Classifying Admission to College in US

Prepared by Anita

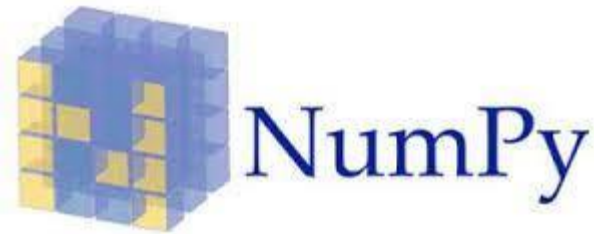
Objective

- For the Education Department to know the total number of applications successfully accepted and visas processed by correctly classify the relevant result which is the successful admission from the dataset.

Methodology

- Data Cleaning
- Exploratory Data Analysis
- Feature Selection
- Model Selection
- Model Evaluation

Tools Used



Data Collection

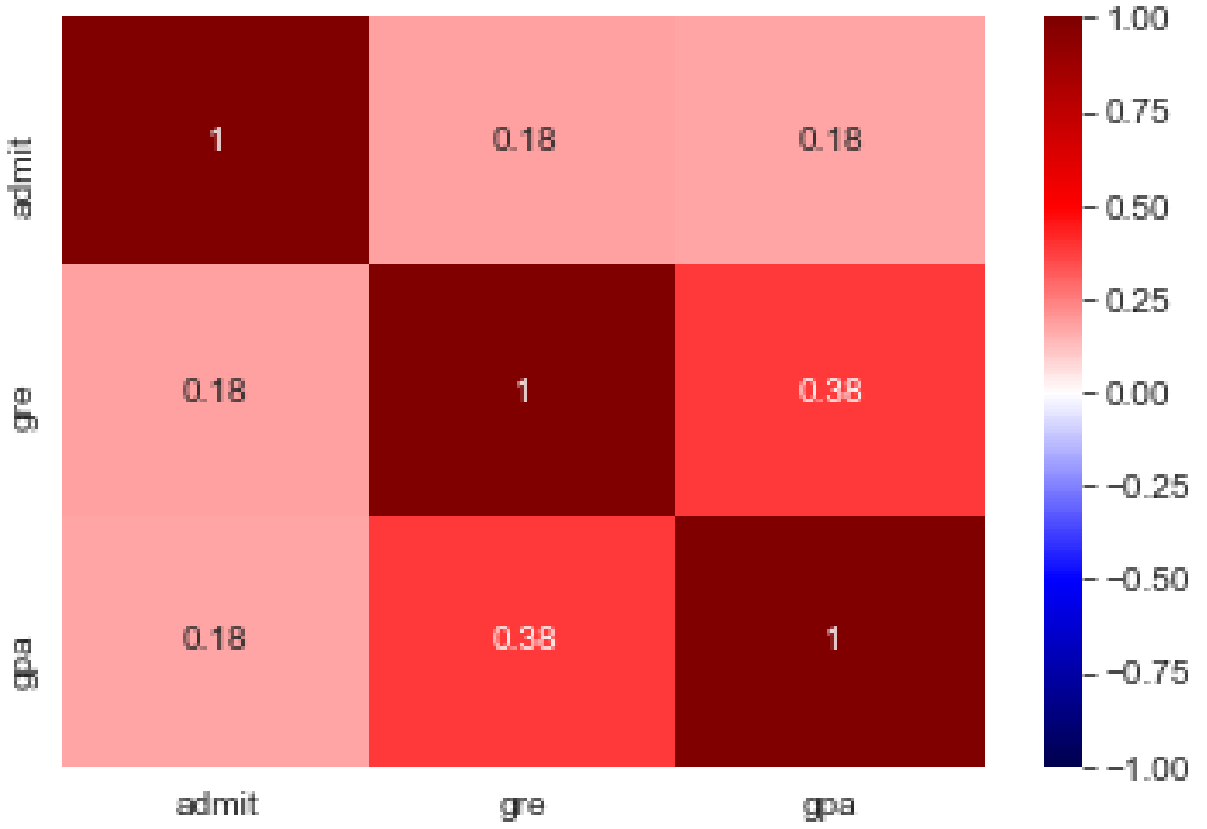
- From www.kaggle.com/datasets
- Dataset description:
  - GRE: Graduate Record Exam Scores
  - GPA: Grade Point Average
  - Rank: refers to the prestige of the undergraduate institution. Institutions with a rank of 1 have the highest prestige, while those with a rank of 4 have the lowest.
  - Admit: is a binary target where 1 indicates that student is admitted and 0 indicates that student is not admitted.
  - SES: refers to socioeconomic status; 1-low, 2-medium, 3-high.
  - Gender: 0 -> Female, 1 -> Male
  - Race: 1 -> Hispanic, 2 -> Asian, 3 -> African-American

Data Cleaning

- 400 rows
- 7 columns:
  - admit (target)
  - gre (numerical feature)
  - gpa (numerical feature)
  - ses (categorical feature)
  - race (categorical feature)
  - rank (categorical feature)
  - gender (categorical feature)

