MBA Decision Data Analytics Capstone Project

Ethan Tornga

Northwood University

MIS 4300: Data Analytics Capstone

Professor Marc Beaty

May 5th, 2025

Research Topic

For this capstone project I decided to research the domain of education, as I have familiarized myself with education throughout high school and college and I was interested to research any insights that help understand the decision making of students. After researching different education topics and having discussions about my peers, I decided that a relevant and interesting analysis project could be derived from the question of what factors have the most influence over a student's decision to pursue higher education, which in this case is a MBA degree. With the MBA being "the most popular graduate degree in the United States since 2011", according to Harvard Business School, I decided that this would be a great opportunity to look into. Additionally, between online research and in-person interviews, I had an idea of what factors may cause a student to pursue the degree or not, so I wanted to do actual data analysis to see if those ideas were actually true, such as the potential job prospects and pay increases.

Dataset

After much research I elected to analyze a dataset called MBA Decision that I found from the website Kaggle. This dataset includes ten thousand different students' responses to different key factors that play a role in their decision, as well as their decision itself, about whether to pursue their MBA degree. The data includes nineteen different columns of numerical and categorical factors. Some of those factors that are included in the data are age, gender, undergraduate GPA, work experience, location preference, and reason for MBA. One of the reasons that I chose this dataset because of the large sample size of ten thousand students allowed me to have a lot of data to analyze and draw conclusions from, which I hoped would lead me to the best results to answer my overarching question of what factors play the most

important roles in a student's decision in pursuing their MBA. I also chose this dataset because it directly aligned with my research question and it had many different fields to test.

Data Management

I broke down my data management into two parts: that which I did in SQL and that done in Python. In SQL I dropped columns that were not of use to my analysis and added two calculated fields that I hypothesized would give me interesting insights. I also created five different KPIs that I used to not only better understand my data, but also to use for visualizations to help others view the data. In Python, the first thing I did was check for any missing values, which there were not any. Then I ran code to detect any outliers using z-scores, which there also were not any that were more than three standard deviations away from the mean. Finally, I encoded all non-numerical variables with dummy variables so that they could be more easily used in some of my later Python models.

SQL

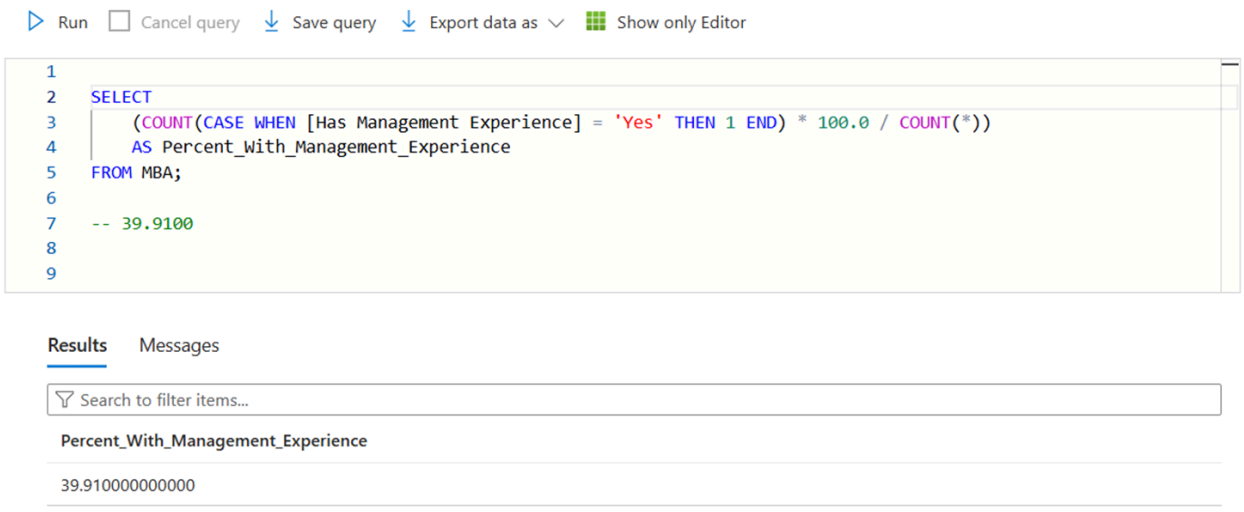
As I mentioned above, I handled some of my data cleaning through the use of SQL. I utilized Microsoft Azure's SQL Database to do my cleaning, which also led to me using the service to create a pipeline from my data (housed in a .csv file) to the SQL database. Once that was successfully set up, I was able to perform my initial cleaning. First, I deleted two columns that I felt had too much personal variance, because there was not a set scale on how to judge them and depending on the person, they may view them on the scale differently even if they believed the same idea. These two columns were Entrepreneurial Interest and Networking Importance, and the code used to remove them is shown below:

```sql
SELECT COLUMN_NAME, DATA_TYPE, IS_NULLABLE
FROM INFORMATION_SCHEMA.COLUMNS
WHERE TABLE_NAME = 'MBA';

ALTER TABLE MBA
DROP COLUMN [Entrepreneurial Interest], [Networking Importance];

