Who Will Get the Final Rose?

Predicting the winner of The Bachelor with machine learning models

By Anna Ramer, Monica Dahl, Natasha Fidler, and Ramya Subramanian

What is The Bachelor?

- Reality Dating Show
- One Bachelor with 20-30 contestants vying to win
- Continuing week to week is marked with roses
- The winner is the contestant who the Bachelor proposes to in the final episode

Data Cleaning

Data Cleaning

- First step is to process the raw CSV data
- Data was collected from a GitHub repo from a similar project from 2021
- Limitations:
  - This data is not complete, some seasons are missing, and since it was done several years ago, it is not up to date with the current show
  - There are only a few pieces of hard data that can be gathered about the participants, and not all can be put into an ML model

Data Cleaning

- Height and Occupation columns were dropped (height was mostly not filled and occupation was too hard to make into a numerical data type for use in ML modeling)
- Hometown was changed to only hold the state or country (if outside the US)

1	Name	Age	Occupation	Hometown	Height	ElimWeek	Season
2	Amanda Marsh	23	Event Planner	Chanute, Kansas			1
3	Trista Rehn	29	Miami Heat Dancer	Miami, Florida		6	1
4	Shannon Oliver	24	Financial Management Consultant	Dallas, Texas		5	1
5	Kim	24	Nanny	Tempe, Arizona		4	1
6	Cathy Grimes	22	Graduate Student	Terra Haute, Indiana		3	1
7	Christina	28	Attorney	Bonita, California		3	1
8	LaNease Adams	23	Actress	Los Angeles, California		3	1
9	Rhonda	28	Commercial Real Estate Agent	Woodward, Oklahoma		3	1

Data Cleaning

- Height column was dropped (very little data was present)
- Hometown was changed to state or country (if outside the US)
- The age was used to calculate age difference between each contestant and their bachelor, and whether or not the contestant was younger
- Each contestant's hometown was compared to their bachelor to see if it matched or not

1	Name	Age	Hometown	Height	Season
2	Alex Michel	32	Charlottesville, Virginia		1
3	Aaron Buerge	28	Butler, Missouri		2
4	Jesse Palmer	34	Toronto, Ontario		5
5	Lorenzo Borghese	34	Milan, Italy		9
6	Andy Baldwin	30	Lancaster, Pennsylvania		10
7	Brad Womack	35	Austin, Texas		11

Data Cleaning

	Name	Age	ElimWeek	Season	State/Country	Age Difference	Younger	Same Home
0	Amanda Marsh	23	0	1	Kansas	9	1	0
1	Trista Rehn	29	6	1	Florida	3	1	0
2	Shannon Oliver	24	5	1	Texas	8	1	0
3	Kim	24	4	1	Arizona	8	1	0
4	Cathy Grimes	22	3	1	Indiana	10	1	0
5	Christina	28	3	1	California	4	1	0
6	LaNease Adams	23	3	1	California	9	1	0
7	Rhonda	28	3	1	Oklahoma	4	1	0
8	Alexa	27	2	1	California	5	1	0
9	Amy	28	2	1	New York	4	1	0

