Speed Dating

Eduardo Gonzalez

Business Problem

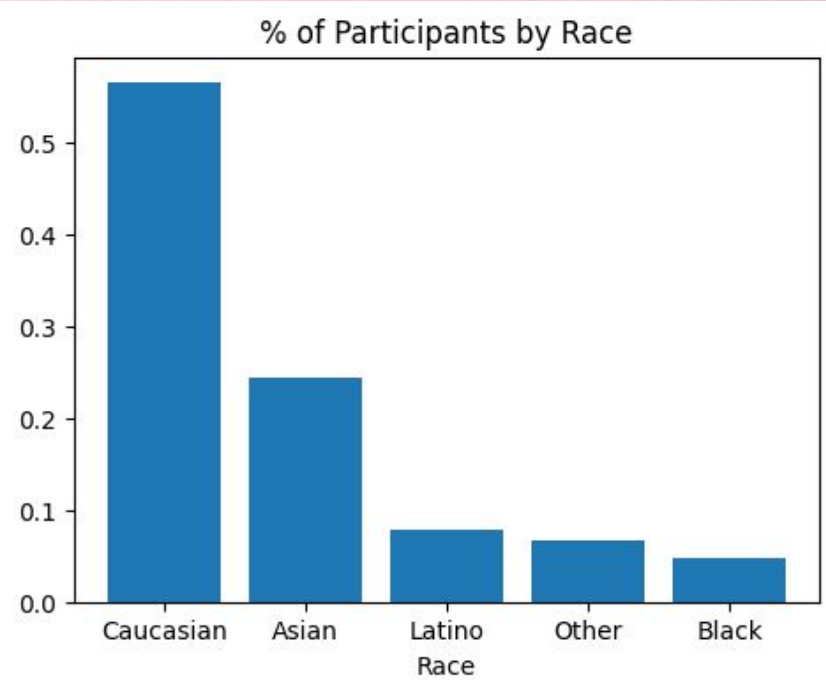
Dating Advice / Freemium app

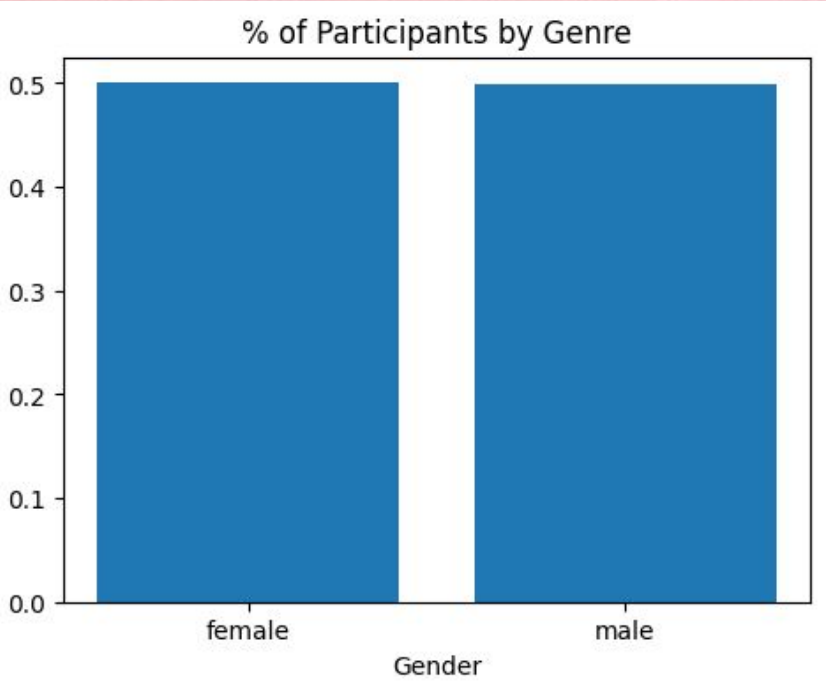
- What's the best model to predict whether if a couple will match or not?
- What hyperparameters should be optimized to improve accuracy?
- What are the factors that determine if a couple will match or not?

