MULTICLASS CLASSIFICATION AND PRICE PREDICTION OF DRY BEANS

Final Project (Group 3)

STAT 602, Spring 2022

Tahmid Alam

Kyle Downs

Md Mominul Islam

Anu Sangotayo

https://food.unl.edu/article/how-cook-dry-beans-scratch

PROJECT OUTLINE

- Background Study
- Project Objective & Selection Criteria
- Data Set & Introduction
- Data Exploration
- Feature Selection
- Statistical Model Development
- Cultivation Data Exploration
- Statistical Model Accuracy Summary
- Project Conclusion
- References

BACKGROUND STUDY

Dry bean cultivation using mixed species of seeds is practiced in Turkey and Asian countries

Final products contain different species of seeds as populations have different genotypes

When the cultivations are released to the market without being separated by species, the market value decreases immensely [2]

Because of the wide range of genetic diversity of dry bean, seed classification is essential for marketing and production

To establish a sustainable agricultural system, Koklu et al. [1] provided method for obtaining uniform seed varieties from crop production

Separating the best seed variety from the mixed dry bean population is important to maintain the market value, otherwise, depreciation occurs

[1] Computers and Electronics in Agriculture 174 (2020): 105507.
[2] Journal of Agriculture and Food Sciences 26.1 (2012): 15-26.

PROJECT OBEJCTIVE

- Selecting the best seed species is one of the main concerns for both bean producers and the market.
- Primary goal is to develop an automated method that predicts the value of a harvest from a 'population cultivation' taken from a single farm that has been presented at the market.

SELECTION CRITERIA

- First, we will compare different classification algorithms and recommend the best one to predict prices from unknown samples.
- The algorithm selection will be based on the algorithm that produces the highest accuracy in classifying the bean type.
- The algorithm with the highest classification accuracy will serve to minimize the market value cost of classification errors between dry bean varieties.

DATA SET AND INTRODUCTION

- There are two datasets used in this project; The **'labeled' and 'unlabeled' datasets**.
- The labeled dataset contains the classes of dry beans, which will be used to train the various machine learning (ML) models, hence also referred to as the training dataset.
- The training dataset contains 3000 observations and 8 variables. It also has 6 levels (Classes): BOMBAY, CALI, DERMASON, HOROZ, SEKER, and SIRA - These classes represent the different varieties of dry beans. Each class has 500 observations.
- The unlabeled dataset is made up of the combination of three separate samples, namely, Sample A, B, and C. There are 7 variables, not including the class variable.

DESCRIPTIVE STATISTICS OF DRY BEANS

Statistical Distribution of Features of Dry Beans

	Count	Mean	Standard Deviation	Median	Minimum	Maximum	Range
Area	3000	69874.98	49578.52	48714.50	20645.00	251320.00	230675.00
Perimeter	3000	1012.24	347.75	941.90	384.17	2164.10	1779.93
MajorAxisLength	3000	362.05	124.52	332.90	161.52	740.97	579.45
MinorAxisLength	3000	225.19	73.35	202.73	106.00	473.39	367.39
Eccentricity	3000	0.76	0.10	0.77	0.30	0.94	0.64
ConvexArea	3000	70944.12	50382.27	50807.50	8912.00	259965.00	251053.00
Extent	3000	0.75	0.05	0.77	0.57	0.85	0.28

- The variables, Area and Convex Area, had the largest range for all four datasets.
- There are large differences in the range of variables, the variables with larger ranges can dominate over those with small ranges which may lead to biased results.

