Final Report:
Predicting the Winner from the Great British Baking Show
By: Matt Gargiulo

Problem Statement

The Great British Bake Off (or Great British Baking Show if you live in the US) 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(link). Each Season sees roughly 12 contestants battle it out in the kitchen whipping up delicious bakes in order to impress the two judges, Paul Hollywood and Prue Leith. Furthermore, each week has a different theme, requiring the bakers to be versatile in their skill set, including, but not limited to, knowledge on how to make cakes, biscuits, vegan desserts, tarts, and much more! In order to move on in the competition, each baker must bake a signature challenge, a technical challenge, and finally, a showstopper challenge. The winner each week is then deemed the star baker. The show progresses until there are only three contestants left; this is known as 'Final's Week'.

From a strategy standpoint, the best game plan is obvious; just win star baker every week and you're guaranteed to win the show. Of course that is easier said than done, but what if we could predict who would win star baker? What if we could determine which columns (henceforth known as features) of a dataset are more important to focus on? What if certain recipes or even buzz words within the recipe were biased towards winning - wouldn't the bakers want to know if using recipes with the word orange in it, adding meringue to every bake, or even including raspberries would better their chances of winning star baker?

Data Wrangling

Our data comes from two sources:

1. An article from medium.com (here) takes a similar approach to analyzing data from the show. In the article, the author supplies a dataset that I've repurposed for my own analysis. This data will be referred to as the GBBO dataset.
2. A R package - bakeoff - by Alison Hill, Chester Ismay and Richard Lannone. This package contained information on weekly challenges including recipes used the show stopper and signature challenges, as well as numerical scores given during the technical challenges. This data will be referred to as the Challenges dataset.

Data Features

GBBO Dataset	Challenges Dataset
Season: The season of the show Judge: Whether the season was judged by Mary or Prue Week_Number: The number of the week Week_Name: The theme corresponding to the week Baker: The first name of the baker Gender: The gender associated with each baker, note this dataset denotes M/F Age: Age of the baker Signature_Handshake: If the baker received a handshake following their signature bake Technical_Rank: Score following the technical challenge Showstopper_Handshake: If the baker received a handshake following their showstopper bake Favorite: Whether the baker is deemed a favorite going into the showstopper challenge Least_Favorite: Whether the baker is deemed a least favorite going into the showstopper challenge Star_Baker: Winner of the week's star baker award Eliminated: Whether the baker was eliminated Competed: Whether the baker competed that week Winner: Whether the baker won the Season	Season: The season of the show Week_Number: The number of the week Baker: The first name of the baker Result: Whether the baker was is in the competition or eliminated Signature: The signature baker recipe Technical_Rank: Score following the technical challenge Showstopper: The showstopper bake recipe

Colors indicate columns that were joined on when combine the individual data frames

GBBO Dataset

	Season	Judge	Week Number	Week Name	Baker	Gender	Age	Signature Handshake	Technical Rank	Showstopper Handshake	Favorite	Least Favorite	Star Baker	Eliminated	Competed	Winner
0	Series 1	Mary	1	Cake	Annetha	F	30	0	2.0	0	1.0	0	0	0	1	0
1	Series 1	Mary	1	Cake	David	M	31	0	3.0	0	0.0	1	0	0	1	0
2	Series 1	Mary	1	Cake	Edd	M	24	0	1.0	0	0.0	0	0	0	1	1
3	Series 1	Mary	1	Cake	Jasminder	F	45	0	NaN	0	0.0	0	0	0	1	0
4	Series 1	Mary	1	Cake	Jonathan	M	25	0	9.0	0	0.0	0	0	0	1	0

This dataset comprises 1256 rows of data across 16 features. While the show is moving on to season 13, this dataset only covers seasons 1 through 11. Over 11 seasons, the Great British Bake Off has seen 132 bakers, 68 of whom identify as female and 64 who identify as male. The series has also seen bakers range in age from the youngest of 17 years old, to the oldest of 71 years old. Additionally, Paul Hollywood is known for selectively giving out handshakes after exceptional bakes. He has given out 31 handshakes after signature challenges, and 4 after showstopper challenges.

Unfortunately, the show's first season had a different format - handshakes weren't given out, multiple bakers were eliminated in the first two weeks, and technical scores weren't reported. Therefore season 1 was deleted from the dataset.

