Northeastern University

ALY-6015
INTERMEDIATE ANALYTICS

BY INSTRUCTOR
ROY WADA

MODULE-3 PROJECT REPORT
SUBMITTED BY

ITI ROHILLA

On Jan-30-2024

SUMMARY

The College dataset, drawn from the 1995 US News and World Report, encompasses statistics from 777 American colleges, offering a comprehensive view with 18 variables. This dataset provides valuable insights into the characteristics of these institutions, including application numbers, acceptance rates, tuition costs, and faculty qualifications. Initial exploratory data analysis (EDA) revealed a mix of private and public colleges, showcasing significant variations across different features.

The EDA process included generating summary statistics and visualizations, allowing for a deeper understanding of the dataset's patterns. A correlation matrix and density plot further unveiled relationships and distributions, providing key insights into the dataset's characteristics and paving the way for subsequent modeling.

For predicting whether a college is private, logistic regression was employed, achieving an accuracy of 84.07% on the training data. Evaluation metrics such as sensitivity (70.78%) and specificity (89.29%) provided a comprehensive view of the model's performance. The confusion matrix and additional metrics offered insights into the model's ability to correctly identify private and public colleges, highlighting its strengths and areas for improvement.

The model's performance was extended to a test set, resulting in an accuracy of 82.68%. Metrics such as sensitivity, specificity, and others were carefully examined, emphasizing the model's capability to generalize to unseen data. This step reinforced the reliability of the logistic regression model in making accurate predictions beyond the training set, contributing to its credibility and applicability.

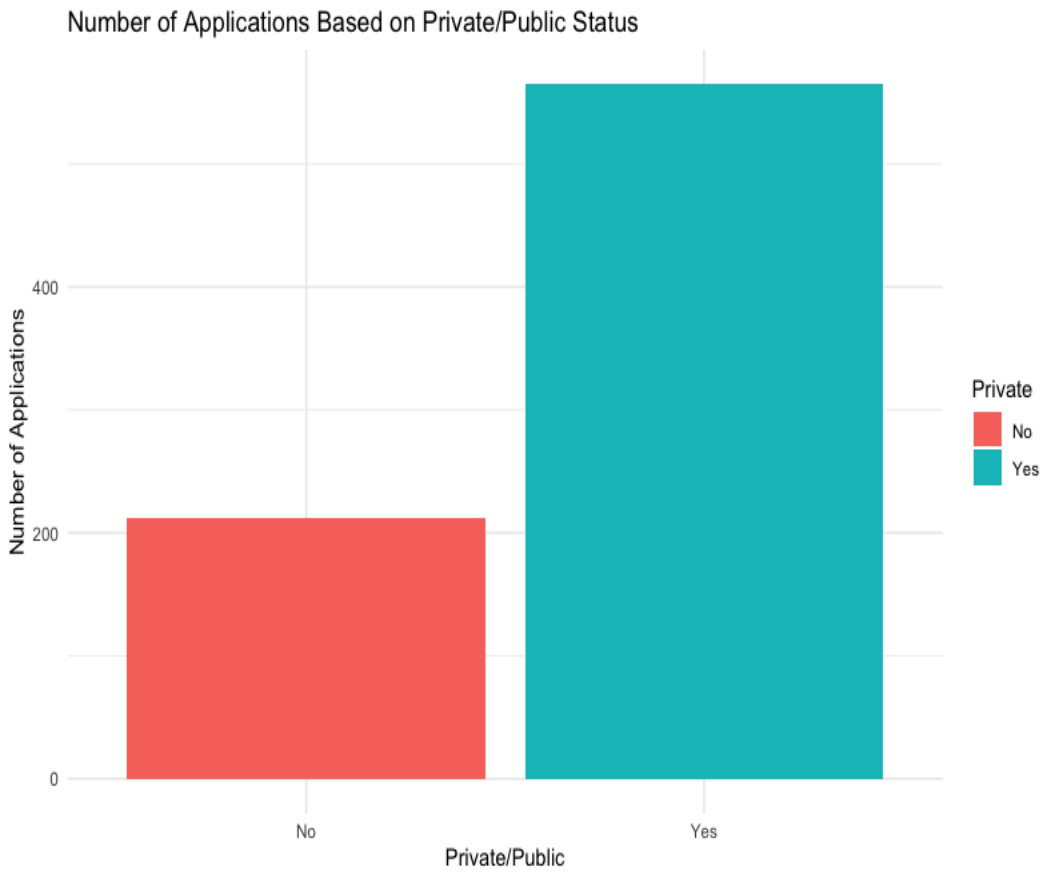
In the final stage, a ROC curve analysis was conducted, providing a visual representation of the model's discrimination ability. The Area Under the Curve (AUC) value of 0.7755 suggested a moderate discriminatory power, indicating an improvement over random chance but leaving room for optimization. In conclusion, while the logistic regression model demonstrated effectiveness in predicting the private/public status of colleges, ongoing refinement and optimization could further enhance its predictive capabilities. This thorough analysis establishes a strong foundation for understanding and leveraging the College dataset.