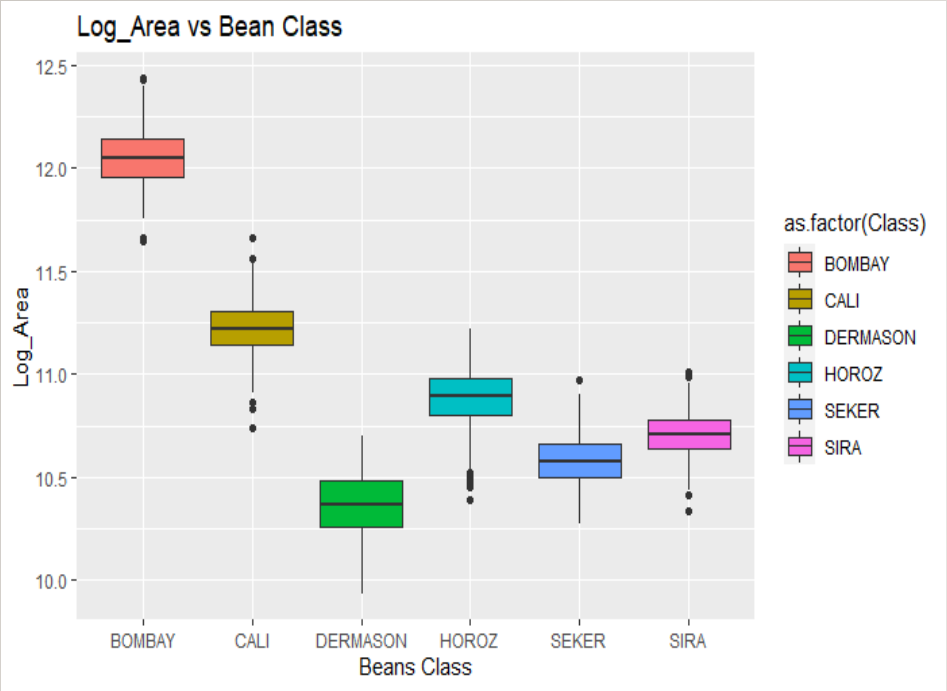
DESCRIPTIVE STATISTICS OF DRY BEANS

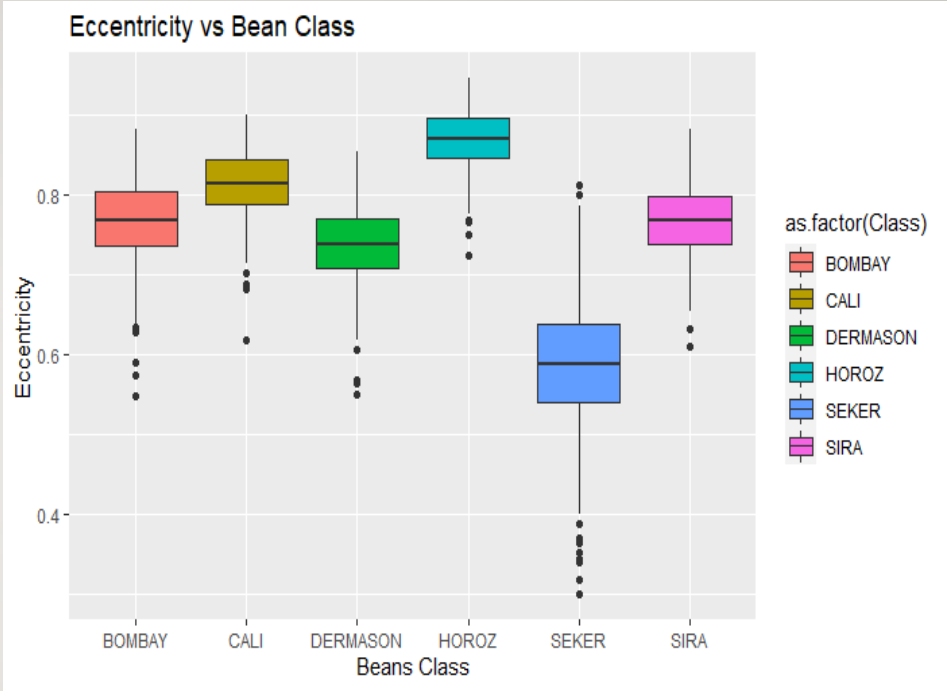
Statistical Distribution of Features of Dry Beans

	Count	Mean	Standard Deviation	Median	Minimum	Maximum	Range
Area	3000	69874.98	49578.52	48714.50	20645.00	251320.00	230675.00
Perimeter	3000	1012.24	347.75	941.90	384.17	2164.10	1779.93
MajorAxisLength	3000	362.05	124.52	332.90	161.52	740.97	579.45
MinorAxisLength	3000	225.19	73.35	202.73	106.00	473.39	367.39
Eccentricity	3000	0.76	0.10	0.77	0.30	0.94	0.64
ConvexArea	3000	70944.12	50382.27	50807.50	8912.00	259965.00	251053.00
Extent	3000	0.75	0.05	0.77	0.57	0.85	0.28

- Variables with larger ranges may benefit through feature engineering.
- In these results, the mean area for the dry beans is 69,875, and the median area is 48,714.5. The data appears to be skewed to the right, which explains why the mean is greater than the median.

