Introduction

Each graduate application season, the Katz Graduate School of Business must strike a delicate balance in its marketing – how can it structure its outreach to target prospective candidates who are most likely to pursue an MBA, such that it maximizes its enrollment yield per unit expenditure? Using historical data containing information on individuals who indicated interest in the MBA program, I analyzed the characteristics that made prospective students more likely to pursue an MBA. Katz can utilize targeted advertisement and outreach initiatives to more directly reach people who match the profiles of these past candidates.

After testing several models including an OLS regression, a lasso regression, a k-nearest neighbor classifier, and a Naïve Bayes classifier, I found that Katz can predict a prospective candidate's likelihood to pursue an MBA with moderate predictive power. According to the Naïve Bayes, characteristics like working in consulting, obtaining loan funding, valuing networking and entrepreneurship, and others can be effective predictors of pursuing an MBA. I also found that my classifier would not produce any significant gender disparities in targeted advertising and enrollment, based on the Naïve Bayes results and chi-squared tests.

Data

The file "mba_decision_dataset.csv" contains 20 columns of information for 10,000 observations of respondents in a Katz survey of students who may be interested in pursuing an MBA. The dataset includes numerical and categorical variables: age, gender, undergraduate major, undergraduate GPA, years of work experience, management experience, current job title, current salary and post-MBA salary expectation, standardized test scores, undergraduate university ranking, interest in networking and entrepreneurship, MBA funding source, desired post-MBA role, location preference, motivation for pursuing an MBA, and on-campus versus online format. The target variable is a column stating whether the individual pursued an MBA.

I began cleaning the data into a usable format by encoding categorical and boolean variables into binary numerical variables. Management experience became 'mgmt' where 1 indicates management experience and 0 indicates none; gender became 'female' where 1 indicates a woman and 0 indicates a man; location preference became 'international' where 1 indicates international and 0 indicates domestic; format became 'online' where 1 indicates online and 0 indicates on-campus; and the target variable became 'pursued' where 1 indicates that the individual pursued an MBA and 0 indicates that they did not. I used one-hot encoding to create new numerical binary dummy variables for each categorical column with several possible values. One example was current job title, with values 'analyst,' 'consultant,' and others.

Methodology

Using the cleaned dataset full of numerical columns and encoded binary dummies, I began with a simple OLS regression (Table 2) of all the independent variables on the response 'pursued.' I used a VIF test to check for collinearity – ensuring that there aren't any linear relationships among the explanatory variables. I found multiple collinearity issues (Table 1), so I eliminated the constant in the OLS regression to account for them. The regression had an R-squared of 0.002, indicating very weak predictive power. I moved on from this method.

The next method was a lasso regression. This technique is more sophisticated than the typical OLS because it uses regularization to shrink values and prevent overfitting. I scaled the features, split the data into training and test groups, and used cross-validation to determine the best alpha (denoting the amount of shrinkage). The regression (Table 3) produced a mean squared error of 0.24 and coefficients of mostly zeroes, which weren't very helpful for prediction.

I tried two machine learning techniques next – the k-nearest neighbor (KNN) and Naïve Bayes classifiers. The KNN classifier compares observations to other similar observations in the dataset. I separated the data into testing and training subsets and trained the KNN model on 70% of the data, testing it on the remainder and achieving an accuracy of about 53% (Table 4). The Naïve Bayes classifier is another technique that assumes conditional independence for each observation. I separated, trained, and tested the model, and got about 57% (Table 5).

I cross-validated the KNN and Naïve Bayes, testing different values of k-neighbors to evaluate the fit of the model. The best accuracy came from k=9, at 54%, which is still less than the Naïve Bayes. The mean cross-validated accuracy for Naïve Bayes was 57.7% (Table 6).

Naïve Bayes had emerged as the best predictor, but an important question remained – would this predictor create disparities between demographics like men and women because one was deemed more likely to pursue an MBA? An initial measurement of the proportion of men and women who pursued an MBA generated a miniscule difference of 0.008 (Table 10). Is it statistically significant? I utilized the Naïve Bayes (Table 7) and a chi-squared test (Table 9) to determine that there is no significant gender disparity in the predictive model.

