizations (such as bar charts, boxplots and correlation heatmap) to have better understanding the placement distribution and its association with academic preferences, work experiences, MBA specialization and other key factors.

Next, after obtaining better understanding of the dataset, further preprocessing is conducted, including handling outliers (using IQR method) and perform feature engineering where it creates new variable, education_score (weighted average of SSC, HSC, and degree percentages, with weights of 0.3, 0.3 and 0.4 respectively). Categorical variables are encoded into numerical formats to ensure compatibility with the models. Next, feature selection is being performed to identify the most impactful variables to be used in the models. Prior performing the model, the target variable will be split into training (70%) and testing (30%) to ensure accuracy and fair outcomes.

In the Modeling Phase, logistics regression is used as the primary model with forward feature selection adding variables incrementally: education score, work experience, MBA Specialization and high school streams. The models are then evaluated using metrics like accuracy, precision, recall, AUC-ROC and VIF.

In the last stage, the Result Interpretation Phase involves analyzing logistics regression coefficients to assess the impact of predictors like academic scores and work experience on placement likelihood. Visualization such as ROC curve, confusion matrix and feature importance charts guide the selection of the best performing model, providing actionable insights into placement success factors.

Data Preprocessing

Data Cleaning

Preprocessing began with data cleaning to ensure accuracy and reliability for analysis where columns with unique or null values were identified, with no missing or special character issues detected and next categorical variables, like gender, work experience, and status, were validated for consistent levels, ensuring compatibility for machine learning models. Overall, this preprocessing step ensured the dataset was well-structured, reliable, and ready for further analysis.

Handling Missing Values and duplication

Next missing values in the dataset was identified, where just one feature, “salary” had 67 missing values. After careful observation these missing values are not placed, therefore the messing values were imputed using a 0 and the outcome indicated 0 missing values for all the columns. This dataset did not have any duplicate values.

```
## Variables with missing values before handling:

## salary
##     67

##
## Variables with missing values after handling:

## named numeric(0)
```

Exploratory Data Analysis (EDA) for preprocessing

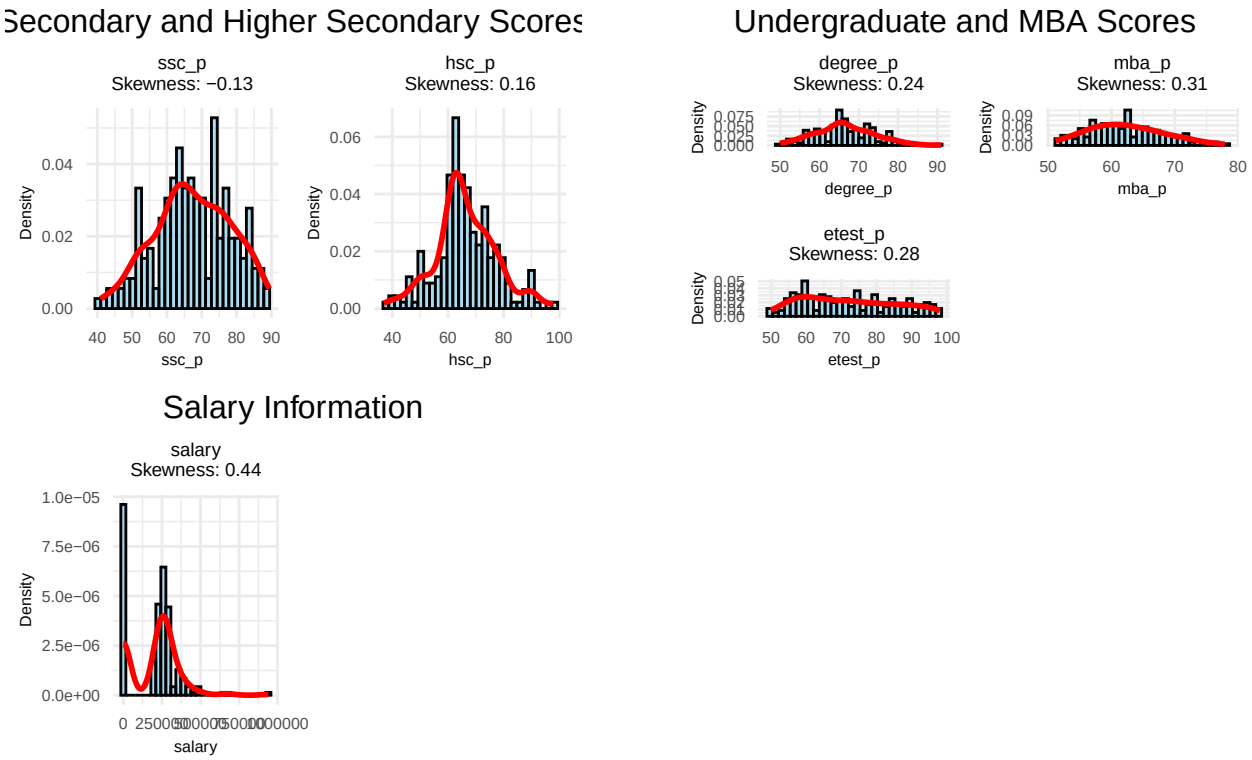
Basic Statistics for Numerical Variables

```
##      sl_no           ssc_p           hsc_p          degree_p        etest_p
##  Min.   :  1.0   Min.   :40.89   Min.   :37.00   Min.   :50.00   Min.   :50.0
##  1st Qu.: 54.5   1st Qu.:60.60   1st Qu.:60.90   1st Qu.:61.00   1st Qu.:60.0
##  Median :108.0   Median :67.00   Median :65.00   Median :66.00   Median :71.0
##  Mean   :108.0   Mean   :67.30   Mean   :66.33   Mean   :66.37   Mean   :72.1
##  3rd Qu.:161.5   3rd Qu.:75.70   3rd Qu.:73.00   3rd Qu.:72.00   3rd Qu.:83.5
##  Max.   :215.0   Max.   :89.40   Max.   :97.70   Max.   :91.00   Max.   :98.0
##      mba_p           salary
##  Min.   :51.21   Min.   :     0
##  1st Qu.:57.95   1st Qu.:     0
##  Median :62.00   Median :240000
##  Mean   :62.28   Mean   :198702
##  3rd Qu.:66.25   3rd Qu.:282500
##  Max.   :77.89   Max.   :940000
```

The dataset provides crucial information such as students’ academic performance, test scores, and placement outcomes which provide insights into the which factors that influence campus placement success. Ssc_p and hsc_p scores exhibit a similar pattern where the average scores are 67.33% and 66.33% respectively and median of 67% and 65% respectively. The same pattern can also be seen for the degree scores (degree_p) and MBA scores (mba_p) where the mean is 66.37 and 62.28% while the median is 66% and 62% respectively. However, the employment test (etest_p) scores demonstrate strong performance, with an average of 72.1% and median of 71%.

Placement outcomes, as measured by salary data, reveal a significant variability ranging from minimum of 0 (unplaced student) to a maximum of 940,000. The average salary of placed students is 198,702, while median salaries are 240,000. This variability suggests that factors beyond academic such as employment experience may possibly influence the placement success. Overall, these trends offer valuable insights into the factors influencing campus places successfully.

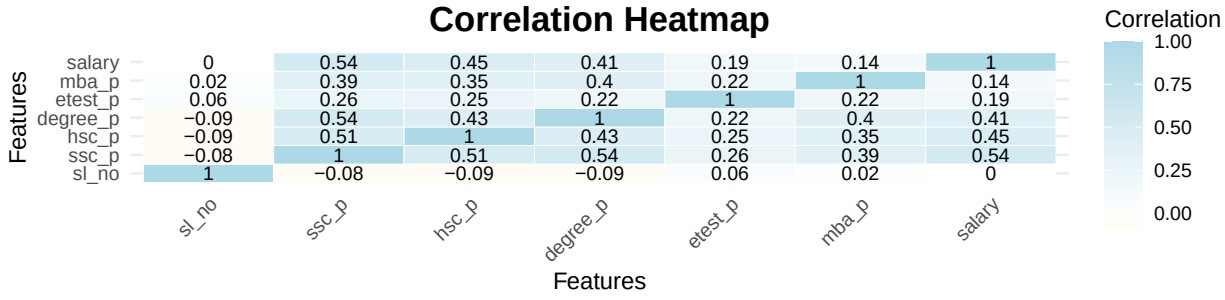
Distribution Overview of Numerical Variables



Academic scores for higher secondary (hsc_p), undergraduate (degree_p), MBA (mba_p) and employment test (etest_p) have slightly positive skewed (0.16, 0.24, 0.31 and 0.28 respectively) which indicates most students performed moderately to well, with fewer students scoring very high or very low.

On contrast, secondary scores (ssc_p) are slightly negative skewed which means more students scored higher compared to those scores low. This suggests at this level, most students have a strong academic foundation.

The salary distribution also shows a positive skewness with 0.44 suggest that most students have lower salaries, with a smaller group earned significant higher amounts. The highest amount is zero representing those unplaced, while a peak around the median salary for placed students. This highlights the variability in placement outcomes and the importance of factors influencing the salary.