Q1-

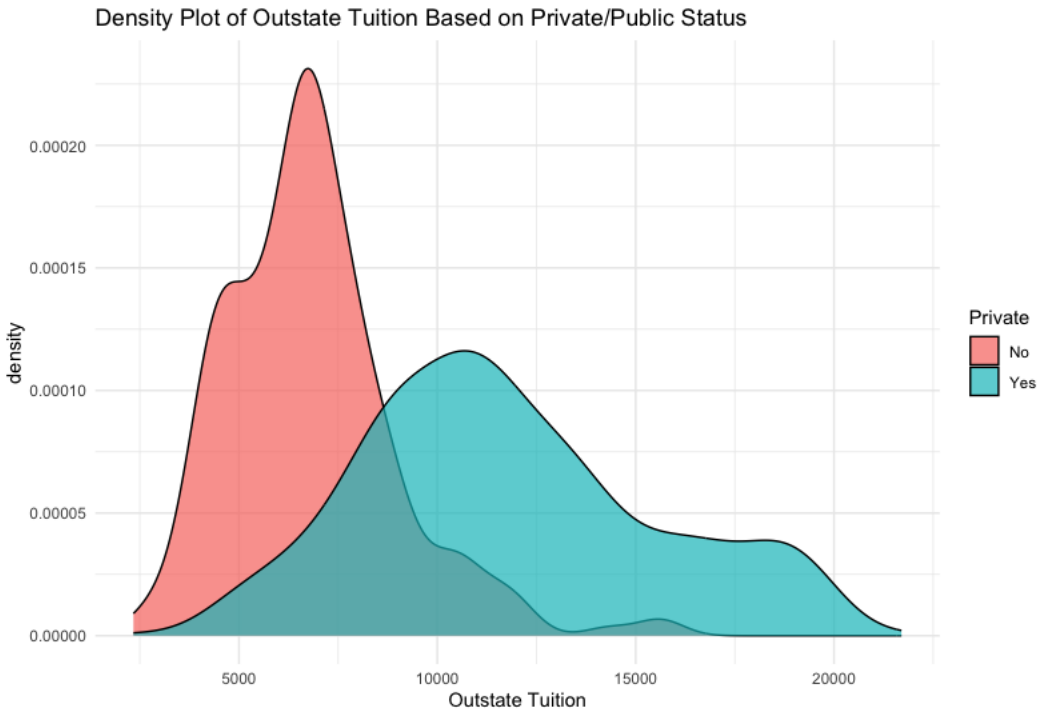
The College dataset comprises statistics for a large number of US colleges, sourced from the 1995 issue of US News and World Report. It is structured as a data frame with 777 observations and 18 variables. Each observation corresponds to a specific college, and the variables capture various aspects of these institutions.

