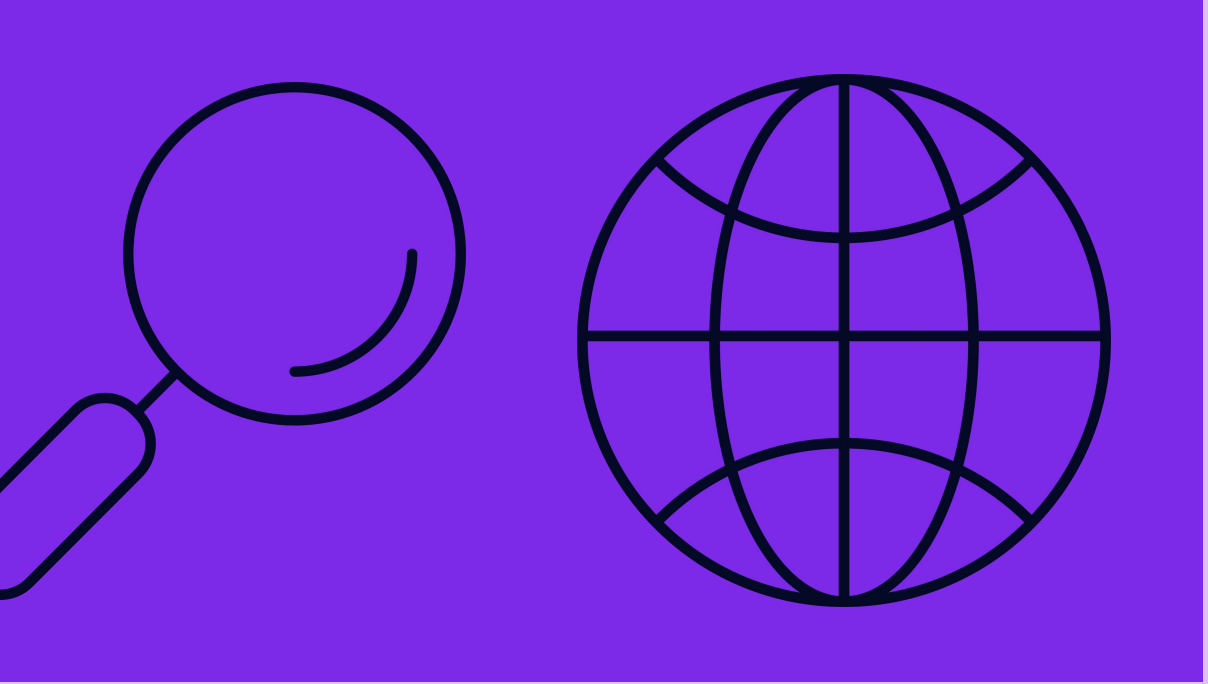
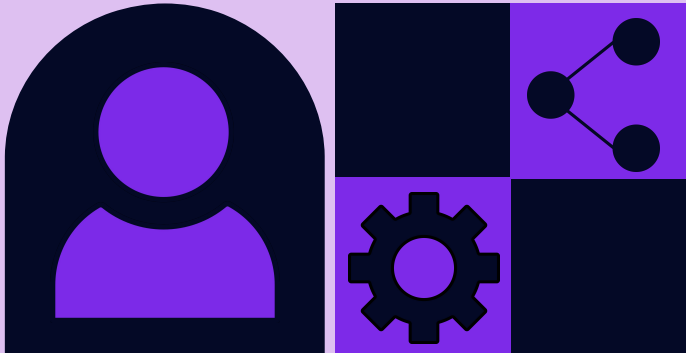


Adela Cho

Data Science Program Recommendation System



Problem Statement

Many prospective students to Master's degrees in Data Science have difficulty navigating the large landscape of Data Science programs to find the optimal program that fits their preference in location, cost, and duration. In order to provide students with insights as to what scope of options in Data Science programs exist, we want to provide prospective students with a comprehensive recommendation system to see all availabilities and popular types of existing programs.

Assumptions/Hypothesis

- The demographic of Master's programs are often times older, the age range in the US being between 22-28 years old. During this time, many students are transitioning into monumental life changes, whether it be a career change, getting a married and/or starting a family, or having to financially support themselves independently. These changes may lead to students desiring stability in location, having a tighter timeline to accommodate their circumstances or having a strict budget.
- This recommendation system will help prospective Data Science Master's students to choose the best Master's program for their location, budget, duration, etc preferences.