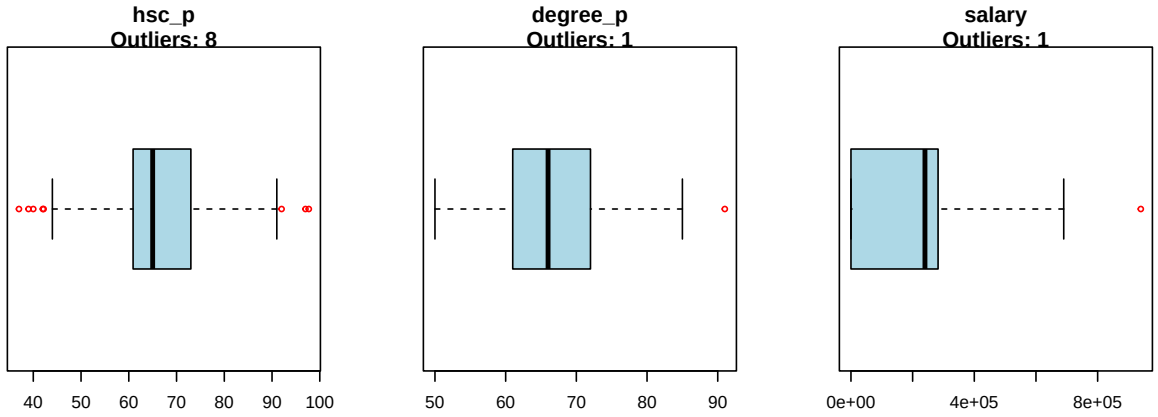
Correlation heatmap matrix indicates that the relationship between salary and SSC is the highest correlation with 0.54 followed by HSC and degree (0.45 and 0.41 respectively indicating strong academic performance tends to be associated with higher salary outcomes. MBA score shows a weak correlation of 0.14 which suggests that academic performance at this stage plays a small role in the placement. On the other hand,

employment test also reflects weak correlation of 0.19 indicates that this test may not have significant impact on the salary outcomes.

The heatmap of ANOVA P-Values (figure shown in Appendix B) illustrates the relationship between numerical and categorical variables. It can be observed that status has a strong relationship with SSC, HSC, degree and salary where p values are 0 which means that high academic performance at those stages are likely to secure placement status. On contrast, MBA and employment test shows weaker relationships with 0.2614 and 0.0617 respectively indicates that these qualifications may not significantly impact on the placement status. This heatmap further strengthen the statement in correlation matrix.

Handling Outliers

Handling Outliers During the Basic Statistics and EDA process the histogram distribution for numerical variables, we identified some outliers in the numerical variables , this was further identified using the Interquartile Range Method (IQR), whereby formular is Lower Bound = Q1−1.5×IQR and Upper Bound=Q3+1.5×IQR and any values that follow greater than upper bound and lower than the lower bound was identified as outliers and count is displayed and visualised using boxplots.




```
## No outliers detected in the dataset.
```

Outliers for the detected columns were handled using the IQR methods where any value below the lower bound is capped at the lower bound and any value above the upper bound is capped at the upper bound where this method prevented any data lost. Then the remaining outliers were detected where the count indicates absence of outliers and reconfirmed through the box plot visualizations (see the visualisation in apppendix for full detail).

Feature Engineering A correlation heatmap created during the EDA process revealed high correlation between the education metrics hence the need to create a new feature called education_score was identified and this were created by combining 3 variables namely ssc_p. Hsc_p and degree_p using a weighted formula. education_score = 0.3 × ssc_p + 0.3 × hsc_p +0.4 × degree_p.

Encoding This process involves converting categorical variables to a numerical format. This was achieved by first applying Binary encoding to binary categories which were ‘status’, ‘gender’ and ‘work_experience’

and dummy variables that create separate columns for multi-category variables for ‘specialisation’ and ‘hsc_s’ columns was used.

Feature Selection The feature selection process is focused on identifying the most relevant variables to ensure interpretability and model effectiveness. Correlation matrix revealed strong relationships between the academic scores leading to the creation of education_scores which captures the overall performance. Salaries and mba_p are critical for understanding the relationship between academic and placement outcomes. Features such as specialisation, hsc_s and workex revealed patterns that could influence placement success when earlier visualized in the distribution of categorical variables.

```
##  [1] "education_score"        "work_experience"        "specialization_Mkt_Fin"
##  [4] "specialization_Mkt_HR"  "edu_work_interaction"   "hsc_science"
##  [7] "hsc_commerce"           "hsc_arts"               "salary"
## [10] "mba_p"                  "status"
```

The interaction between academic performance and work experience revealed and interaction term edu_work_interaction which captures the combined impact of the factors on placement success. Considering the correlations between the features and target variable mainly, the features retained in shown in the figure.

Status is the target variable.

The EDA informed the feature selection as it highlighted patterns, correlations and distribution of different features that relate to placement outcomes. This reinforced the selection of the relevant predictors.

Model Result

Splitting the Dataset

After successful encoding stage, the data was then divided into training set (70%) and testing set (30%) using the reproducible random seed (set.seed(123)) in R tools to ensure that the target variable is balanced in order to maintain the consistency of the model and maintain fairness in the outcomes.The 151 observations in the training set will be used to run the model, while 64 observations in the testing set will be used for model evaluation process and the distribution of the target variable (Placed vs. Not Placed) is then displayed using a bar chart where it shows that the proportions of Placed and Not Placed candidates are consistent across both subsets, minimizing potential sampling bias. Refer appendix D for visualisation.

Regression Model

Baseline Logistic Regression Model: Education Score as a Predictor of Placement

To begin with education_score is compared with the target variable (status) using logistic regression model and the model indicated a coefficient of 0.3089 for the variable, indicating that for every 1-unit increase in education_score, the odds of being placed increase by approximately 36.2% (odds ratio: 1.36). The significance level of education_score is less than 0.001, that is, the P value is less than 0.001, indicating that education_score has a significant impact on status. Model’s residual deviance decreased significantly from 187.27 (null deviance) to 106.60, confirming education_score improved the model’s fit and AIC value

is shown as 110.6 which could be used to compare with the other models used and the lowest AIC value will indicate the overall best model.

```
##
## Key Findings from Logistic Regression Model:

## - Coefficients:

##                    Estimate Std. Error   z value     Pr(>|z|)
## (Intercept)     -19.1712592 3.31539137 -5.782503 7.359743e-09
## education_score   0.3088961 0.05174907  5.969113 2.385471e-09

##
## - P-values:

##     (Intercept) education_score
##    7.359743e-09    2.385471e-09

##
## - Model Deviance: 106.6012

## - Null Deviance: 187.2713

## - AIC: 110.6012
```

For full model details, kindly refer to appendix C.

Regression Model 02: Incorporating Work Experience

For the second model work_experience was added alongside initial predictor education_score and the results indicated that both variables significantly influence placement status for students. The coefficient for education_score increased slightly to 0.32097, indicating that for every 1-unit increase in education_score, the odds of being placed increase by approximately 37.8% where this predictor remains highly significant with a p-value less than 0.001. work_experience emerged as a strong predictor with a coefficient of 1.58743, showing that candidates with prior work experience have nearly 4.89 times higher odds of being placed compared to those without work experience where this effect is statistically significant with a p-value of 0.0123.

Model 2 shows improvement compared to the baseline as the residual deviance decreased from 106.6 in Model 1 to 99.3 in Model 2, indicating that the addition of work_experience explains more variability with related to student's placement status. The models AIC value also reduced from 110.6 in Model 1 to 105.3 in Model 2, further supporting that Model 2 is a better fit and the log-likelihood also increased to -49.65, reflecting models' quality improvement.

```
##
## Key Findings from Logistic Regression Model 2:

## - Coefficients:
```

```
##                    Estimate Std. Error   z value     Pr(>|z|)
## (Intercept)     -20.3473807 3.65692293 -5.564072 2.635519e-08
## education_score   0.3209732 0.05655679  5.675236 1.384975e-08
## work_experience   1.5874326 0.63377161  2.504739 1.225417e-02

##
## - Odds Ratios:

##     (Intercept) education_score work_experience
##    1.456280e-09    1.378469e+00    4.891175e+00

##
## - P-values:

##     (Intercept) education_score work_experience
##    2.635519e-08    1.384975e-08    1.225417e-02

##
## - AIC Value: 105.3043

## - Log-Likelihood: -49.65214
```

Regression Model 03: Adding MBA Specialization

The specialization_Mkt_Fin was included alongside other features used in model 2, to run the model 3 and the results indicated specialization_Mkt_Fin does not appear to have a statistically significant compared to the other 2 variables which was continously shown as strong predictors even in model 3 where a 1-unit increase in education_score raises the chances of placement by 37.2%, making it a strong and highly significant predictor (p-value < 0.001), similarly, having work experience increases placement odds by 4.6 times, showing its importance in determining placement success.

Coefficient for specialization_Mkt_Fin is 0.43839, which translates to an odds ratio of 1.55, suggesting a modest increase in placement odds for candidates specializing in Marketing-Finance and despite this, the p-value of 0.3901 indicates that this effect is not statistically significant when compared.