```

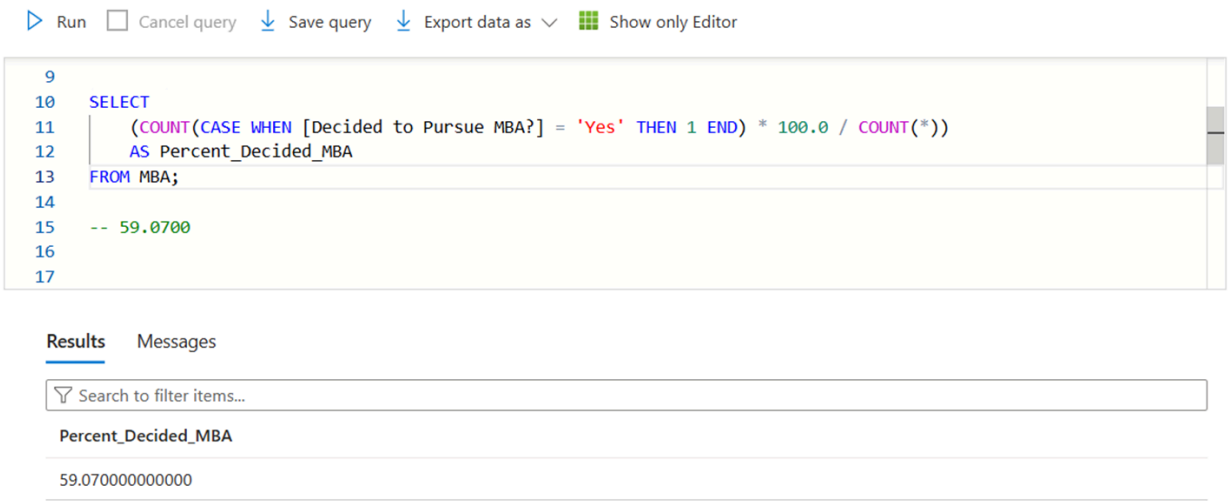
After those were dropped, I then added two columns to the dataset. One of the calculated columns was the expected salary difference that a student expected after an MBA, which was calculated by subtracting their current annual salary by their expected post-MBA salary. The other was a percentage of life worked columns that divided the student's age by the number of years that they have worked. These are shown below:

```sql
ALTER TABLE MBA
ADD [Salary Difference] AS ([Expected Post-MBA Salary] - [Annual Salary (Before MBA)]);


ALTER TABLE MBA
ADD [Percent of Life Worked] AS (CAST([Years of Work Experience] AS FLOAT) / CAST(Age AS FLOAT));
```