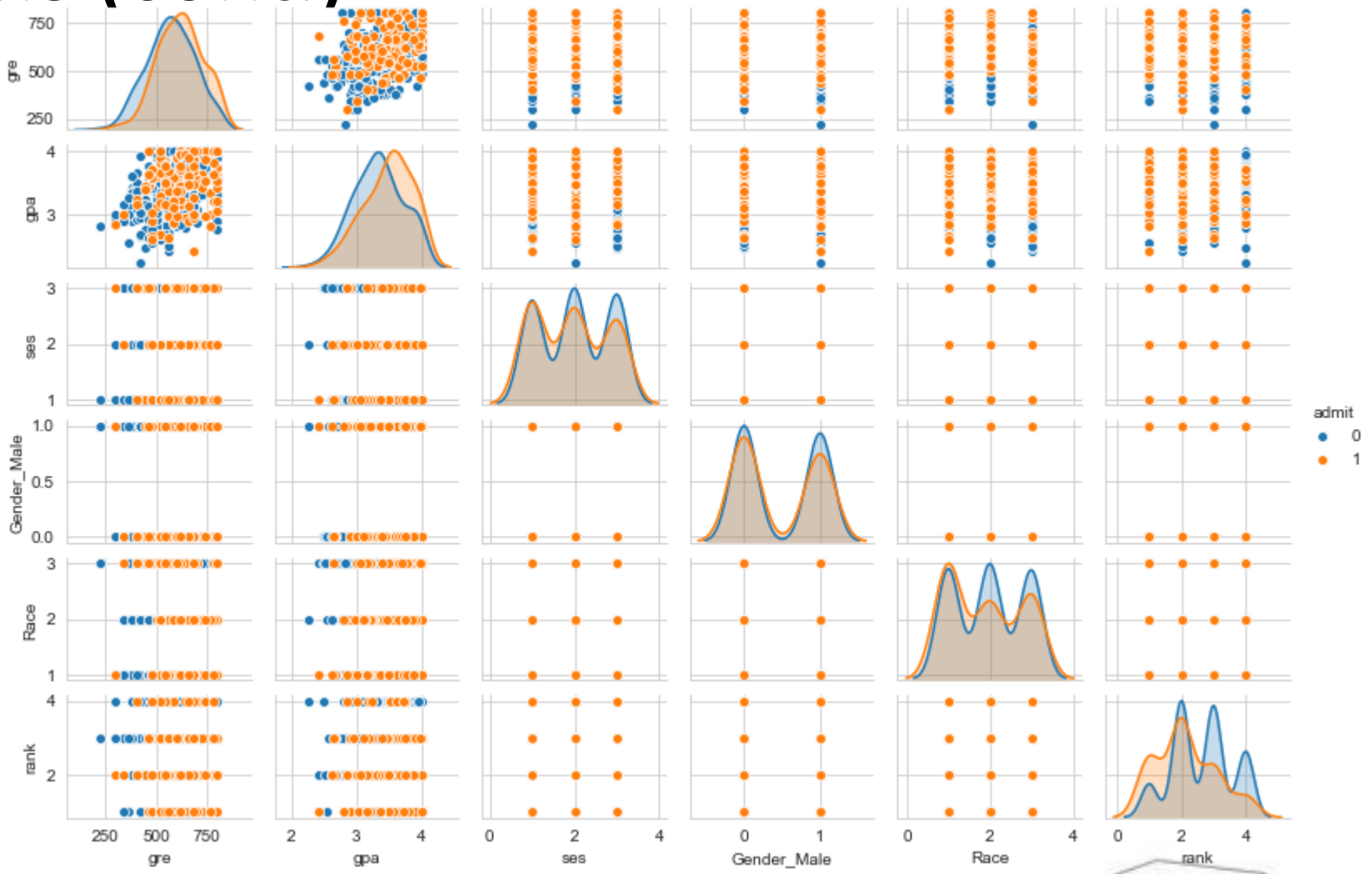
Data Analysis

- To view the collinearity between feature to feature.
- Observed collinearity between gre and gpa, therefore gpa feature is removed.