Results

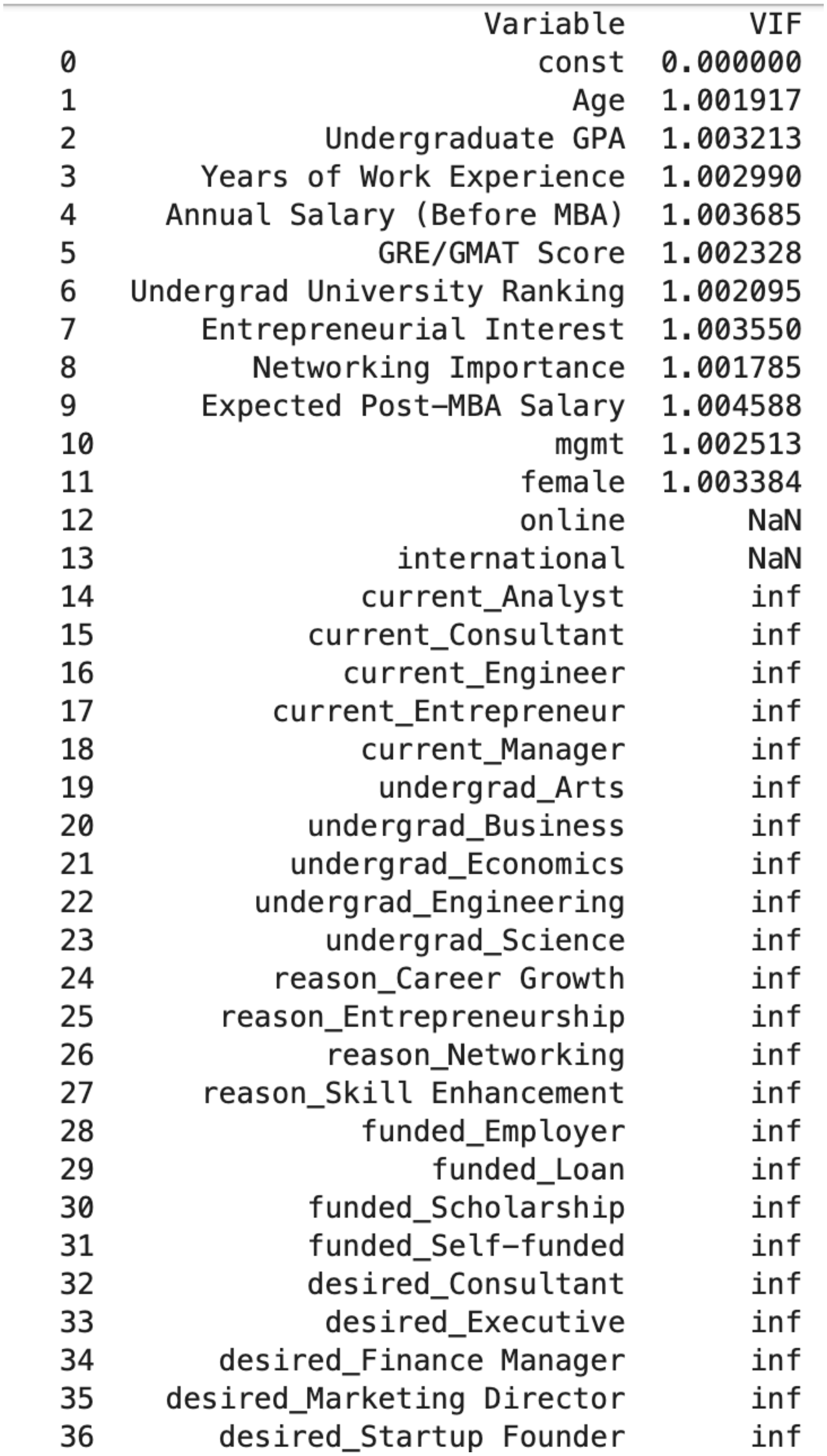
Using the results from the Naïve Bayes classifier, I ranked the predictors based on how influential they were in determining whether the individual in the observation pursued an MBA (Table 8). Predictors like undergraduate university ranking, age, and others came out on top, but the one-hot encoded variables provide more insight as to the profile of a successful candidate. It shows that Katz should target prospective students who can secure MBA funding from loans, value networking and entrepreneurship as career goals, majored in economics as undergrads, and currently work as consultants. These factors, among others, positively impacted the likelihood of pursuing an MBA, compared to the negative impact of factors like having management experience, majoring in business, valuing career growth, and working as analysts.

