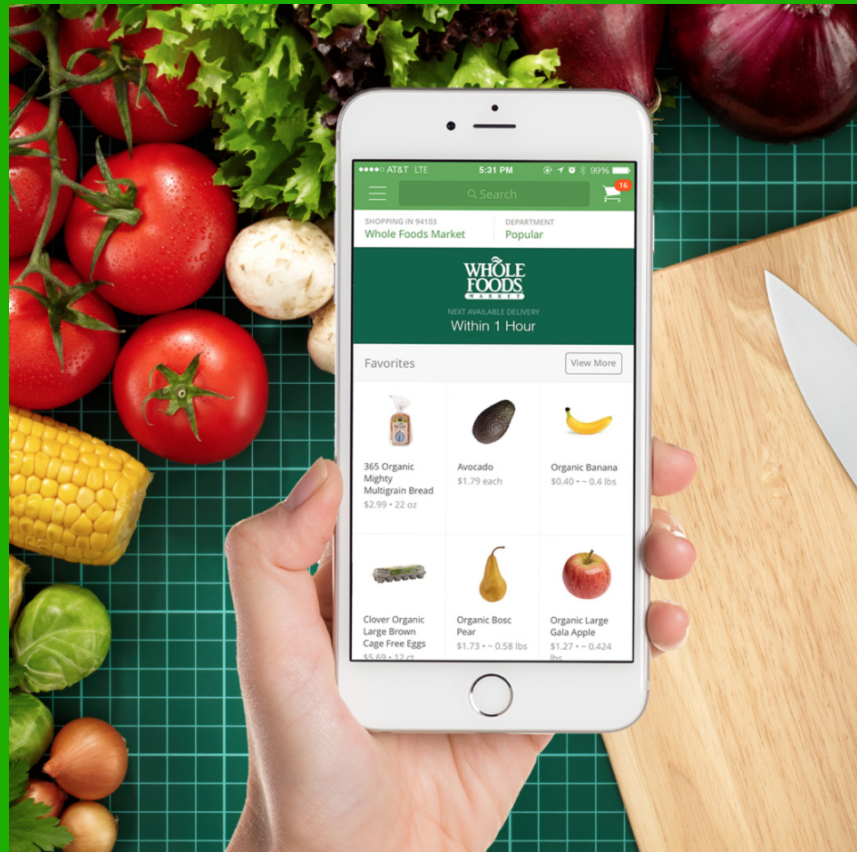
Instacart Open Source Data

On-demand grocery shopping which delivers to your door from a variety of stores

- 3 million orders
- 206,209 users
- 49,468 unique products



Instacart Open Source Data

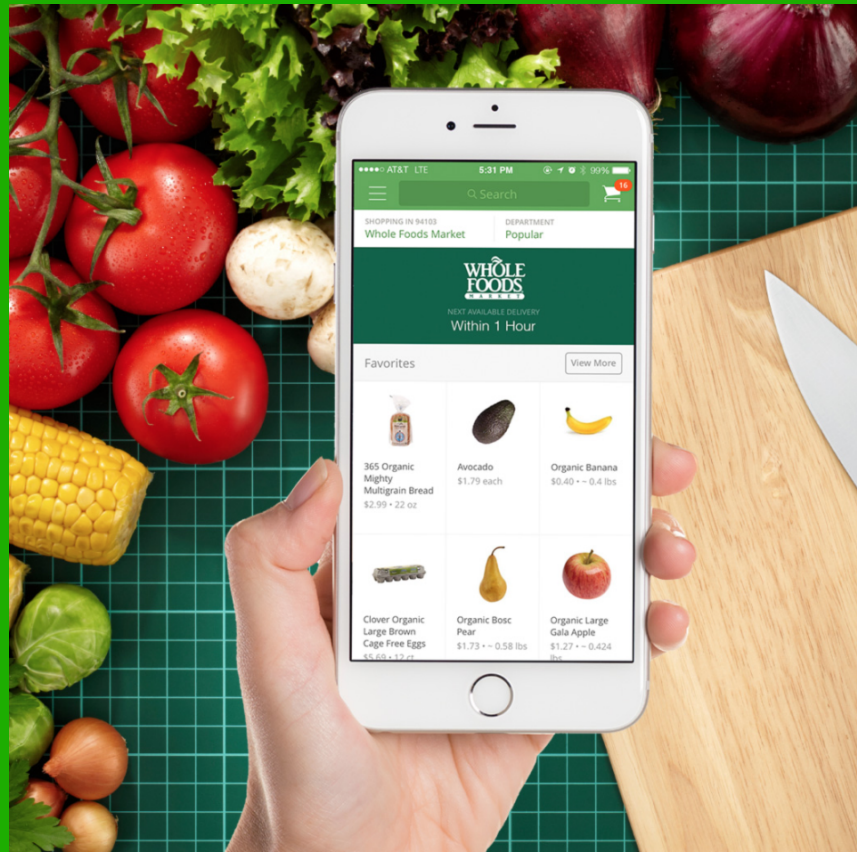
On-demand grocery shopping which delivers to your door from a variety of stores

