Data Mining Team Project

Andrew Morse
Mark Roberts

Shuting Zang
Paul Whitson

Business Case

Food Processing Industry:

- ❖ $100 B in United States annually
- ❖ Example: conversion of raw fruit into juice, canned fruit, purees, jams, etc.

Globalization of supply chains

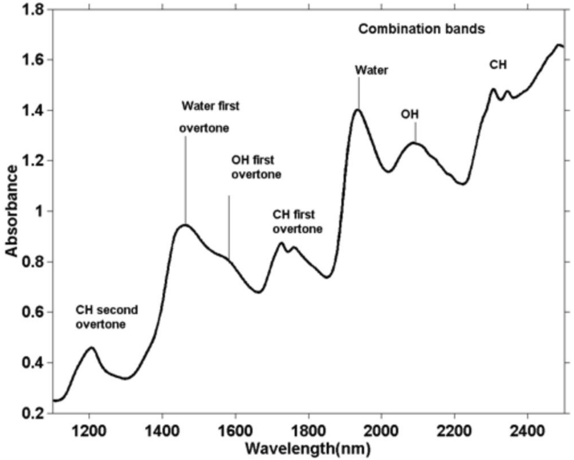
Increasing concerns about food safety

- ❖ Melamine in milk / infant products (China, 2008)
- ❖ Beech-Nut brand "apple juice" found to contain no actual apple juice (U.S., 1988)
- ❖ E. coli and other contamination of meat, vegetables, prepared foods

→ **Need to verify identity and purity of food shipments !**

Analysis Method

- Foods are complex mixtures of sugars, proteins, fiber, vitamins, esters and other flavor compounds
- 983 samples of fruit purees were analyzed
  - 351 pure strawberry;
  - 632 “adulterated” with other fruits and juices
- Analyzed by near-infrared spectroscopy (NIR):
  - Absorbance of sample at each of 235 different wavelengths of light is measured
  - Absorbance at a particular wavelength corresponds to presence of particular chemical bonds (C=C, C-N, etc.)
  - IR has the advantage that it does not require extensive sample preparation
- 1 physical sample -> vector of 235 measurements
- **Goal: use NIR spectroscopy data to distinguish pure strawberry samples from other materials**

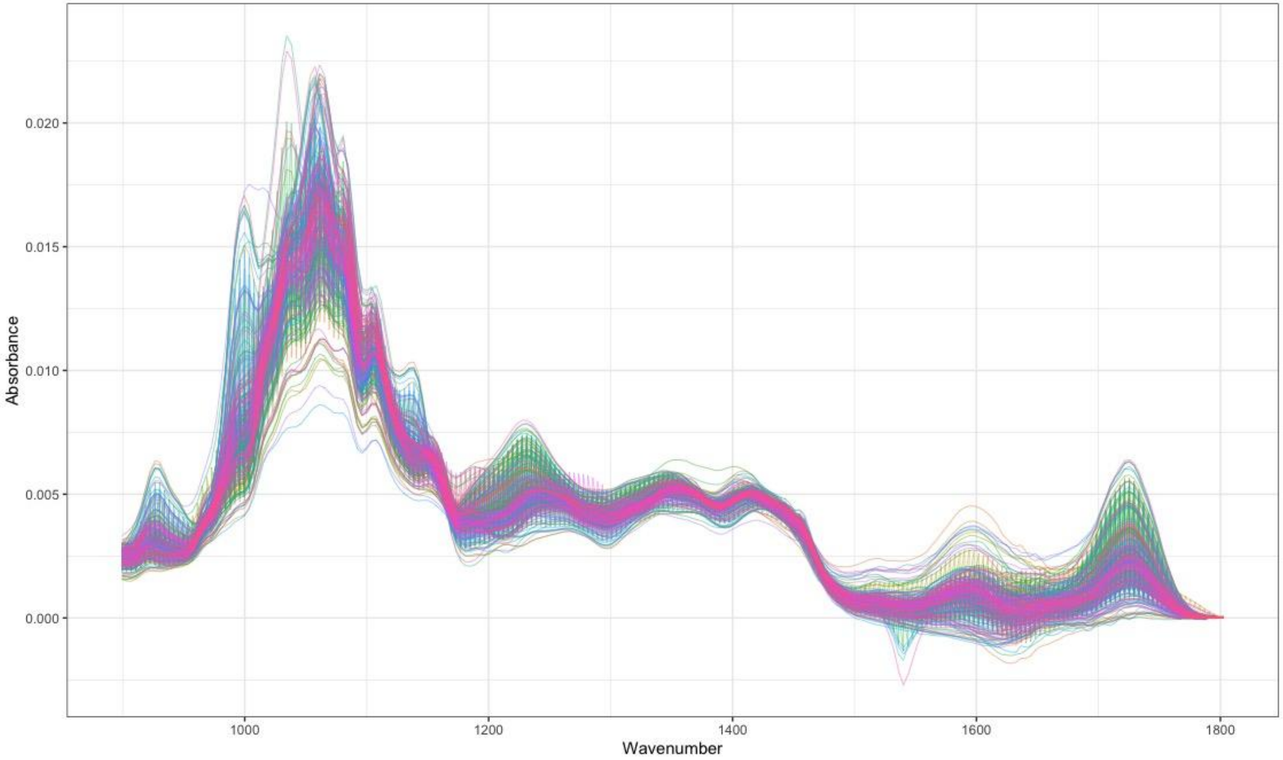


Data source: “Use of Fourier Transform Infrared Spectroscopy and Partial Least Squares Regression for the Detection of Adulteration of Strawberry Purees.” J K Holland, E K Kemsley and R H Wilson, *J Sci Food Agric* 1998, 76, 263E269

Analysis Plan

<u>Data Description</u>

- ❖ 983 observations
- ❖ 235 features (wavelengths) per sample
- ❖ 1 binary response feature
  - ➢ Strawberry
  - ➢ Non-Strawberry