The Data

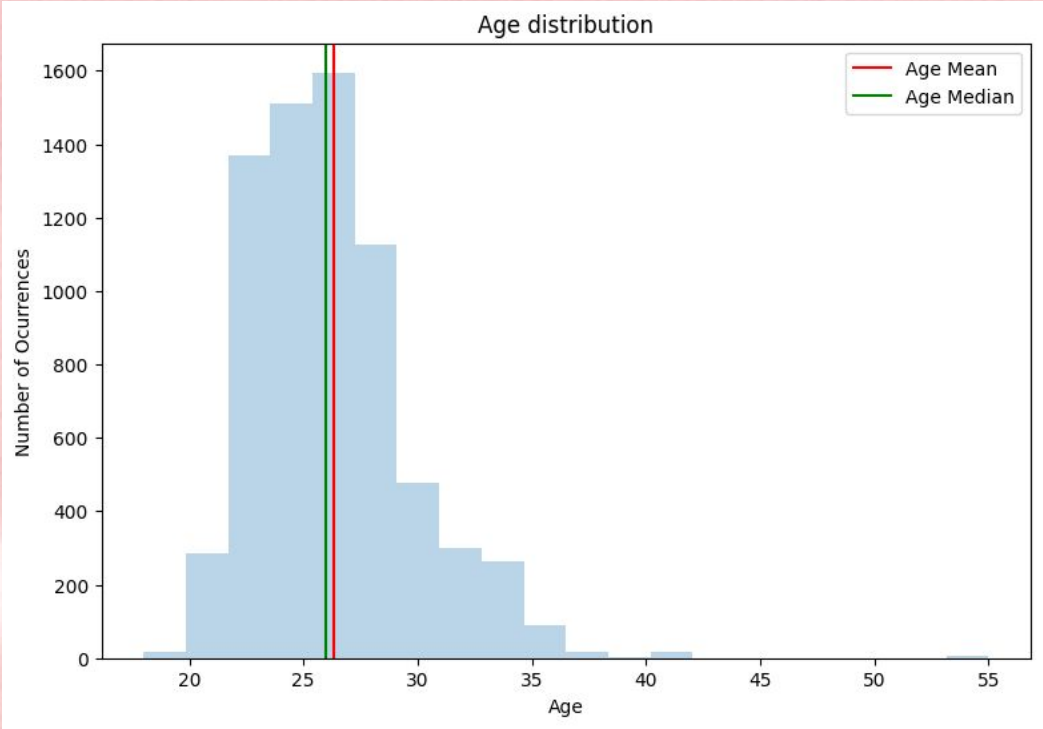
- Speed Dating From Columbia University
- 8378 Rows / 123 columns
- Data about Speed Dating encounters and individual stats and preferences
- Very Unbalanced Data 83/17

The Data


% of Participants by Race
0.5
0.4
0.3
0.2
0.1
0.0
Caucasian
Asian
Latino
Other
Black
Race

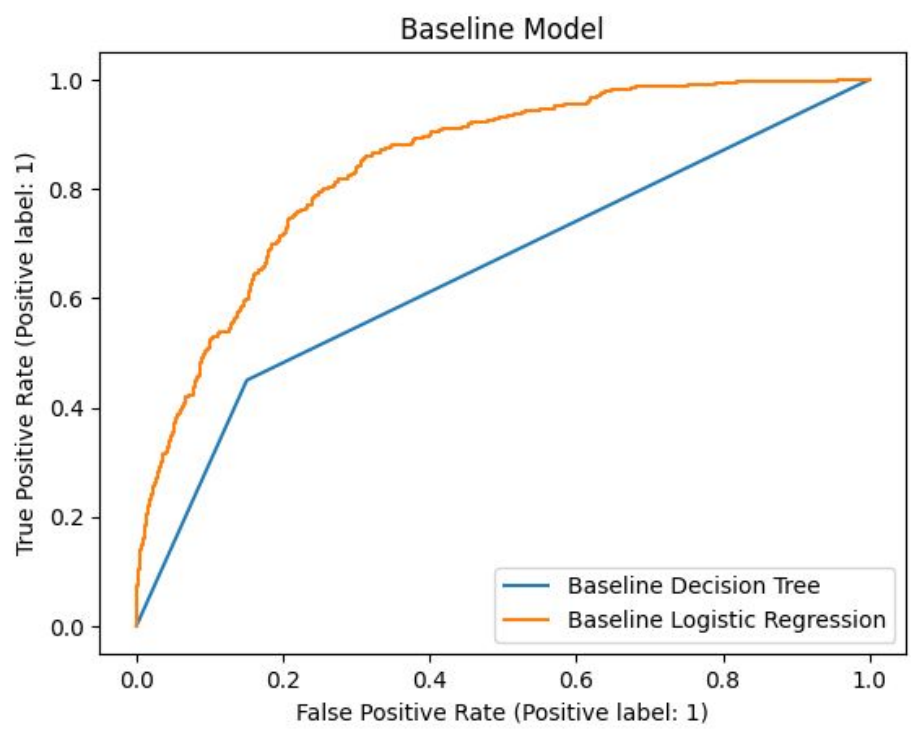

% of Participants by Genre
0.5
0.4
0.3
0.2
0.1
0.0
female
male
Gender

The Data


Age distribution
Age Mean
Age Median
Number of Ocurrences
Age
0
200
400
600
800
1000
1200
1400
1600
20
25
30
35
40
45
50
55

Baseline Models

- Our metric is accuracy.
- The baseline models have the exact same accuracy: 0.775
- The baseline decision tree tends to overfit



Hyperparameter Tuning

- Create a ModelValidator class
- Iterates and cross-validates every possible combination of hyperparameters

```
validator_pol_int.scores(both_models=True)
```

	Decision Tree	Logistic Regression
Accuracy train	0.796	0.807
Accuracy test	0.807	0.757
Recall train	0.738	0.840
Recall test	0.644	0.726
Precision train	0.834	0.788
Precision test	0.486	0.413
F1 train	0.783	0.813
F1 test	0.554	0.527
CV results	0.795	0.801