Gender was also an important question – if my model said that men or women are inherently more likely to pursue MBAs, there could be a disparity in the individuals Katz targets in its outreach and potentially a gender imbalance in enrollment. Though there was a slight difference in the proportion of women vs. men who decided to get MBAs (women marginally more than men), the chi-squared test did not provide sufficient evidence to reject the null hypothesis that there is no disparity. Taking the mean difference of predicted probabilities from the Naïve Bayes classifier I fitted, I found that the female variable had no significant influence on the target variable. Thus, we can conclude that in this dataset, there is no statistically significant difference between gender in enrollment yield, and thus minimal risk of creating a disparity.

If there was an imbalance, I could adjust the predictive model to handle gender distinctly from other independent variables. One option is stratified sampling – we could remove the effect of differences in gender proportions by weighting men and women according to their respective proportion when splitting data into training and testing sets. We could monitor the disparity in the cross-validation process to ensure the proportion difference in the raw data does not leak into the results of the predictive model.

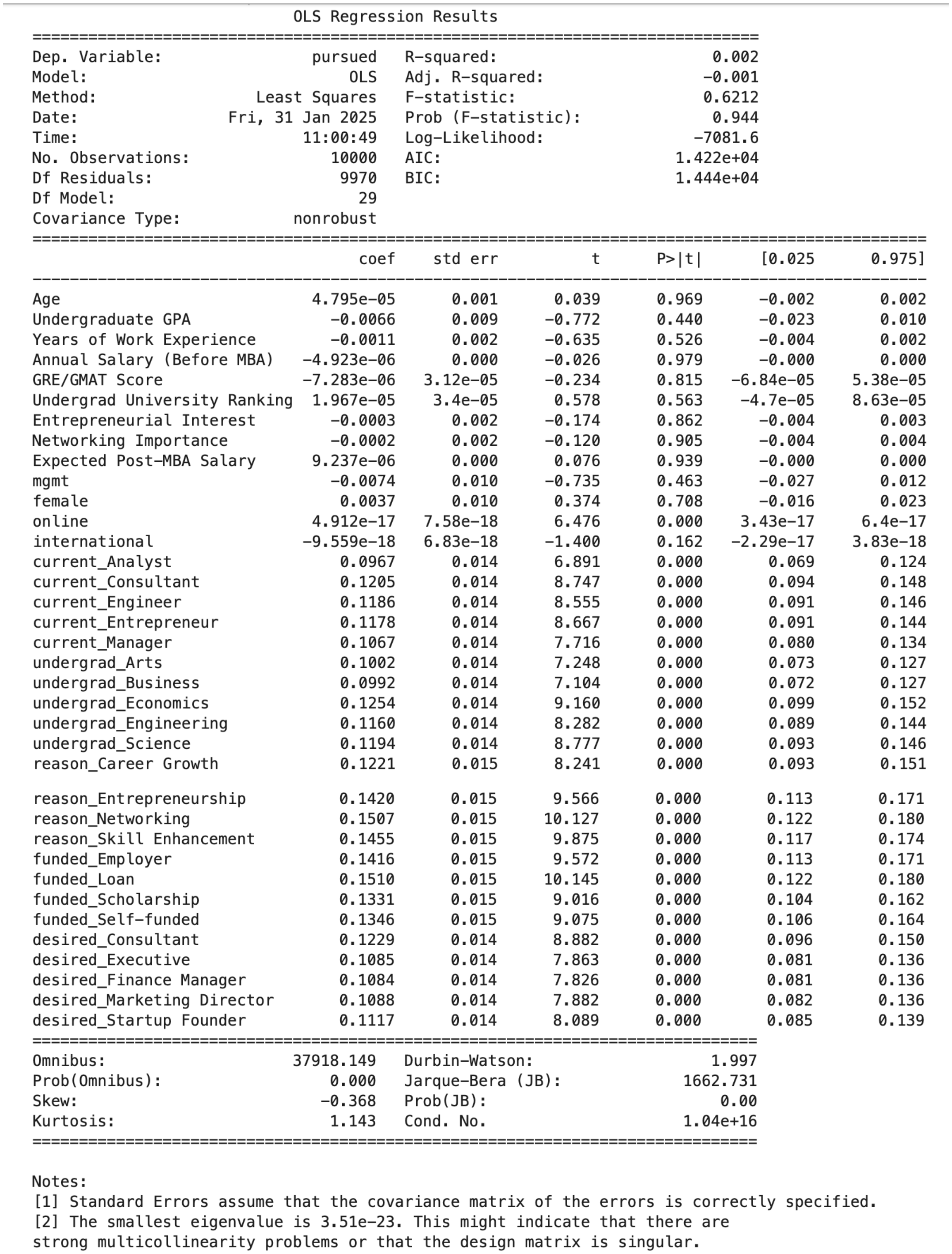
Appendix

Table 1: VIF test for regressor collinearity



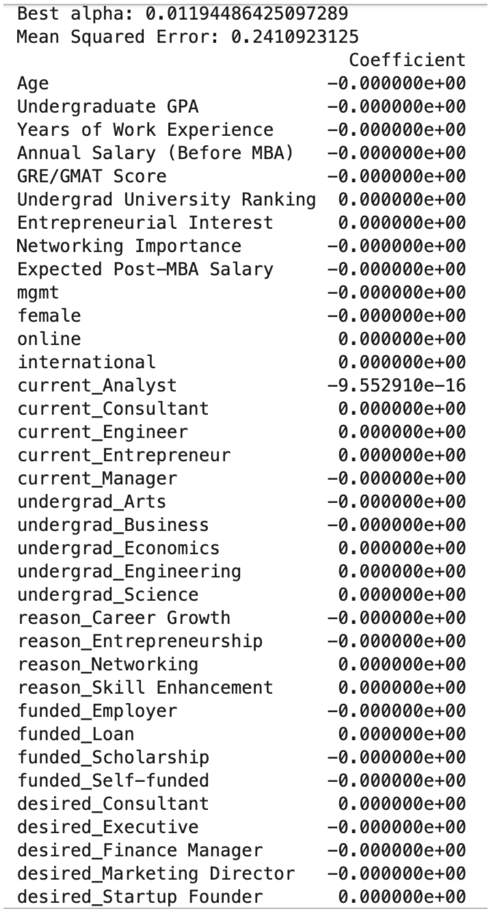
```
                         Variable       VIF
0                           const  0.000000