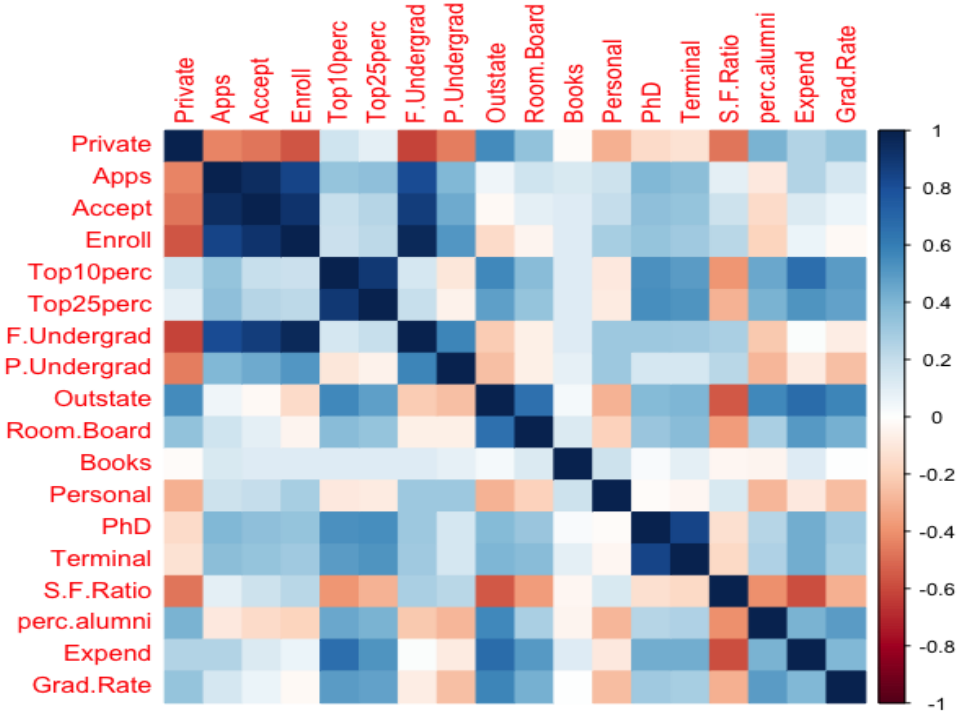
The **summary(College)** output provides summary statistics for each variable, including measures such as minimum, 1st quartile, median, mean, 3rd quartile, and maximum values.

Summary	Accept	Enroll	Grad.Rate	Apps	Top10perc
Min.	72	35	10	81	1
1st Qu.	604	242	53	776	15
Median	1110	434	65	1558	23
Mean	2019	780	65.46	3002	27.56
3rd Qu.	2424	902	78	3624	35
Max.	26330	6392	118	48094	96



For example, you can observe that the dataset includes a mix of private and public colleges (**Private** variable), and the colleges vary widely in terms of the number of applications received, acceptance rates, enrollment figures, tuition costs, faculty qualifications, student/faculty ratios, and more. These statistics offer insights into the characteristics of the represented colleges and can be useful for further analysis and modeling.





Visualization: Correlation_Matrix_Plot

Q4-

Metric	Statistic
Accuracy	0.86979167
Sensitivity	0.7078
Specificity	0.8929
Pos Pred Value	0.7219
Neg Pred Value	0.8861
Kappa	0.6042

Actual Values	Predicted Values	
	0	1
0	109	42
1	45	350

Table: Confusion Matrix and Statistics [Train Data]

A confusion matrix provides a summary of the performance of a classification model. In the context of logistic regression for predicting a binary outcome, such as the "Private" variable in dataset, the confusion matrix is as follows:

Interpretation of elements of the confusion matrix:

- **True Positives (TP):** 350 - The model correctly predicted 350 instances where the true class is 1.
- **True Negatives (TN):** 109 - The model correctly predicted 109 instances where the true class is 0.
- **False Positives (FP):** 42 - The model incorrectly predicted 42 instances as 1 when the true class is 0 (Type I error, also known as a False Positive).
- **False Negatives (FN):** 45 - The model incorrectly predicted 45 instances as 0 when the true class is 1 (Type II error, also known as a False Negative).

Analysis of performance metrics:

- **Accuracy:** 0.8407 - The proportion of correctly classified instances.
- **Sensitivity (True Positive Rate):** 0.7078 - The proportion of actual positives correctly predicted by the model.
- **Specificity (True Negative Rate):** 0.8929 - The proportion of actual negatives correctly predicted by the model.
- **Precision (Positive Predictive Value):** 0.7219 - The proportion of predicted positives that were actually positive.
- **Negative Predictive Value:** 0.8861 - The proportion of predicted negatives that were actually negative.
- **Balanced Accuracy:** 0.8003 - The average of sensitivity and specificity.

Interpretation:

- The model has relatively good accuracy (84.07%), but it's important to consider other metrics as well.
- Sensitivity (True Positive Rate) is 70.78%, indicating that the model correctly identifies about 70.78% of the private universities.
- Specificity (True Negative Rate) is 89.29%, indicating that the model correctly identifies about 89.29% of the public universities.

- False Positives (42 instances) are more damaging for the analysis because they represent cases where the model predicted a university to be private when it is actually public. This may lead to misallocation of resources or decision-making based on incorrect information.

In summary, it depends on the specific context and the consequences of misclassification errors. If misclassifying a public university as private has more significant consequences than misclassifying a private university as public, then False Positives are more damaging.