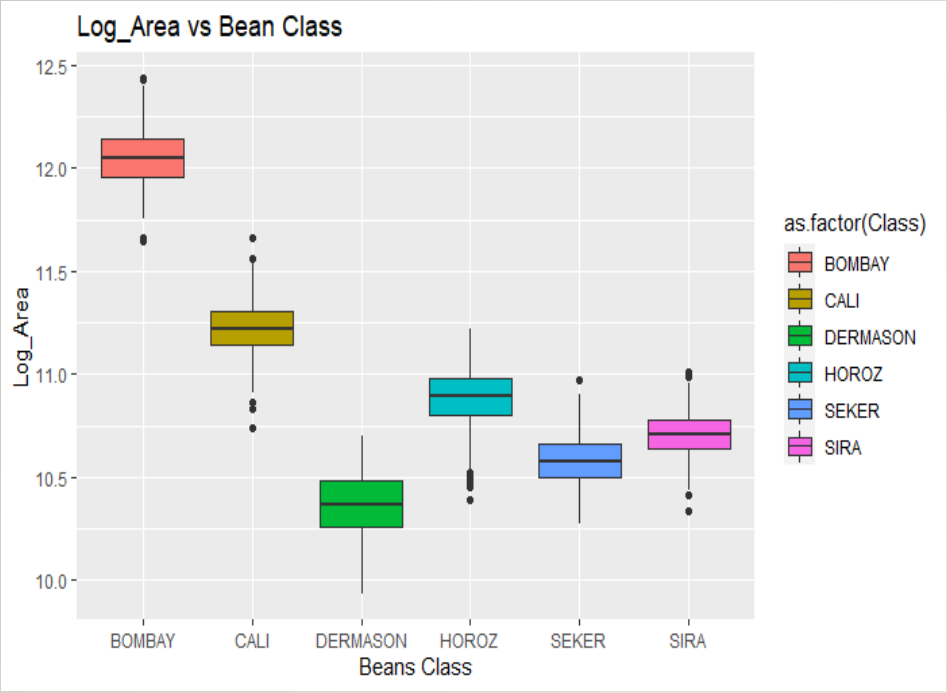
DATA EXPLORATION (BOXPLOTS)

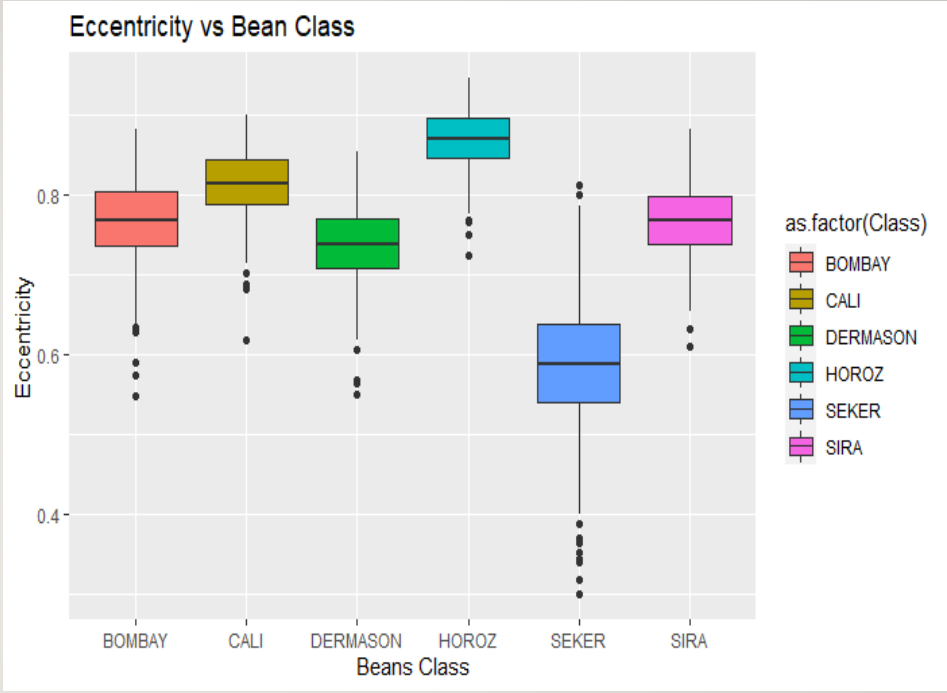




- A boxplot is a standardized way of displaying the distribution of data based on a five number summary ("minimum", first quartile (Q1), median, third quartile (Q3), and "maximum").
- Plotted variables 'Log_Area' and 'Eccentricity' with respect to each class.
- The range is the highest for the 'BOMBAY' class for 'Log_Area' and 'SEKER' for 'Eccentricity'.