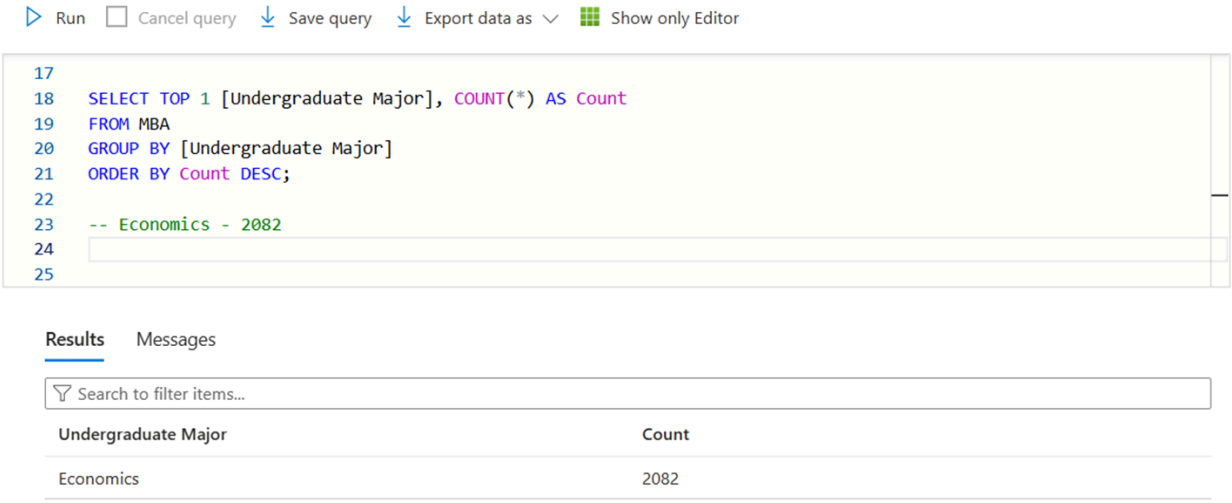
My final SQL initiative was to create five different KPIs that could help me have a more in depth look at my data. The first of these was to determine how many students had management experience, which was calculated to be 39.91%.



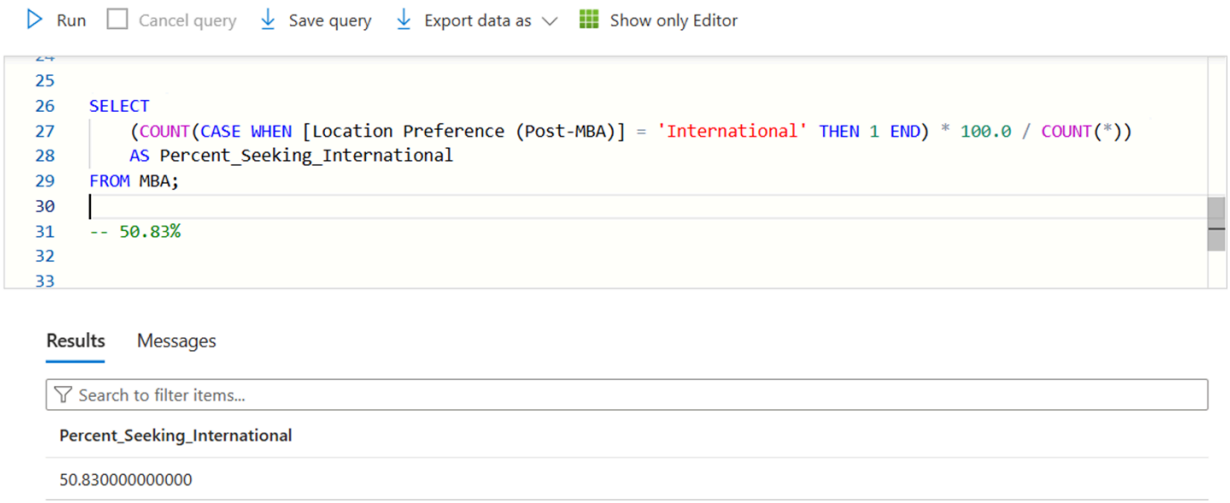
The second KPI was to see what percentage of students surveyed actually decided to pursue their MBA degree, which ended up being 59.07% of the data, or roughly 5900 students.