Decision Tree

Decision Tree

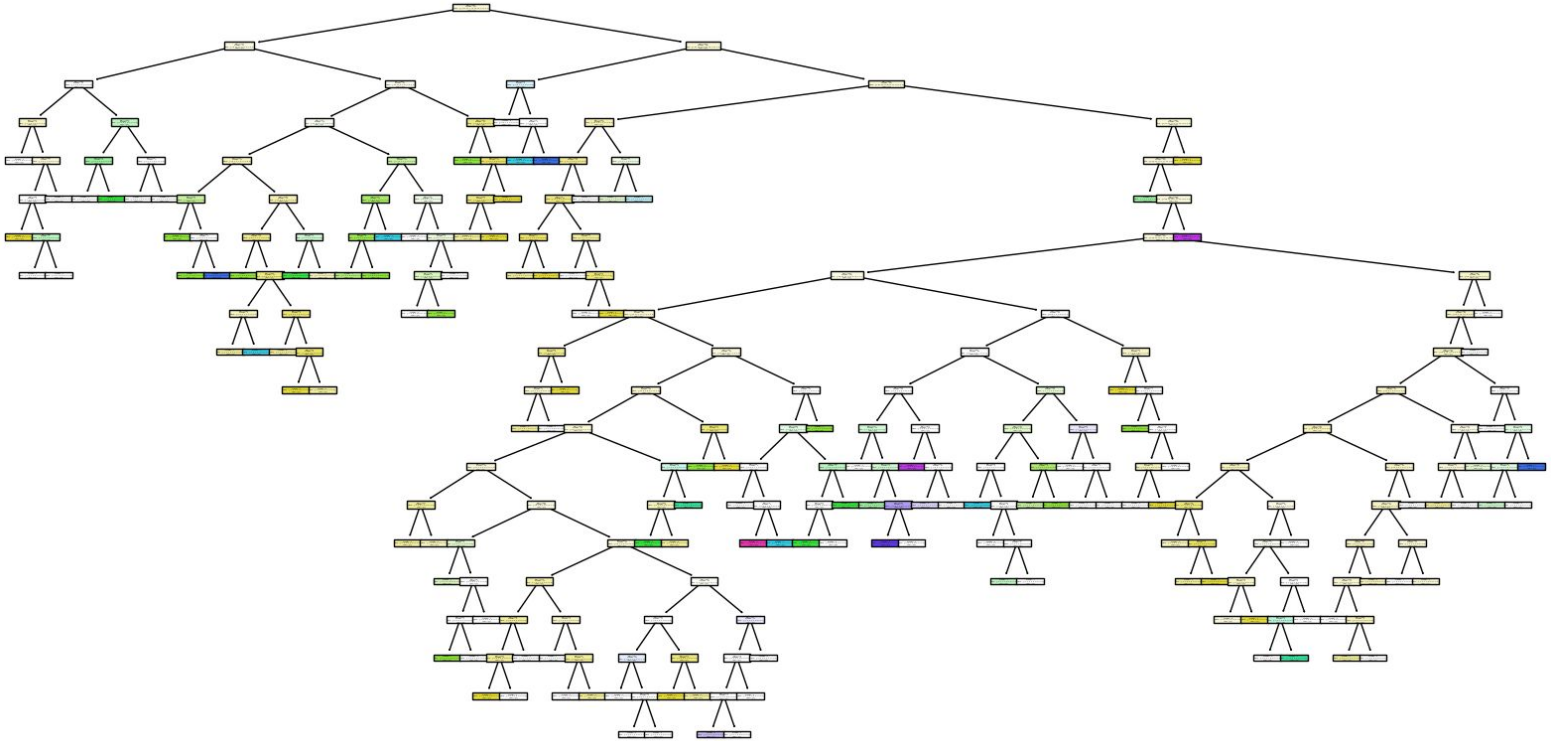
Features (X variable):

- Age
- Season
- Age Difference
- Younger
- Same Home

Target (y variable):

- Elimination Week

Decision Tree



Decision Tree

```
Accuracy: 0.25882352941176473
```

Logistic Regression

Logistic Regression

For this model, a column was added to mark if a contestant was the overall winner or not

Features (X variable):

- Age
- Age Difference
- Younger
- Same Home

Target (y variable):

- Winner

Logistic Regression

Accuracy: 0.9622641509433962

Training Data Score: 0.9588607594936709
Testing Data Score: 0.9622641509433962

	Prediction	Actual
0	0	0
1	0	0
2	0	0
3	0	0
4	0	0
...	...	...
101	0	0
102	0	0
103	0	1
104	0	0
105	0	0