1                             Age  1.001917
2                 Undergraduate GPA  1.003213
3       Years of Work Experience  1.002990
4     Annual Salary (Before MBA)  1.003685
5                  GRE/GMAT Score  1.002328
6   Undergrad University Ranking  1.002095
7        Entrepreneurial Interest  1.003550
8           Networking Importance  1.001785
9        Expected Post-MBA Salary  1.004588
10                           mgmt  1.002513
11                         female  1.003384
12                         online       NaN
13                  international       NaN
14                current_Analyst       inf
15             current_Consultant       inf
16               current_Engineer       inf
17           current_Entrepreneur       inf
18                current_Manager       inf
19                 undergrad_Arts       inf
20             undergrad_Business       inf
21            undergrad_Economics       inf
22          undergrad_Engineering       inf
23              undergrad_Science       inf
24           reason_Career Growth       inf
25        reason_Entrepreneurship       inf
26              reason_Networking       inf
27      reason_Skill Enhancement       inf
28                funded_Employer       inf
29                    funded_Loan       inf
30             funded_Scholarship       inf
31             funded_Self-funded       inf
32             desired_Consultant       inf
33              desired_Executive       inf
34        desired_Finance Manager       inf
35     desired_Marketing Director       inf
36        desired_Startup Founder       inf
```