- 3 million orders
- 206,209 users
- 49,468 unique products

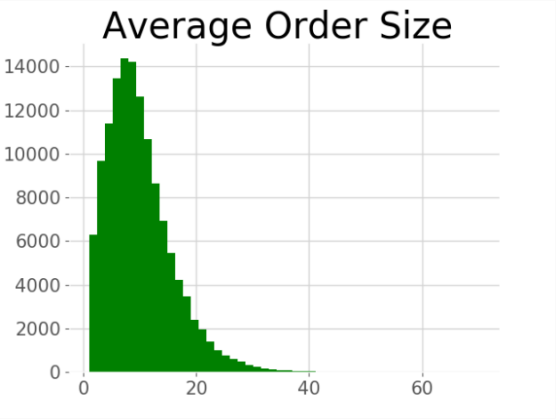
Kaggle Competition

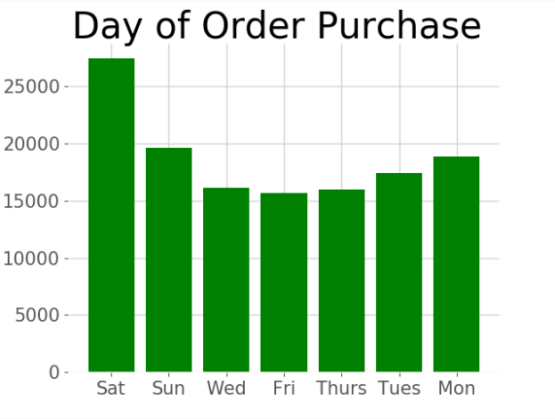
Goal: Predict products that will be purchased again by the user in their next order.

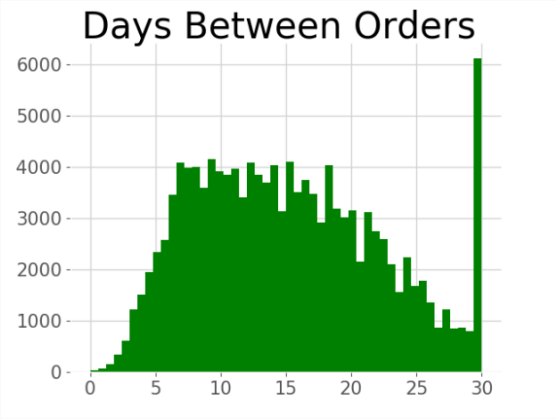
Top Leaderboard score: 40.8%

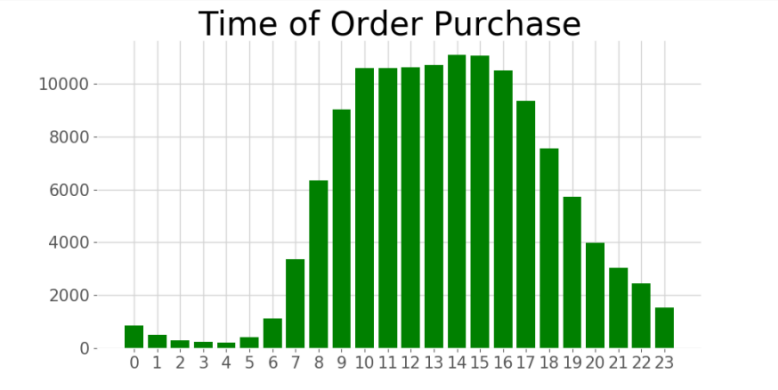


Features









- Product reorder rate for user
- Size of the order with reorders
- Most common days of product order
- Total unique items ordered by user
- Days passed between product purchase
- Product organic vs. non-organic
- Etc ...