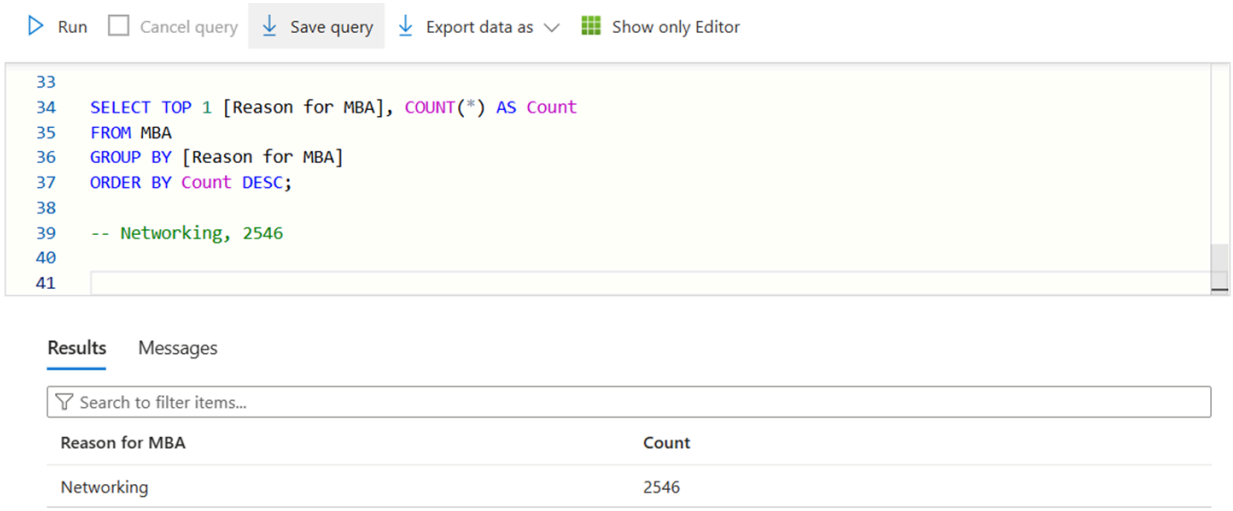
Then I looked to see what the most popular undergraduate major was out of the students surveyed, which was Economics.



The fourth KPI that I created was to see what percentage of students were looking to pursue their MBA internationally, which is a growing trend among students. This showed that 50.83% of students are seeking that avenue.

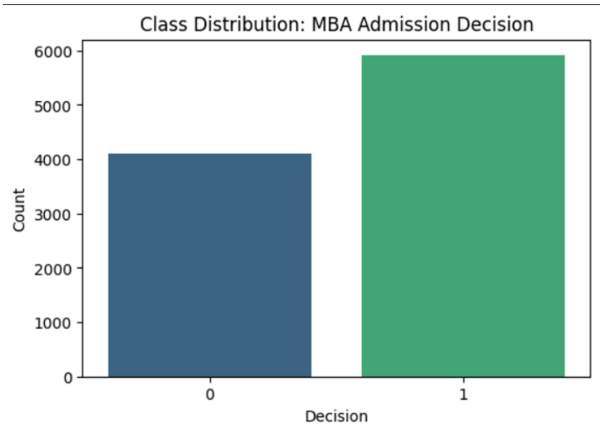


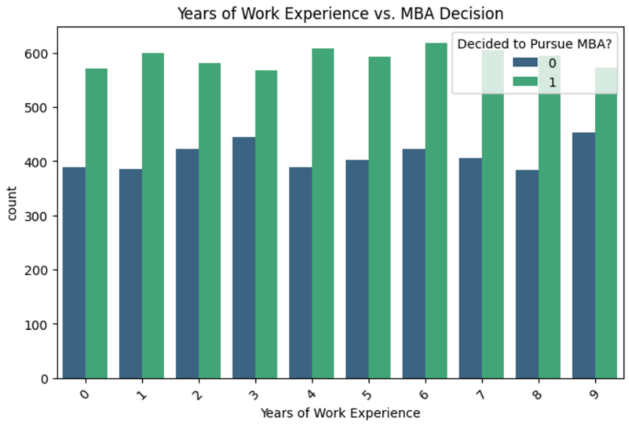
The final KPI I created was to see what was the most popular reason for wanting to pursue an MBA, and Networking was found to be the number one reason.