Table 2: OLS regression output



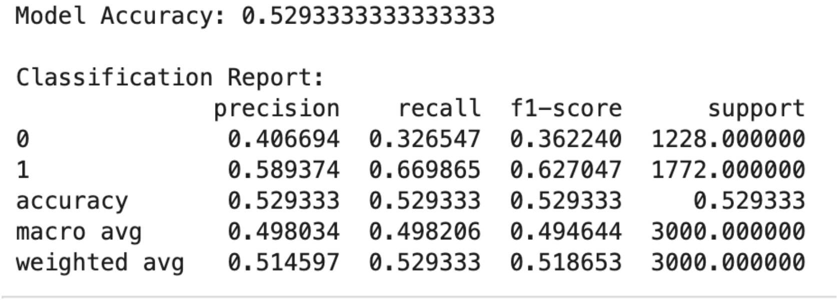
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                pursued   R-squared:                       0.002
Model:                            OLS   Adj. R-squared:                 -0.001
Method:                 Least Squares   F-statistic:                    0.6212
Date:                Fri, 31 Jan 2025   Prob (F-statistic):              0.944
Time:                        11:00:49   Log-Likelihood:                -7081.6
No. Observations:               10000   AIC:                         1.422e+04
Df Residuals:                    9970   BIC:                         1.444e+04
Df Model:                          29
Covariance Type:            nonrobust
================================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Age                         4.795e-05      0.001      0.039      0.969      -0.002       0.002
Undergraduate GPA             -0.0066      0.009     -0.772      0.440      -0.023       0.010
Years of Work Experience      -0.0011      0.002     -0.635      0.526      -0.004       0.002
Annual Salary (Before MBA)  -4.923e-06      0.000     -0.026      0.979      -0.000       0.000
GRE/GMAT Score              -7.283e-06   3.12e-05     -0.234      0.815   -6.84e-05    5.38e-05
Undergrad University Ranking  1.967e-05    3.4e-05      0.578      0.563     -4.7e-05    8.63e-05
Entrepreneurial Interest      -0.0003      0.002     -0.174      0.862      -0.004       0.003
Networking Importance         -0.0002      0.002     -0.120      0.905      -0.004       0.004
Expected Post-MBA Salary    9.237e-06      0.000      0.076      0.939      -0.000       0.000
mgmt                          -0.0074      0.010     -0.735      0.463      -0.027       0.012
female                         0.0037      0.010      0.374      0.708      -0.016       0.023
online                      4.912e-17   7.58e-18      6.476      0.000    3.43e-17     6.4e-17
international               -9.559e-18   6.83e-18     -1.400      0.162   -2.29e-17    3.83e-18
current_Analyst                0.0967      0.014      6.891      0.000       0.069       0.124
current_Consultant             0.1205      0.014      8.747      0.000       0.094       0.148
current_Engineer               0.1186      0.014      8.555      0.000       0.091       0.146
current_Entrepreneur           0.1178      0.014      8.667      0.000       0.091       0.144
current_Manager                0.1067      0.014      7.716      0.000       0.080       0.134
undergrad_Arts                 0.1002      0.014      7.248      0.000       0.073       0.127
undergrad_Business             0.0992      0.014      7.104      0.000       0.072       0.127
undergrad_Economics            0.1254      0.014      9.160      0.000       0.099       0.152
undergrad_Engineering          0.1160      0.014      8.282      0.000       0.089       0.144
undergrad_Science              0.1194      0.014      8.777      0.000       0.093       0.146
reason_Career Growth           0.1221      0.015      8.241      0.000       0.093       0.151