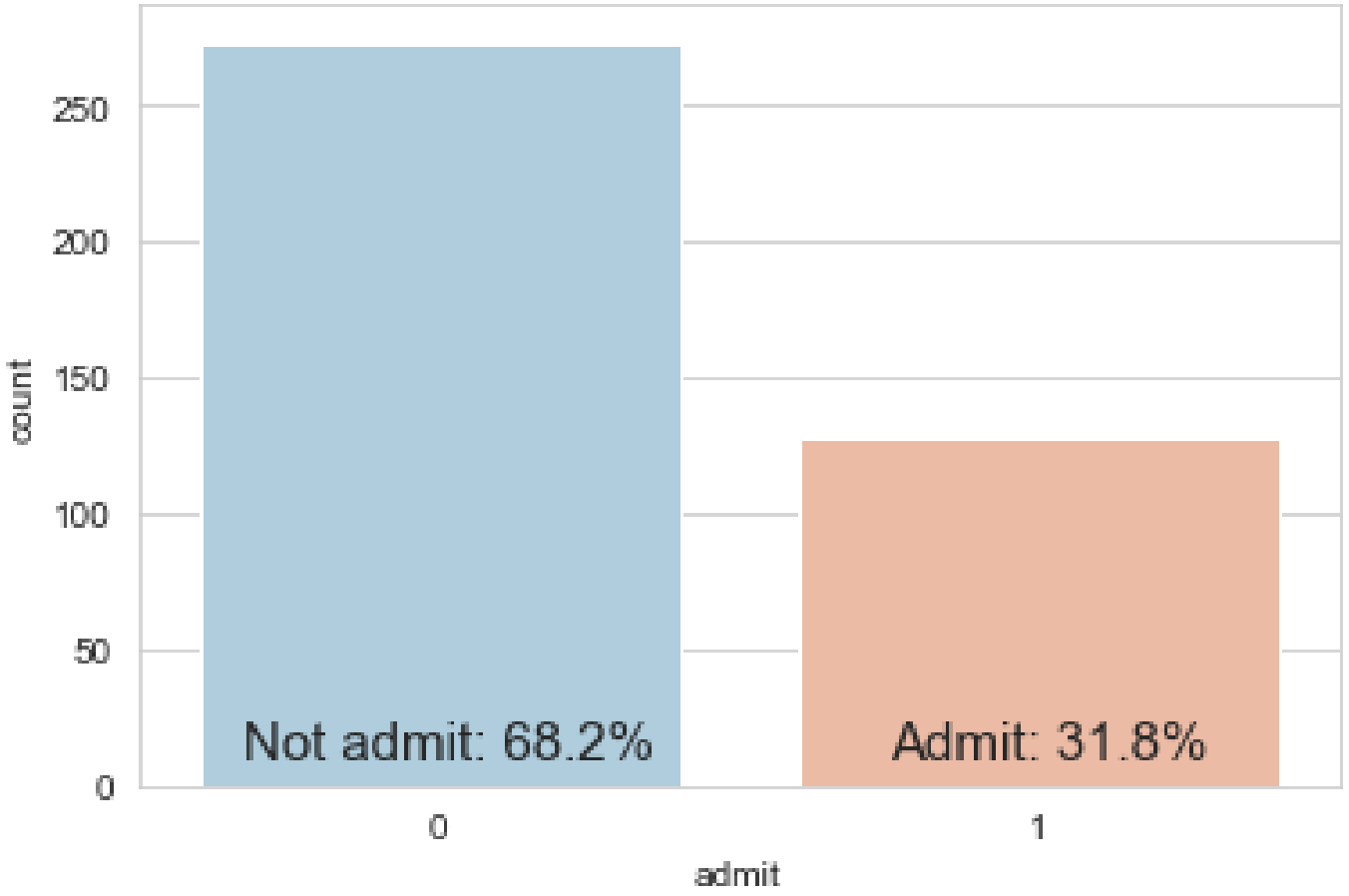
Data Analysis (cont.)

- To view the distribution plot of the feature by the target.



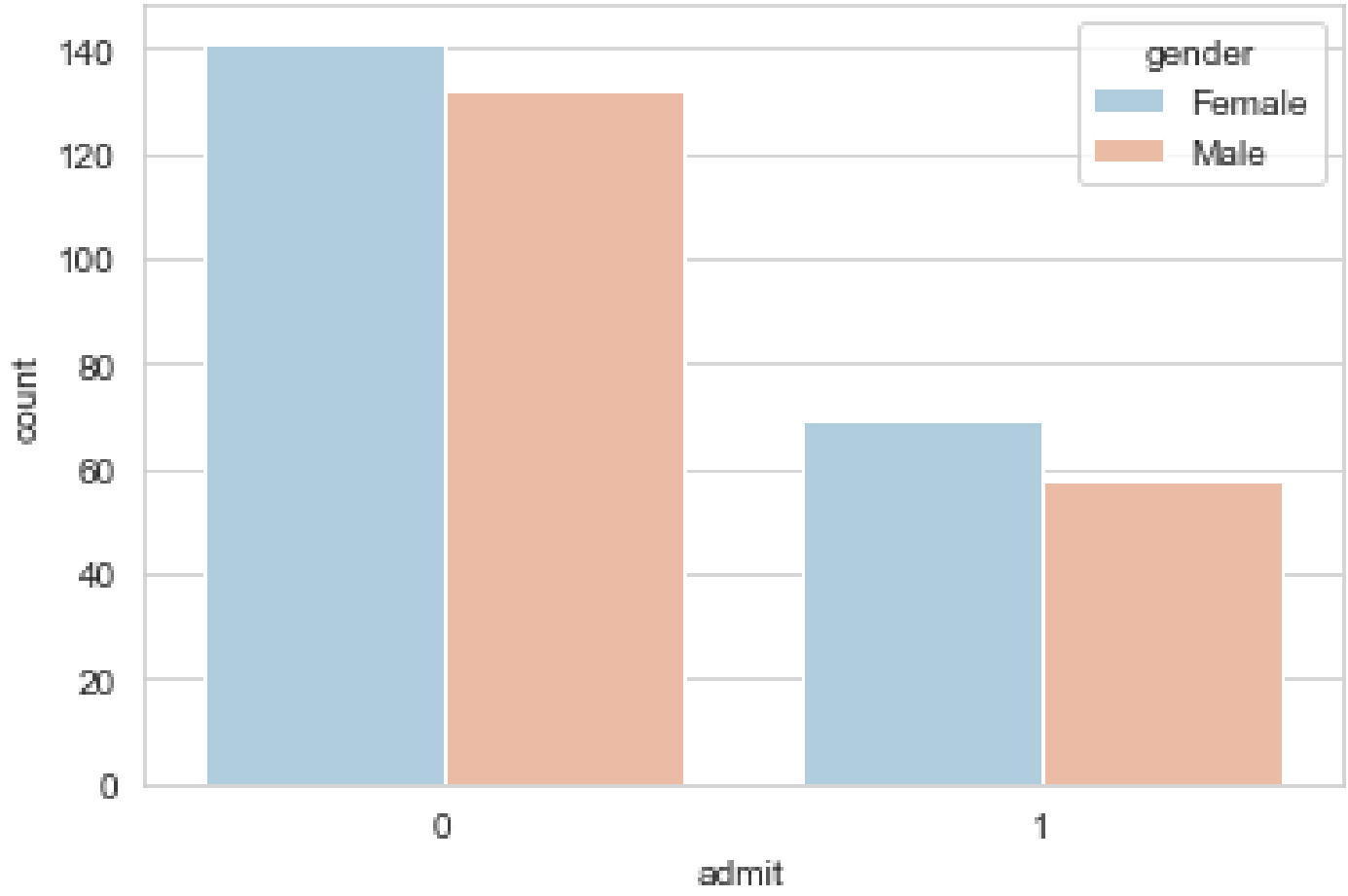
Data Analysis (cont.)