The model's residual deviance decreased slightly to 98.57, and the AIC value is 106.57, showing marginal improvement compared to the second model's AIC of 105.30. The log-likelihood is -49.28, reflecting a similar level of model fit. The Variance Inflation Factor (VIF) values for all predictors are close to 1, confirming the absence of multicollinearity and the stability of the model.

```
##
## Key Findings from Logistic Regression Model 3:

## - Coefficients and P-values:

##                            Estimate     Pr(>|z|)
## (Intercept)            -20.2542083 3.934968e-08
## education_score          0.3164413 2.879137e-08
## work_experience          1.5262282 1.723602e-02
## specialization_Mkt_Fin   0.4383913 3.901005e-01
```

```
##
## - Odds Ratios:

##            (Intercept)        education_score        work_experience
##           1.598487e-09           1.372236e+00           4.600791e+00
## specialization_Mkt_Fin
##           1.550211e+00

##
## - VIF Values:

##        education_score        work_experience specialization_Mkt_Fin
##               1.075149               1.082470               1.009314

##
## - AIC: 106.5654

## - Log-Likelihood: -49.28272
```

Model 4: Evaluating Placement Predictors with Educational and Professional Features

Two new features were added to the model 4 which are high school streams (hsc_science, hsc_commerce) and MBA performance (mba_p) alongside the original predictors used in model 3 and the results confirmed education_score and work_experience remain important for the 4th consercative time. A 1-unit increase in education_score raises the chances of placement by 65.8%, and candidates with work experience are 15 times more likely to be placed than those without it and interestingly, mba_p (MBA performance) shows a significant negative relationship, where higher MBA scores reduce the odds of placement by 25.3%.

Few non-significant features was identified by the model which are high school streams and specialization in Marketing-Finance, suggesting they do not strongly impact placement outcomes and in terms of model fit, Model 4 shows significant improvement with an AIC of 87.50 and a residual deviance of 73.50, making it the best-fitting model so far and provides deeper insights into placement predictors.

```
##
## Call:
## glm(formula = status ~ education_score + work_experience + specialization_Mkt_Fin +
##     hsc_science + hsc_commerce + mba_p, family = binomial, data = train_data)
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -13.96945    5.04323  -2.770  0.00561 **
## education_score          0.50556    0.09754   5.183 2.18e-07 ***
## work_experience          2.73305    0.88025   3.105  0.00190 **
## specialization_Mkt_Fin   0.35535    0.61729   0.576  0.56485
## hsc_science             -0.62876    1.34298  -0.468  0.63965
## hsc_commerce            -0.68802    1.34630  -0.511  0.60932
## mba_p                   -0.29160    0.07057  -4.132 3.59e-05 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 187.271  on 150  degrees of freedom
## Residual deviance:  73.499  on 144  degrees of freedom
## AIC: 87.499
##
## Number of Fisher Scoring iterations: 7

## Model 4 AIC: 87.4989

## Log-Likelihood for Model 4: -36.74945

## Odds Ratios for Model 4:

##            (Intercept)        education_score        work_experience
##           8.573263e-07           1.657908e+00           1.537971e+01
## specialization_Mkt_Fin            hsc_science           hsc_commerce
##           1.426674e+00           5.332521e-01           5.025698e-01
##                  mba_p
##           7.470680e-01

## VIF Values for Model 4:

##        education_score        work_experience specialization_Mkt_Fin
##               1.770564               1.585253               1.075233
##            hsc_science           hsc_commerce                  mba_p
##               4.945481               5.145087               1.948613
```

Model 5; Interaction Effects Model

Model 5 was also performed introducing the Interaction terms between education_score and work_experience to study about the potential combined effects, however, the interaction term was not statistically significant ($p = 0.778$) where the model suffered from high multicollinearity (VIF values exceeding 150) and a higher AIC (108.49) compared to Model 4 (87.49). Model 5 was excluded from the main analysis as a result of this and the full details of Model 5 are included in Appendix A.

Model Evaluation and Predicted Probabilities

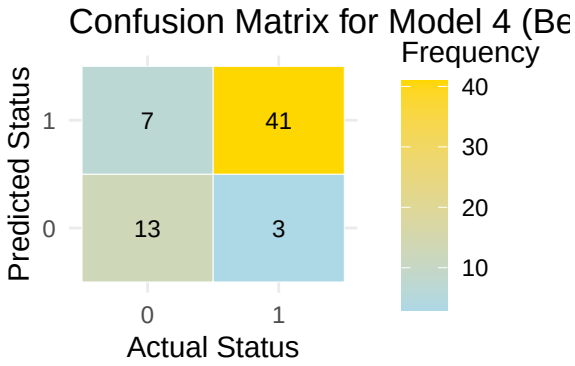
Model Comparison Using AIC

According to the Akaike Information Criterion (AIC) values table, model 4 have the lowest AIC values (87.4989), making it the best fit model among all the models used and features such as high school streams and MBA performance in Model 4 improves model performance without overfitting, compared to simpler models.

```
##             df      AIC
## basic_model  2 110.6012
## model_2      3 105.3043
## model_3      4 106.5654
## model_4      7  87.4989
```

Confusion Matrix Summary for Model 4

Overall, the confusion matrix plot indicates that Model 4 is well performed where it correctly predicted 41 candidates as placed (True Positives) and 13 candidates as not placed (True Negatives). However, it made mistakes by predicting 7 candidates as placed when they were not (False Positives) and 3 candidates as not placed when they were actually placed (False Negatives). Higher number of true positive value indicates that model does a good job of identifying candidates who are placed, however, the 7 non-placed candidates were miscalculated as placed leaving for improvements.

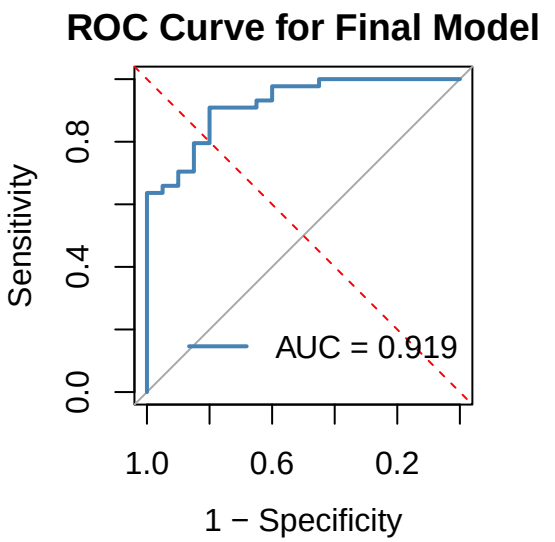


In addition to the confusion matrix, predicted and actual placement proportions were compared using a bar plot. The results show that the predicted values are quite similar to the actual values for both ‘Placed’ and ‘Not Placed’ categories. A detailed chart of this comparison is included in the appendix B

ROC Curve and AUC Analysis for Model 4

The model differentiates between placed and non-placed candidates across various threshold values where area Under the Curve (AUC) is 0.919, indicating that Model 4 has excellent discriminative ability and the model is highly effective at distinguishing between non-placed and placed as the AUC value close to 1 for the model.

The diagonal red line represents the performance of a random classifier (AUC = 0.5), and the ROC curve for Model 4 lies significantly above this line, confirming its robustness and this supports the conclusion that Model 4 performs well in addressing the business objective of predicting placement success when compared with the other simpler models tested and studied.

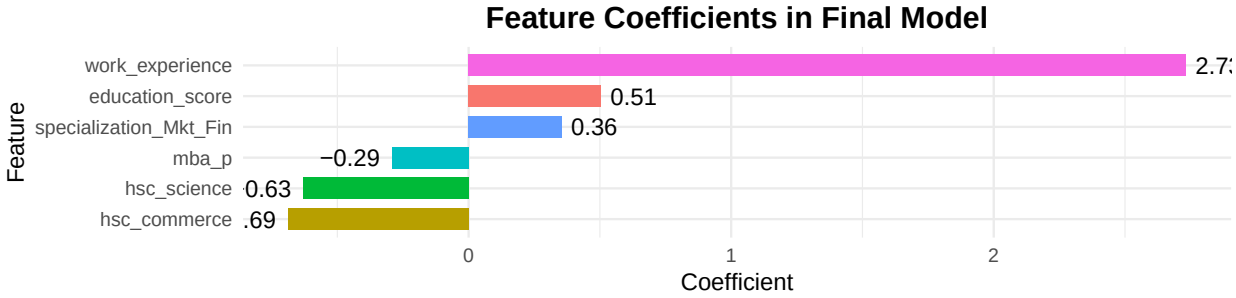



```
## AUC: 0.919
```

Feature Importance and Coefficients Analysis for Model 4

Highest positive coefficient is found for work_experience (2.7), indicating that prior work experience significantly increases the likelihood of being placed and education_score also plays an important role, with a positive coefficient of 0.51, highlighting that higher education scores are linked with the placements in the job market.

On the other hand features like Specialization in Marketing and Finance has a smaller positive impact with a coefficient of 0.36, MBA performance (mba_p) a negative coefficient (-0.29) and the coefficients for hsc_science and hsc_commerce indicate minimal contributions and do not significantly affect the placement with a coefficient value of 63 and 69 respectively. Looking at this it is much easier to understand what features could potentially influnec the outcome of a placement and build strategies accordingly.



Predicted Probabilities Distribution

The density plot in the appendix shows how well the model predicts placement where candidates likely to be placed have predicted probabilities close to 1, while those not placed have lower probabilities. This shows the model can separate placed and not placed candidates using features like education score and work experience where the results match the ROC curve, with a high AUC of 0.919, confirming Model 4's strong prediction ability.

Analysis of Misclassified Observations in Placement Prediction

The jitter plot for misclassified Observations in Placement Prediction is used and it illustrates the classification effect of the model. The model has a high classification accuracy for samples with extreme prediction probabilities (close to 0 or 1). Most of the misclassified samples are concentrated in the middle area of the prediction probability where the model's discrimination ability in the middle probability area still has room for improvement.