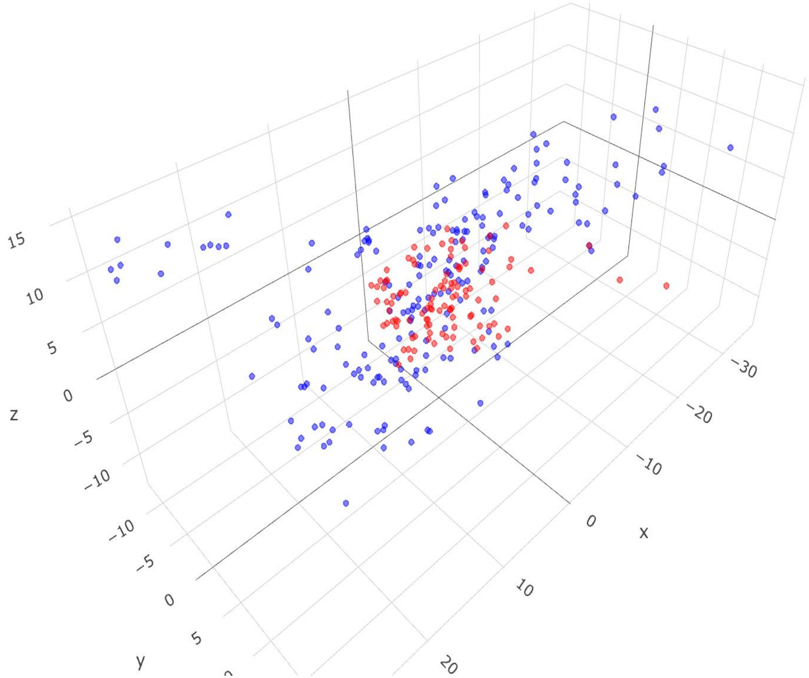
Analysis Plan

Sampling Method

- ❖ Randomly split dataset into Train and Holdout sets
- ❖ Used 70%-30% split
- ❖ Saved Train & Holdout data as individual .csv files to ensure all models were being tested against the same criteria

Goals of Analysis

- ❖ Utilize several different classification modeling techniques and assess accuracy
- ❖ Combine best techniques into an ensemble model
- ❖ Assess strength of ensemble model



Model Fitting

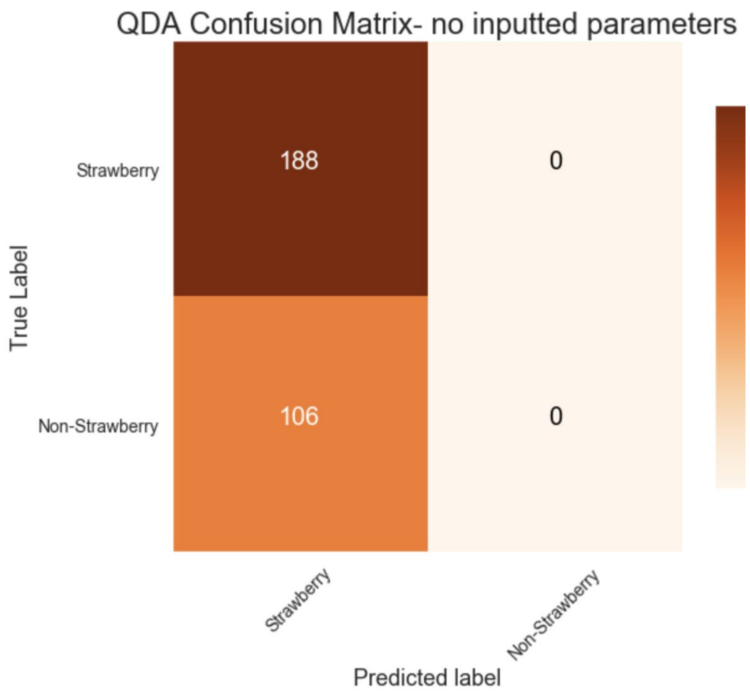
Partial Least Squares (PLS)

- ❖ Similar to Principal Components Analysis (PCA), but PLS is supervised (i.e., includes response variables)
- ❖ PLS creates components that BOTH explain variation in predictors AND maximize the relationship between predictors and response
- ❖ PLS normally uses a numerical response; because the response was binary in this case, this analysis used a logit link function to predict the binary response variable (combination of PLS and GLM: R library plsRglm)
- ❖ Performance improved as more components were used, but beyond 10 components there was some evidence of overfitting

	Training Data			Test Data		
# PLS Components	False Positives	False Negatives	Total Accuracy	False Positives	False Negatives	Total Accuracy
3	8.60%	8.30%	92%	11.30%	8.00%	90%
10	1.60%	1.35%	99%	4.70%	2.10%	97%
20	0%	0%	100%	0.50%	1.60%	99%

Quadratic Discriminant Analysis

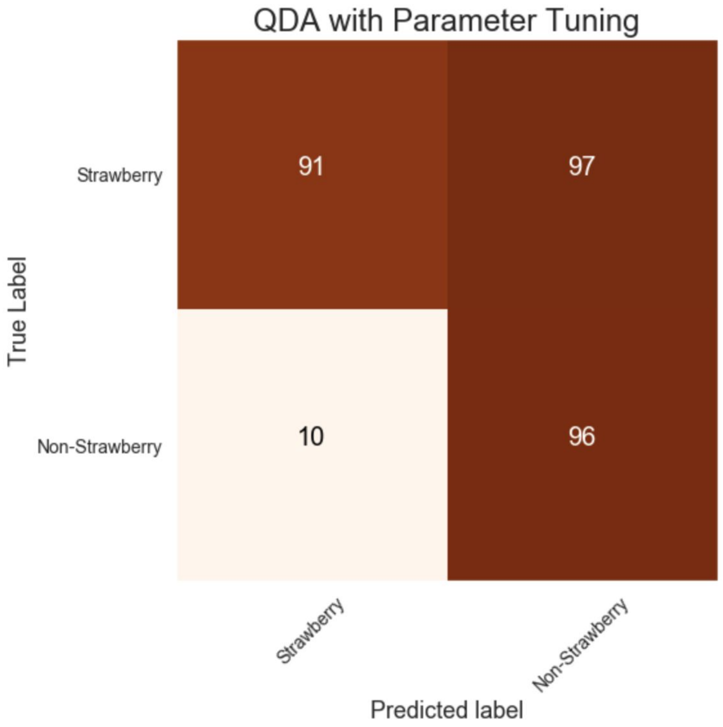
- Assumes different covariance matrix for each group (2)
- Accuracy score: **.639**
- Predicted every observation as Strawberry
  - Due to overfitting on training data
  - Tuning required