Python

For the Python programming aspect of this project, I connected my GitHub repository where all of my data was stored and made a connection so that I could work with the data in Google Colab. Then, as mentioned above, I finished cleaning and preprocessing the data and all of the variables. After that was complete, I created some graphs such as a Seaborn Pair Plot and some bar graphs to help initially view the data and develop a better understanding of how it is broken down by category.





Next, I created a 70/30 train test split to create models with. Then, I computed the MSE and RMSE, which came out to 0.2419 and 0.4919, respectively.

```
#Mean Squared Error for evaluation of model
print("MSE: ", mse)
print("RMSE: ", np.sqrt(mse))
```

```
MSE:  0.24193221142858715
RMSE:  0.49186605029071395
```

Afterwards, I created a multiple regression model evaluation, which showed that the top three performing models for this data were Lasso, Ridge, and Linear Regression models.

```
#Displaying RMSE
df = pd.DataFrame.from_dict(errors, orient='index').sort_values(0)
df.columns=['RMSE']
display(df)
```

	RMSE
Lasso	0.492336
Ridge	0.493100
LinearRegression	0.493101
DecisionTreeRegressor1	0.493243
DecisionTreeRegressor3	0.495236
NuSVR	0.499193
DecisionTreeRegressor5	0.500358
KNeighborsRegressor10	0.514723
DecisionTreeRegressor10	0.534992
SVR	0.547929
KNeighborsRegressor3	0.567618

I also created a classification model and KNN model, neither of which had accuracies above 60%.

	precision	recall	f1-score	support
0	0.41	0.16	0.23	1205
1	0.60	0.84	0.70	1795
accuracy			0.57	3000
macro avg	0.50	0.50	0.46	3000
weighted avg	0.52	0.57	0.51	3000

Accuracy: 0.569

```
3NN Accuracy: 0.5096666666666667
kNN: 0.51
Prediction time for k-NN: 0.32 seconds
```

Finally, I created a feature importance table in which undergraduate class ranking, expected salary, and GRE/GMAT score were found to be the most important in determining whether a student would pursue an MBA.

```
[44] # Create a DataFrame to display feature importances
     feature_importance_df = pd.DataFrame({
         'Feature': filtered_columns,
         'Importance': filtered_importances
     })

     # Sort the DataFrame by importance in descending order
     feature_importance_df = feature_importance_df.sort_values(by='Importance', ascending=False)

     # Display the feature importances
     feature_importance_df
```

	Feature	Importance
9	Undergrad University Ranking	0.098983
12	Expected Post-MBA Salary	0.098890
8	GRE/GMAT Score	0.098879
6	Annual Salary (Before MBA)	0.098585
16	Salary Difference	0.097551

Other methods were performed such as PCA Analysis, hierarchical clustering, and R-squared analysis, among others. These were not included because either their results were not statistically significant or they had perfect multicollinearity in the results.

Data Visualization (Power BI)

The first page of visualizations that I made in my dashboard is a summary of some key statistics that can give a basic understanding of some of the key insights of my data, such as the breakdown of the decision to pursue an MBA based on age and location preference. Then I created different pages dedicated to some of the most important features from my analysis. The first two are age and gender, respectively, because they are very common features and easy for anyone to understand. The next three are the features deemed most important from my Python analysis: undergraduate class ranking, expected salary, and GRE/GMAT score. These can be seen in the following screenshots, respectively:

Age
21 34

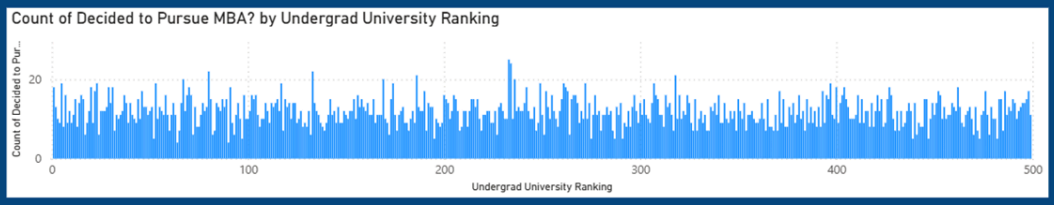
Undergraduate Class Ranking

Gender
- [ ] Female
- [ ] Male
- [ ] Other

Current Job Title	Average of Undergrad University Ranking
Entrepreneur	244.30
Manager	244.94
Engineer	245.89
Consultant	249.31
Analyst	250.76
Total	**247.04**