Discussion

Logistic model's statistical evidence clearly indicated that education performance and work experience were critical predictors for campus placements for example, in Model 4, education score had a coefficient of 0.50556 and work experience had a coefficient of 2.73305, indicating that students with prior work experience were 15 times more likely to be placed compared to those without. These findings could help universities to develop new strategies to handle student with lower education scores by providing extra academic support and implementing internship opportunities and on the other hand employers can focus on candidates with higher education scores and work experience during their recruitment processes which saves much time and give the possibility to choose the best candidate for the job.

The developed model, with an AUC of 0.919, demonstrates strong predictive accuracy, making it a reliable tool for placement teams and the confusion matrix highlights the model's effectiveness, correctly predicting 41 students as placed (True Positives) and 7 as not placed (True Negatives). Key thing to consider is that model misclassified 13 not placed students as placed (False Positives) and 3 placed students as not placed (False Negatives) leaving areas for improvement but models ability to predict majority of outcomes accurately this could be implemented in the business environment for better outcomes. This model could be used to identify the students who need extra support by the educational institutes, academic stakeholders can use the insights to refine their curriculum, focusing on skills and experiences that directly improve placement rates and employers can benefit from data-driven hiring decisions, ensuring they select candidates most likely to succeed in their roles.

Limitations

The dataset used in this study had some gaps, as it didn't include factors like soft skills, extracurricular activities, or interview performance, which are important for placement outcomes. Adding these variables could make the predictions more accurate and reliable. Confusion matrix demonstrated strong accuracy in identifying placed candidates but struggled with misclassifying some non-placed candidates as placed (false positives) which should be considered during further analysis. Another limitation is that data might not represent all student groups equally, especially underrepresented ones and using techniques like oversampling or using class-weighted models could be used to further address the class imbalance issues in logistic model and using only logistic regression meant other advanced models, like Gradient Boosting or Neural Networks, weren't explored, and if used could improve the analysis allowing comparison to accurately choose the best model.

Conclusion

This study identified key factors affecting campus placement, focusing on academic performance and work experience and the logistic regression model showed that education score and work experience are the most

important predictors of placement success, with work experience significantly increasing placement chances (odds ratio ~15). AUC of 0.919 indicates that model performed well also indicating strong reliability in predicting placements. These results provide useful insights for academic institutions and placement teams to improve student employability highlighting the importance of academic achievements and internships in influencing placements where the visual tools, like the confusion matrix and ROC curve, helped explain the model's performance and offered clear, actionable insights for stakeholders to enhance decision-making.

For future work, adding more variables like extracurricular activities, soft skills, and interview performance can help improve predictions. Addressing misclassifications, especially false negatives or ensemble techniques could make the model even more accurate. Expanding the model's application to other industries or institutions would increase its usefulness. In summary, this study provides a strong framework for understanding and improving placement strategies, helping academic stakeholders make data-driven decisions and support students better.

References

Kesavaraj, G. and Pattnaik, M. (2012) 'A study on the effectiveness of campus recruitment and selection process in IT industries,' in Lecture notes in mechanical engineering, pp. 745–757. https://doi.org/10.1007/978-81-322-1007-8_71.

Bhargavi, S. and Yaseen, A. (2016) Leadership styles and organizational performance, Strategic Management Quarterly. journal-article, pp. 87–117. https://doi.org/10.15640/smq.v4n1a5.

Class of 2023 starting salaries show considerable growth (2024). https://www.naceweb.org/job-market/compensation/class-of-2023-starting-salaries-show-considerable-growth.

Caballero, C.L. and Walker, A. (2010) 'Work readiness in graduate recruitment and selection: A review of current assessment methods,' Journal of Teaching and Learning for Graduate Employability, 1(1), pp. 13–25. https://doi.org/10.21153/jtlge2010vol1no1art546.

Andrews, J. and Higson, H. (2008) 'Graduate Employability, 'Soft Skills' versus 'Hard' Business knowledge: A European study,' Higher Education in Europe, 33(4), pp. 411–422. https://doi.org/10.1080/03797720802522627.

First destinations for the college class of 2020 (2020). https://naceweb.org/job-market/graduate-outcomes/first-destination/class-of-2020.

York, T. et al. (2015) 'Defining and measuring academic success,' Practical Assessment, Research & Evaluation, 20, p. 2. https://www.researchgate.net/publication/278305241.

MCCANN, M. and HEWITT, M., 2023. Academic performance and work placements: does academic performance influence the decision to complete a work placement? Higher Education, Skills and Work-Based Learning, 13 (1), pp. 97-112. ISSN 2042-3896

De Araujo, P. and Murray, J. (2010) 'Estimating the effects of dormitory living on student performance,' SSRN Electronic Journal [Preprint]. https://doi.org/10.2139/ssrn.1555892.

Galbraith, D. and Mondal, S. (2020) The potential power of internships and the impact on career preparation. https://eric.ed.gov/?id=EJ1263677.

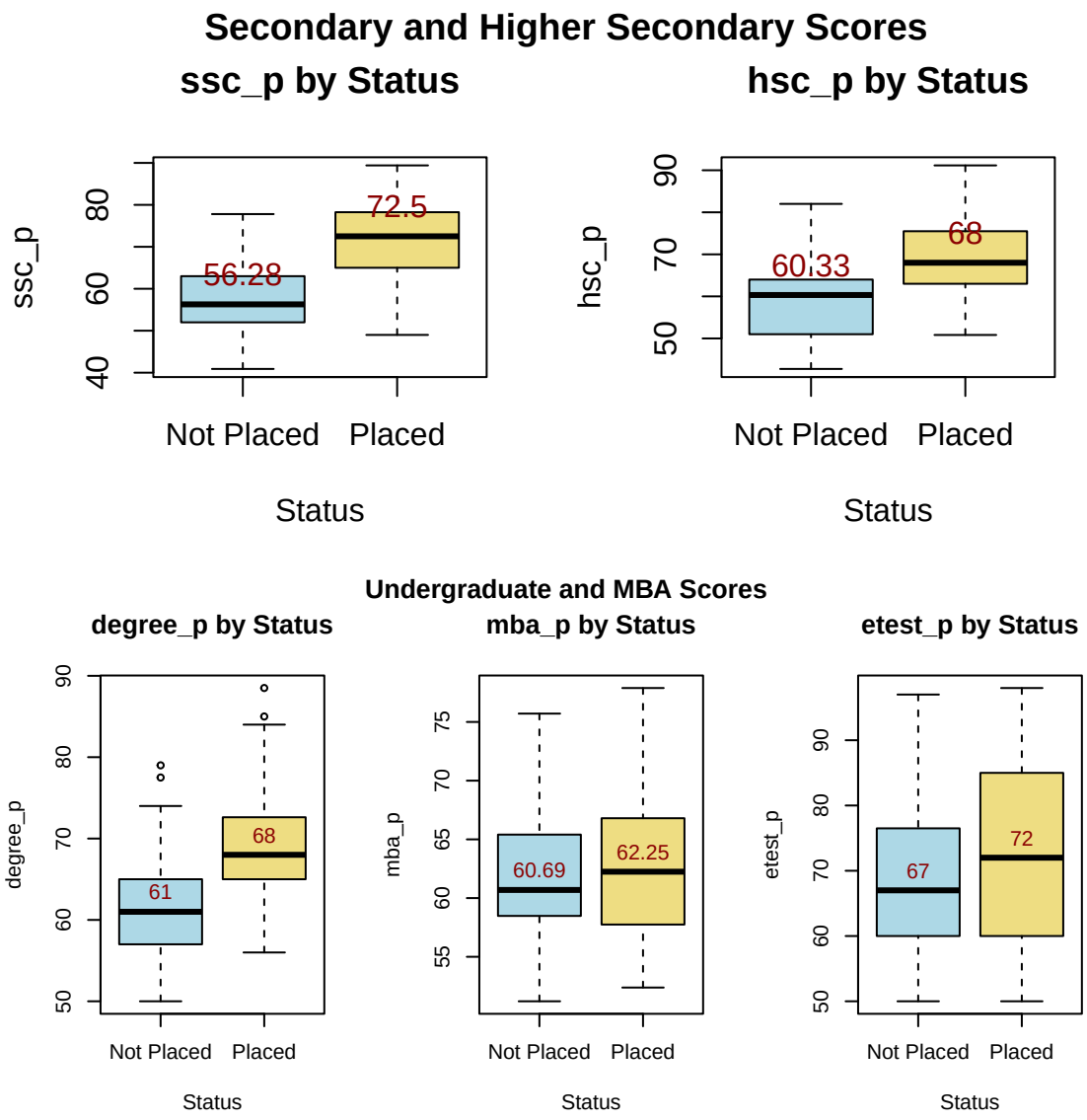
Caballero, C., Walker, A. and Fuller-Tyszkiewicz, M. (2011) 'The Work Readiness Scale (WRS): Developing a measure to assess work readiness in college graduates,' Journal of Teaching and Learning for Graduate Employability, 2–2, pp. 41–54.

Ellis, C. et al. (2017) 'Living the Post-University life: Academics talk about retirement,' Qualitative Inquiry, 1–14.

Additional Works

In this section, numerical and categorical will be further analyzed against the placement status. The purpose is to dive deeper into the factors that might influence the placement status for the students.