QDA with Parameter Tuning

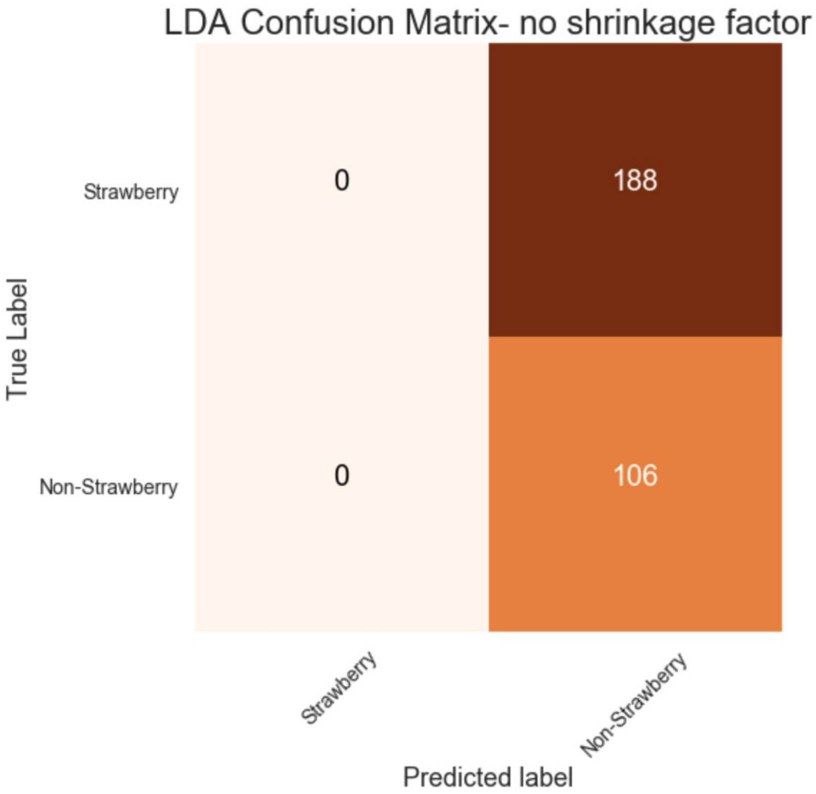
- ❖ Shrinkage parameter introduced to correctly adjust for overfitted data
  - ➢ Cross validated in test data
  - ➢ Picked optimal level (.95)

- ❖ Accuracy score: **0.636**



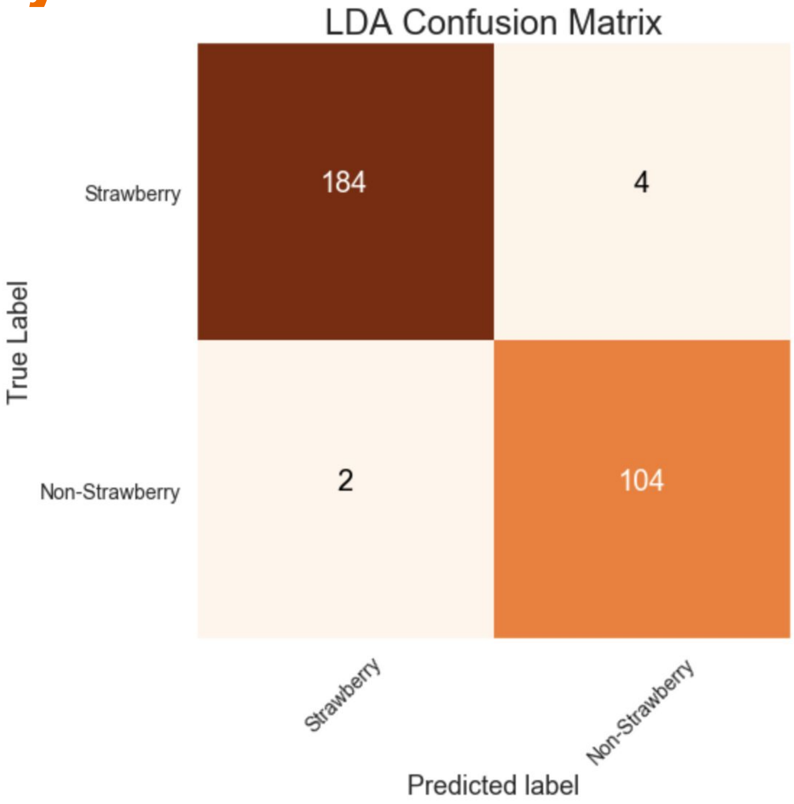
Linear Discriminant Analysis LDA

- ❖ Defines linear combination of features to classify a response variable
- ❖ Assumes comprehensive covariance matrix between all groups
- ❖ First attempt using **no "shrinkage" factor: accuracy score =.36**
  - ➢ Incorrectly predicts all observations to be non-strawberry



