PREDICTING STAR BAKER FROM THE GREAT BRITISH BAKE OFF

Presented by Matt Gargiulo

Capstone 2 Final

About GBBO

- The Great British Bake Off has been captivating audiences for 12 seasons with the 13th season just around the corner.
- In 2022, the show’s first episode was watched by nearly 11 million people in the UK alone, setting a record for the Channel 4 network that has stood since 1985.
- Each Season sees roughly 12 contestants battle it out in weekly themed baking challenenges including a signature, technical, and showstopper challenge.
- Worst baker each week is eliminated until a winner emerges!

Our Data

GBBO Dataset

- Sourced from an article in medium.com.
- This dataset comprises 1256 rows of data across 16 features.
- Covers seasons 1 through 11.

Challenges Dataset

- Sourced from an R package titled bakeoff.
- This dataset comprises 1136 rows of data across 7 features.
- Mainly focuses on the recipes from the showstopper and signature challenges.

Our Data

GBBO Dataset

#	Column	Non-Null Count	Dtype
0	Season	1256 non-null	object
1	Judge	1256 non-null	object
2	Week Number	1256 non-null	int64
3	Week Name	1256 non-null	object
4	Baker	1256 non-null	object
5	Gender	1256 non-null	object
6	Age	1256 non-null	int64
7	Signature Handshake	1256 non-null	int64
8	Technical Rank	771 non-null	float6
9	Showstopper Handshake	1256 non-null	int64
10	Favorite	1256 non-null	float6
11	Least Favorite	1256 non-null	int64
12	Star Baker	1256 non-null	int64
13	Eliminated	1256 non-null	int64
14	Competed	1256 non-null	int64
15	Winner	1256 non-null	int64

Challenges Dataset

#	Column	Non-Null Count	Dtype
0	series	1136 non-null	int64
1	episode	1136 non-null	int64
2	baker	1136 non-null	object
3	result	710 non-null	object
4	signature	703 non-null	object
5	technical	696 non-null	float64
6	showstopper	688 non-null	object

Our Data
Cleaning and Merging

- The datasets were merged on Season, Week_Number and Baker.
- over 400 null rows were removed from the GBBO dataset.
- Due to mismatching seasons, our final DataFrame contains data from Seasons 2-10.
- Final DataFrame contains 668 rows and 16 featuers.

Season	category
Week_Number	category
Week_Name	category
Baker	category
Gender	category
Age	int64
Signature_Handshake	int64
Technical_Rank	int64
Showstopper_Handshake	int64
Favorite	float64
Least_Favorite	int64
Star_Baker	int64
Eliminated	int64
Winner	int64
signature	object
showstopper	object

Exploratory Data Analysis

Things to Explore:

- How Age affects Star Baker
- How Gender affects Star Baker
- Distribution of handshakes
- Relationship between the number of Star Baker's a baker recieves and the winner for that season.

Exploratory Data Analysis

Gender

Takeways:

- Gender was kept at the M/F designation.
- Female bakers won Star Baker 50% more often than males.
- The show only had four more female bakers than males.