106 rows × 2 columns

Neural Network

Neural Network: Data Preprocessing, Feature Selection, and Target

---

Again a column was added to track if a contestant won

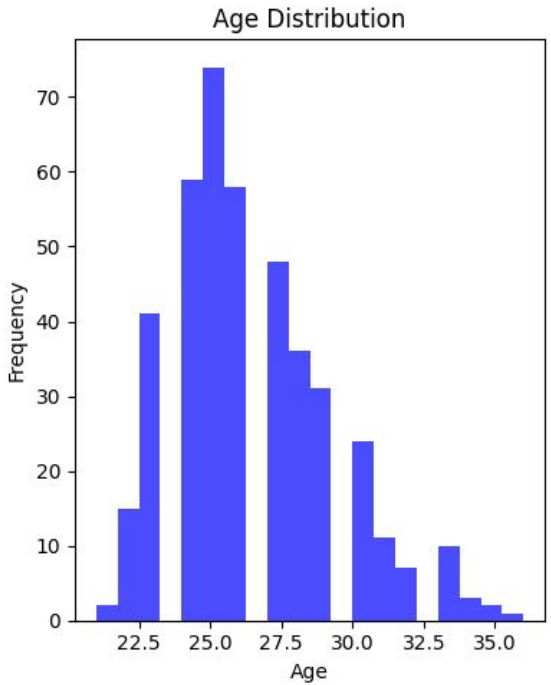
Features (X variable):

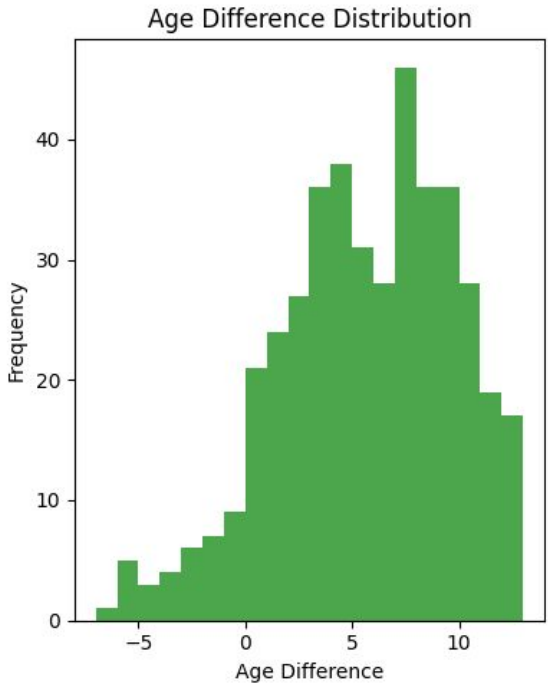
- Age
- Age Difference
- Same Home

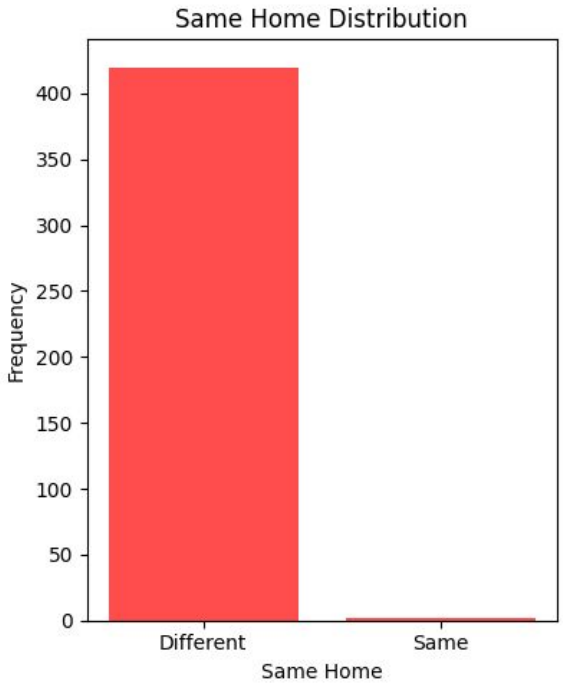
Target (y variable):