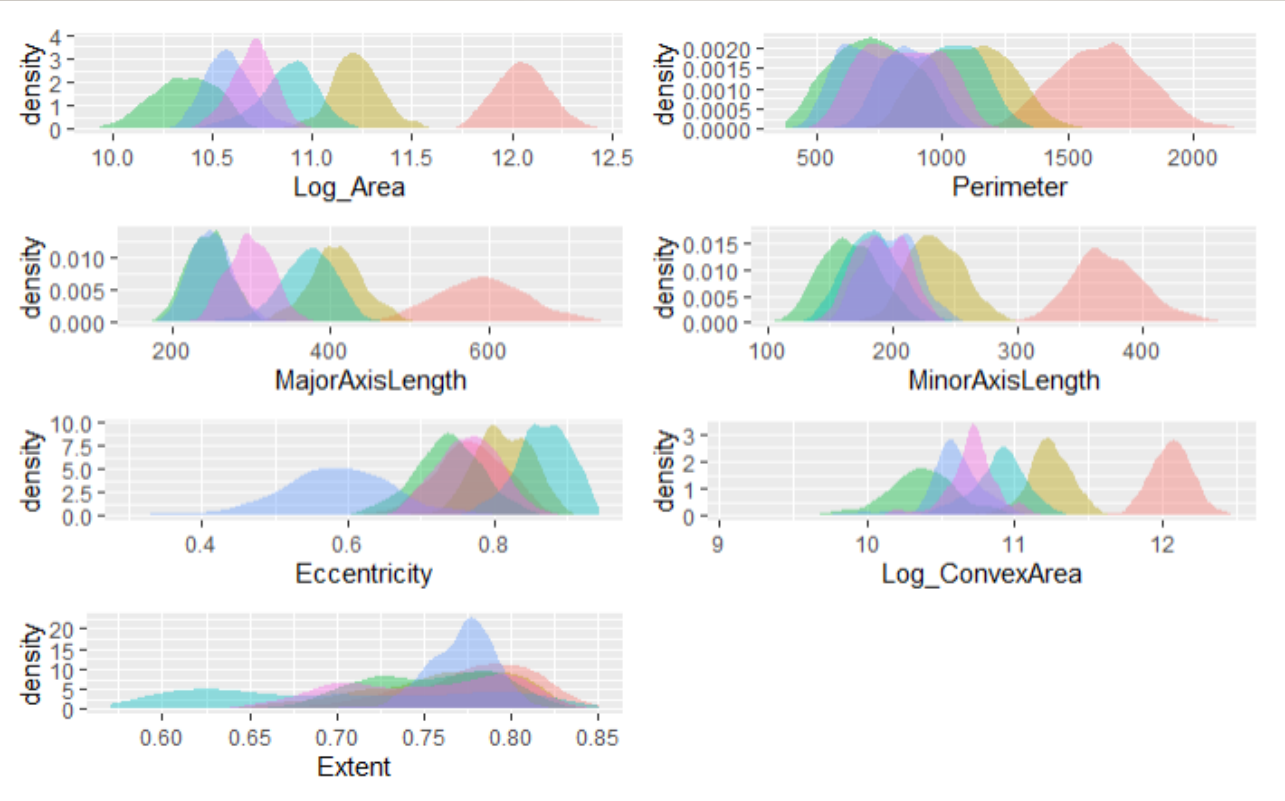
DATA EXPLORATION (BOXPLOTS)