COMPARING MODELS
Bayesian GLM
Random Forest
Logistic Regression
Binary Classification
Product reordered: yes or no?

Bayesian GLM

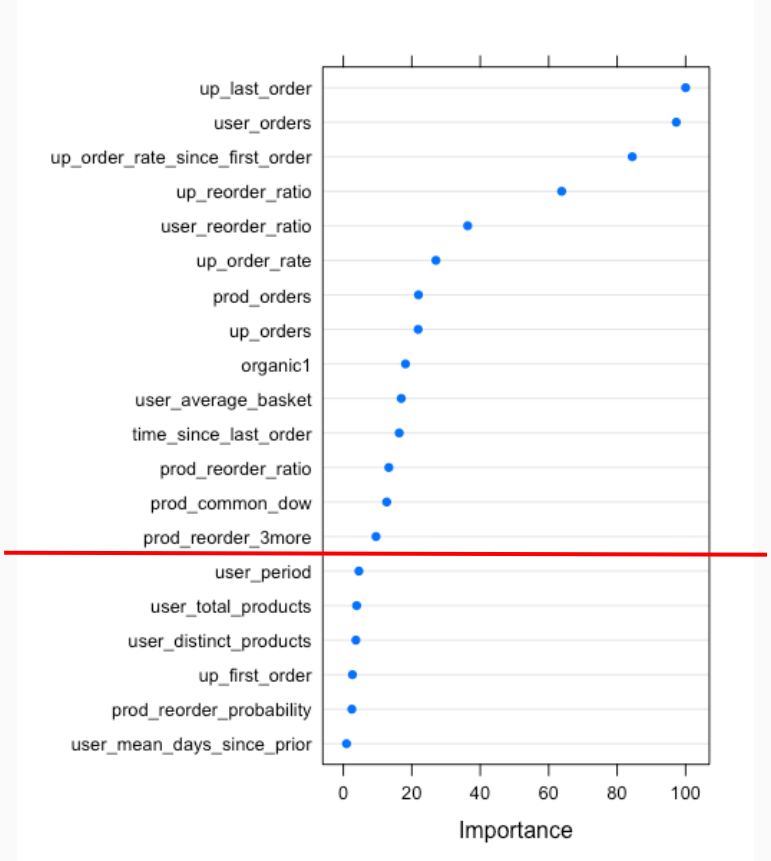
caret::train(method="bayesglm")

$$P(theta | D) = P(D | theta) * P(theta) / P(data)$$

Assumes data is fixed, and parameters are random, presenting the question as what is the probability of the parameters given the data at hand.

Bayesian GLM

caret::train(method="bayesglm")