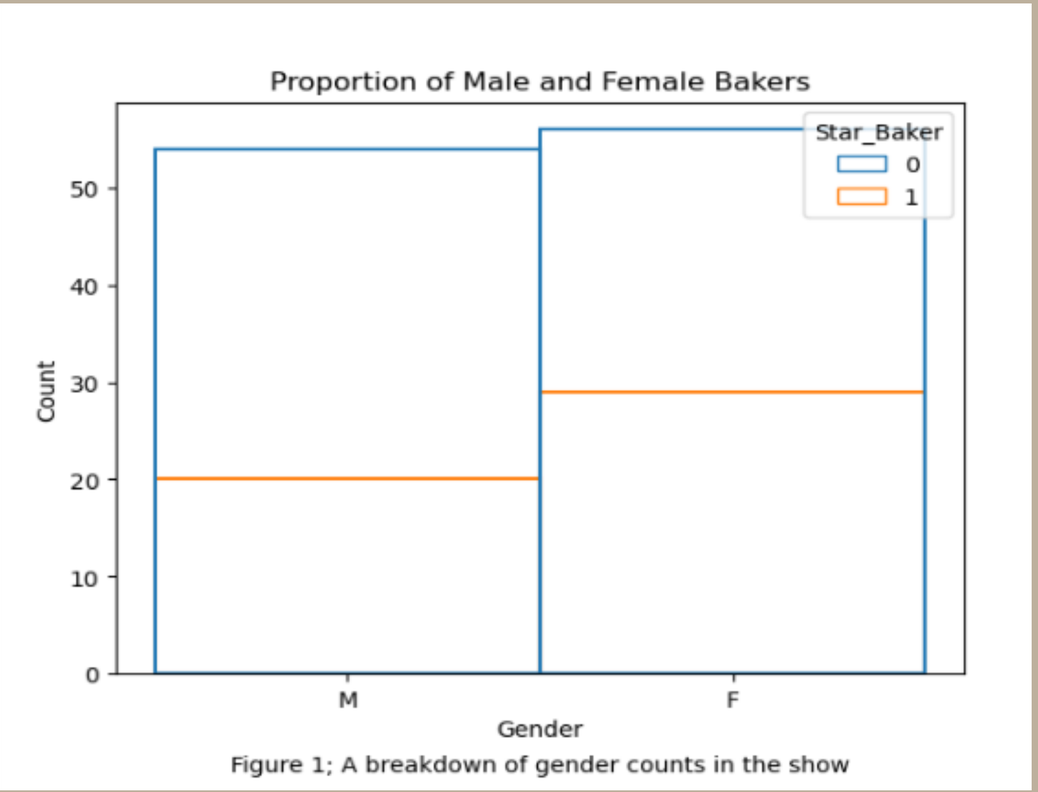

Figure 1; A breakdown of gender counts in the show

Exploratory Data Analysis

Age

Takeways:

- Age wasn't a factor in deciding star baker.
- Great to see older contestants participate.