reason_Entrepreneurship        0.1420      0.015      9.566      0.000       0.113       0.171
reason_Networking              0.1507      0.015     10.127      0.000       0.122       0.180
reason_Skill Enhancement       0.1455      0.015      9.875      0.000       0.117       0.174
funded_Employer                0.1416      0.015      9.572      0.000       0.113       0.171
funded_Loan                    0.1510      0.015     10.145      0.000       0.122       0.180
funded_Scholarship             0.1331      0.015      9.016      0.000       0.104       0.162
funded_Self-funded             0.1346      0.015      9.075      0.000       0.106       0.164
desired_Consultant             0.1229      0.014      8.882      0.000       0.096       0.150
desired_Executive              0.1085      0.014      7.863      0.000       0.081       0.136
desired_Finance Manager        0.1084      0.014      7.826      0.000       0.081       0.136
desired_Marketing Director     0.1088      0.014      7.882      0.000       0.082       0.136
desired_Startup Founder        0.1117      0.014      8.089      0.000       0.085       0.139
==============================================================================
Omnibus:                    37918.149   Durbin-Watson:                   1.997
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1662.731
Skew:                          -0.368   Prob(JB):                         0.00
Kurtosis:                       1.143   Cond. No.                     1.04e+16
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 3.51e-23. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

Table 3: Lasso regression output



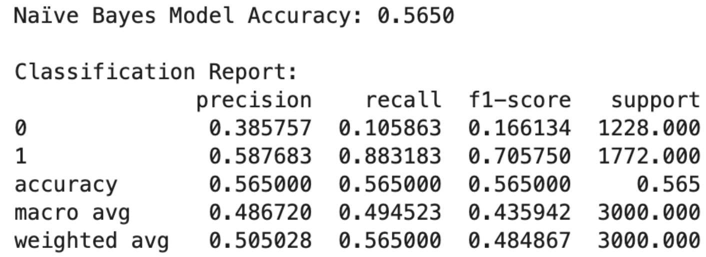
```
Best alpha: 0.01194486425097289
Mean Squared Error: 0.2410923125
                                Coefficient
Age                           -0.000000e+00
Undergraduate GPA             -0.000000e+00
Years of Work Experience      -0.000000e+00
Annual Salary (Before MBA)    -0.000000e+00
GRE/GMAT Score                -0.000000e+00
Undergrad University Ranking   0.000000e+00
Entrepreneurial Interest       0.000000e+00
Networking Importance         -0.000000e+00
Expected Post-MBA Salary      -0.000000e+00
mgmt                          -0.000000e+00
female                        -0.000000e+00
online                         0.000000e+00
international                  0.000000e+00
current_Analyst               -9.552910e-16
current_Consultant             0.000000e+00
current_Engineer               0.000000e+00
current_Entrepreneur           0.000000e+00
current_Manager               -0.000000e+00
undergrad_Arts                -0.000000e+00
undergrad_Business            -0.000000e+00
undergrad_Economics            0.000000e+00
undergrad_Engineering          0.000000e+00
undergrad_Science              0.000000e+00
reason_Career Growth          -0.000000e+00
reason_Entrepreneurship       -0.000000e+00
reason_Networking              0.000000e+00
reason_Skill Enhancement       0.000000e+00
funded_Employer               -0.000000e+00
funded_Loan                    0.000000e+00
funded_Scholarship            -0.000000e+00
funded_Self-funded            -0.000000e+00
desired_Consultant             0.000000e+00
desired_Executive             -0.000000e+00
desired_Finance Manager       -0.000000e+00
desired_Marketing Director    -0.000000e+00
desired_Startup Founder        0.000000e+00
```

Table 4: k-nearest neighbor accuracy and classification report



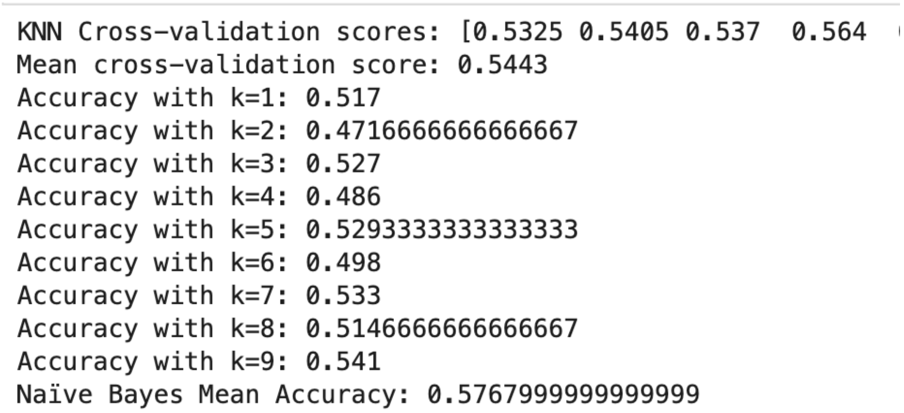
```
Model Accuracy: 0.5293333333333333

Classification Report:
              precision    recall  f1-score      support
0              0.406694  0.326547  0.362240  1228.000000
1              0.589374  0.669865  0.627047  1772.000000
accuracy       0.529333  0.529333  0.529333     0.529333
macro avg      0.498034  0.498206  0.494644  3000.000000
weighted avg   0.514597  0.529333  0.518653  3000.000000
```