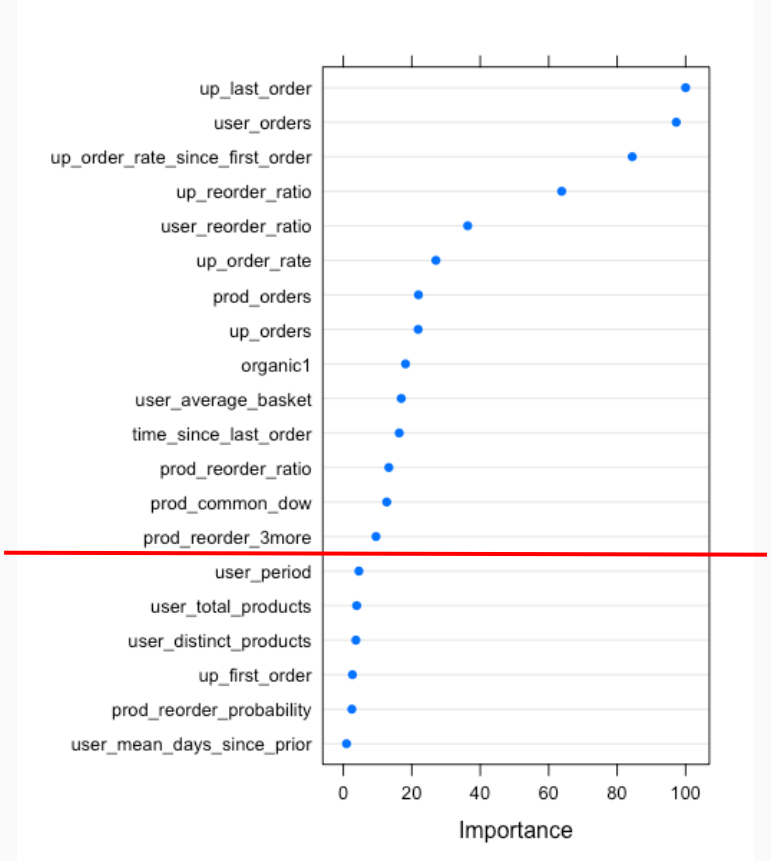
25 features

Reduced to 14 features

Avg F_1 Score: 25.8%

Bayesian GLM

caret::train(method="bayesglm")



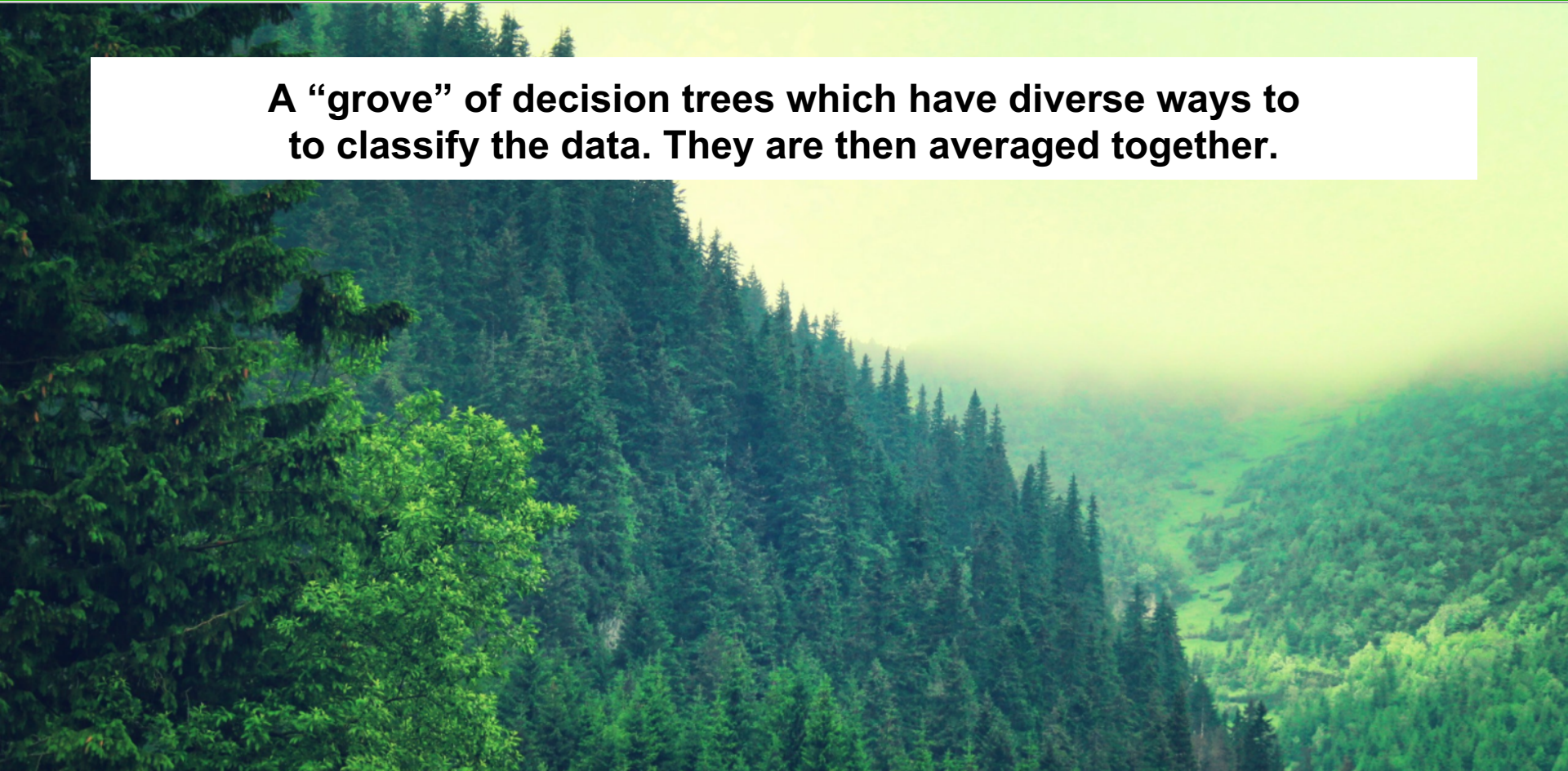
25 features
Reduced to 14 features
Avg F_1 Score: 25.8%