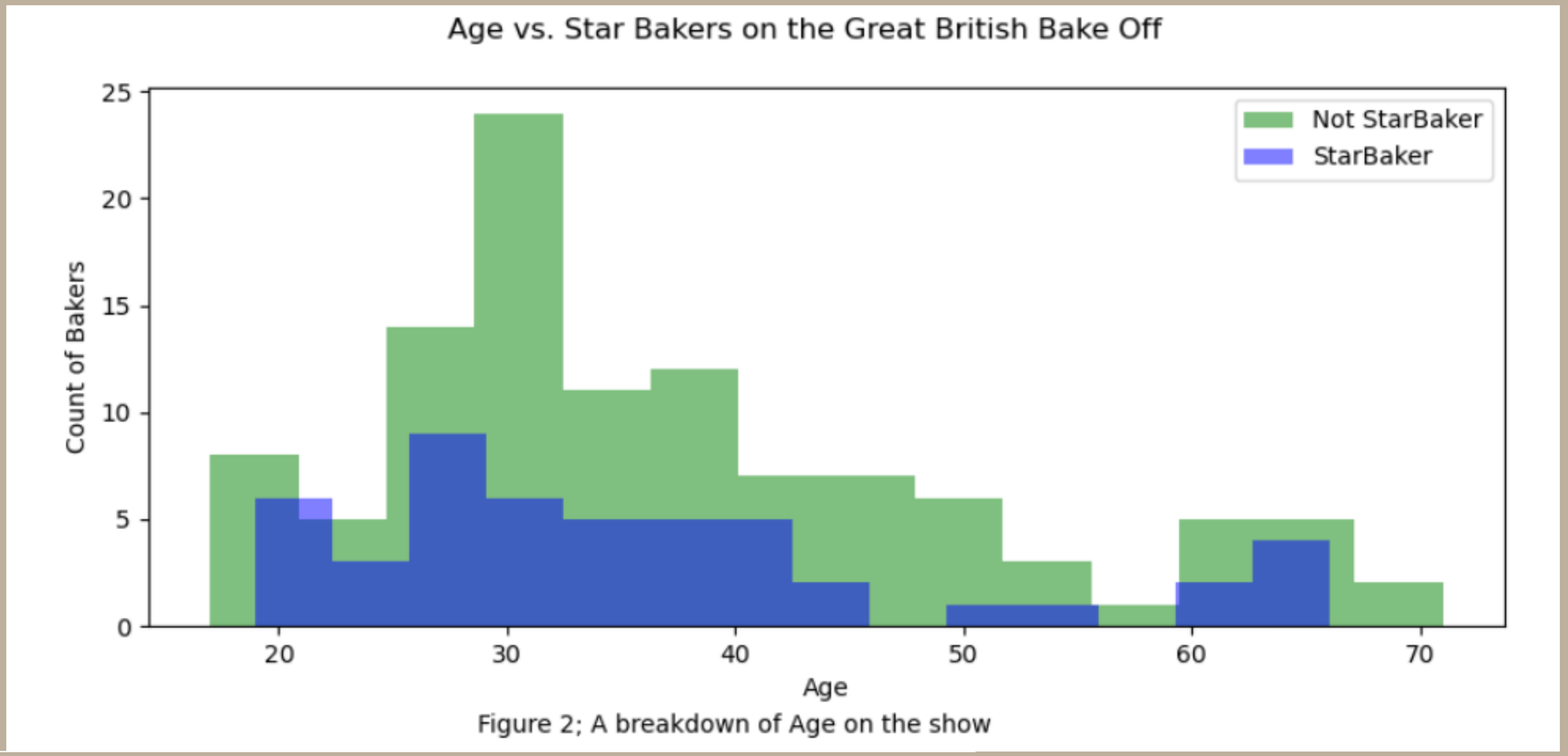

Figure 2; A breakdown of Age on the show

Exploratory Data Analysis

Handshakes

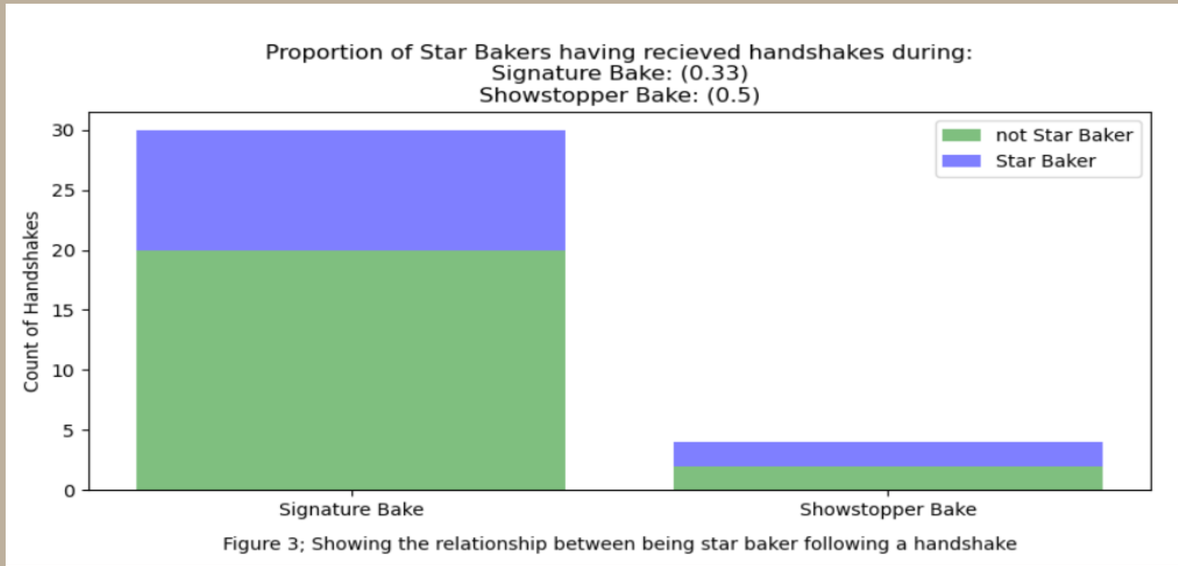

Figure 3; Showing the relationship between being star baker following a handshake