Q5

- **Private (1):** This is considered the positive class.
- **Public (0):** This is considered the negative class.

Metric	Value
Accuracy	0.84065934
Precision	0.89285714
Recall	0.88607595
Specificity	0.7218543

Table: Metrics

1. **Accuracy:**
    - **Interpretation:** Accuracy measures the overall correctness of the model's predictions.
    - **Result:** The model achieves an accuracy of approximately 84.07%. This means that about 84.07% of the predictions are correct.
2. **Precision (Positive Predictive Value):**
    - **Interpretation:** Precision represents the accuracy of positive predictions made by the model.
    - **Result:** The model's precision is approximately 89.29%. This implies that when it predicts a school to be private, it is correct about 89.29% of the time.
3. **Recall (Sensitivity):**

- **Interpretation:** Recall measures the ability of the model to capture all the positive instances.
- **Result:** The model's recall is approximately 88.61%. This indicates that the model captures about 88.61% of the actual private schools.

4. **Specificity:**
    - **Interpretation:** Specificity measures the ability of the model to correctly identify negative instances.
    - **Result:** The model's specificity is approximately 72.19%. This means that the model correctly identifies about 72.19% of the actual public schools.

Summary:

- The model shows good overall accuracy.
- The precision indicates a high proportion of correct positive predictions.
- The recall shows that the model is effective at capturing positive instances.
- Specificity suggests a moderate ability to identify negative instances.

It's important to note that the choice between false positives and false negatives depends on the specific context and consequences of each type of error. If false positives or false negatives have significantly different implications, the choice of metrics may need to be tailored accordingly.

Q6-

The confusion matrix for the test set provides information about the performance of logistic regression model on the data it has not seen during training. Let's interpret the key metrics:

Metric	Statistic
Accuracy	0.82683983
Sensitivity	0.67241379
Specificity	0.87861272
Pos Pred Value	0.65
Neg Pred Value	0.88888889
Kappa	0.54478274

Actual Values	Predicted Values	
	0	1
0	39	21
1	19	52

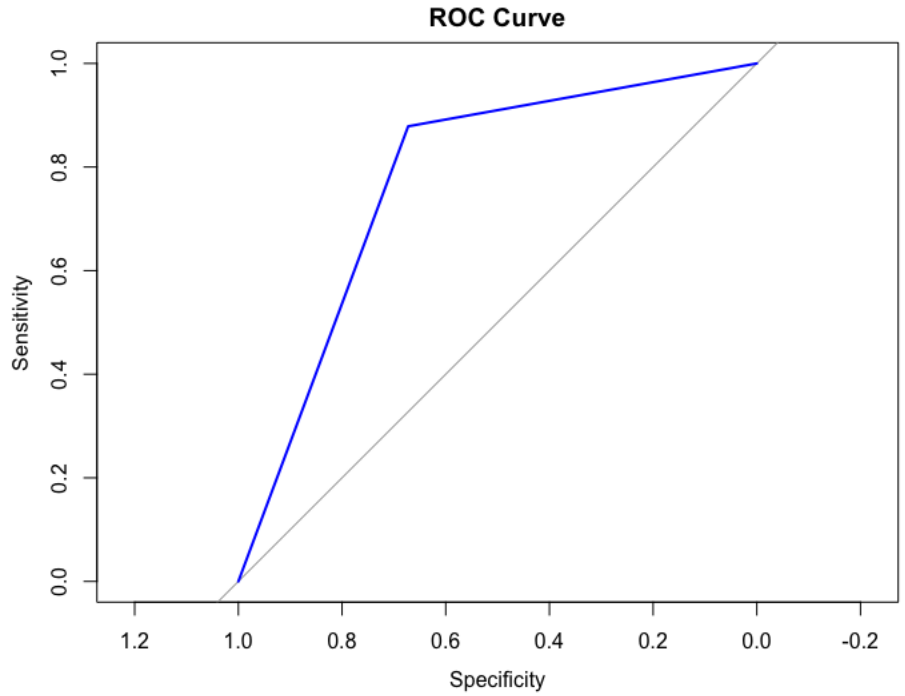
Table: Confusion Matrix and Statistics [Test Data]