- The plot of the target, i.e. admission.



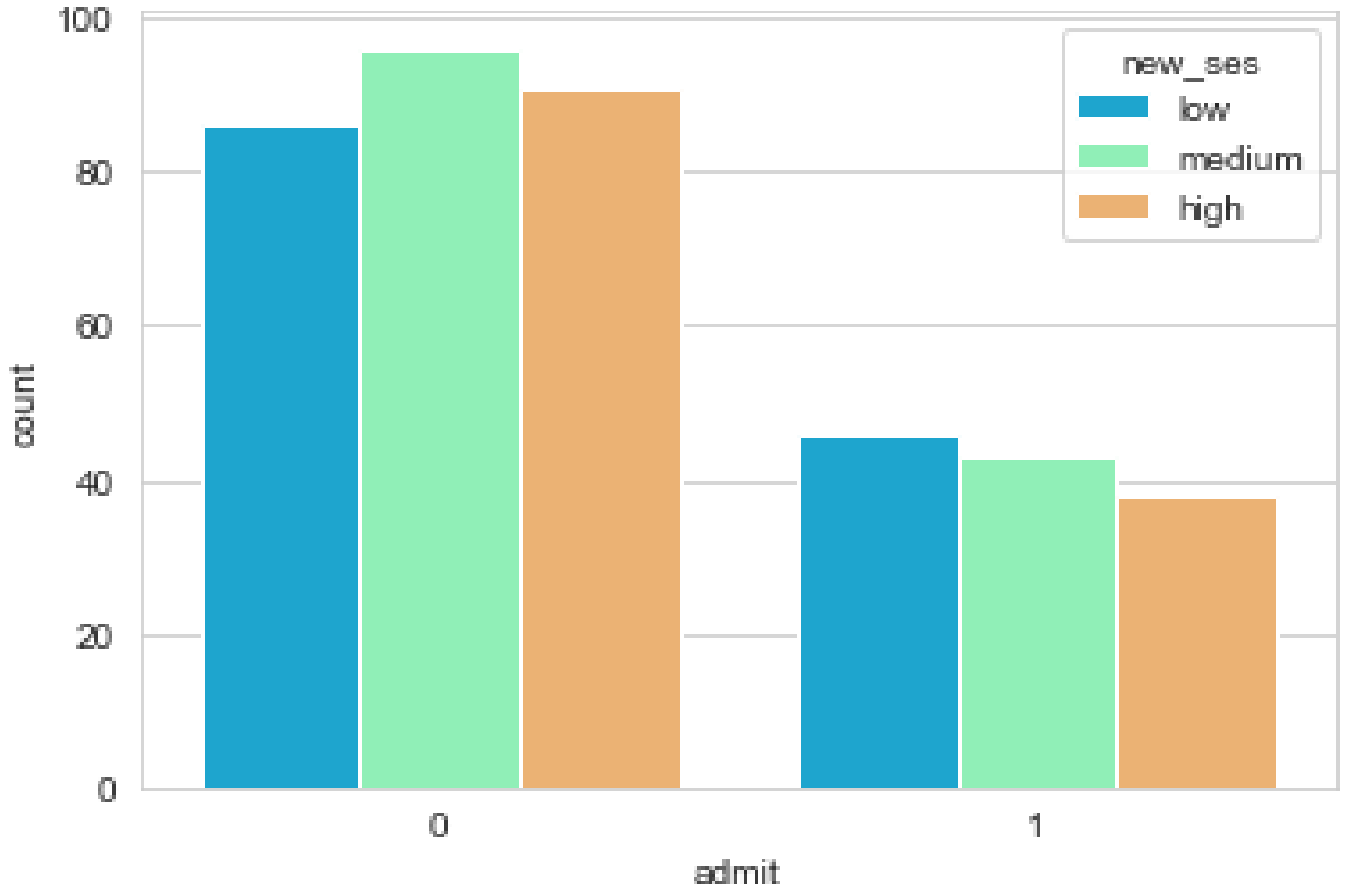
Data Analysis (cont.)

- The plot of the target, i.e. admission by gender.