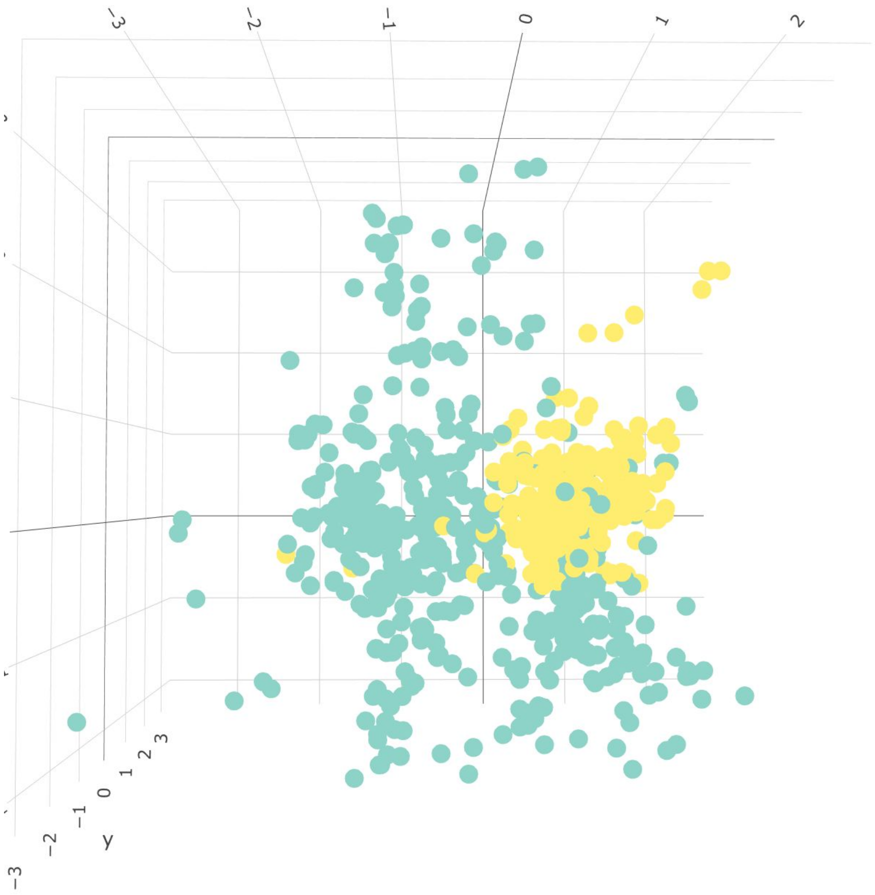
Linear Discriminant Analysis

- ❖ Tune parameter: include shrinkage factor to account for overfitting on training data
- ❖ Updated accuracy score **0.979**
  - ➢ Sensitivity: **.989**
  - ➢ Specificity: **.962**



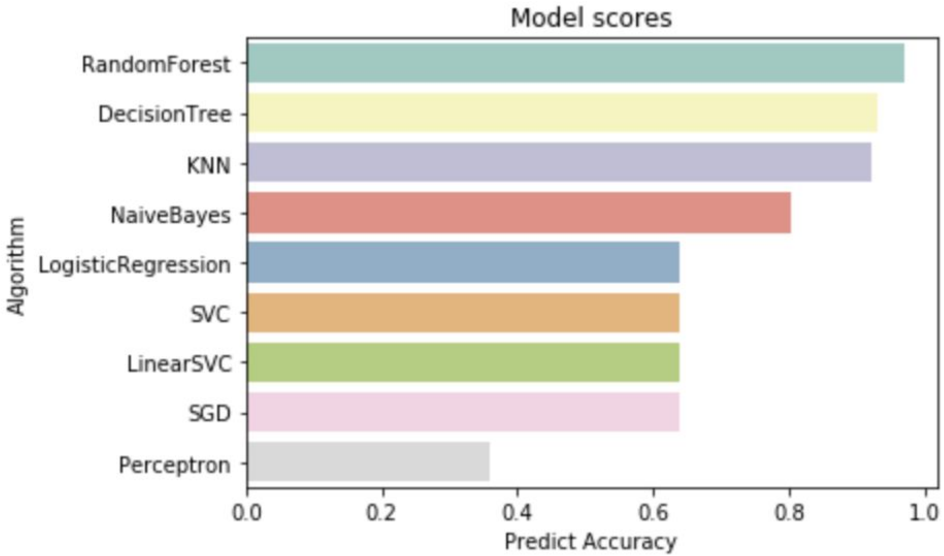
LDA VS QDA

- Data appears to have linear splits between groups
  - LDA creates linear separations
  - QDA creates quadratic separations
  - QDA may have introduced too much flexibility
- Therefore, better fit with LDA makes sense



Models and Visualization

As we have “strawberry”, which is binary type, as the predictor, we first tried nine commonly used models and check their accuracy.



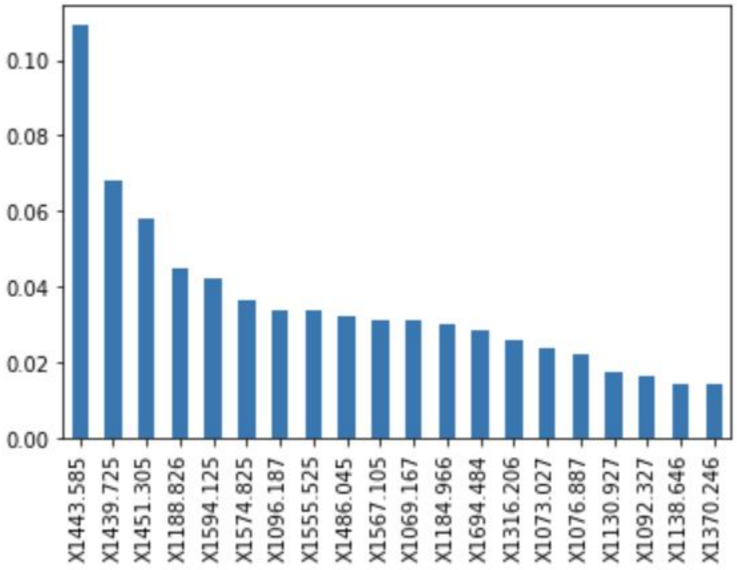
We use accuracy scores to compare models performance.

$$\boldsymbol{Accurancy} = \frac{\boldsymbol{True\ Positive} + \boldsymbol{True\ Negative}}{\boldsymbol{Total}}$$

Among these nine models, Random Forest gives the best performance.