Challenges dataset

	series	episode	baker	result	signature	technical	showstopper
0	1	1	Annetha	IN	Light Jamaican Black Cakewith Strawberries and...	2.0	Red, White & Blue Chocolate Cake with Cigarell...
1	1	1	David	IN	Chocolate Orange Cake	3.0	Black Forest Floor Gateaux with Moulded Chocol...
2	1	1	Edd	IN	Caramel Cinnamon and Banana Cake	1.0	NaN
3	1	1	Jasminder	IN	Fresh Mango and Passion Fruit Hummingbird Cake	NaN	NaN
4	1	1	Jonathan	IN	Carrot Cake with Lime and Cream Cheese Icing	9.0	Three Tiered White and Dark Chocolate with Alm...

This dataset, comprised 1136 rows of data across 7 features, containing information on the showstopper, technical and signature challenges for the show from seasons 1-10. The showstopper and signature features are the recipe names that bakers used for these challenges. Season 1 was removed to match the GBBO dataset.

Merged Dataframe
I merged the two datasets on the colored columns listed above. The original GBBO dataset contained a unique row for each baker of each episode. This included rows for the baker even after they were eliminated from the competition. This resulted in over 400 rows of null data in our dataset that were removed. Additionally, I removed

superfluous and/or duplicate columns, and shrank the final dataset to 668 rows and 16 features. Finally, the Challenges dataset didn't have data on season 11. Therefore I removed Season 11 data from the GBBO dataset .

Exploratory Data Analysis

As mentioned early, a baker ultimately wins the competition if they receive star baker on the final week. Therefore it is important to view features of the dataset with respect to star baker. From figure 1 on the right, it is evident that female bakers won star baker 50% more often than male contestants.

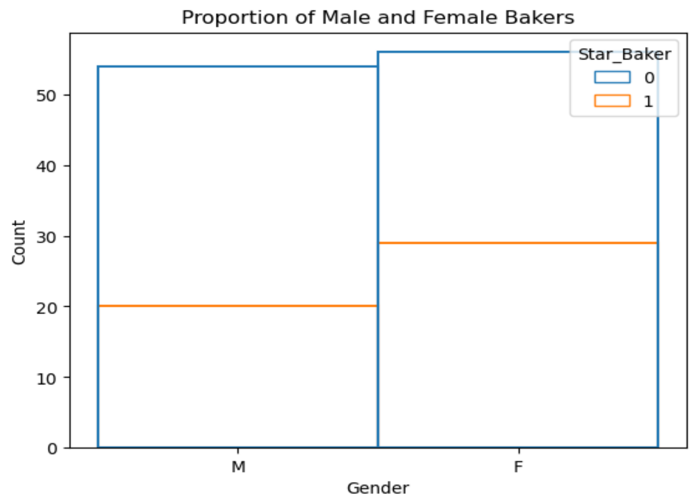

Figure 1; A breakdown of gender counts in the show

Similarly, I reviewed age's relationship with star baker. From Figure 2 below, being of a certain age does not improve the probability of winning star baker. I expected the number of young contestants on the show, yet the number of contestants over the age of 60 was very surprising.