- Winner

Neural Network: Visualizing Feature Distributions







Neural Network: Model Definition and Summary

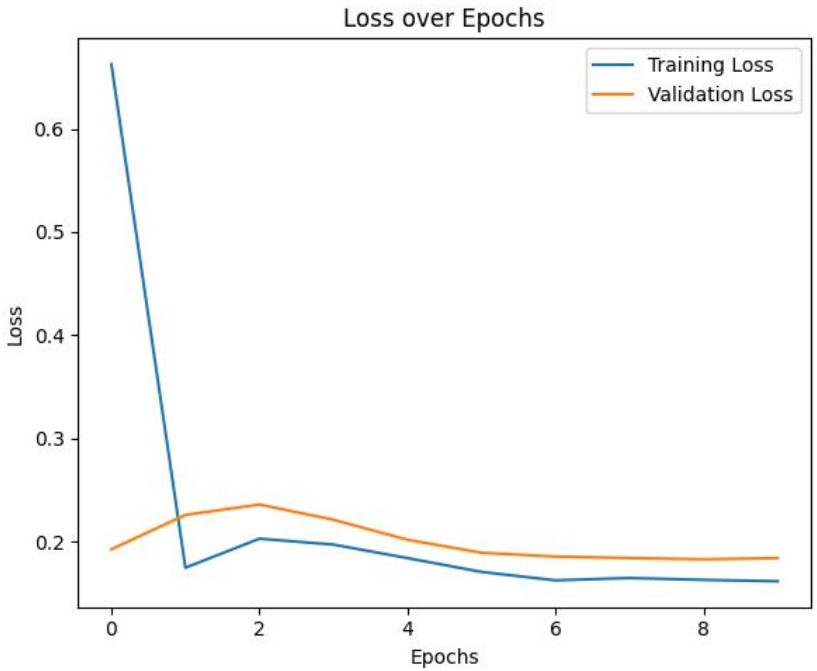
Model Summary:

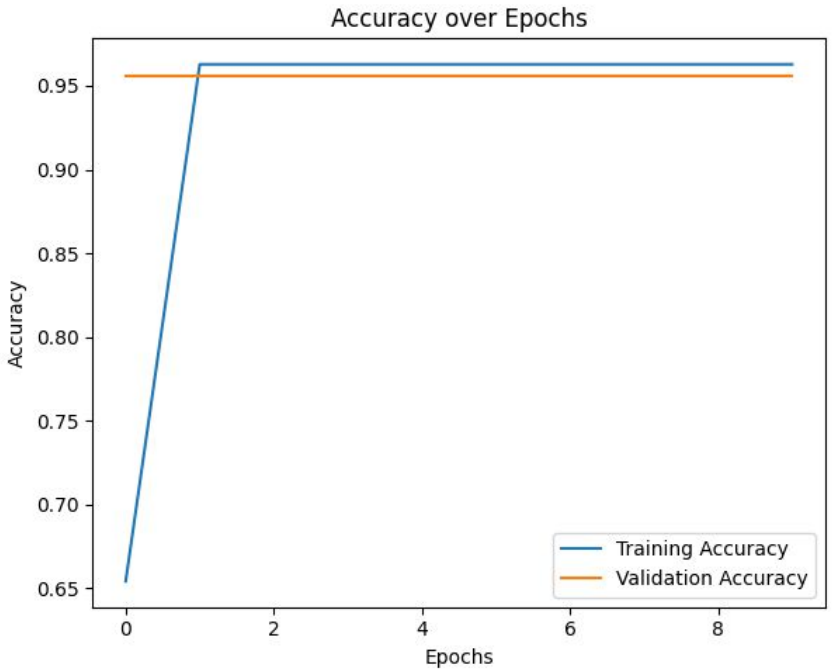
Hidden Layers: 2
Output Layer Activation: Sigmoid
Optimizer: Adam
Loss Function: Binary Cross-Entropy

```
Model Summary:
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 64)                256

 dense_1 (Dense)             (None, 64)                4160

 dense_2 (Dense)             (None, 1)                 65

=================================================================
Total params: 4481 (17.50 KB)
Trainable params: 4481 (17.50 KB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

Neural Network: Model Training

We trained the model with the training data.





Neural Network: Evaluation and Predictions

We evaluated the model on the testing data.