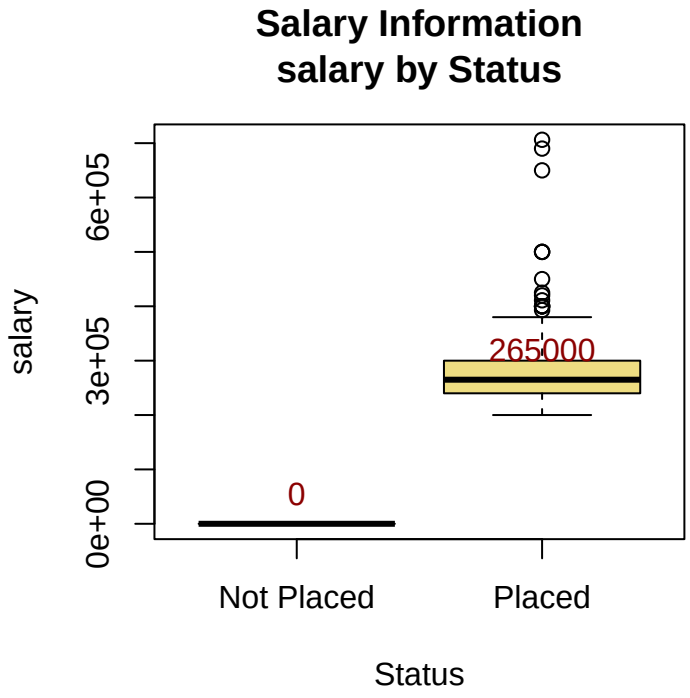
Additional EDA: Numerical Features Against Status



In section EDA, it was determined that early education such as SSC and HSC play a critical role in the placement status. In the above boxplot, it shows that students who were placed tend to have a higher median score compared to not placed where students were placed has a median score of 72.5 while not place is 56.28. This indicates that academic achievement in the early stages plays a critical role in securing placement. Additionally, the wider range of scores among placed students shows that even students with varied performance can succeed if they meet certain benchmarks. For the degree, it shows a slight difference between placed and not placed students with median of 68 and 61 respectively. This might show the importance of consistent academic achievement during undergraduate studies and early education.

Meanwhile, MBA shows a slight difference with placed is 62.25 and not placed is 60.69 indicating that MBA still play a role but may not have much impact compared to other factors.

On the other hand, despite having big differences between the median in employment test, the upper and whisker shows that the maximum and minimum has only slight difference. This further supports the previous statement in section EDA where it states that it does not play a significant role in the placement.



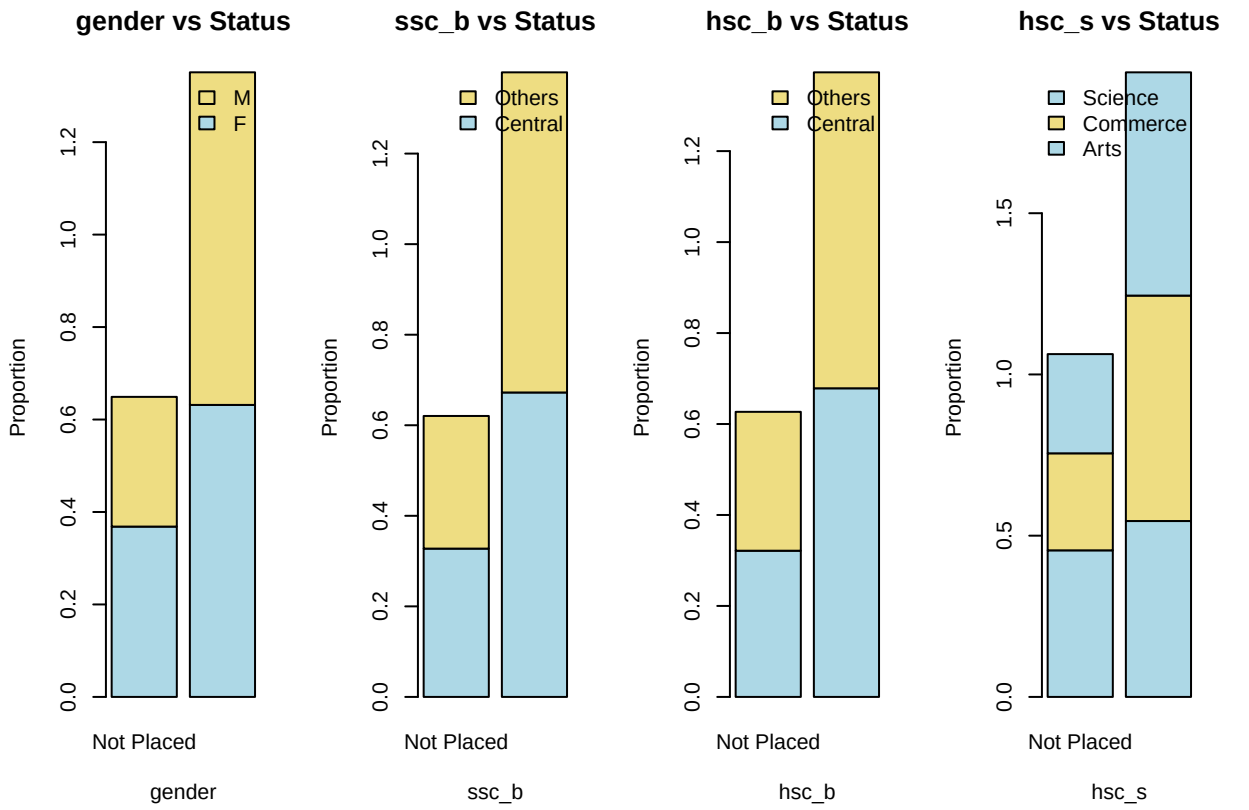
In Salary Information boxplot provides a comparison of salaries between placed and not placed. For the students that are not placed, it is consistent with 0 as these students did not secure employment.

While for the placed students, the median salary is 265,000 and it can be observed that most students earn salaries close to the median. However, it can be observed that there are points above the upper whiskey that indicate that there are students that secure significantly higher salaries which might be exceptional cases such as specialized roles.

The contrast between the two groups is clear as it not only shows that it leads to earning opportunities but also potential financial growth. The financial growth can be seen in the outliers in the placed group. This boxplot highlights the importance of securing placement for financial success after graduation.

In summary, the insights can be used to emphasize the value of placement and invest in strategies to enhance placement opportunities.

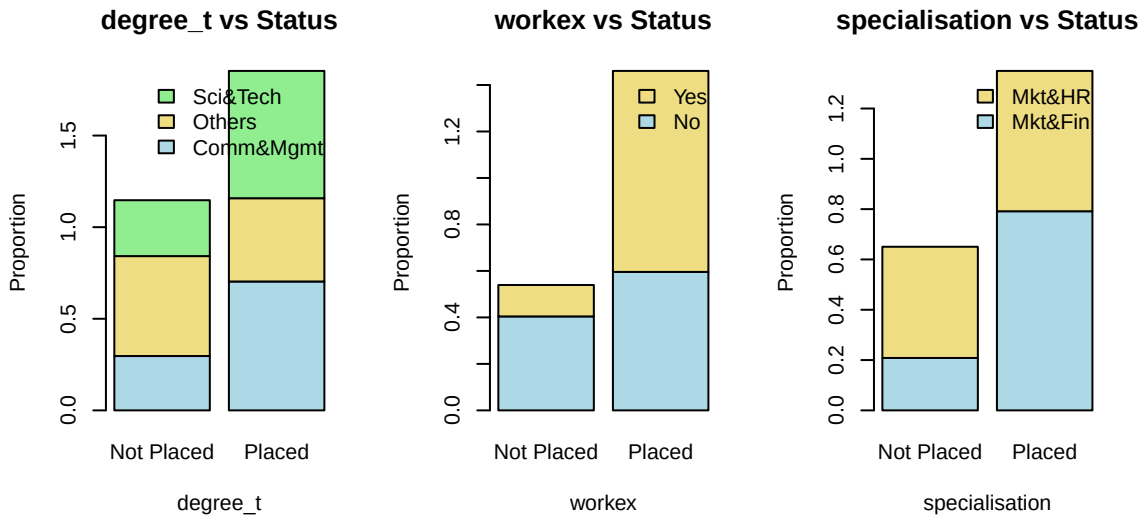
Additional EDA: Categorical Features Against Status



Boxplot above provides the relationship of the demographic information with the placement status. In the first plot shows that, it can be observed that male student makes up the larger portion of placed students compared to the placed student, however, for the not placed, the distribution is relatively balanced between the two genders. This might suggest female students might face greater challenge in achieving placement success due to the different in opportunities of preferences based on gender.

Next, the ssc_b vs status plot shows among the students that were placed were mostly from ‘Others’ board compared from the ‘Central’. However, among the not placed, both ‘Others’ and ‘Central’ were more evenly distributed. This indicates that students from ‘Others’ boards tend to perform better in terms of securing placements. Similar patterns were found in the hsc_b vs status plot. This shows students from ‘Others’ boards for both SSC and HSC are more successful in placement compared to those from the ‘Central’ board.

In the last boxplot, HSC specializations were divided into 3 which are Science, Commerce and Arts. Overall, it can be observed that most placed students are from ‘Science’ and ‘Commerce’ with a minority from ‘Arts’. Among the students not placed, the distribution is more balanced with ‘Arts’ being the dominant specialization. This might suggest that the employers prefer candidates with either ‘Science’ or ‘Commerce’ background, possibly due to their business nature.



In the boxplot above, it provides the comparison between the degree type, working experience and specialization against the placement.

It can be observed that in the first plot that those with 'Sci & Tech' and 'Comm&Mgmt' are the majority in the placed and on the other hand, 'Others' category has the highest amount in not placed. This highlights that type of degree plays a crucial role in the placement.