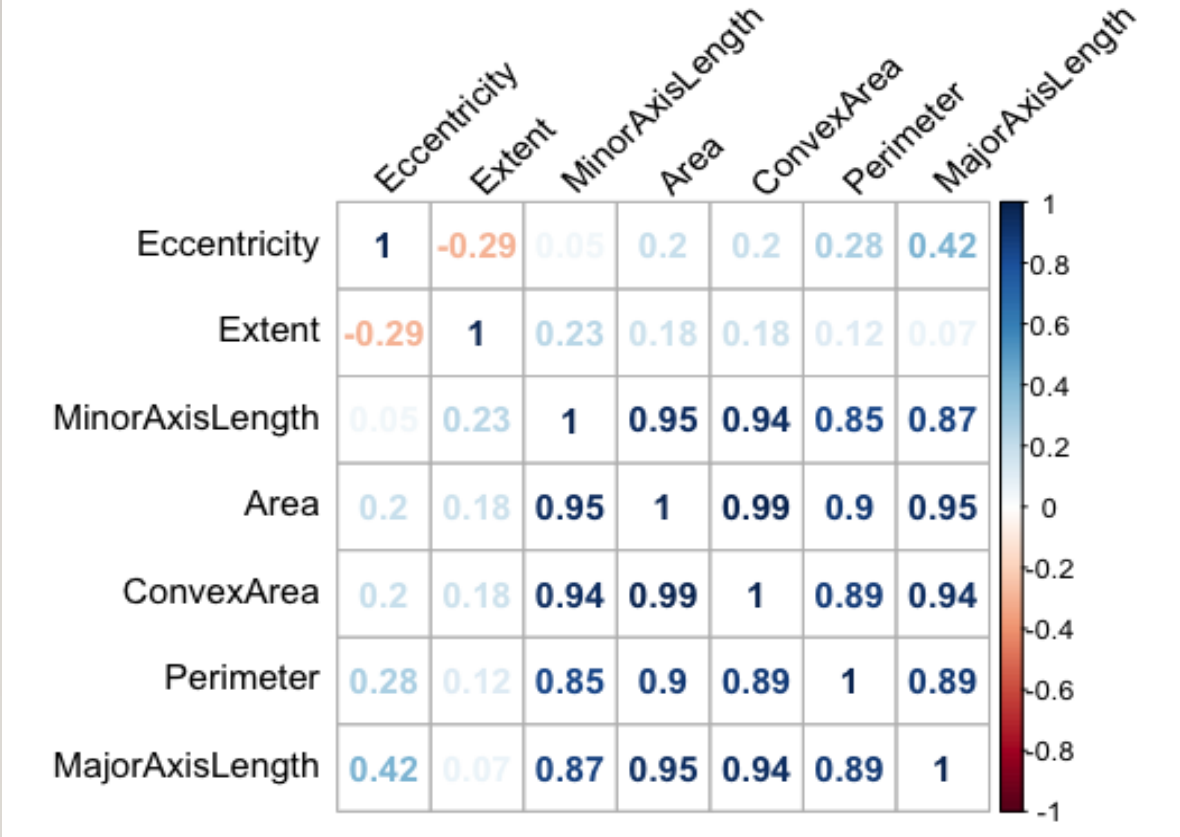
- All of the classes have similar variability except for these two class.
- These inference is justified when we built our model.
- The median value of the class ‘DERMASON’ is lowest for ‘Log_Area’ while class ‘SEKER’ is lowest for ‘Eccentricity’.

DATA EXPLORATION (DENSITY PLOTS)