Undergrad University Ranking	Average of Age
1	28.62
2	27.11
3	27.38
4	26.43
5	27.35
6	28.77
7	27.77
8	27.35
9	28.05
10	26.70
11	28.00
Total	**27.49**

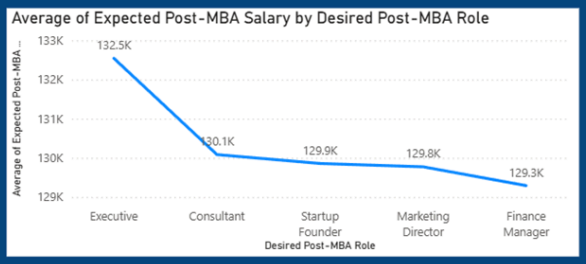
MBA Funding Source	Average of Undergrad University Ranking
Employer	243.99
Loan	244.17
Self-funded	249.99
Scholarship	250.15
Total	**247.04**

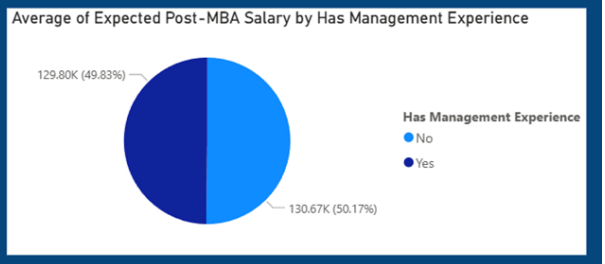


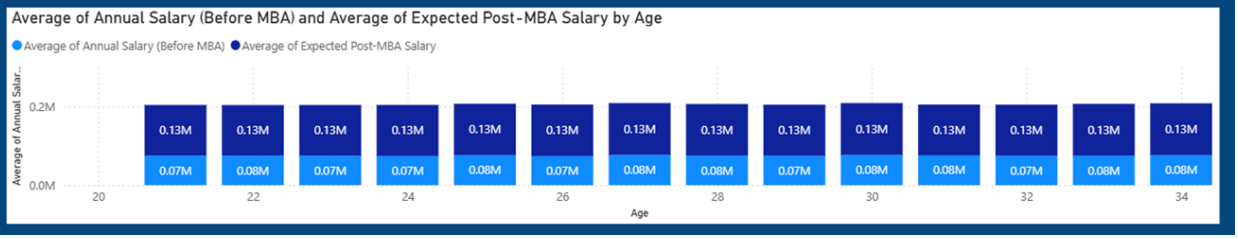
Age
21 34

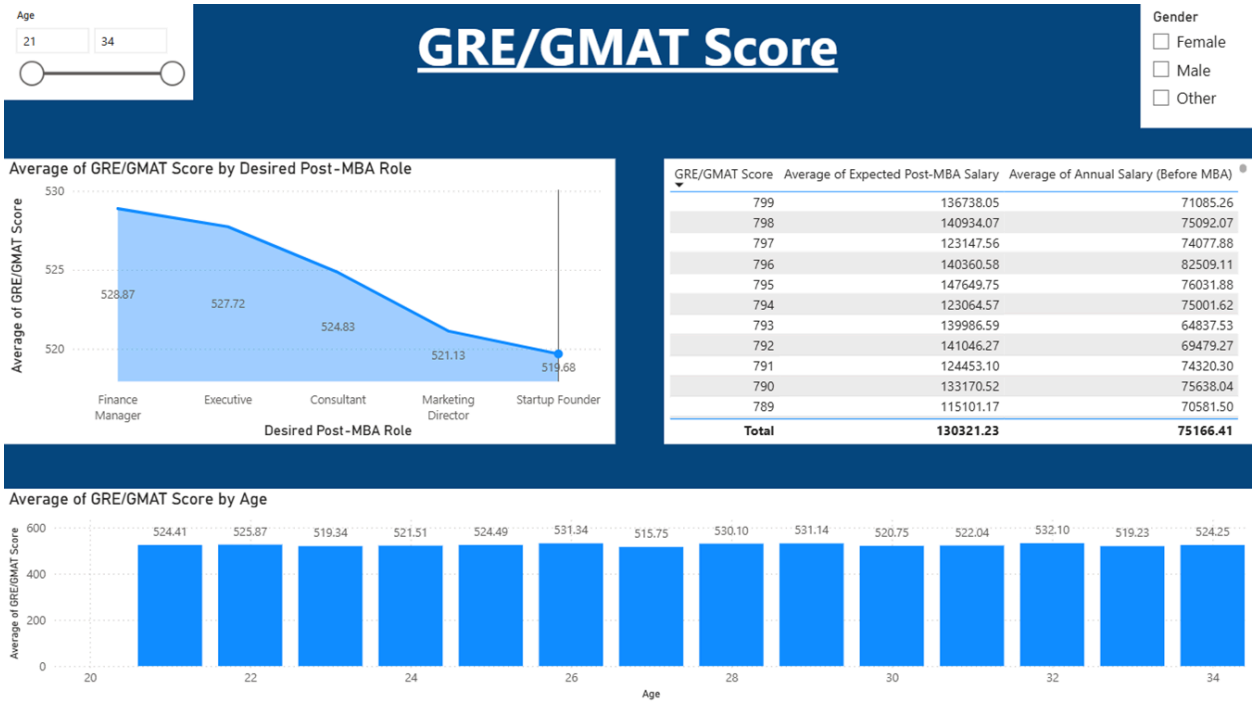
Expected Salary

Gender
- [ ] Female
- [ ] Male
- [ ] Other









Interpretation of Results

As was mentioned earlier, none of the models performed particularly well based on accuracy scores. While the MSE and RMSE were close to zero, the low accuracy of the other models created were not ideal results. The best performing model created was the Random Forest classifier model that had an accuracy of 56.9%. When the multiple models were created, the top performing models based on RMSE, which were all less than 0.5, were the Lasso, Ridge, and Linear Regression models. The biggest takeaways I was able to have was from the feature importance, which helped tell more of the story. While none of the features stood out a lot, there were many different features that were considered, so by nature their values would be a lesser piece of the puzzle. The undergraduate class ranking, expected salary after MBA, and GRE/GMAT scores were found to be the most important features, which I was able to analyze further to gain more insights in PowerBI. I also created multiple confusion matrices in which the

actual performed better than the predicted in all cases. In my data visualization step, I was able to analyze additional variables by one another, such as age and gender and see the breakdown of the data by them. For full transparency, there were not any standout variables throughout this analysis, as there was a fairly even importance and statistical split for all variables within this dataset, which can be seen in the dashboard.

Actionable Recommendations

Based on my analysis, I recommend that schools looking to increase the number of students that enroll in their MBA program to try to focus on some of the important features from this analysis that are within their scope and that they can control. The most important feature I found was the undergraduate class ranking, so finding ways to encourage top students to pursue an MBA is a good path to produce more students. Additionally, providing better learning opportunities and tutoring could help raise a student's class ranking and change their perspective about continuing school. Another important feature that should be noted is expected salary. Colleges should take advantage of this by advertising and explaining the salary differences that happen when receiving an MBA and how the job landscape changes with the added qualifications. Students, and people in general, respond to monetary changes so this could be something that can really drive students' decisions. Finally, the GRE/GMAT score metric is important because it is a judge on whether a student is ready for the next level of schooling. By providing resources such as study guides and learning sessions, schools can show a commitment to a student's learning and encourage them to pursue further degrees with them.

References

Cote, C. (2022, November 8). 7 Benefits of Getting an MBA: HBS Online. Business Insights Blog. https://online.hbs.edu/blog/post/benefits-of-getting-an-mba

Fenner, M. E. (2022). Machine Learning with Python for Everyone. Pearson Addison-Wesley Professional.