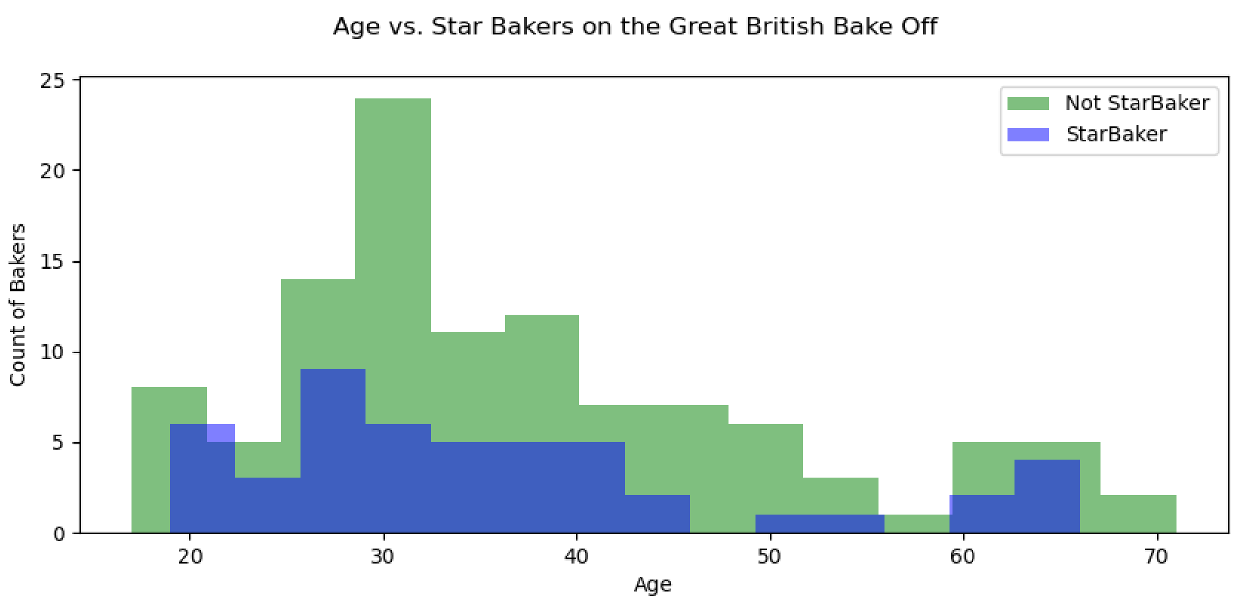

Figure 2; A breakdown of Age on the show

Furthermore, Paul's handshakes have become something of legend on the show. One-on-one interviews consistently show bakers on the verge of tears whenever they receive a handshake. Therefore, I wanted to investigate how handshake counts differ following the signature and showstopper challenges. As shown in Figure 3, whenever a baker receives a handshake following a showstopper or signature challenge, it is very unlikely (it has only happened twice, once being during Finals week) that the baker will get sent home that week. Additionally, Figure 3 reveals that bakers are much more likely to receive a handshake during the signature challenge than the showstopper challenge with only four showstopper handshakes ever being distributed. 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.

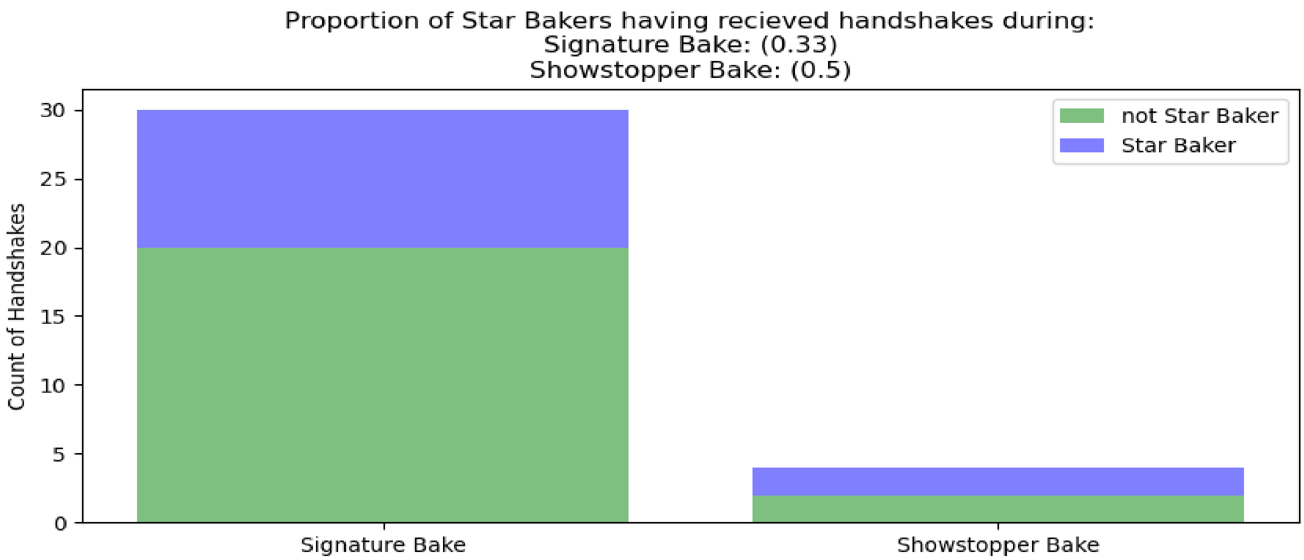

Figure 3; Showing the relationship between being star baker following a handshake

Naturally, I was curious about the discrepancy in the number of handshakes distributed during showstopper challenges. Figure 4 below shows that 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.