Feature Importance from Random Forest

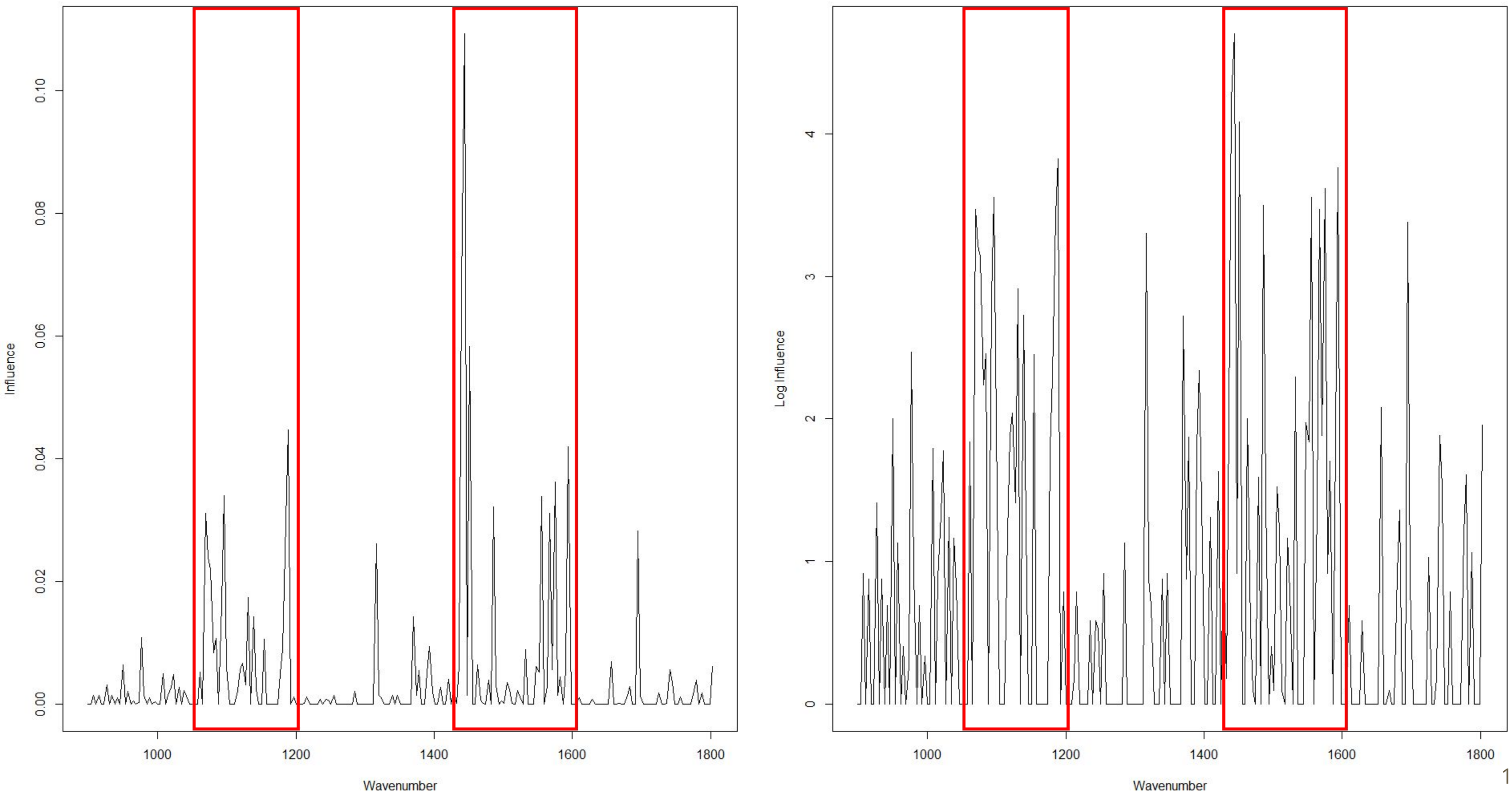
As Random Forest yields to the best performance among nine models we previously ran, we select the top 20 importance features for further analysis



	Feature_Importance
X1443.585	0.1093
X1439.725	0.0679
X1451.305	0.0583
X1188.826	0.0447
X1594.125	0.0420
X1574.825	0.0362
X1096.187	0.0340
X1555.525	0.0339
X1486.045	0.0321
X1567.105	0.0311

Here, we find that features as X1443.585, X1439.725, and X1452.305 are important for doing correct prediction of strawberry group.