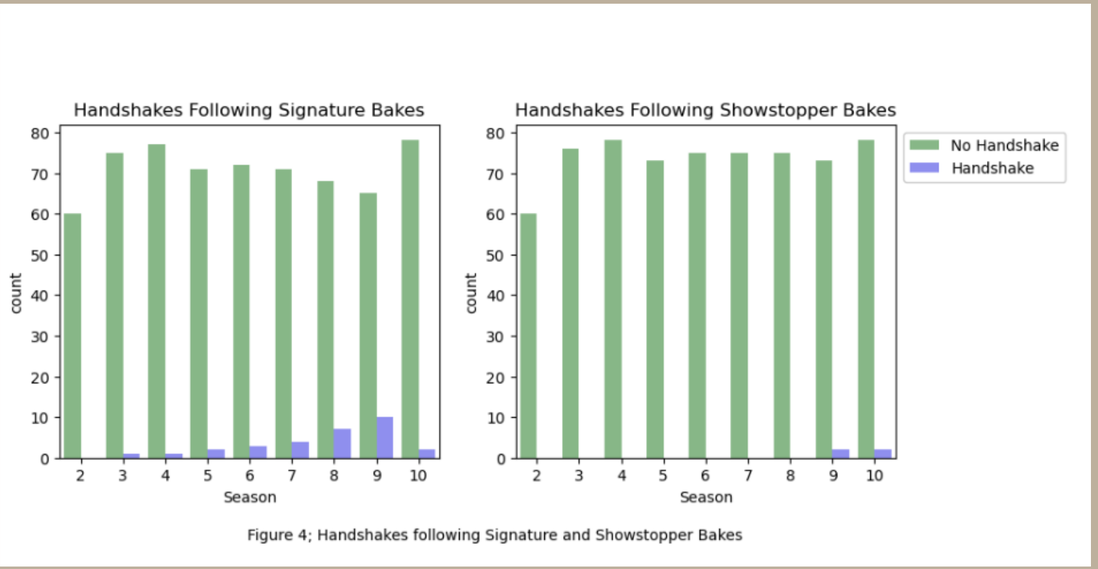

Figure 4; Handshakes following Signature and Showstopper Bakes

Exploratory Data Analysis

Handshakes

Takeways:

- Handshakes rarely lead to baker elimination.
- Bakers receiving a handshake during their signature challenge have a ⅓ chance to be that week's star baker while showstopper handshake recipients have a ½ chance to win the award.
- Paul didn't start giving them out until season 9 of the show. Season 9 also seemed like a big season for signature handshakes, with over 10 given.
-

Exploratory Data Analysis

Star Baker

Takeways:

- (Richard) won star baker five times yet did not win that season! Season 5's winner (Nancy) only accumulated two star bakers.
- The correlation between star baker and winning the show at only 0.438.
-

