Its Cuffing Season!

A random tree and bagging solution to finding a match ;)

What user attributes and self-presentation features best predict high matchability on OkCupid?

Data Description

- Around 60,000 user profiles + 31 variables
- High Dimensionality with mixed types (continuous, categorical and ordinal)
- Messy dataset; missing values and different essay length

Demographic

- Age (mean = 32)
- Sex and Orientation
- Height in inches (mean = 68 inches)
- Ethnicity
- Location
- Income (Highly missing)

Lifestyle Preferences

- Body type descriptions
- Diet Preferences
- Languages
- Drinking, Smoking, Drug habits
- Education level
- Job Category
- Religion and how serious they are about it
- Offspring status
- Pet preferences
- Astrological sign

Engagement Indicator

- Essay prompts (10)

Data Preparation

1. Feature Engineering - Matchability Proxity

- Pseudo target based on research
- Three intermediate features

1. **Curation Effort**: Number of essays completed (0 to 10)
2. **Profile Completeness**: Number of fields filled
3. **Bio Words**: Total number of words across all essay responses

- Normalize to 0-1

Matchability score = 0.4*(curation scaled) + 0.3*(completeness scaled) + 0.3 *(bio words scaled)

Top 30% = High Matchability

Middle 40% = Delete

Bottom 30% = Low Matchability

Data Preparation

2. Data Leakage Prevention

Exclude engineered features

Models learn from the original demographic and lifestyle variables only

3. Train-Test Split

36,000 final observations

70/30 Train Test Split to compare across 3 models

- Sufficient training data for stability, and reliable estimates

4. Missing-Value Imputations

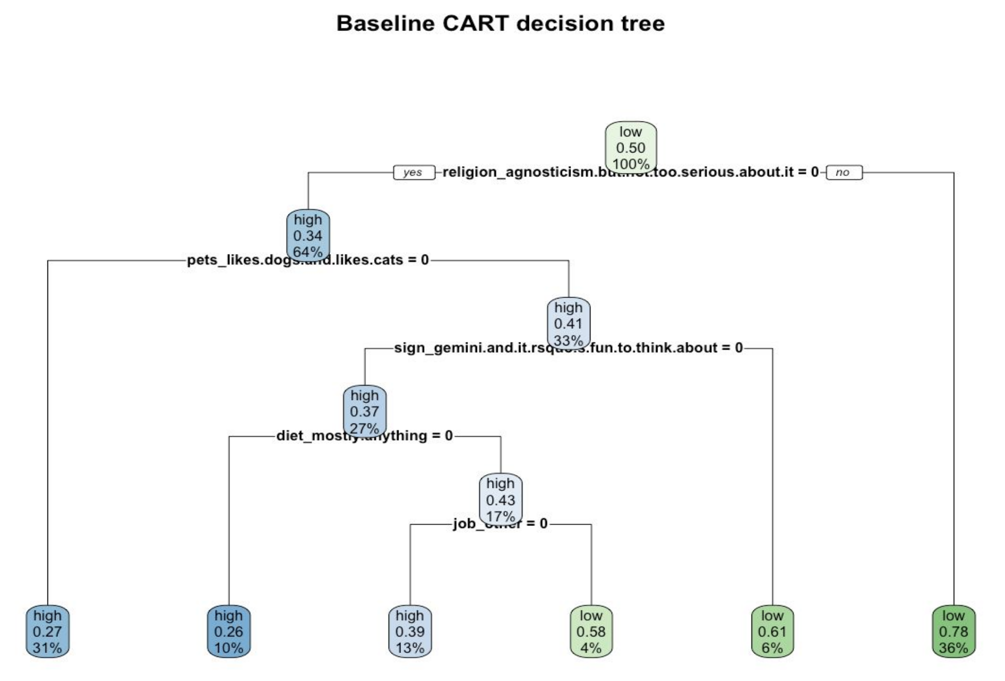
- Numeric variables -> KNN imputations with K=5
- Categorical variables -> Mode imputation