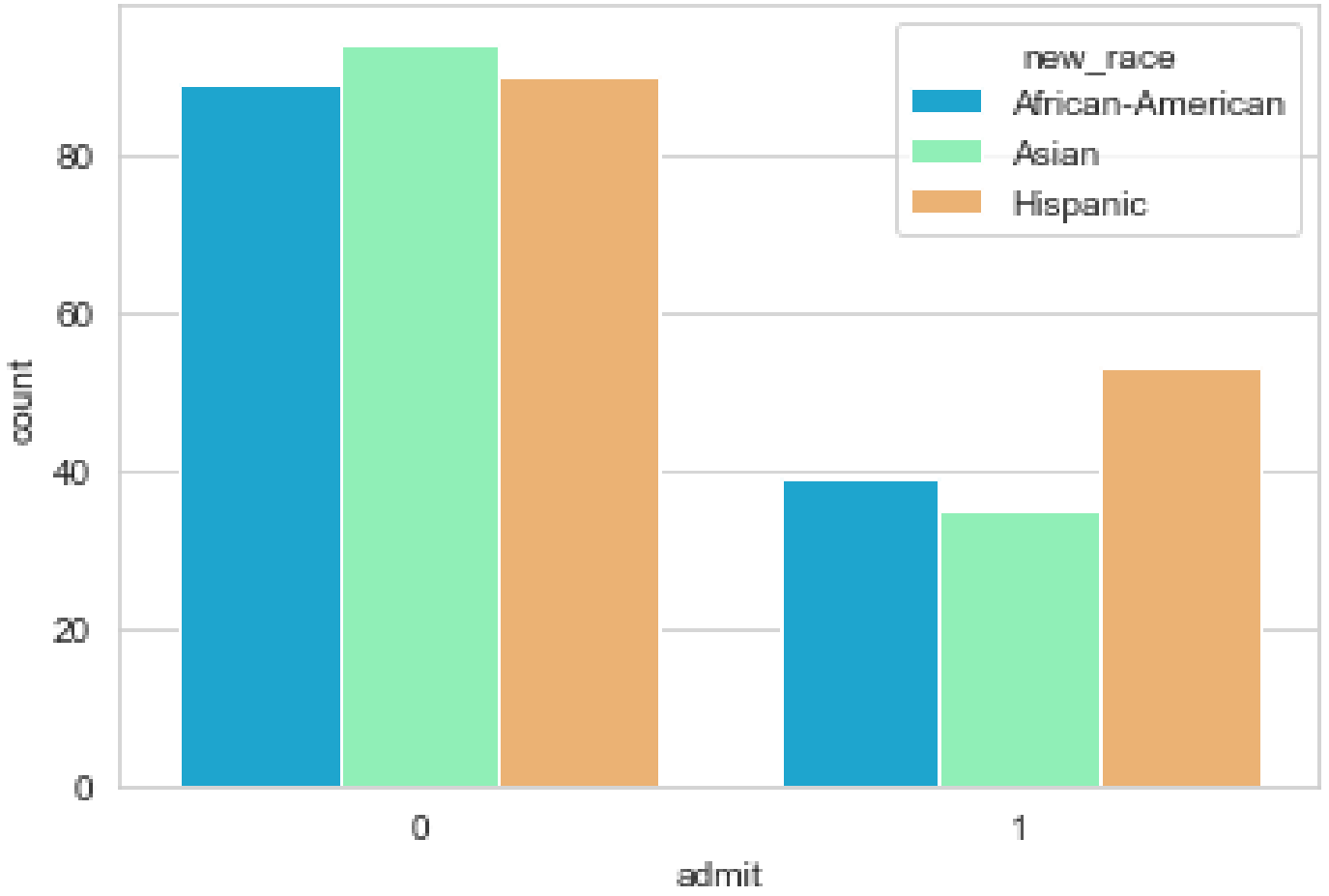
Data Analysis (cont.)

- The plot of the target, i.e. admission by ses.



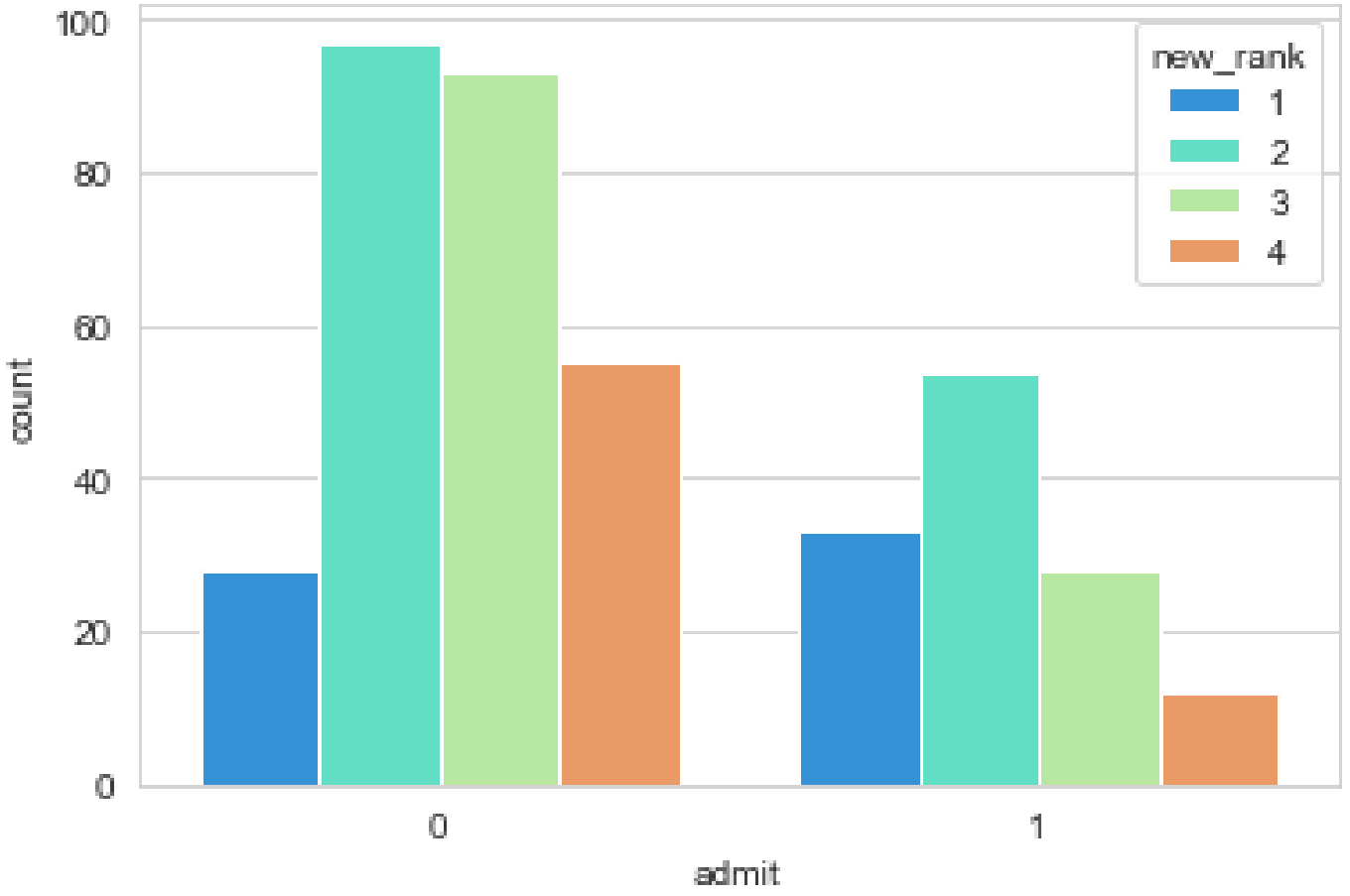
Data Analysis (cont.)