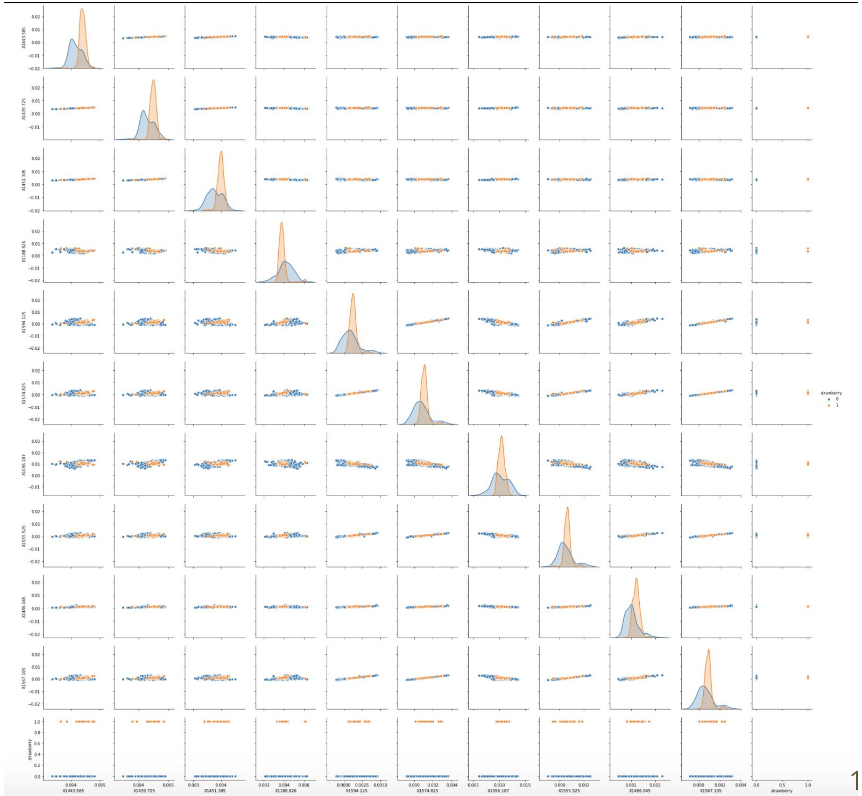
We could suggest the company to do spectrum analysis based on these features but not all of them to reduce cost.



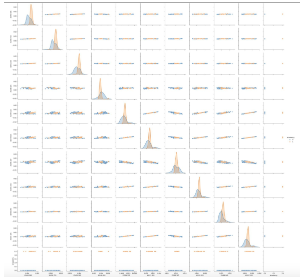
Influence
0.00
0.02
0.04
0.06
0.08
0.10
Wavenumber
1000
1200
1400
1600
1800
Log Influence
0
1
2
3
4
Wavenumber
1000
1200
1400
1600
1800

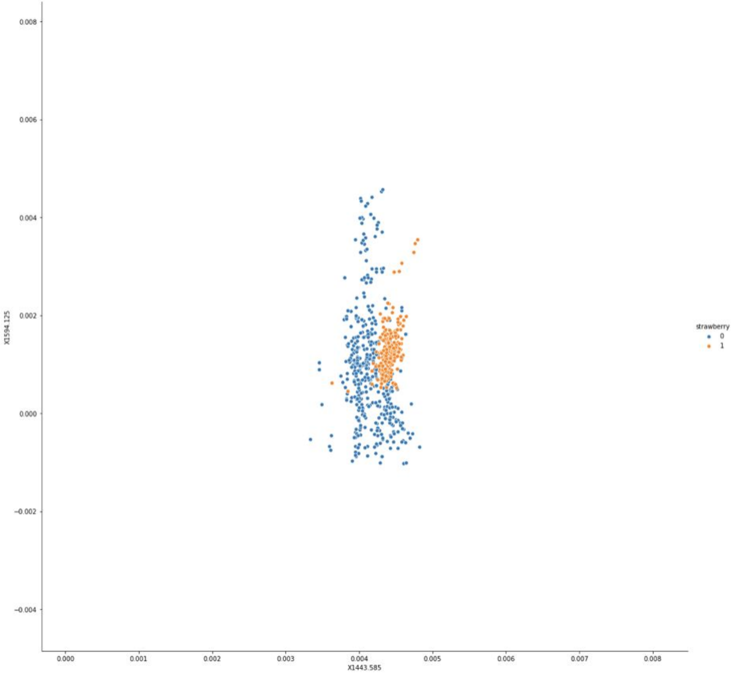
Pairplot for Top10 Important Features

- ❖ Check differences for whether it is strawberry (colored, blue is yes and orange is no) among important features
  - ➢ Some show significant differences
- ❖ Diagonal line lists distributions for each variables/features
  - ➢ Group colored by blue spreads wider than orange group
  - ➢ Orange group has higher peaks



Pairplot for Top10 Important Features





Select one example with most important features as x-axis and y-axis from previous plot:

X-axis: X1443.585

Y-axis: X1594.125

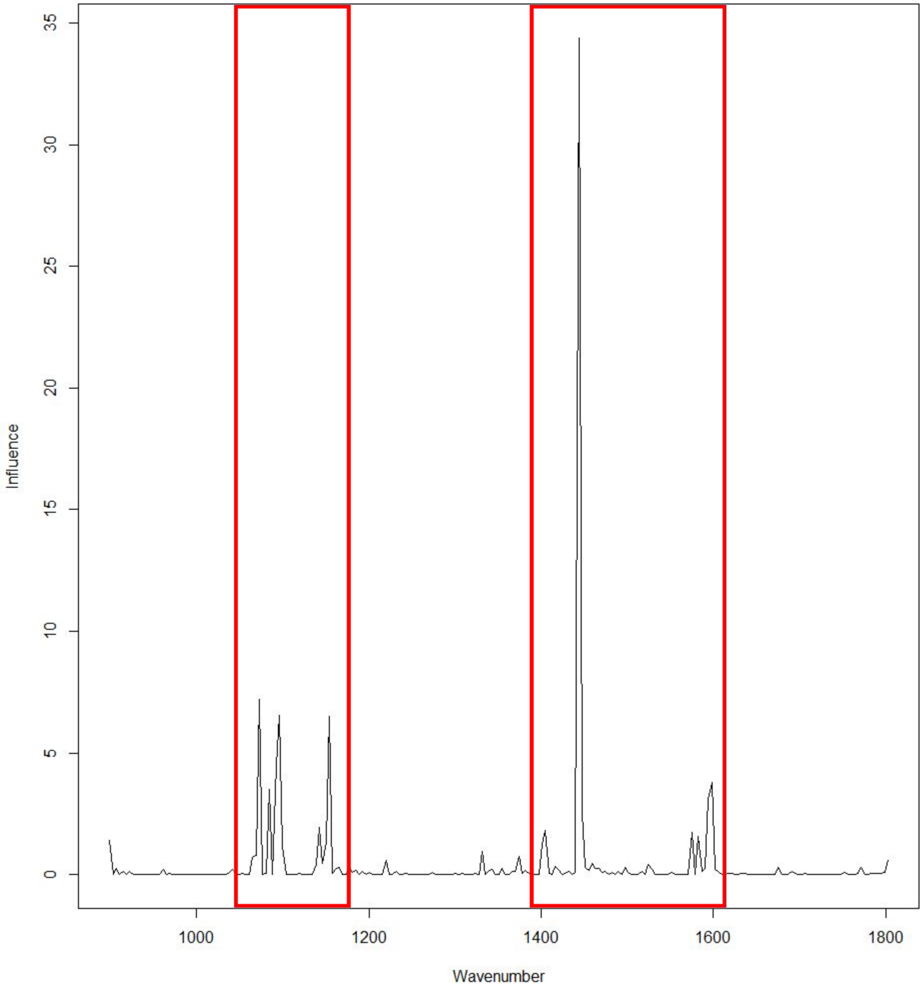
Colors:

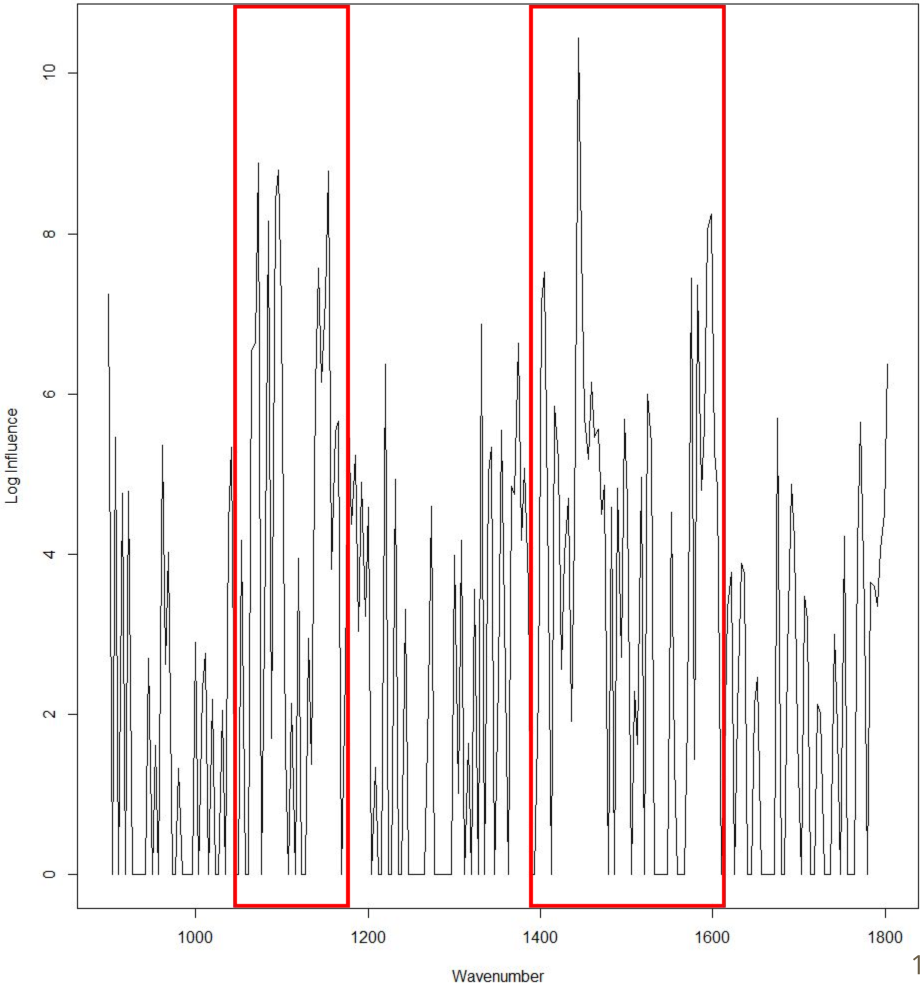
Blue is 0, not strawberry

Orange is 1, strawberry

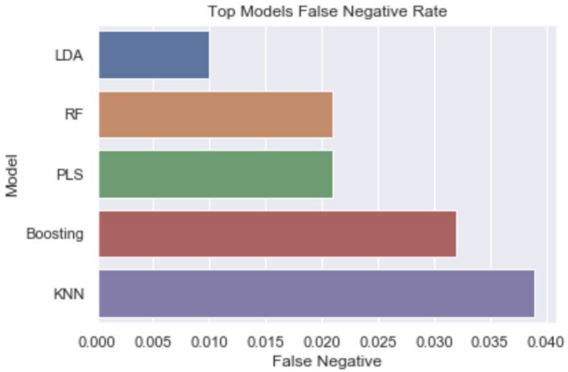
Boosting: Gradient Boosting Machines

- Gradient boosting using `gbm()` in R, utilizing Friedman algorithm
- GBM repeatedly models on residuals, approximating gradient descent process
- Model features selected based on feature importance after running GBM with all predictors
- Results:
  - With PCA (11 features): 96.6% accuracy, 3.8% false positive rate, 3.2% false negative rate
  - Without PCA (21 features): 95.6% accuracy , 5.7% false positive rate, 3.7% false negative rate