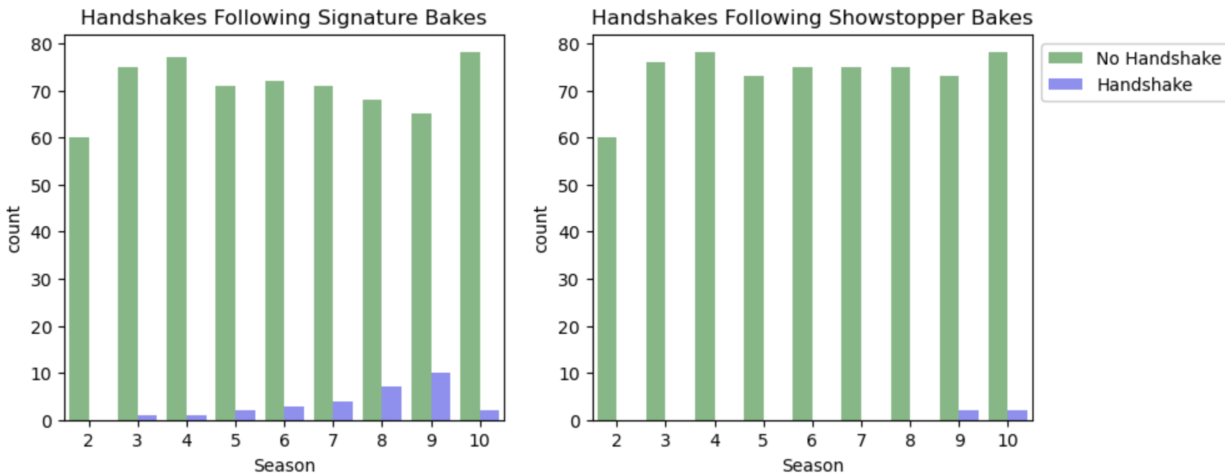

Figure 4; Handshakes following Signature and Showstopper Bakes

Finally, I wanted to see if there was any relationship between the number of star baker awards won per baker per season and winning the show. On the left, I am showing the number of times the winner of each season won star baker. The right image shows the number of bakers grouped by the number of times they were star baker each season. As you can see, in season 5, one contestant (Richard) won star baker five times yet did not win that season! Season 5's winner (Nancy) only accumulated two star bakers. This explains the correlation between star baker and winning the show at only 0.438.

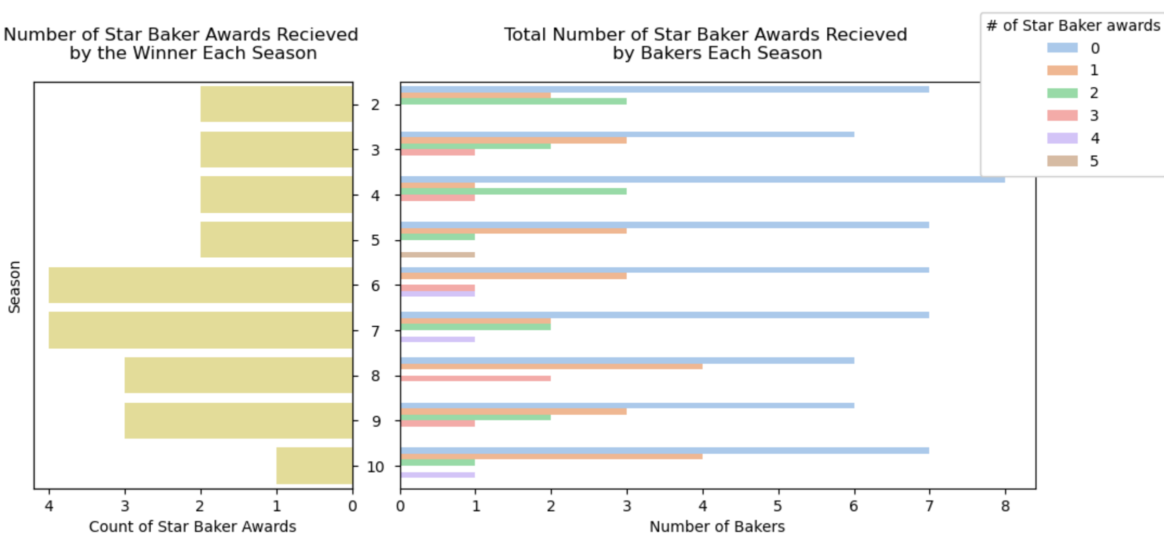

Figure 5; Exploring the number of star baker awards accumulated by bakers per season and the number of star bakers recieved by the shows winner