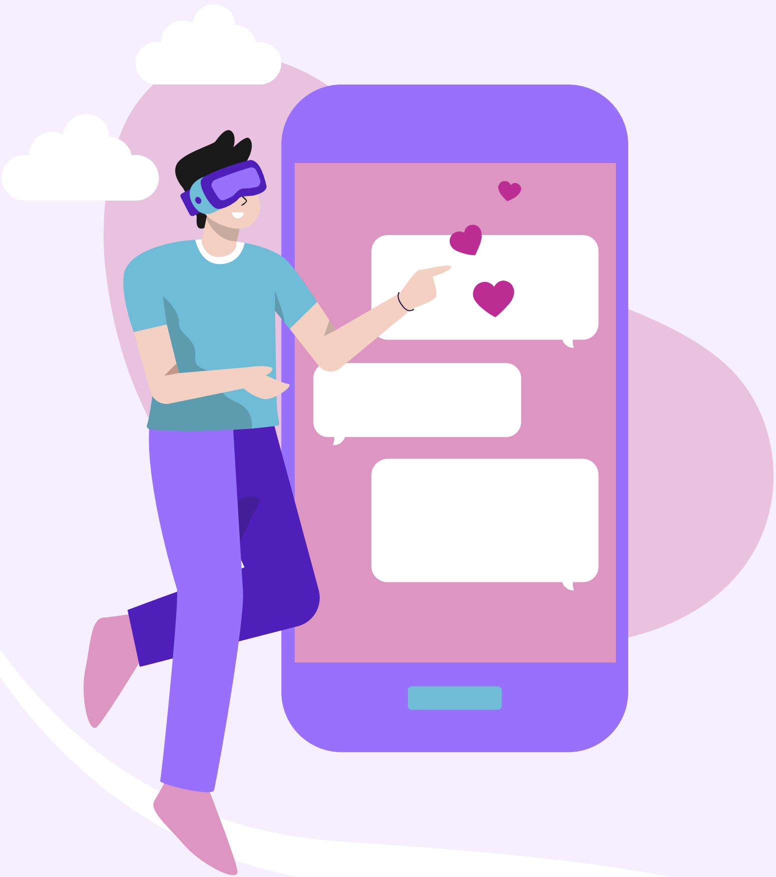
01

Analyses

CART Decision Tree

- Used for benchmarking
- Known for simplicity and interpretability
- Limitations:
    - Overfitting
    - Instability
    - Sensitive to Noise