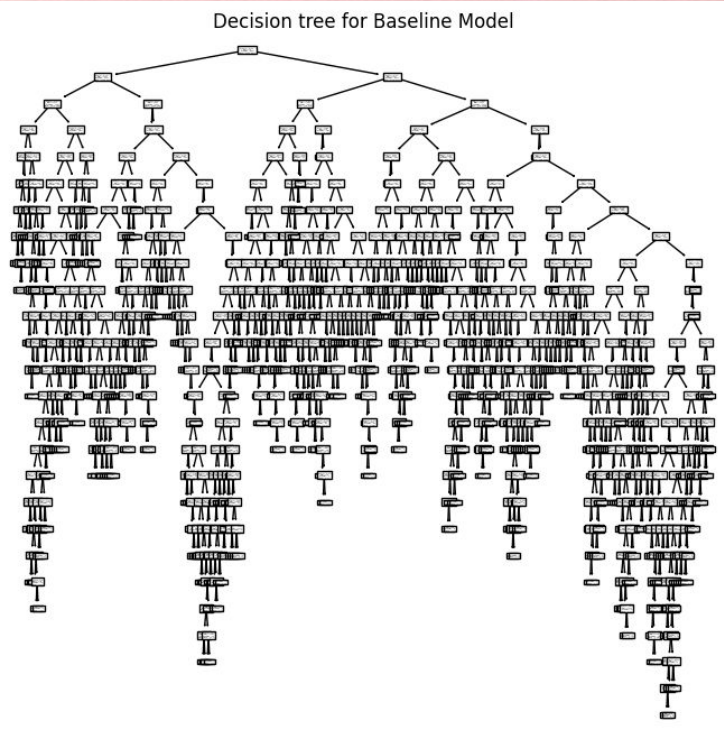
```
[76]: validator_pol_int.top_hp_lr

[76]: [({'C': 100,
         'max_iter': 100000,
         'penalty': 'l1',
         'random_state': 420,
         'solver': 'liblinear',
         'tol': 1e-05},
        0.8006306108760706,
        0.7574901074053138)]

[77]: validator_pol_int.top_hp_dt

[77]: [({'criterion': 'gini',
         'max_depth': 2,
         'min_samples_leaf': 1,
         'min_samples_split': 2,
         'random_state': 420},
        0.7947274533071921,
        0.8072357263990956)]
```

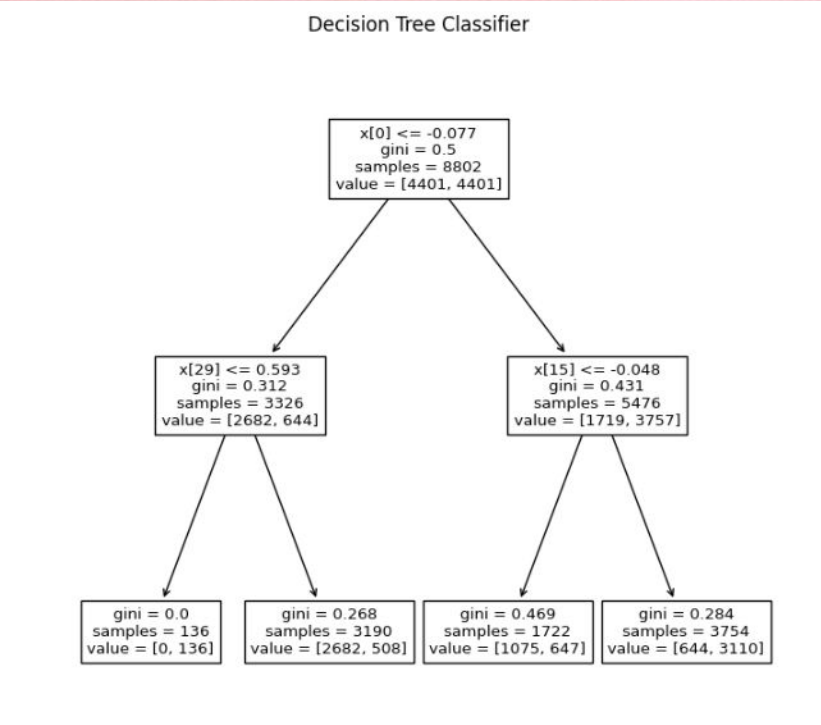
The Best Model

- Accuracy: from 0.775 to 0.81
- AUC from 0.65 to 0.76
- Most important features:
  - Attractiveness
  - Likeness
  - Did they met before?
  - Shared Interest

- Reduce Overfitting
- Simplify the model to analyze just the important features without losing accuracy
- Adding Polynomial Features

The Best Model





Recommendations

- Create workshops for grooming and style
- Classes to break the ice to avoid the effect of meeting someone before.
- Create the premium tier to make people more likely to match

With this recommendations and our model, we can definitely scale our dating advising business and release our matching app, in which premium users will be more likely to find people to match with.

Next Steps

- Create subsets by race or gender to analyze individual demographic factors
- Figure how to collect data from users directly from their phones for analysis and prediction
- Create personalized recommendations based on individual preferences.

Questions?

Eduardo Gonzalez Linkedin Github