Exploratory Data Analysis

Figure 1) General Information about DS Program Dataset

```
Basic Information about the Dataset:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 443 entries, 0 to 442
Data columns (total 7 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Subject Name         443 non-null    object
 1   University Name      443 non-null    object
 2   Per Year Fees        428 non-null    float64
 3   About Program        443 non-null    object
 4   Program Duration     443 non-null    object
 5   University Location  443 non-null    object
 6   Program Name         443 non-null    object
dtypes: float64(1), object(6)
memory usage: 24.4+ KB
None
```

Figure 2) Descriptive Statistics

	Subject Name	University Name	Per Year Fees	About Program	Program Duration	University Location	Program Name
count	443	443	436	443	427	443	443
unique	199	276	339	434	22	163	40
top	Data Science	Arizona State University	26,655 EUR / year	Data Analysis and Research Psychology (Online)...	1 year	Online	M.Sc. / Full-time / On Campus
freq	135	13	10	2	258	124	160

Exploratory Data Analysis

Figure 3) Distribution of DS Program Tuition Fees Per Year

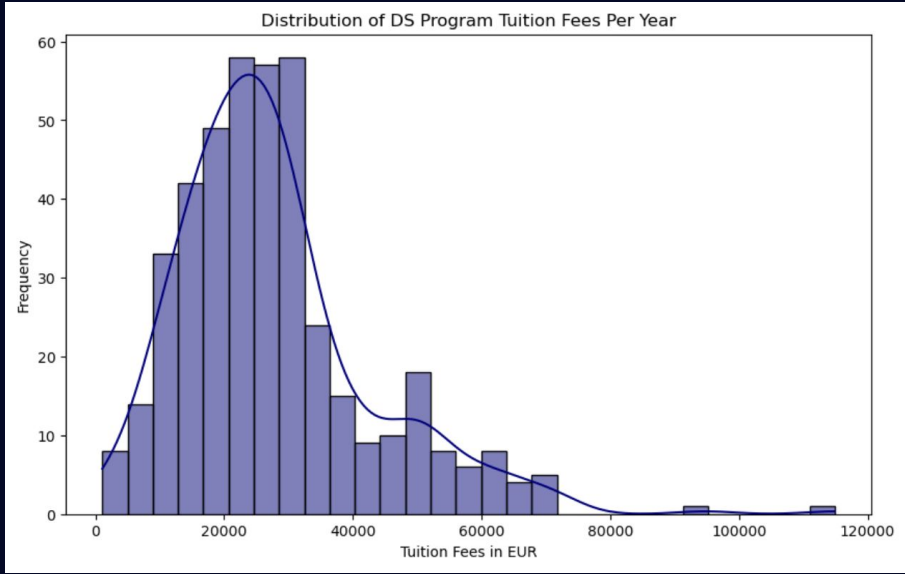
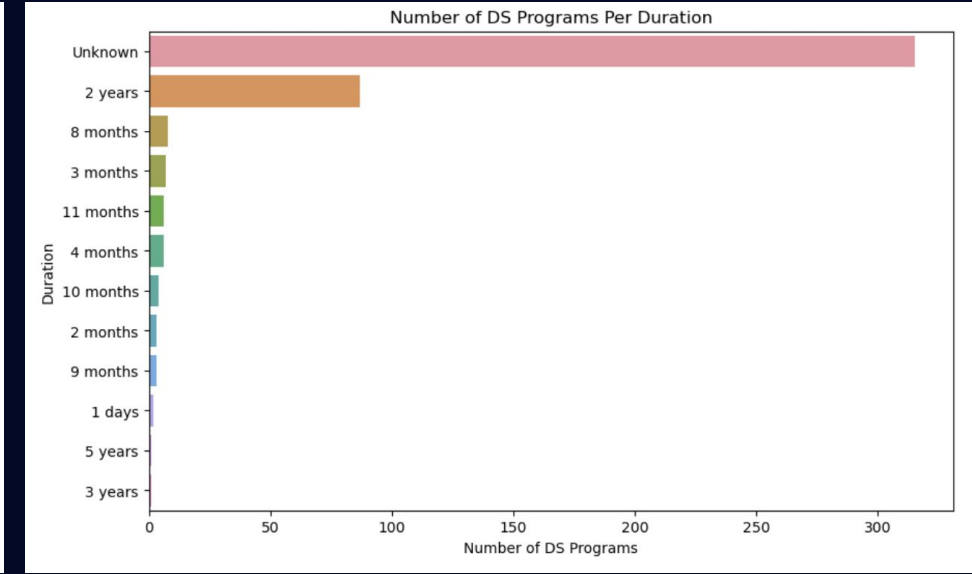
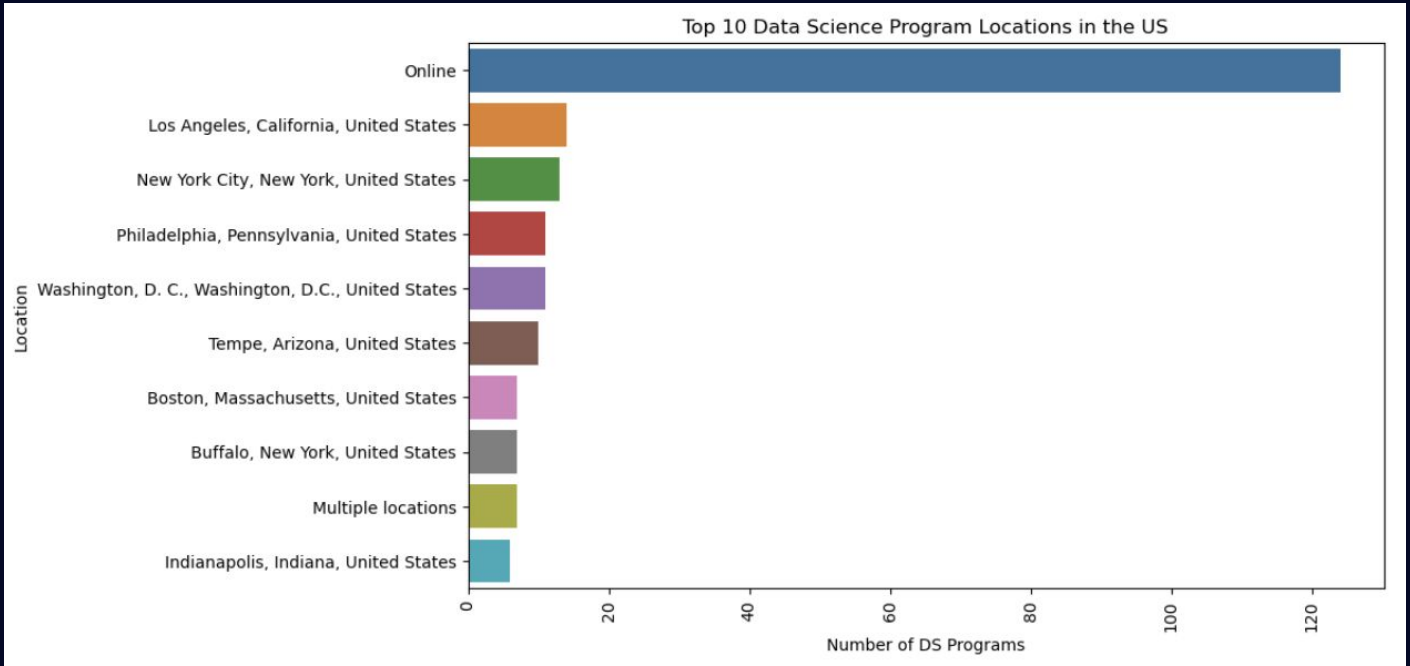