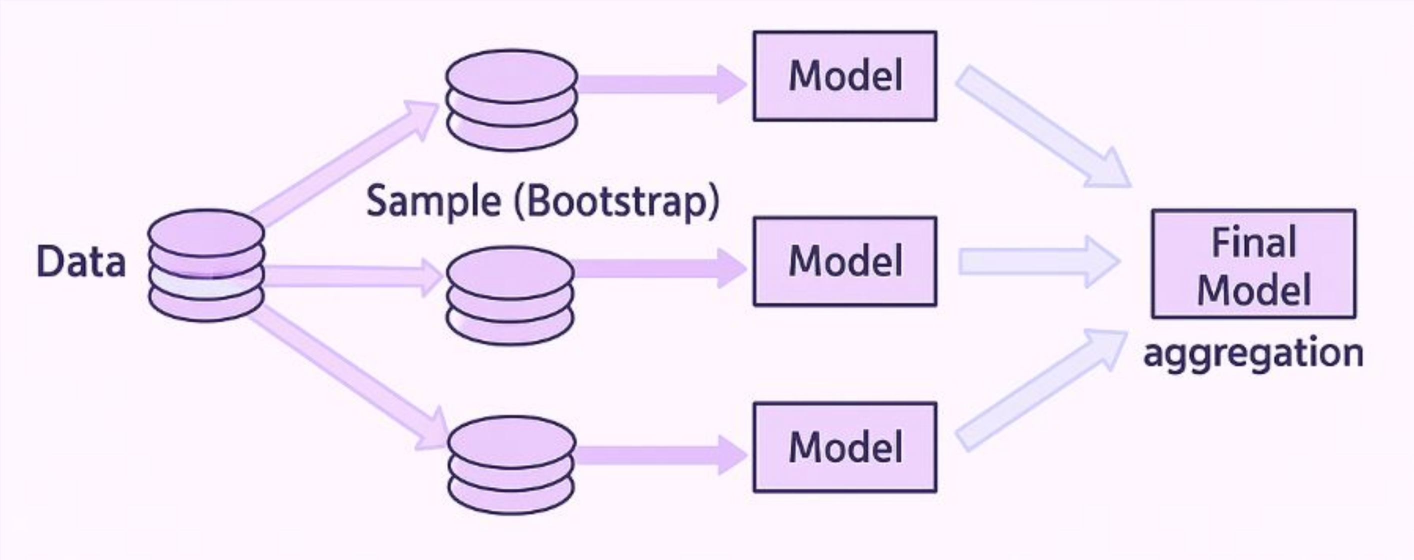
Bagging

Bagging (Bootstrap Aggregating)

Overview

1. CART trees are **high-variance**: small changes in the data → big changes in the tree
2. Bagging reduces variance by **training many trees**
3. Each tree sees a **slightly different dataset**, because of bootstrap sampling
4. Aggregation (majority vote) gives **more stable + more accurate predictions**

Bagging process



Model setup

1

200 bootstrap CART trees

2

mtry = p (all predictors available at each split → classical bagging, Breiman)