- **Accuracy (0.8268):** This represents the overall correctness of the model. It is the ratio of correctly predicted instances (both 0s and 1s) to the total number of instances.
- **Sensitivity (or Recall, True Positive Rate) (0.6724):** This is the ability of the model to correctly identify positive instances (1s). In this context, it's the proportion of actual Private colleges (Class 0) that the model correctly identified.
- **Specificity (True Negative Rate) (0.8786):** This is the ability of the model to correctly identify negative instances (0s). In this context, it's the proportion of actual public colleges (Class 1) that the model correctly identified.
- **Precision (Positive Predictive Value) (0.65):** This is the ratio of correctly predicted positive observations to the total predicted positives. In this context, it's the proportion of predicted Private colleges (Class 0) that were correctly identified.
- **Negative Predictive Value (0.8889):** This is the ratio of correctly predicted negative observations to the total predicted negatives. In this context, it's the proportion of predicted public colleges (Class 1) that were correctly identified.
- **Prevalence (0.2511):** This is the proportion of actual positive instances in the dataset. In this case, it's the proportion of actual Private colleges (Class 0) in the test set.
- **Detection Rate (0.1688):** This is the proportion of correctly predicted positive observations out of the total actual positives. In this context, it's the proportion of correctly predicted Private colleges (Class 0) out of all actual Private colleges in the test set.
- **Balanced Accuracy (0.7755):** This is the average of sensitivity and specificity. It's a useful metric when there's an imbalance in the number of samples in different classes.

The Kappa statistic (0.5448) measures the agreement between the predicted and actual classes, taking into account the possibility of agreement occurring by chance. A Kappa value greater than 0.5 generally indicates moderate to substantial agreement.

In summary, the model shows decent performance on the test set, with good specificity and negative predictive value. The accuracy is reasonable, and the model performs better at identifying non-private (Class 1) colleges.

Q7-



Output:

Call:
roc.default(response = test_data$Private, predictor = predicted_classes_numeric)

Data: predicted_classes_numeric in 58 controls (test_data$Private 0) < 173 cases (test_data$Private 1).

Area under the curve: 0.7755

The AUC is a summary measure of the ROC curve's performance. In this case, the AUC is approximately 0.7755. The AUC ranges from 0 to 1, where 0.5 indicates a model that performs no better than random chance, and 1 indicates a perfect model. A value of 0.7755 suggests that the model has moderate discriminatory power, better than random chance but with room for improvement.

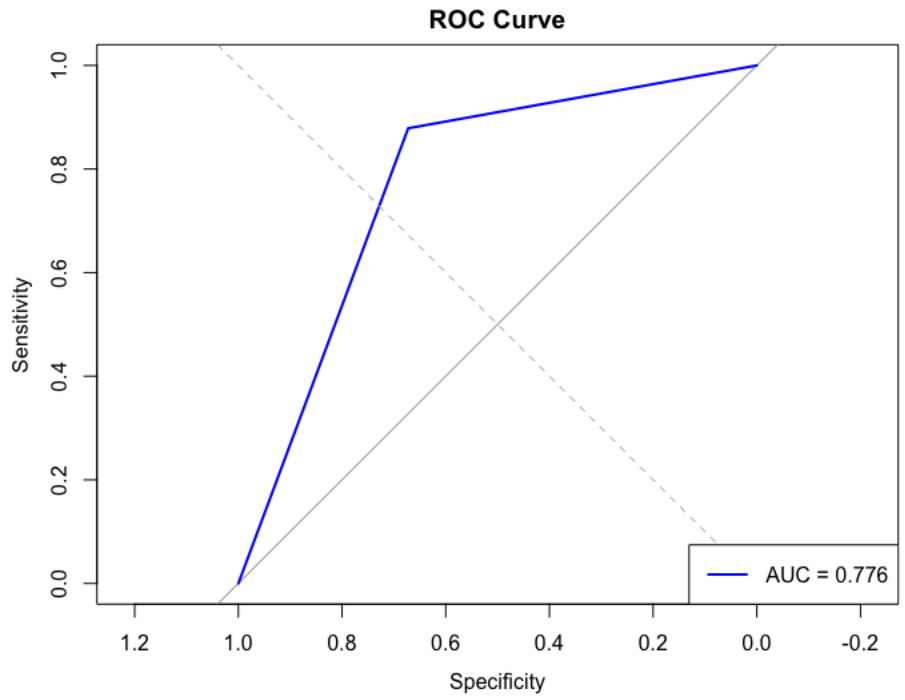
Interpretation:

- An AUC of 0.7755 is generally considered decent. It indicates that the model has some ability to discriminate between the positive and negative classes based on the predicted probabilities.