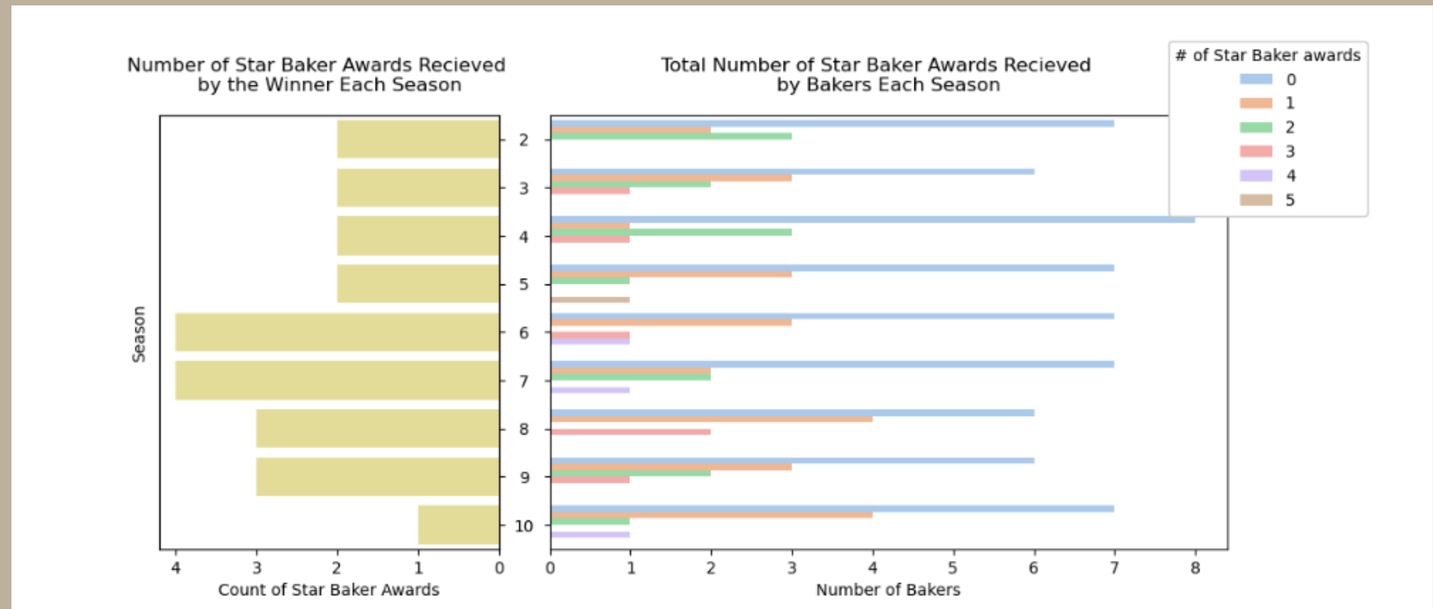

Figure 5; Exploring the number of star baker awards accumulated by bakers per season and the number of star bakers recieved by the shows winner

Advanced Analysis

Feature Engineering

Methods used include:

- StandardScalar() to scale numerical features.
- OneHotEncoder() to transform categorical features.
- TfidVectorizer to split and count text features.

Advanced Analysis

Feature Engineering

Additional Notes:

- Our target class (star baker) was severely unbalanced. (6.58 :1)
- Resampling techniques were analyzed for performance including SMOTE(), SMOTETomek() and TomekLinks()

Advanced Analysis

Model Creation

Models Tested:

- Random Forest (w/o CV)
- Random Forest (w/ CV)
- Logistic Regression (w/o CV)
- Logistic Regression (w/ CV)
- Logistic Regression (w/ CV and feature reduction)
- Gradiant Boosting (w/ CV)
- KNN (w/o CV)
- KNN (W/ CV)

Advanced Analysis
Evaluating Models

Models were evaluated on their F1 Scores as well as their variance.

KNN with Cross Validation shows greatest performance as well as a very low variance.

	Random Forest w/o CV	Random Forest (w/CV)	Logistic Regression (w/o CV)	Logistic Regression (w/ CV)	Logistic Regression (w/ CV & Lasso)	Gradient Boosting (w/CV)	KNN (w/o CV)	KNN (w/CV)
Training	1.000000	0.996287	0.929455	0.929455	0.926980	0.965347	0.897277	1.000000
Testing	0.875622	0.885572	0.845771	0.845771	0.845771	0.850746	0.796020	0.895522
F1	0.561404	0.561404	0.617284	0.617284	0.617284	0.605263	0.549451	0.676923
Variance	0.124378	0.110715	0.083684	0.083684	0.081209	0.114600	0.101257	0.104478

Advanced Analysis
Evaluating Models

- A Confusion Matrix can be used to visualize the performance
- 180 correctly classified the Star Baker target features
- A recall score of 0.81, signaling that our model was not over tuned