3

Uses **Out-of-Bag** (OOB) error for internal validation

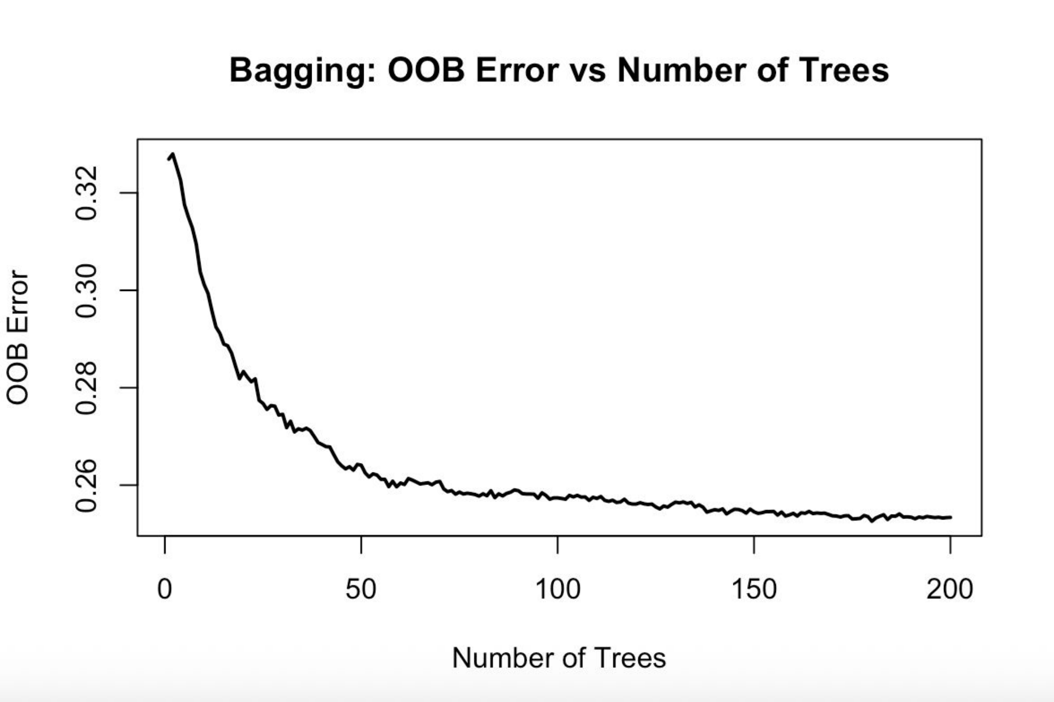
4

No pruning → each tree grows deep

5

Removes **instability** from a single tree

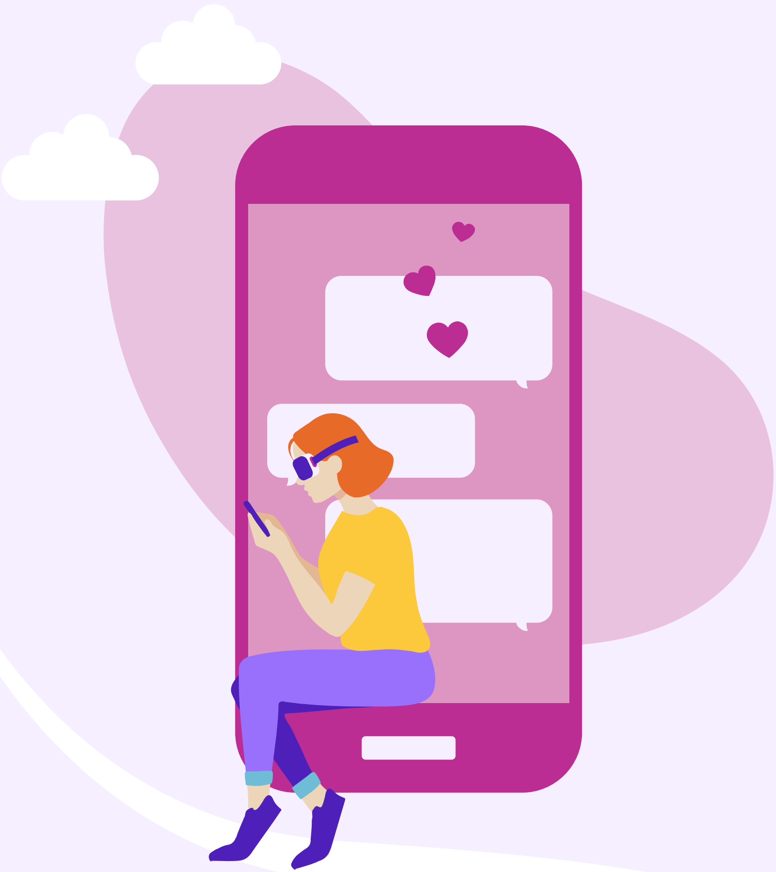
OOB Error plot



Bagging results

	OOB error	Accuracy	Precision	Recall
Bagging	0.253	0.752	0.720	0.824

	Truth high	Truth low
Pred high	4445	1731
Pred low	950	3665



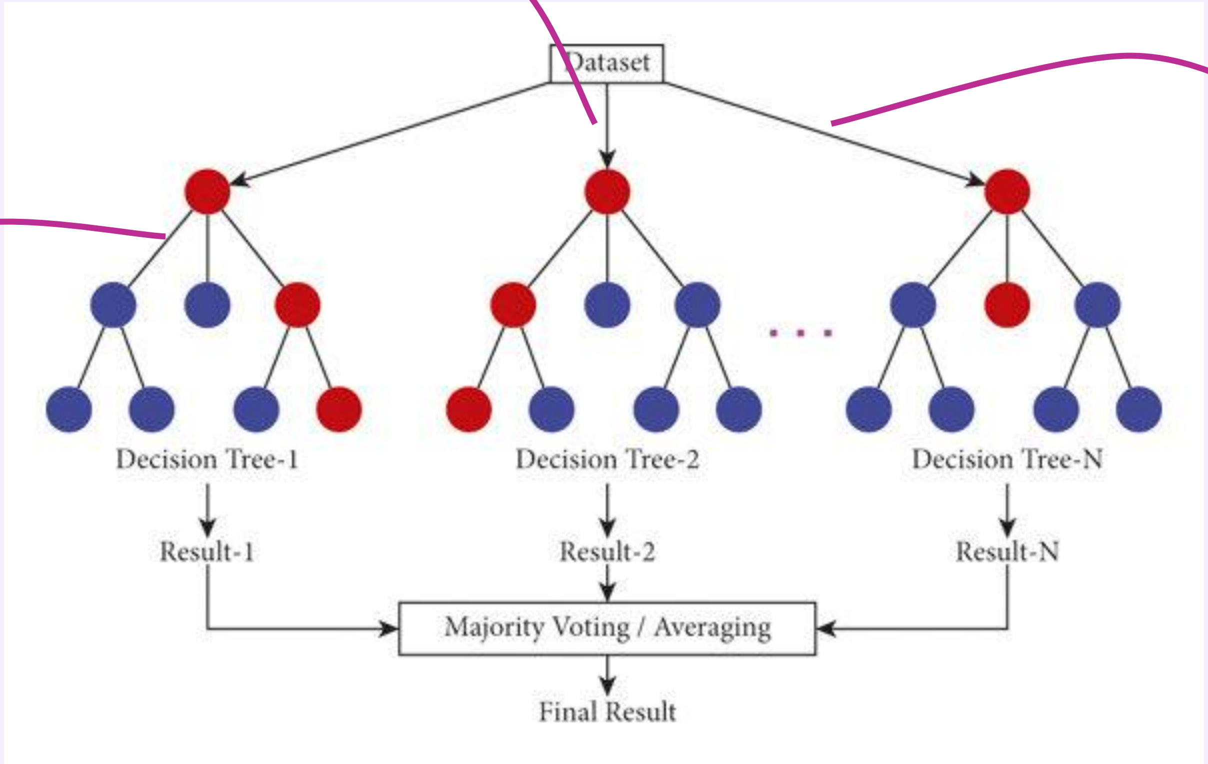
Random Forests

Random Forests

Why RF works

- Many decision trees work together
- Each tree sees a different bootstrapped sample
- At each split- tree chooses a random subset of predictors
- Final Prediction → Majority Vote for Classification
- Reduces overfitting + stabilizes high variance trees
- Excels in high dimensional, messy, and categorical heavy data
- Trees grow deep because pruning isn't used and depth increases diversity across trees.