Kaggle Leaderboard Score:

20.9%

Random Forest

scikit-learn: RandomForestClassifier



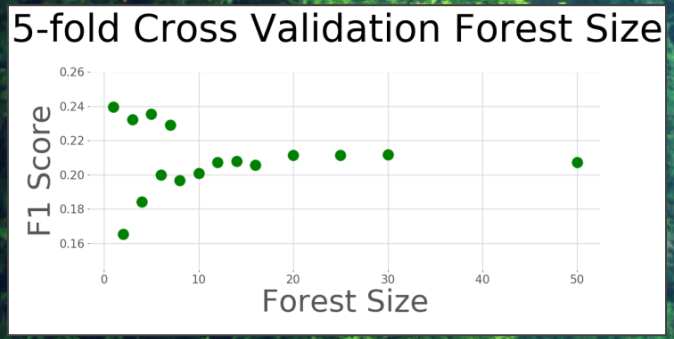
A “grove” of decision trees which have diverse ways to to classify the data. They are then averaged together.

Random Forest

scikit-learn: RandomForestClassifier

A “grove” of decision trees which have diverse ways to to classify the data. They are then averaged together.

16 features
5 trees in the forest
Avg F_1 Score: 20.9%

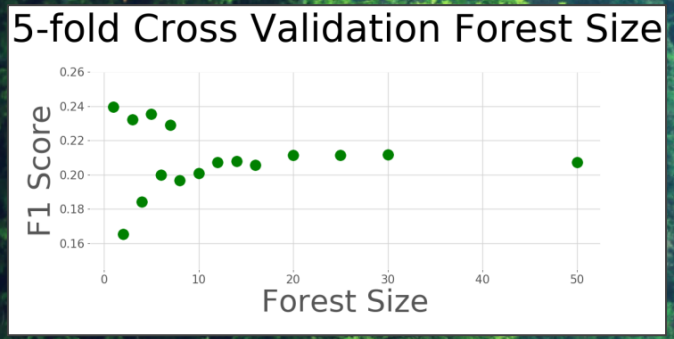


Random Forest

scikit-learn: RandomForestClassifier

A “grove” of decision trees which have diverse ways to to classify the data. They are then averaged together.

16 features
5 trees in the forest
Avg F_1 Score: 20.9%



Kaggle Leaderboard Score:

17.4%

Logistic Regression

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$$

$$\ell(\theta) = \sum_{i=1}^{m} y_i \log(h_\theta(x_i)) + (1 - y_i) \log(1 - h_\theta(x_i))$$

Goal: find the vector θ to maximize $\ell(\theta)$

Five features:

- Fraction of past orders containing product
- Product popularity
- User's reorder rate
- How soon product has been added to the cart in the past
- Average number of days since the previous order

F_1 mean score: 23%

Future Directions

- Improve efficiency to train on more data
- Refine feature space
- Test different algorithms (e.g. FFM)
- Refine parameters for these algorithms