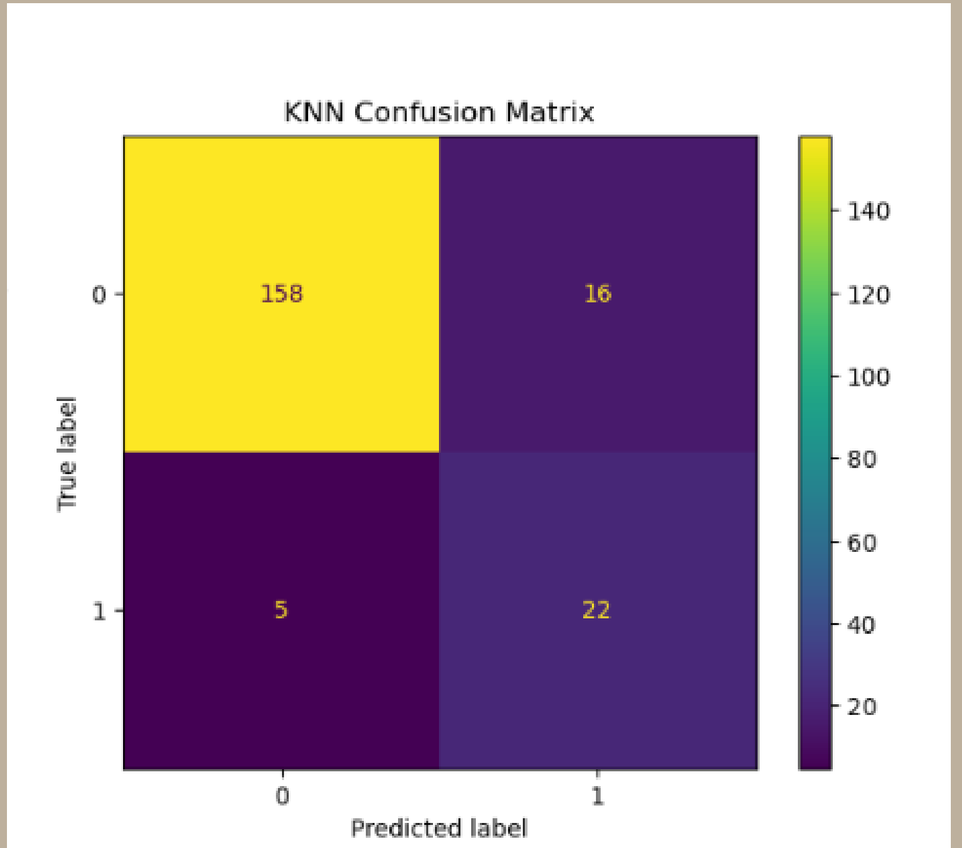

Figure 8; Confusion Matrix for the KNN Model

Advanced Analysis
Finding the Important Features

- Used Lasso Regression and Recursive Feature Reduction to find important Features
- Coefficients that can be positive or negative
- Positive coefficients are will push a model to a class value of 1, or star baker
- Negative coefficients lead the model to a 0

Finding the Important Features

- Bakers should use recipes with the word "caramel" during the signature challenge, and "lime" during the showstopper challenge

- Doing well during Tarts week is very important

	RFE		Lasso Regression	
	Features	Coef	Features	Coef
0	cat__Week_Name_Final	3.129128	cat__Week_Name_Final	5.487518
1	text_sig__caramel	1.357940	text_sig__caramel	1.549812
2	text_show__lime	1.334346	text_show__lime	2.098971
3	cat__Week_Name_Tarts	1.334249	cat__Week_Name_Tarts	1.413356
4	num__Favorite	1.311275	num__Favorite	1.394483
5	text_show__rose	1.258322	text_show__rose	1.898597
6	text_show__orange	1.147124	text_show__orange	1.596396
7	text_show__lemon	0.483759	NaN	NaN
8	NaN	NaN	text_show__cream	1.666845
9	NaN	NaN	text_sig__pie	0.545792
10	NaN	NaN	text_show__mirror	0.520642
11	NaN	NaN	cat__Week_Name_Dessert	0.436915
12	NaN	NaN	text_show__hazelnut	0.408552
13	NaN	NaN	cat__Week_Name_Bread	0.379705
14	NaN	NaN	cat__Week_Name_Pie	0.368409
15	NaN	NaN	text_sig__with	0.344581
16	NaN	NaN	num__Signature_Handshake	0.306773
17	NaN	NaN	text_sig__and	0.289899
18	NaN	NaN	num__Showstopper_Handshake	0.155185
19	NaN	NaN	text_sig__chocolate	0.126173
20	NaN	NaN	text_sig__loaf	0.092111

Conclusion

- There is much work to be done
- Future models will run models for each individual weekly theme
- Remember that this analysis was done with an extremely small dataset (roughly 650 rows of data)
- I will be looking forward to seeing how this model changes as we add more seasons of data to the model

THANK YOU

Presented by Matt Gargiulo