In the second plot, it compares the working experience with the placement. It is obvious that those with working experience are likely to get placement compared to students with no working experience. It can also be seen that only a quarter of not placed is with work experience. This demonstrates that experience also might be a strong determinant of placement success, which is likely due to employers' preference to those that have practical skills and industry experience.

Lastly, in the final plot shows the relationship between MBA specialization with the placement status. In general, students specializing in 'Mkt&Fin' have a higher success rate in the placement compared to 'Mkt&HR'. This is likely due to the higher demand for skills in finance-related roles.

Appendix

Appendix A: Structure of Dataset

Dataset consists of 215 rows and 15 columns where each row represents a unique student.

Variable	Description
Gender	Gender of the student
SSC Percentage (ssc_p)	Numerical variable for secondary school performance percentage
HSC Percentage (hsc_p)	Numerical variable for higher secondary school percentage
Degree Specialization (degree_t)	Categorical variable showing student's field of study
Placement Status (status)	Categorical variable showing whether student has been placed or not
Salary (salary)	Numerical variable for offered salary of the placed students

Appendix B: Distribution Overview of Categorical Data

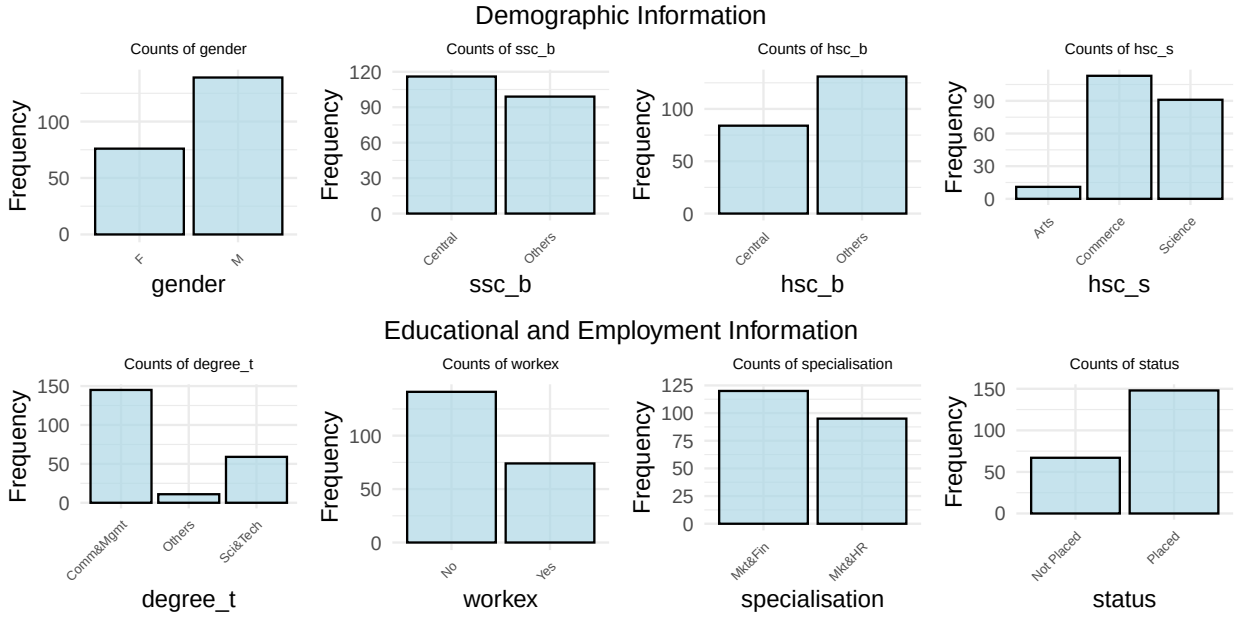
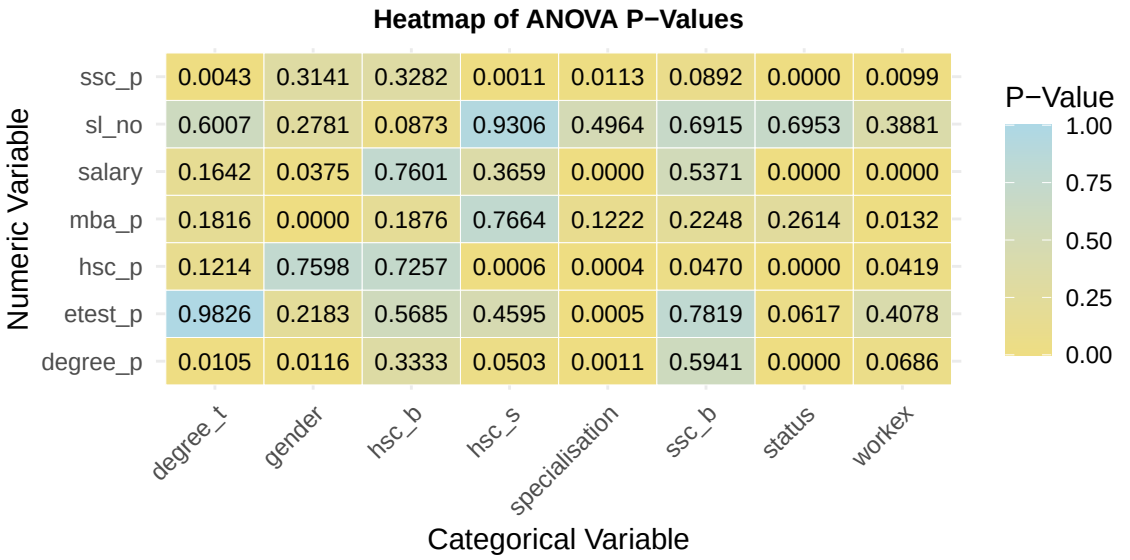


Figure above shows the distribution of the categorical variables in the dataset which offers information such as demographic, educational and employment of the students.

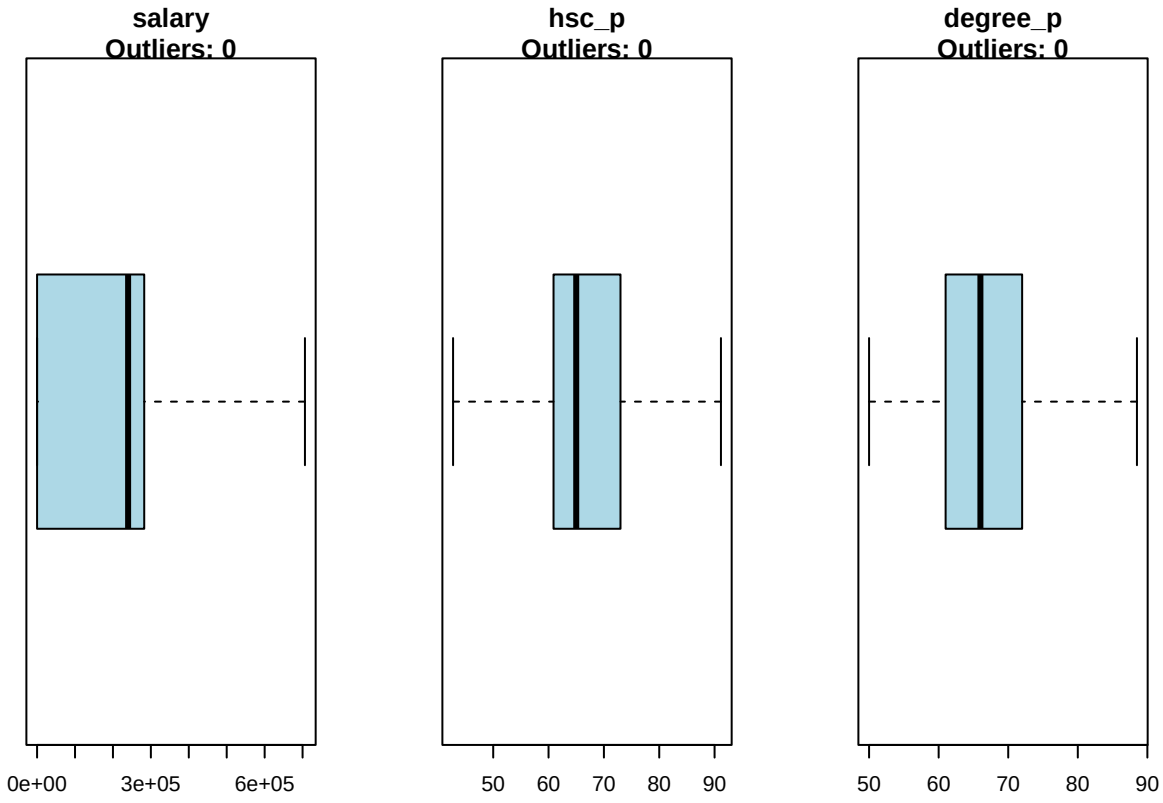
It can be observed that the gender distribution shows an imbalance with higher numbers of male students compared to females. This inconsistency may reflect the trends in the field of study or the placement rates. The distribution of SSC (10th grade) (ssc_b) boards is almost evenly split between the “Central” and “Others” indicates that students come from diverse educational backgrounds. However, for HSC (12th grade) (hsc_b) reflects that “Others” has a higher portion compared to the “Central”. This may suggest a shift toward non-central boards for higher secondary education. The majority of HSC (12th grade) students opt for “Commerce” or “Science” with less of 30 students choosing “Arts” which indicates strong inclination towards business and scientific education during this stage.