Another bootstrap sample
random set of predictors



Bootstrap
sample
random set of
predictors

Also another
bootstrap
sample with
different
random set of
predictors

Model setup

1

25k training profiles + **163** total processed features

2

mtry = sqrt(p) with tuning. Random subset of predictors available at each split

3

Uses **Out-of-Bag** (OOB) error for internal validation

200 trees because OOB plateaued at 180-200

4

No pruning → each tree grows deep

5

Removes **instability** from a single tree + reduces **variance** from bagging

Random Forest Results

Internal Estimate

23.72%

OOB Error

76.3%

OOB Accuracy

Test Set

0.770

Accuracy

0.732

Precision

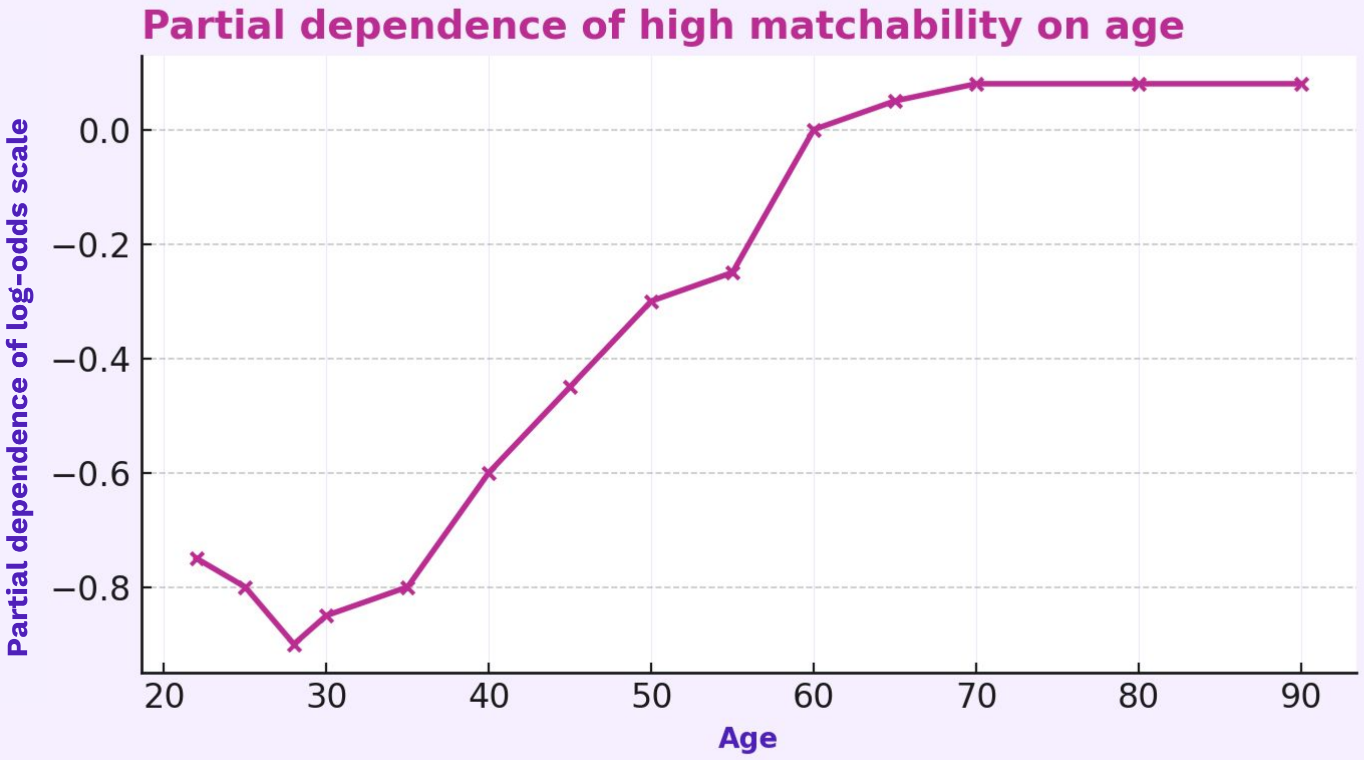
0.851

Recall

Confusion Matrix

	Truth high	Truth low
Pred high	4535	1659
Pred low	860	3737

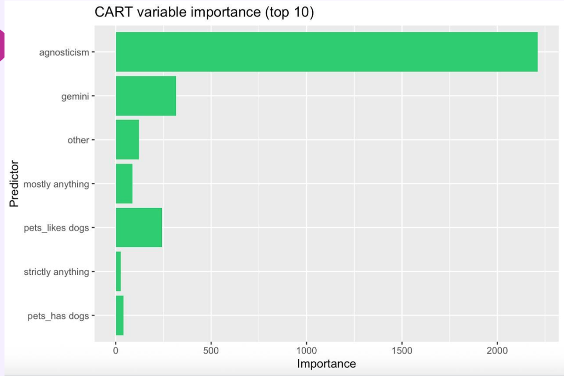
Partial Dependence Plot - Age

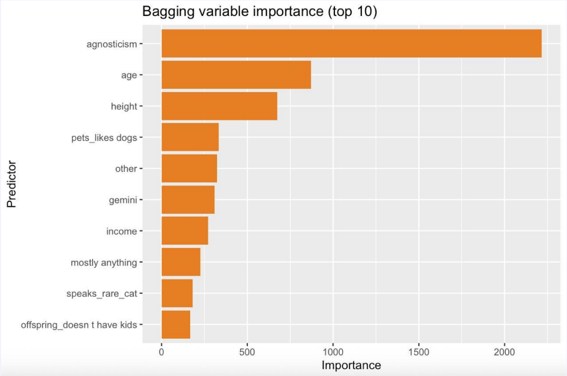


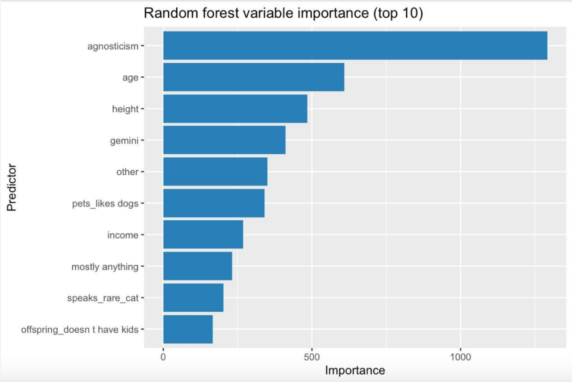
02

Model Comparisons

Variable Importance Plots







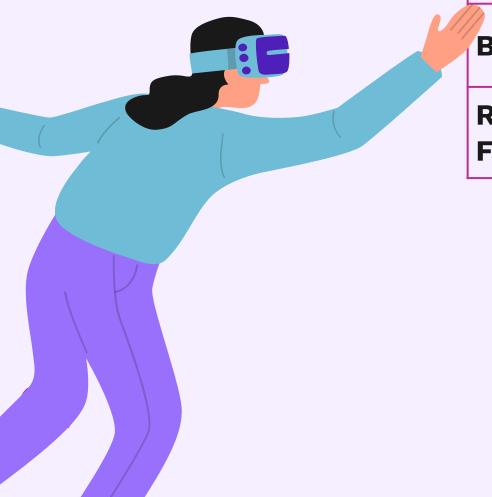
- **Sensitive** to dominant dummies
- Focuses heavily on 1-2 splits

- Smoothes out **instability**
- Variable importance more **balanced**

- **Best** overall stability
- More **robust** because of feature sampling (mtry)

Evaluation Metrics

	Accuracy	Precision	Recal	Notes
CART	0.720	0.703	0.761	Simple and Interpretable