Table 5: Naïve Bayes accuracy and classification report



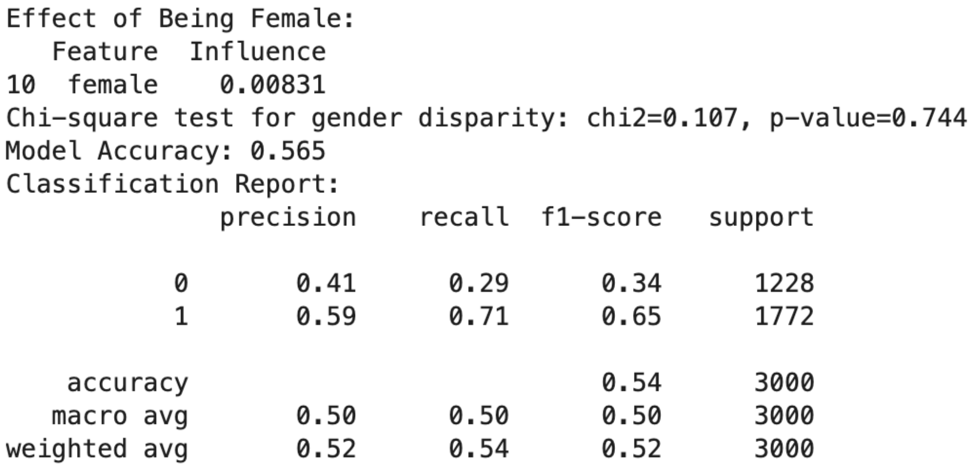
```
Naïve Bayes Model Accuracy: 0.5650

Classification Report:
              precision    recall  f1-score   support
0              0.385757  0.105863  0.166134  1228.000
1              0.587683  0.883183  0.705750  1772.000
accuracy       0.565000  0.565000  0.565000     0.565
macro avg      0.486720  0.494523  0.435942  3000.000
weighted avg   0.505028  0.565000  0.484867  3000.000
```

Table 6: Cross-validation of KNN and Naïve Bayes classifiers



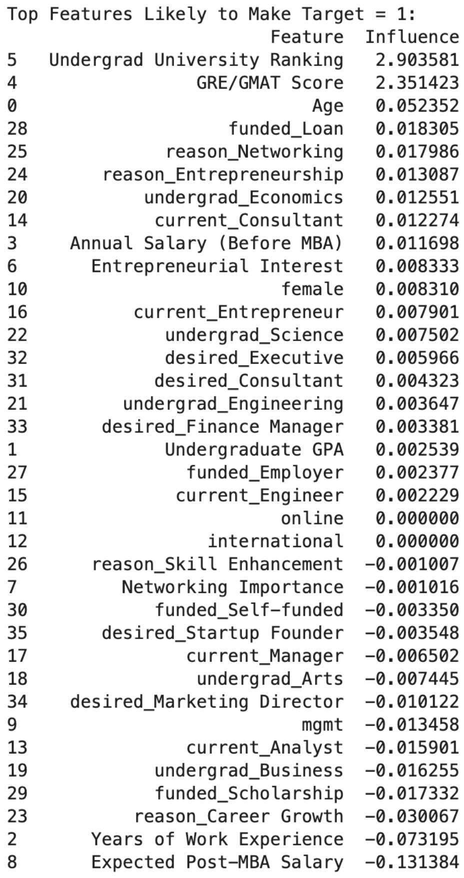
```
KNN Cross-validation scores: [0.5325 0.5405 0.537  0.564
Mean cross-validation score: 0.5443
Accuracy with k=1: 0.517
Accuracy with k=2: 0.4716666666666667
Accuracy with k=3: 0.527
Accuracy with k=4: 0.486
Accuracy with k=5: 0.5293333333333333
Accuracy with k=6: 0.498
Accuracy with k=7: 0.533
Accuracy with k=8: 0.5146666666666667
Accuracy with k=9: 0.541
Naïve Bayes Mean Accuracy: 0.5767999999999999
```