For degree education, majority of the students pursued degree in “Comm&Mgmt” and “Sci&Tech” and with less than 30 students opt for “Others”. This is consistent with higher secondary school education backgrounds. While for MBA, “Mkt&Fin” is preferable compared to “Mkt&HR”, which could be to the possibility higher demand or better placement opportunities in finance related roles. This suggests that students are strategically aligning their education with market trends and career prospects.

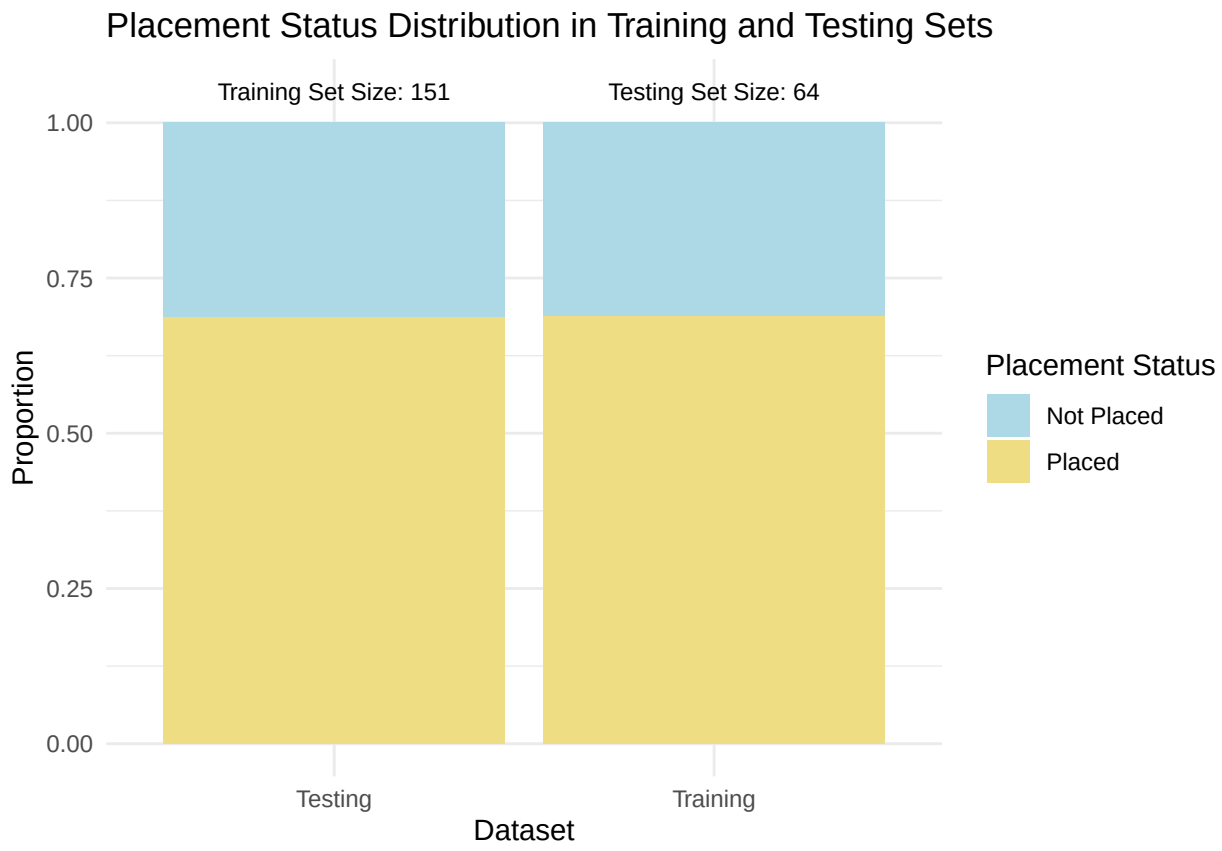
For working experience, it can be observed that majority of students does not have experience with only small portion has working experience. This factor possibly may influence the placement outcomes as employers might prefer those with working experience as they have the exposure to the field on top of the education background. However, placement status shows that most of the students were successfully placed, while remaining remained unplaced which indicates working experience might not be the major influence to determine the placement status.



Appendix C: Result of handling outliers



Appendix D: Splitting the Dataset



Appendix E: Baseline Logistic Regression Model: Education Score as a Predictor of Placement

```
##
## Call:
## glm(formula = status ~ education_score, family = binomial, data = train_data)
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -19.17126    3.31539  -5.783 7.36e-09 ***
## education_score   0.30890    0.05175   5.969 2.39e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 187.27  on 150  degrees of freedom
## Residual deviance: 106.60  on 149  degrees of freedom
## AIC: 110.6
##
## Number of Fisher Scoring iterations: 6
```

Appendix F: Regression Model 02: Incorporating Work Experience

```
##
## Call:
## glm(formula = status ~ education_score + work_experience, family = binomial,
##     data = train_data)
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -20.34738    3.65692  -5.564 2.64e-08 ***
## education_score   0.32097    0.05656   5.675 1.38e-08 ***
## work_experience   1.58743    0.63377   2.505   0.0123 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 187.271  on 150  degrees of freedom
## Residual deviance:  99.304  on 148  degrees of freedom
## AIC: 105.3
##
## Number of Fisher Scoring iterations: 6

## [1] "AIC Value for Model 2: 105.304288522019"

## [1] "Odds Ratios:"

##     (Intercept) education_score work_experience
##    1.456280e-09    1.378469e+00    4.891175e+00

## [1] "Log-Likelihood: -49.6521442610096"
```

Appendix G: Regression Model 03: Adding MBA Specialization

```
##
## Call:
## glm(formula = status ~ education_score + work_experience + specialization_Mkt_Fin,
##     family = binomial, data = train_data)
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -20.25421    3.68678  -5.494 3.93e-08 ***
## education_score          0.31644    0.05703   5.549 2.88e-08 ***
## work_experience          1.52623    0.64083   2.382   0.0172 *
## specialization_Mkt_Fin   0.43839    0.51009   0.859   0.3901
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 187.271  on 150  degrees of freedom
## Residual deviance:  98.565  on 147  degrees of freedom
## AIC: 106.57
##
## Number of Fisher Scoring iterations: 6
```

```
## [1] "VIF Values for Model 3:"
```

```
##        education_score        work_experience specialization_Mkt_Fin
##               1.075149               1.082470               1.009314
```

```
## [1] "Model 3 AIC: 106.565431626324"
```

```
## [1] "Log-Likelihood: -49.282715813162"
```

```
## [1] "Odds Ratios for Model 3:"
```

```
##            (Intercept)        education_score        work_experience
##           1.598487e-09           1.372236e+00           4.600791e+00
## specialization_Mkt_Fin
##           1.550211e+00
```

Appendix H: Interaction Effects Model

```
## Model :
## status ~ education_score + work_experience + specialization_Mkt_Fin +
##     education_score:work_experience
```

```
##
## Call:
## glm(formula = status ~ education_score + work_experience + specialization_Mkt_Fin +
##     education_score:work_experience, family = binomial, data = train_data)
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -20.91153    4.46814  -4.680 2.87e-06 ***
## education_score                  0.32660    0.06908   4.728 2.27e-06 ***
## work_experience                  3.68718    7.69359   0.479    0.632
## specialization_Mkt_Fin           0.44319    0.51114   0.867    0.386
## education_score:work_experience -0.03458    0.12243  -0.282    0.778
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 187.271  on 150  degrees of freedom
## Residual deviance:  98.488  on 146  degrees of freedom
## AIC: 108.49
##
## Number of Fisher Scoring iterations: 6
```

```
## there are higher-order terms (interactions) in this model
## consider setting type = 'predictor'; see ?vif
```

```
## [1] "VIF Values for Model 5:"
```

```
##                 education_score                 work_experience
##                        1.566417                      162.488698
##          specialization_Mkt_Fin education_score:work_experience
##                        1.011351                      158.511385
```