- The plot of the target, i.e. admission by race.



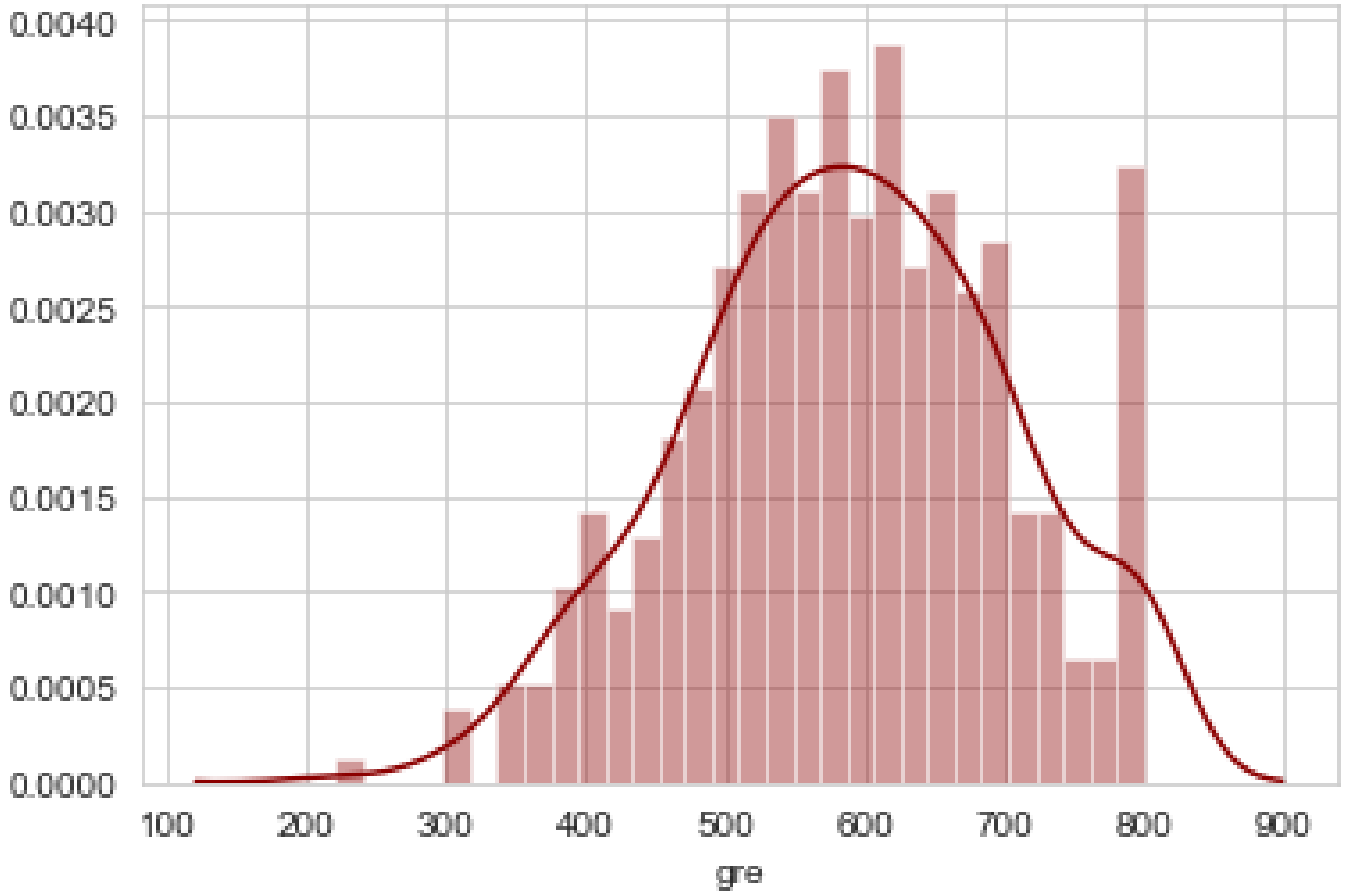
Data Analysis (cont.)