Figure 4) Number of DS Programs Per Duration



Exploratory Data Analysis

Figure 5) Top 10 Data Science Program Locations in the US



Feature Engineering & Transformations

- Check Missing Values
  - “Program Duration” missing values filled with “1 year” (avg duration)
  - Filled remaining missing values like “Per Year Fees”, “University Location”, “Program Duration”, “Program Names
- Standardized Numerical Featured
  - “Per Year Fees” → scale so is mean: 0 and is SD: 1 to normalize data to improve performance
- Convert Categorical values to Numerical values
  - “University Location”, “Program Location”, and “Program Duration” converted to numerical

```
Per Year Fees          0
University Location    0
Program Duration       0
Program Name           0
dtype: int64
   Per Year Fees  University Location  Program Duration  Program Name
0   1.120599e-16                115.0              11.0          39.0
1  -1.744374e+00                115.0               1.0          39.0
2  -4.677339e-01                 49.0              11.0          33.0
3  -6.047212e-01                115.0              11.0          37.0
4  -3.951609e-01                149.0               4.0           9.0
```

Proposed Approaches (Model) w/ check for overfitting/underfitting

- Random Forest Regressor → checks for overfitting through process of averaging out decision trees
- Process
  - Prepared X and y variable and split the data into training/testing set
  - Trained RandomForestRegressor model, initializing with 100 trees.
  - Cross-validation checked for overfitting/underfitting
    - Cross-validation scores (-0.080, -0.127, -0.204, 0.038, -0.104)
      - Negative scores indicated bad generalization and low prediction
      - ~0.038 indicates good fit
    - Average cross-validation score (-0.095) indicates not the best performance
    - RMSE of 50.773 indicates that the model's prediction was off by 50.773

Proposed Solution (Model Selection) w/ regularization

- Ridge Regression → potential to prevent overfitting
  - Cross-Validation Scores: -0.053, -0.026, -0.141, -0.193
    - Scores are less negative in comparison to RF Regressor; ridge regression performs better
  - Average Ridge Cross-Validation Score: -0.0899
    - Does not perform that well but does better than RF Regressor.

* additional model solution & learnings continue in slides 12 & 13

Results (Accuracy) & Learnings from the methodology

- As seen by large RMSE and negative cross-validation scores, RandomForestRegressor and Ridge Regression both don't generalize or perform very well; however, Ridge Regression performs better than Random Forest in comparison
- Ridge Regression works better with simpler datasets so may not have been the best model to implement for a more complex dataset like this as seen by its negative cross-validation scores
- Additional Feature Engineering
    - May need additional features like "size of institution", "number of enrolled students", etc and more, specific descriptors rather than long-winded descriptions of each institution might lead to higher ability to find patterns in the data and thus better performance of the model.

Future Work

- Expand dataset to include more variables to give deeper insight to the popularity of certain institutions' traits and prospective student preferences → lead to better model performance on grasping data patterns
    - ex) class size, Data Science program ranking, qualitative traits (culture), acceptance rate
- Hyperparameter tuning on RFRegressor and Ridge Regression for best parameter for improvement in model performance in overfitting and/or generalization.
- Use of different, more advanced models to capture complex dataset
    - Although RFRegressor and Ridge Regression in general do allow for a certain amount of control in overfitting and flexibility in regularization, it seem to be unable to generalize the dataset very well for this dataset
    - Data might benefit from models like the Gradient Boosting Model (Grid or Randomized Search)

*Future Work (Additional Model)

Further modeling through use of Gradient Boosting Model (Randomized Search) in comparison to Ridge Regression model

- Best Parameters
  - Learning rate: 0.0113, max_depth: 3, min_sample_leaf: 2, min_samples_split: 3, n_estimators: 108, subsample: 0.716
- Cross validation scores: 0.0176, -0.0322, 0.06558, -0.01007, -0.0098
  - The mix of both positive and negative scores show that the Randomized Search model's performance has variability and doesn't show signs of significant overfitting as the scores are significantly smaller.
  - Randomized Search has better generalization in comparison to Ridge Regression negative performance of -0.0900.

Future Work (Additional Model)

- Average cross-validation score: 0.0062
  - Positive score indicates that the model performs better than simple mean predictor
- RSME: 47.6090
  - Predictions do deviate for the Random Search model, The Ridge Regression model did slightly better with a RSME of 46.53

In Conclusion

- Randomized Search is efficient as it reduces amount of computation by sampling a smaller set of combinations and allows for flexibility when it comes to observing different hyperparameters. Although its pitfalls come in the possibility of variability and the necessity of sufficient data, out of all the models, Randomized Search model is the most optimal model to use for a Data Science Program recommender system for its better generalization and lack of overfitting.