Table 7: Chi-squared test for gender disparity



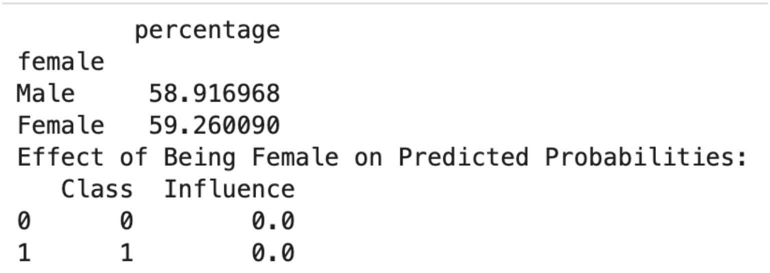
```
Effect of Being Female:
   Feature  Influence
10  female    0.00831
Chi-square test for gender disparity: chi2=0.107, p-value=0.744
Model Accuracy: 0.565
Classification Report:
              precision    recall  f1-score   support

           0       0.41      0.29      0.34      1228
           1       0.59      0.71      0.65      1772

    accuracy                           0.54      3000
   macro avg       0.50      0.50      0.50      3000
weighted avg       0.52      0.54      0.52      3000
```

Table 8: Ranking of feature influence on target variable



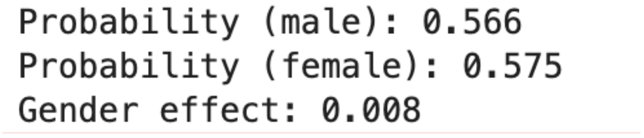
```
Top Features Likely to Make Target = 1:
                         Feature  Influence
5   Undergrad University Ranking   2.903581
4                 GRE/GMAT Score   2.351423
0                            Age   0.052352
28                   funded_Loan   0.018305
25             reason_Networking   0.017986
24       reason_Entrepreneurship   0.013087
20           undergrad_Economics   0.012551
14            current_Consultant   0.012274
3     Annual Salary (Before MBA)   0.011698
6       Entrepreneurial Interest   0.008333
10                        female   0.008310
16          current_Entrepreneur   0.007901
22             undergrad_Science   0.007502
32             desired_Executive   0.005966
31            desired_Consultant   0.004323
21         undergrad_Engineering   0.003647
33       desired_Finance Manager   0.003381
1              Undergraduate GPA   0.002539
27               funded_Employer   0.002377
15              current_Engineer   0.002229
11                        online   0.000000
12                 international   0.000000
26      reason_Skill Enhancement  -0.001007
7          Networking Importance  -0.001016
30            funded_Self-funded  -0.003350
35       desired_Startup Founder  -0.003548
17               current_Manager  -0.006502
18                undergrad_Arts  -0.007445
34    desired_Marketing Director  -0.010122
9                           mgmt  -0.013458
13               current_Analyst  -0.015901
19            undergrad_Business  -0.016255
29            funded_Scholarship  -0.017332
23          reason_Career Growth  -0.030067
2       Years of Work Experience  -0.073195
8       Expected Post-MBA Salary  -0.131384
```

Table 9: Mean difference of predicted probabilities using Naïve Bayes classifier



```
        percentage
female
Male     58.916968
Female   59.260090
Effect of Being Female on Predicted Probabilities:
   Class  Influence
0      0        0.0
1      1        0.0
```

Table 10: Proportion of candidates who pursued, grouped by gender



```
Probability (male): 0.566
Probability (female): 0.575
Gender effect: 0.008
```