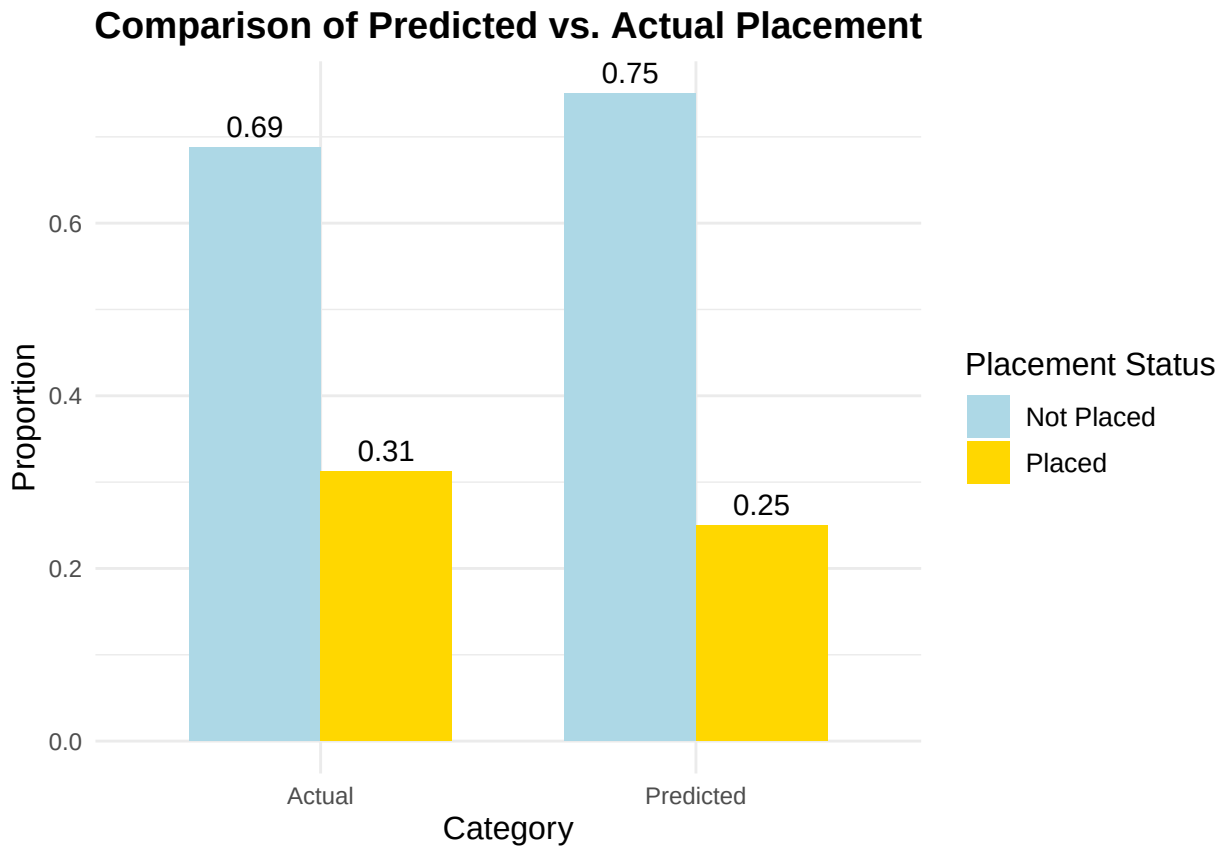
```
## [1] "Model 5 AIC: 108.487950869259"
```

```
## [1] "Log-Likelihood: -49.2439754346294"
```

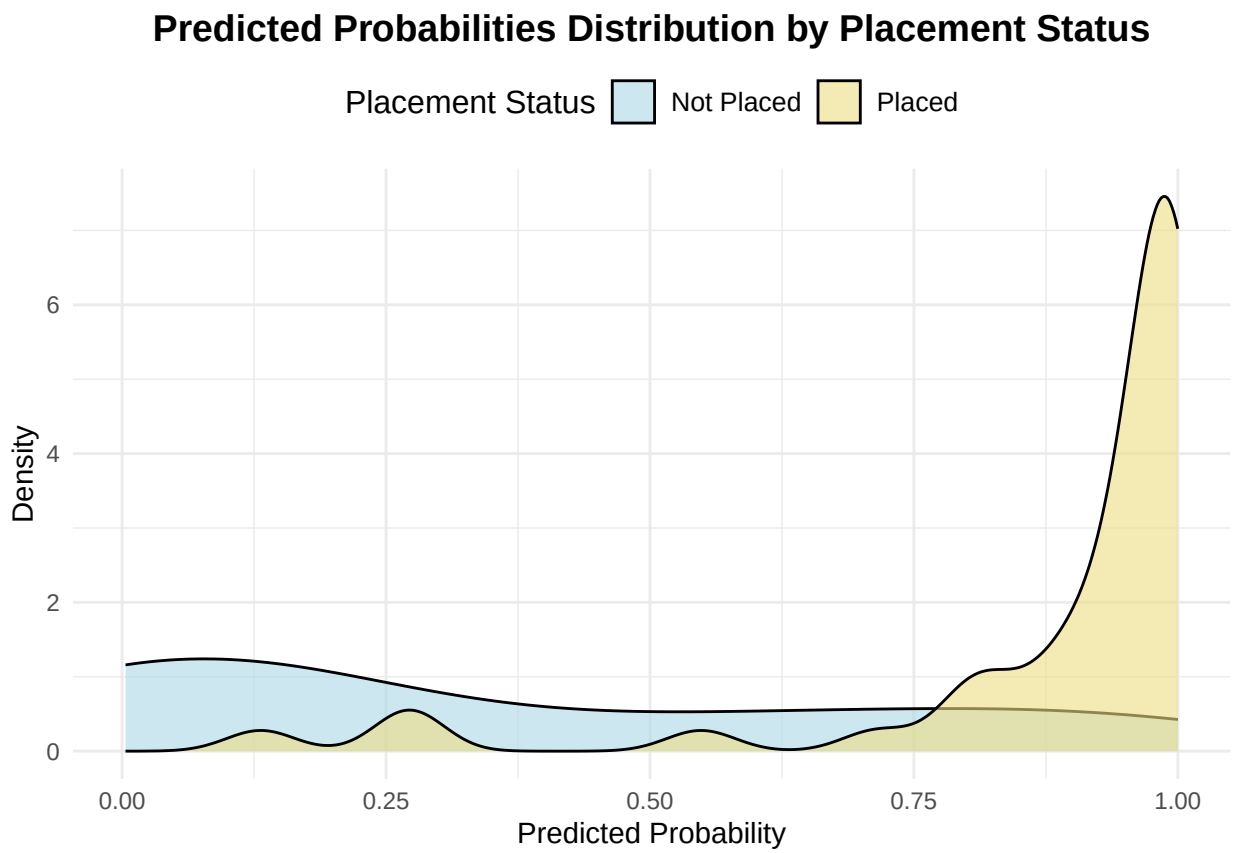
```
## [1] "Odds Ratios for Model 5:"
```

```
##                     (Intercept)                 education_score
##                    8.283977e-10                    1.386243e+00
##                 work_experience          specialization_Mkt_Fin
##                    3.993213e+01                    1.557670e+00
## education_score:work_experience
##                    9.660144e-01
```

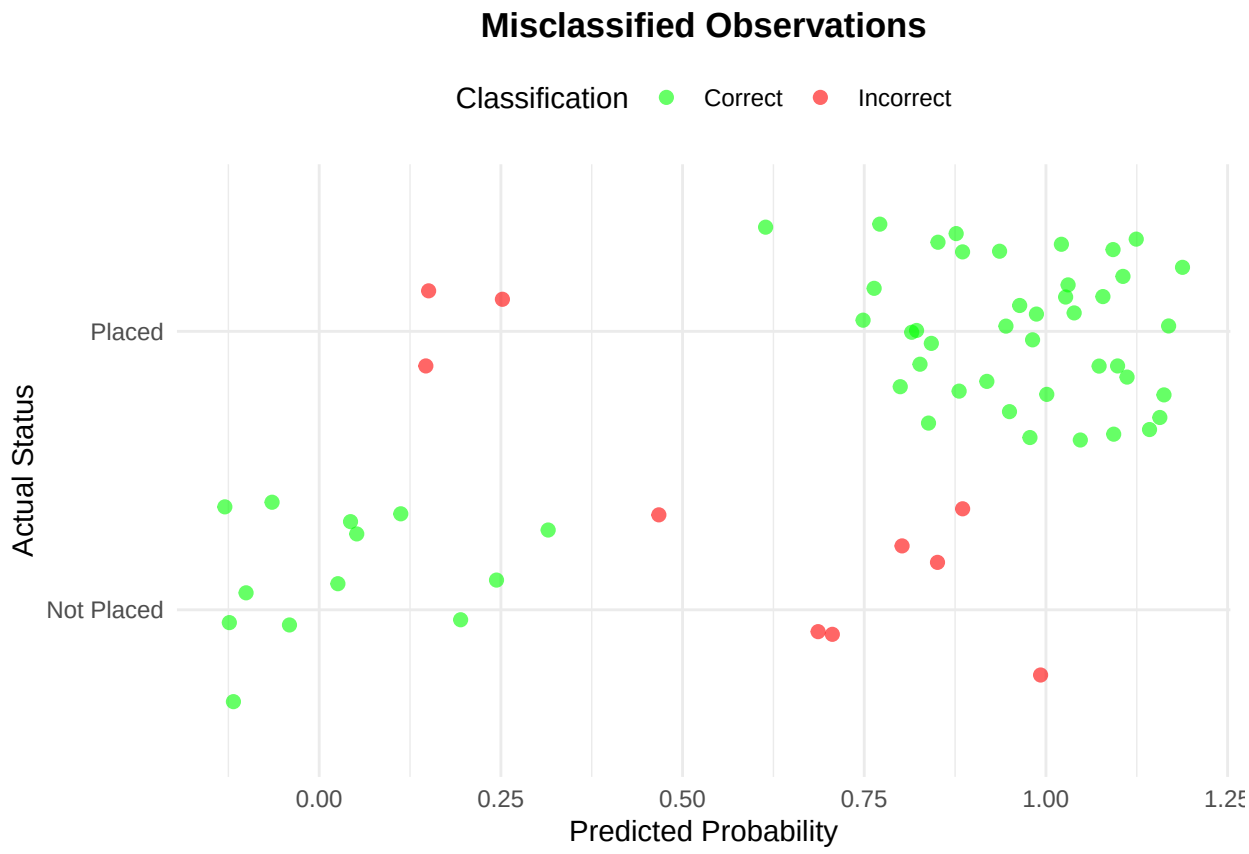
Appendix I: Comparison of Predicted vs. Actual Placement Outcomes for Model 4



Predicted Probabilities Distribution



Analysis of Misclassified Observations in Placement Prediction


Misclassified Observations
Classification
Correct
Incorrect
Placed
Not Placed
Actual Status
0.00
0.25
0.50
0.75
1.00
Predicted Probability