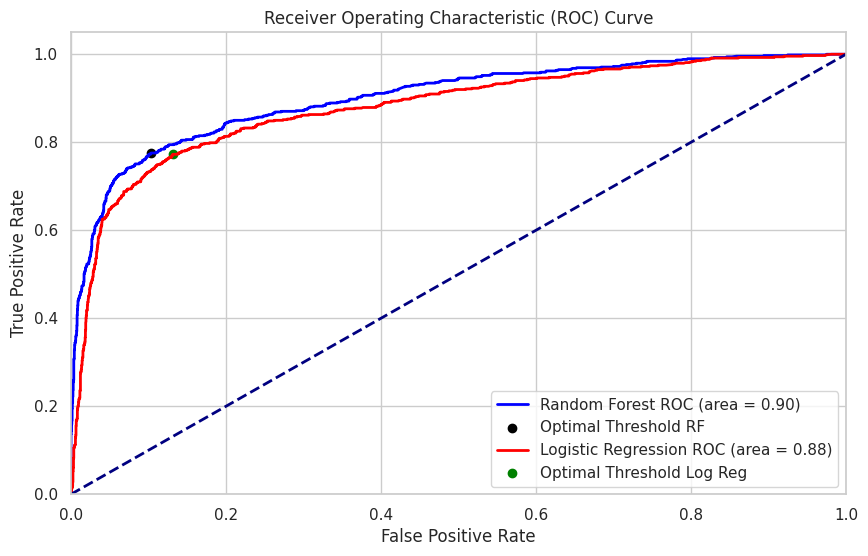
1. **Data Preparation:**
   * **Separate your dataset into two parts:**
     + **Training data (data\_train):** Participants who responded 'Yes' or 'No' to the experience of hiring discrimination.
     + **Prediction data (data\_predict):** Participants who responded 'NA'.
2. **Feature Scaling:**
   * Standardize your features (excluding the target variable) using StandardScaler to ensure that all features contribute equally to the model.
3. **Model Training with Cross-Validation:**
   * **Random Forest:**
     + Use GridSearchCV with RandomForestClassifier.
     + Define a parameter grid (e.g., {'n\_estimators': [100, 200, 300], 'max\_depth': [5, 10, 15]}).
     + Set scoring='roc\_auc' for optimizing the AUC during cross-validation.
   * **Penalized Logistic Regression with Lasso (L1) Penalty:**
     + Use GridSearchCV with LogisticRegression (set penalty='l1' and solver='liblinear').
     + Define a parameter grid for the regularization parameter C (e.g., {'C': [0.01, 0.1, 1, 10, 100]}).
     + Also set scoring='roc\_auc'.
4. **Model Evaluation and Selection:**
   * Compare the AUC scores of the best models from both Random Forest and Penalized Logistic Regression.
   * Select the model with the higher AUC score as your final model.
5. **Fit the Best Model:** After selecting the best model based on the AUC score, fit this model to your training data.
6. **Compute the ROC Curve:** Use roc\_curve from sklearn.metrics to get the true positive rates (TPR), false positive rates (FPR), and thresholds.
7. **Find the Optimal Threshold**: Calculate the sum of sensitivity and specificity for each threshold, and identify the threshold that maximizes this sum.
8. **Use the Optimal Threshold for Prediction:** Use this threshold to convert predicted probabilities into binary predictions.



[Find the optimal threshold probability which gives sensitivity and specifity is maximized for best model]

**Random Forest Predictions (predicted\_rf):**

* **Males (gender = 0)**:
  + About 51.56% predicted as **False** (no hiring discrimination experienced).
  + About 48.44% predicted as **True** (hiring discrimination experienced).
* **Females (gender = 1)**:
  + About 9.09% predicted as **False**.
  + About 90.91% predicted as **True**.

**Logistic Regression Predictions (predicted\_log\_reg):**

* **Males (gender = 0)**:
  + About 53.12% predicted as **False**.
  + About 46.88% predicted as **True**.
* **Females (gender = 1)**:
  + About 12.12% predicted as **False**.
  + About 87.88% predicted as **True**.

**Interpretation:**

* **Random Forest Model**: This model suggests a relatively balanced distribution of hiring discrimination experiences among males, with nearly equal proportions of **True** and **False** predictions. For females, the model predicts a significantly higher experience of hiring discrimination (**True**).
* **Logistic Regression Model**: Similar to the Random Forest model, the Logistic Regression model also predicts a relatively balanced distribution for males, with a slight leaning towards **False** (no discrimination). However, for females, the model predicts a very high rate of experiencing hiring discrimination (**True**).

**Overall Insights:**

* Both models predict a higher rate of hiring discrimination experiences among females compared to males. This is indicated by a larger proportion of **True** predictions for females in both models.
* For males, the predictions are more evenly distributed between **True** and **False**, suggesting a less clear pattern of hiring discrimination experience compared to females.

**Considerations:**

* These results suggest that there might be a difference in the reporting or experience of hiring discrimination between males and females, with females more likely to experience or report discrimination based on these models' predictions.
* It's important to validate these models for accuracy and ensure they are not reflecting biases present in the data.
* The context and methodology of how hiring discrimination was measured and reported are crucial for interpreting these results accurately.
* Caution should be exercised in drawing definitive conclusions from predictive models alone, as they can reflect existing patterns in the data, which may or may not indicate causal relationships.