Advanced Analysis

While this dataset lends itself to many interesting questions, My initial investigation focuses on predicting star baker. The data that we are working with has the following characteristics:

Features	Data Type
Week_Name	category
Gender	category
Age	int64
Signature_Handshake	int64
Technical_Rank	int64
Showstopper_Handshake	int64
Favorite	float64
Least_Favorite	int64
Star_Baker	int64
Eliminated	int64
signature	object
showstopper	object

The sklearn library does require some additional transformations to the data before a model can be generated. Known as feature engineering, I utilized **StandardScaler()** to scale my numerical data, **OneHotEncoder()** to transform the categorical features to 1s and 0s, and finally **TfidVectorizer()** to generate frequency counts of each word of the signature and showstopper recipes. These transformations vastly increase the width of the dataset; going from 11 features to 139. Finally, I needed to deal with our target class imbalance as demonstrated by the pie chart on the left. Only 13% of Star_Baker is labeled as 1 while the remaining amount is 0. Class imbalance can severely hamper a predictive model.

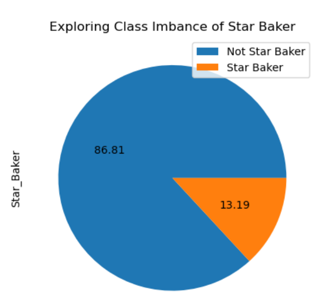

Figure 6; Revealing a high class imbalance within the dataset

Handling Class Imbalance

Several techniques were investigated to deal with class imbalance including, but not limited to, **SMOTE()**, **SMOTETomek()**, and **TomekLinks()**. Using a logistic regression model and the F1 score as our evaluation metric, **SMOTETomek()** increases the performance from 0.45 to 0.52.

Model Creation

I tested four models: random forest, logistic regression, gradient boosting, and KNearestNeighbor (KNN) – both with and without cross validation. Similar to the class resampling technique, I evaluated these models on F1 scores. Additionally, I looked at variance between the training and testing set to see how much my model was over or underfitting the data. Running a confusion matrix on the KNN model, as shown in Figure 8 to the right, shows 180 correctly classified Star_Baker features, with a recall score of 0.81, signaling that our model was not over tuned.

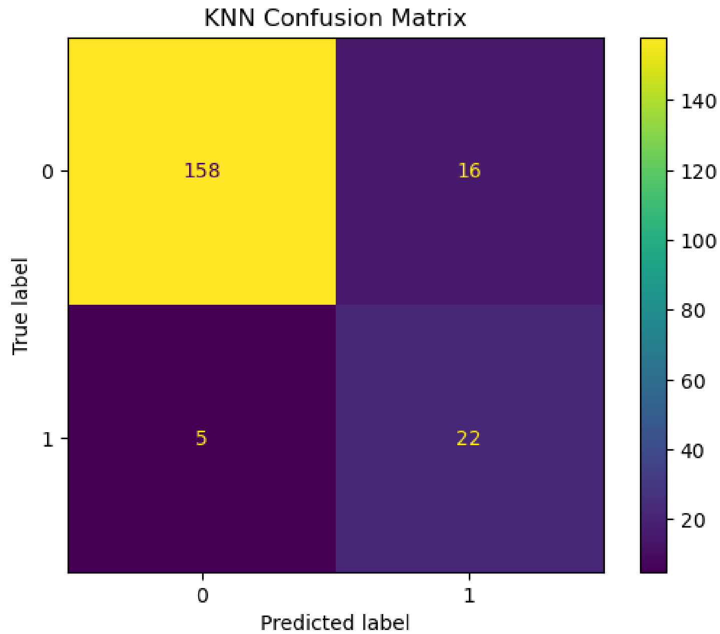

Figure 8; Confusion Matrix for the KNN Model

The results of the model are listed in the table below. Do notice that if I were purely evaluating these models by accuracy then KNN with cross validation would be the winner. Nevertheless, the KNN model with cross validation does have the highest F1 score of 0.68 and the lowest variance of 0.1.

	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

What leads to Star Baker?

The dataset, when fully expanded, has 139 columns. These are not all important for predicting star Baker. I investigated using lasso regression and Recursive Feature

Elimination (RFE) as ways to reduce dimensionality of the dataset. These models both work by eliminating features that are not impactful to the model, resulting in coefficients that can be positive or negative. Positive coefficients are features that will push a model to a class value of 1, or star baker, while negative coefficients lead the model to a 0.