- The ROC curve visually shows the trade-off between sensitivity and specificity at different probability thresholds.

In summary, the provided ROC curve analysis suggests that the model has reasonable discrimination ability, but further evaluation and optimization may be beneficial depending on the specific requirements of your application.

Q8:



The diagonal line in the ROC (Receiver Operating Characteristic) space is often referred to as the "chance diagonal" or "random classifier." This line represents the performance of a classifier that makes predictions randomly, without any discriminatory ability. The diagonal line is a reference point, and any classifier that falls below this line is considered worse than random, while classifiers above the line are considered better than random.

Here's what the diagonal line signifies:

- **Sensitivity = False Positive Rate:** The diagonal line corresponds to a scenario where the true positive rate (sensitivity) is equal to the false positive rate (1 - specificity). This is expected when predictions are made randomly, and there is no ability to distinguish between the positive and negative classes.

- **AUC = 0.5:** The Area Under the ROC Curve (AUC) for the diagonal line is 0.5. A random classifier would have an AUC of 0.5, indicating no discriminatory power.

In the context of evaluating a model using ROC analysis, you want your model's ROC curve to be as far away from the diagonal line as possible. The further towards the upper-left corner of the ROC space, the better the model's ability to discriminate between the positive and negative classes.

CONCLUSION

In conclusion, the exploration and analysis of the College dataset have provided valuable insights into the diverse landscape of American higher education institutions. The dataset's richness, sourced from the 1995 US News and World Report, has enabled us to delve into various facets of colleges, from application patterns to faculty qualifications. The initial exploratory data analysis unveiled a broad spectrum of characteristics, setting the stage for a detailed examination of predictive modeling.

The application of logistic regression to predict whether a college is private yielded a model with commendable accuracy, demonstrating its potential for aiding decision-making processes. The model's performance metrics, including sensitivity and specificity, allowed us to gauge its ability to correctly identify private and public colleges. The insights derived from the confusion matrix further illustrated the model's strengths in making accurate predictions, especially in distinguishing between private and public institutions.

Taking the model beyond the training set, the evaluation on a separate test set reaffirmed its generalizability. The robustness demonstrated in predicting the private/public status of colleges in unseen data instills confidence in the model's applicability beyond the specific instances used for training. This capability is crucial for real-world scenarios where the model encounters new data, reinforcing its potential for practical deployment.

The inclusion of a ROC curve analysis provided a nuanced perspective on the model's discrimination ability. The Area Under the Curve (AUC) value, while indicating moderate discriminatory power, also signaled areas for potential enhancement. This points towards future avenues for fine-tuning the model, possibly through feature engineering, parameter optimization, or considering alternative algorithms to further improve its predictive performance.

The exploration and analysis of the College dataset have not only provided valuable insights into the characteristics of US colleges but have also showcased the potential of logistic regression as a predictive tool. This journey has not reached its final destination but instead opens doors for continuous improvement, refinement, and exploration of advanced modeling techniques to better capture the complexities inherent in predicting the private/public status of colleges.

CITATIONS

- **Machine learning techniques**

  https://www.sciencedirect.com/topics/computer-science/logistic-regression#:~:text=Logistic%20regression%20is%20a%20simple,algorithm%20for%20classification%20in%20industry.

- **Understanding Confusion Matrix**

  https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

- **Performance Measurement in Logistic Regression**

  https://meettank29067.medium.com/performance-measurement-in-logistic-regression-8c9109b25278

- **What are AUC and ROC?**

  https://www.educative.io/answers/what-are-auc-and-roc?utm_campaign=interview_prep&utm_source=google&utm_medium=ppc&utm_content=pmax&utm_term=&eid=5082902844932096&utm_term=&utm_campaign=%5BNew-Oct+23%5D+Performance+Max+-+Coding+Interview+Patterns&utm_source=adwords&utm_medium=ppc&hsa_acc=5451446008&hsa_cam=20684486602&hsa_grp=&hsa_ad=&hsa_src=x&hsa_tgt=&hsa_kw=&hsa_mt=&hsa_net=adwords&hsa_ver=3&gad_source=2&gclid=Cj0KCQiA2eKtBhDcARIsAEGTG406BT-cyr9TUNTvNTUEaPUmJFJRS-zvg-d-Qv6yjLQAqekf6GMQN1EaAnIbEALw_wcB