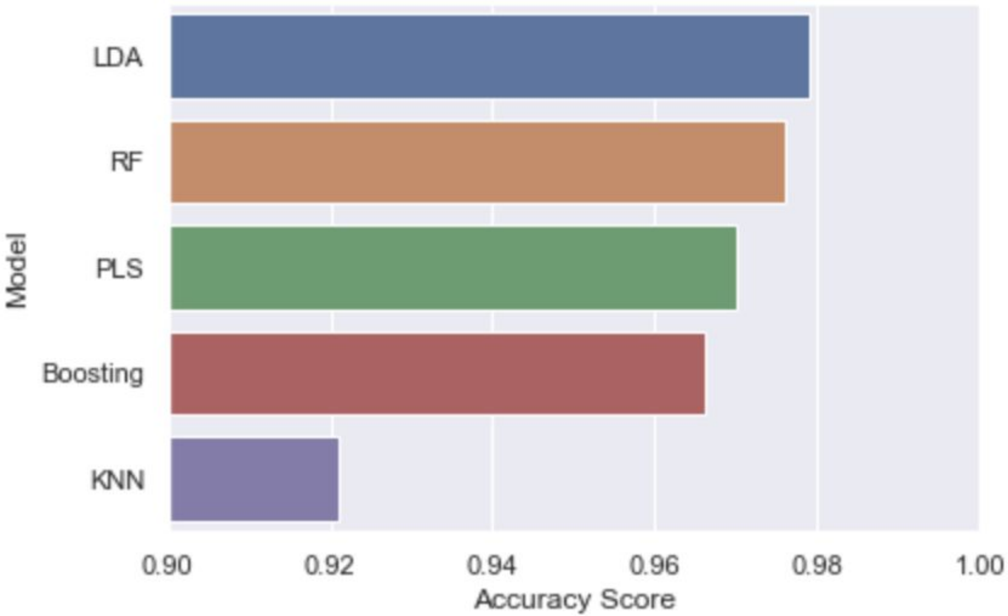

Influence
Wavenumber


Log Influence
Wavenumber

Summary of Best Models

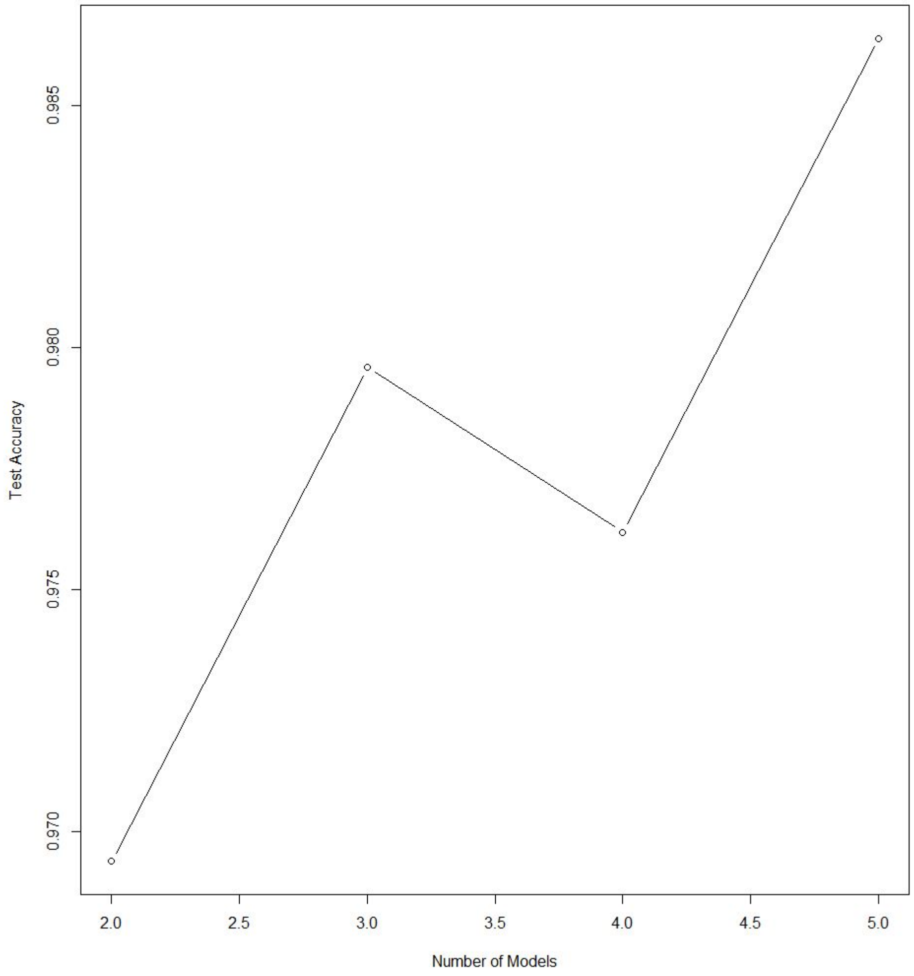


Accuracy Score



Ensemble Model

- ❖ Taking majority of top 5 model predictions yields superior results:
  - ➢ Accuracy: 98.6%
  - ➢ False negative rate: 1.1%
  - ➢ False positive rate: 1.9%
- ❖ Prediction accuracy was cross-validated to choose the best model size



Conclusion

- Using this model, we could predict whether a sample was adulterated or not with 98.6% overall accuracy
- The model correctly identified 99% of samples that were adulterated
- This ensemble model could be deployed in order to ensure product quality
- Suggest to use only part of the spectrum (1050-1200 nm^{-1}, 1400-1600 nm^{-1}) in test to reduce cost

Considerations and Future Suggestions

- May be possible to reduce cost by testing only key features / wavelengths
- Future: replace binary target variable with more detail
  - Regression: % purity of strawberry sample (100% pure, 95% pure, 90%, etc.)
  - And / or classification: distinguish different fruit types (strawberry vs. raspberry, etc.)
- Evaluate application to other agriculture industries (grain, dairy, etc.) and to purity of pharmaceutical or chemical products

Thank You!