The lasso regression does not require me to specify the number of features to include while RFE does. Therefore, I created an RFE grid search plot to determine this value. Figure 9 below shows F1 scores based on an increasing number of features. I decided on proceeding with 20 features.

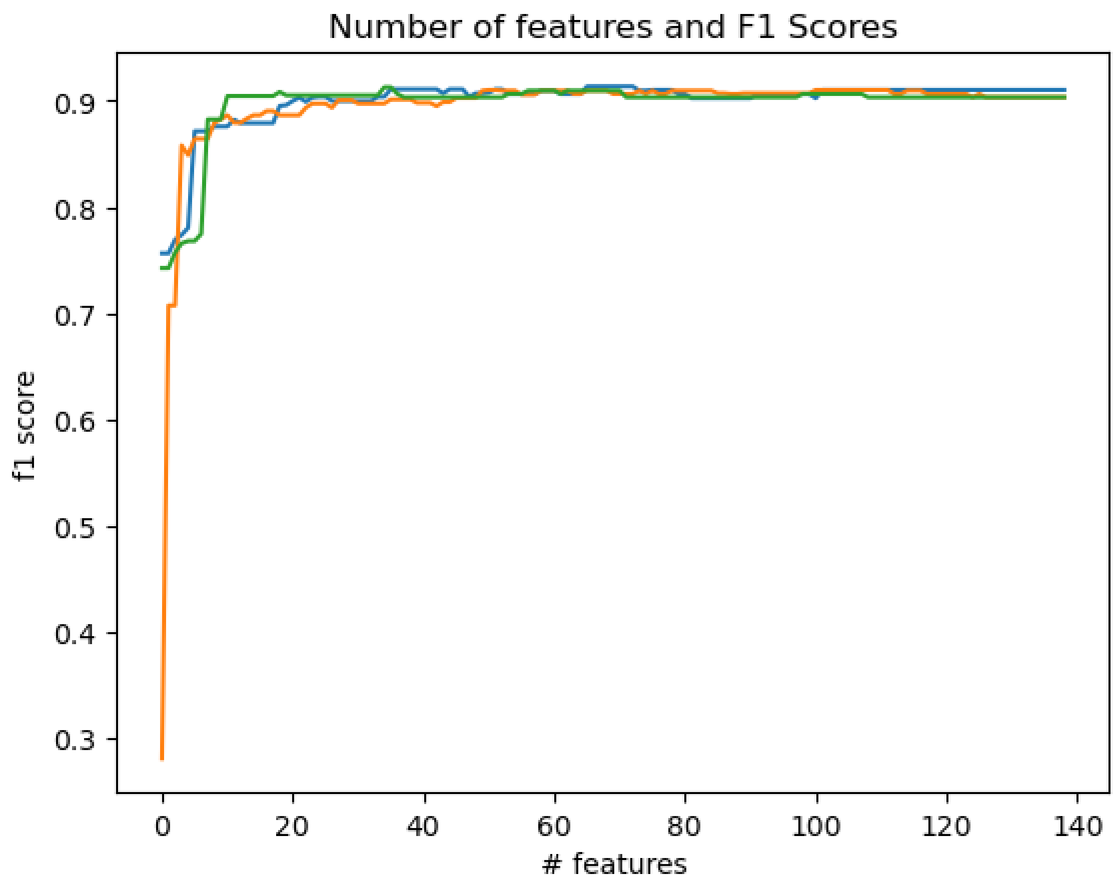

Figure 9; Determining the ideal number of features

The following Table is a print out of the positive coefficients by the two models. The Lasso regression was more liberal in printing out 21 positive values while the RFE only shows 7 positive coefficients. Here is a list of the features that will most likely predict a star baker. Both techniques show the impact of being in the finals on winning star baker. This makes sense since each baker has a ⅓ chance to win. More insightful and potentially useful information are the features with text_sig and text_show tags. These features are individual words from the recipes used during signature and showstopper challenges. The data shows that bakers should use recipes with the word "bread" during the signature challenge, and "pie" during the showstopper challenge. cat_week is a reference to the baking themes each week. This analysis shows that doing well during Tarts week is very important and being labeled a favorite "prior" to the showstopper challenge is a big deal.

	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

While the insights from this exercise are very exciting and impactful, there is much work to be done. According to the rules of the show, each week is independent of one another; a baker's score in the previous week won't help them win future weeks. Therefore, future models will run models for each individual weekly theme. Additionally, it is important to 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.