Bagging	0.752	0.720	0.824	Reduced Variance
Random Forest	0.767	0.732	0.841	Best Overall



CART < Bagging < Random Forests

03

Conclusions

Limitations

General Limitation of Bagging and RF

Reduced Interpretability
No single tree - no clear decision paths

Limitation to our dataset

1. **Pseudo Target**

Proxy - Not real match behaviour

Matchability formula weights are somewhat arbitrary

Future work -> use true number of likes/matches from APIs if possible

2. Generalization Limits

Users in San Francisco, 2012

3. Missing data Assumptions

Income ~ 48% missing
Future improvement with more complete data

3. Data Transformations

Crude essay features

- **Natural Language Processing**

Business Insights and Real Life Applications

For Users

- Fill out most profile fields (shows effort) SELF PRESENTATION THEORY
- Write thoughtful & appropriate essays
- Content also matters!

For Platforms

- Platforms can use matchability scores to segment users

- Can Implement matchability feedback system
- New AI feature

Conclusion

Decision tree

- Unstable & high-variance
    - small sampling changes → very different trees

Bagging

- 200 bootstrap trees - combined predictions through majority
- Improves performance by averaging unstable tree models

Random forest

- Random feature subsampling
- Most robust + generalizable

Key predictors

- Religion
- Age
- Height
- Pets
- Job category

Thank you!
&
Good luck ;)