- The plot of the target, i.e. admission by rank.



Data Analysis (cont.)

- The distribution plot of gre feature.



Data Analysis (cont.)

- Create dummy variables into dataset.
  - SES
  - Race
  - Rank
  - Gender

Feature Selection

- Using Logistic Lasso, no feature is removed.

Feature	Selected
GRE	True
SES_Low	True
SES_Medium	True
Race_Asian	True
Race_Hispanic	True
Rank_2	True
Rank_3	True
Rank_4	True
Gender_male	True

Cross Validation

Model	Accuracy	Precision	Recall	F1 score
GaussianNB	0.643750	0.424542	0.389685	0.403315
LogisticRegression	0.700000	0.582308	0.223502	0.307467
KNN	0.593750	0.295495	0.192397	0.230767
DecisionTree	0.593750	0.368563	0.319321	0.344101
RandomForest	0.621875	0.339876	0.249119	0.319497
SVC	0.675000	0.352381	0.080496	0.123030

GridSearchCV

- Validate the models on the F1 score metric.

Model	Best score	Best parameter
GaussianNB	0.390957	-
LogisticRegression	0.289661	C: 100, penalty: l2
KNN	0.339672	n_neighbors: 3
DecisionTree	0.317440	max_depth: 6
RandomForest	0.300968	n_estimators: 400
SVC	0.344547	C: 10, kernel: rbf

Model Evaluation

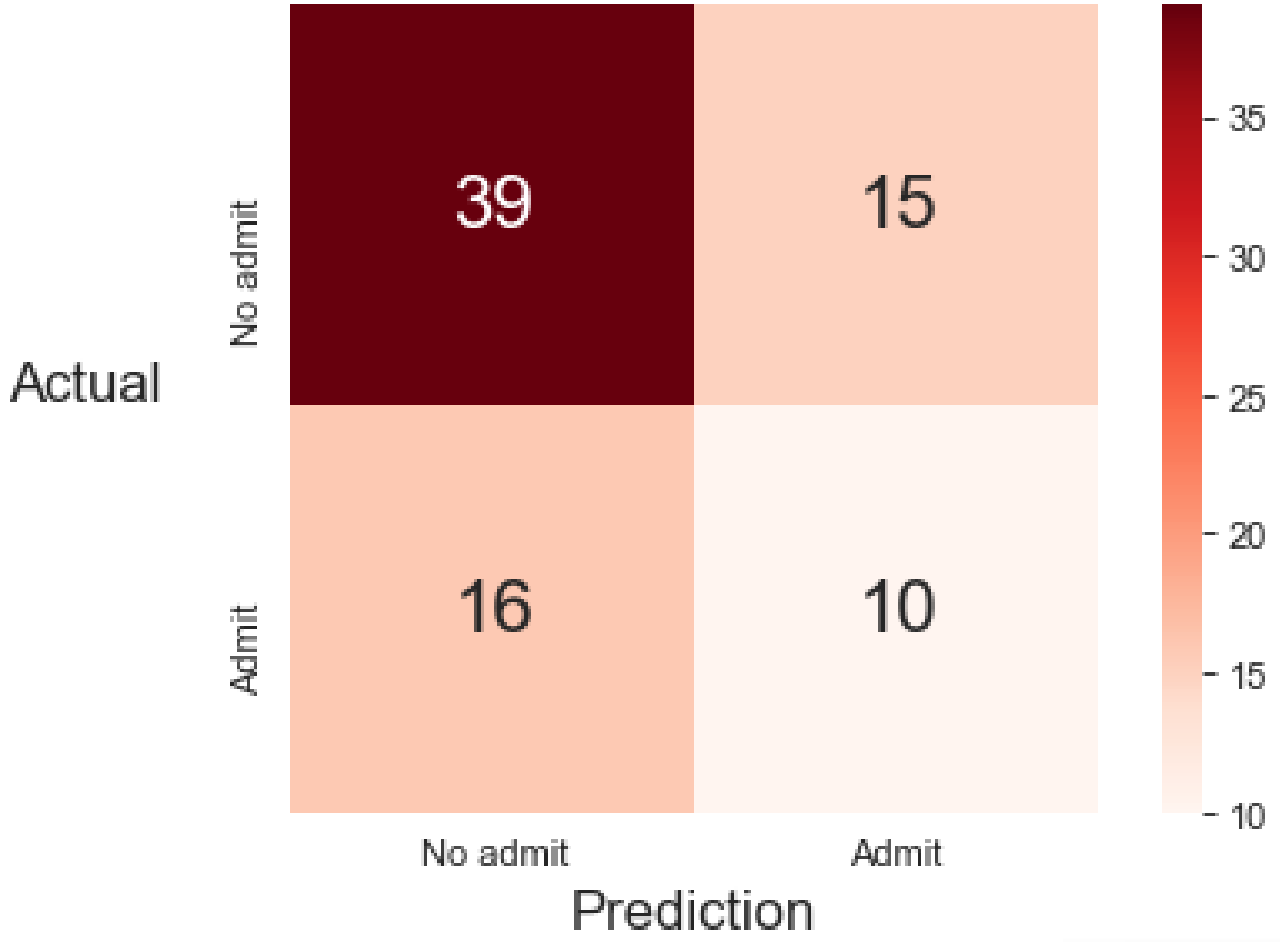
- Model evaluation performed on GaussianNB.