```
3/3 [==============================] - 0s 9ms/step - loss: 0.1868 - accuracy: 0.9529
Test Loss 0.1867581456899643
Test Accuracy 0.9529411792755127
```

Predictions were made on the test set and probabilities were converted to binary outcomes using a threshold of 0.5.

```
Predictions on Test Set
[[0.02528317]
 [0.03863592]
 [0.00580877]
 [0.04563499]
 [0.0349797 ]]
```

Conclusions & Recommendations

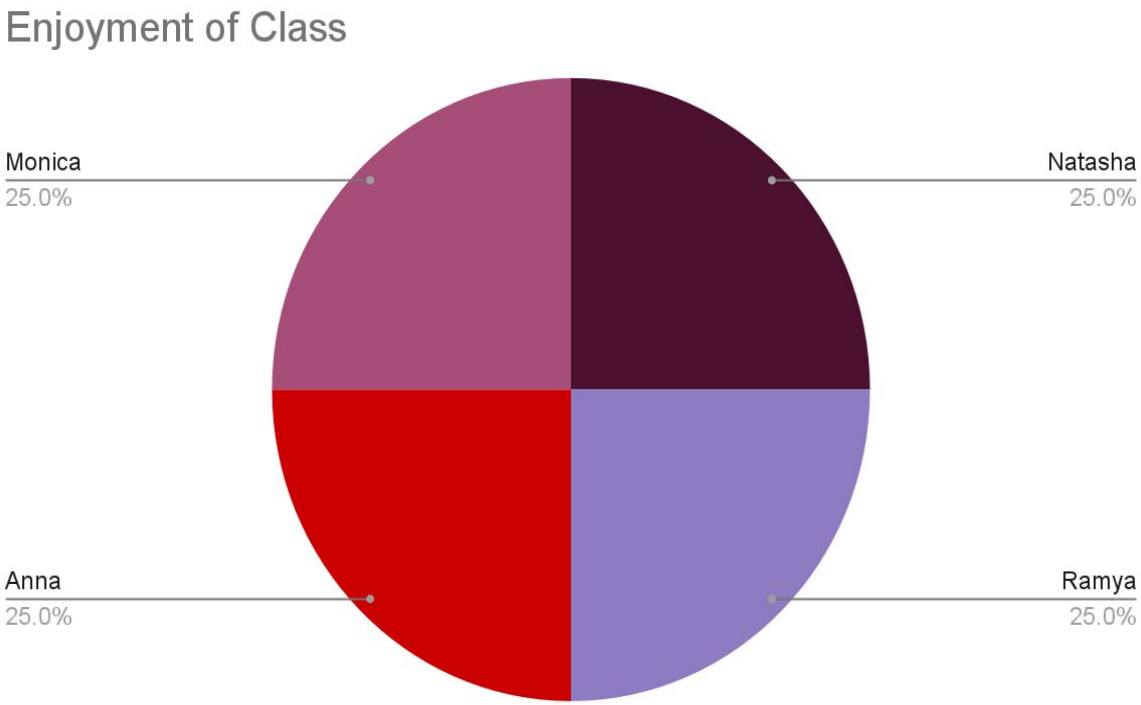
Conclusions

- The logistic regression and neural network models both had very high accuracy for our dataset and would both be good choices for predicting winners based on this available data
- Caveats: This dataset is not complete. There are seasons of the Bachelor not represented, and the categories of data are limited.
- Ideally, the best test would be to use the next season's contestants and see if it predicts the winner

Recommendations

How would we improve if we had more time/resources?

- Collect our own data from websites/other sources to fill in missing seasons
- Collect other pieces of data not represented (height, hair color, income bracket, etc.)
- Spend time encoding the non-numeric types of data to be usable (occupation, personal statements, etc.)
- Add other procedural data such as first impression roses, hometown dates, fantasy dates, etc. which may be good prediction factors but are more complex to collect.


Enjoyment of Class
Monica
25.0%
Natasha
25.0%
Anna
25.0%
Ramya
25.0%

Thank you!