Metric	Score
Accuracy	0.6125
Precision	0.40
Recall	0.3846
ROC AUC	0.5534

```
Classification report for GaussianNB:
              precision    recall  f1-score   support

           0       0.71      0.72      0.72        54
           1       0.40      0.38      0.39        26

    accuracy                           0.61        80
   macro avg       0.55      0.55      0.55        80
weighted avg       0.61      0.61      0.61        80

Confusion Matrix for GaussianNB:
 [[39 15]
 [16 10]]
```

Model Evaluation (cont.)

- Confusion matrix plot:



Model Evaluation (cont.)

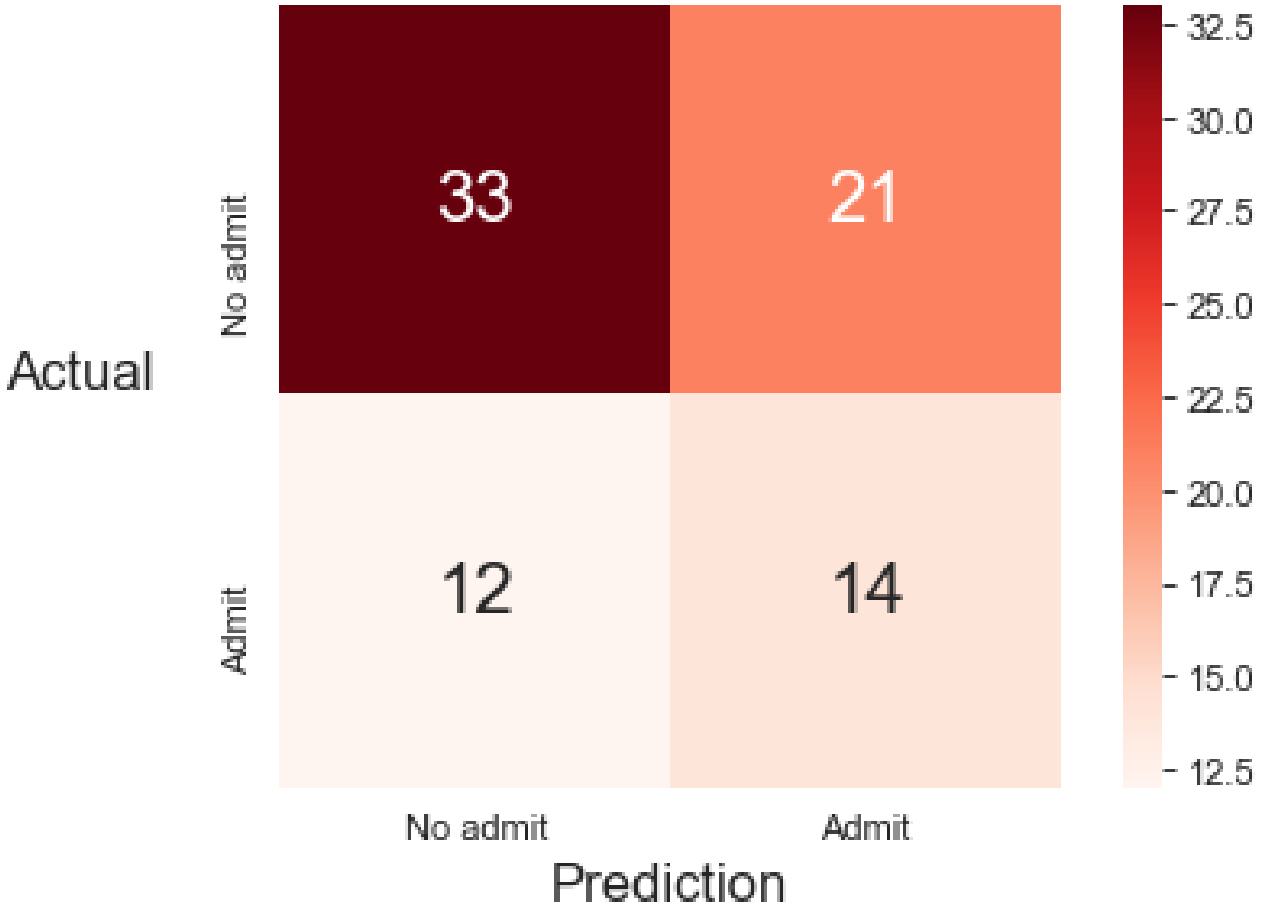
- Tuning the probability threshold to 0.37.

Metric	Score
Accuracy	0.5875
Precision	0.40
Recall	0.5385
ROC AUC	0.5748

```
Classification report for GaussianNB:
              precision    recall  f1-score   support

           0       0.73      0.61      0.67        54
           1       0.40      0.54      0.46        26

    accuracy                           0.59        80
   macro avg       0.57      0.57      0.56        80
weighted avg       0.62      0.59      0.60        80

Confusion Matrix for GaussianNB:
 [[33 21]
 [12 14]]
```

Model Evaluation (cont.)

- Confusion matrix plot:



Conclusions

- The model is still not good as the ratio that the model can correctly predicted positive observations (i.e. admission) to the all observations in actual class is only 0.54.
- And the ratio that the model can correctly predicted positive observations (i.e. admission) to the total predicted positive observations is only 0.40.

THANK YOU