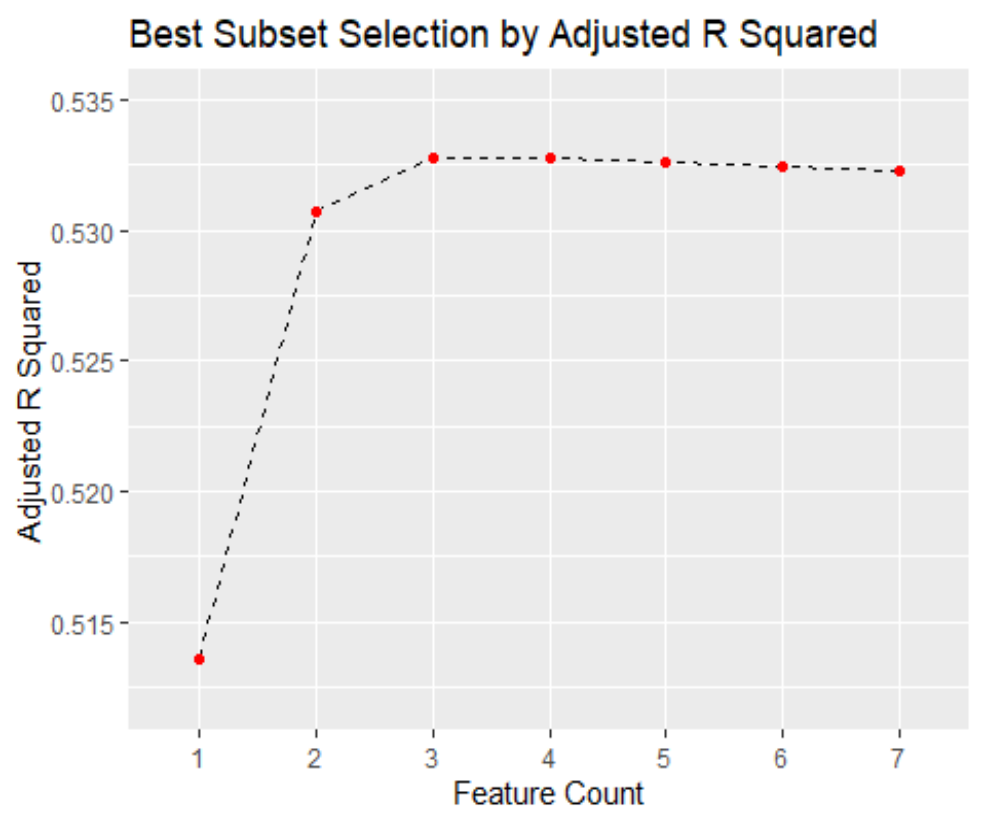
- Some variables have been log-transformed to best display in the above plots.
- Density plots represent the distribution of a numeric variable using the kernel density estimate to show the probability density function of the variable.

DATA EXPLORATION (CORRELATION MATRIX)



- Many of the features are highly correlated and therefore, feature selection will need to be applied in developing the model.

FEATURE SELECTION



- Utilizing Best Subset Selection to select features.
- Each possible subset of features is run in a standard linear model.
- The best model is selected based on **Adjusted R Squared**.
- The features associated with the best model will be selected for use in future model development.
- The best subset contains three features:
  Area, Eccentricity and Extent

Adjusted $R^2 = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}$, where RSS = Residual Sum or Squares, TSS = Total Sum of Squares,
n = Number of Data Points, d = Number of Predictors

FEATURE SELECTION (PART II)

Best Subsets Selection - Feature Inclusion by Feature Count

Feature Count	(Intercept)	Area	Perimeter	Major Axis Length	Minor Axis Length	Eccentricity	Convex Area	Extent
1	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
2	TRUE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
3	TRUE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE
4	TRUE	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE
5	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE
6	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE
7	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE

Feature count three gives us the best output: **Area, Eccentricity and Extent**

STATISTICAL MODEL DEVELOPMENT

- We have divided our labeled data set into **training and test data in a 2:1** ratio.
- We have created confusion matrix for actual and predicted model using
  I. **Linear Discriminant Analysis (LDA)**
  II. **Quadratic Discriminant Analysis (QDA)**
  III. **K-Nearest Neighbors (KNN)**
  IV. **Random Forest**
  V. **Generalized Additive Model (GAM)**
- From our developed models, we have calculated Precision, Recall and Accuracy.
- After comparing the results, we have chosen the model with the highest accuracy as the best model.
- With our best model, we have predicted the market value of a one-pound sample of beans for each cultivation in the unlabeled dataset.

CLASSIFICATION ALGORITHMS

- **LDA Model** is a way to cut down on the number of dimensions in order to solve classification problems.
  - Linear combinations of predictors are used by LDA to try to figure out the class of the observations. *[1: Section 4.4.2]*

- **QDA Model** is the general form of Bayesian discrimination.
  - This is how QDA and LDA classifiers work: They both assume that the observations from each class come from a Gaussian distribution and plug those values into Bayes' Theorem to predict. *[1: Section 4.4.3]*

- **KNN Model** is a non-parametric supervised learning method used for classification and regression.
  - The input in both situations is the k closest training examples in a data collection. Whether k-NN is used for classification or regression determines the outcome. *[2]*

[1] "An Introduction to Statistical Learning with Applications in R Second Edition" by Gareth Jones, Daniela Witten, Trevor Hastie, Robert Tibshirani
[2] https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm

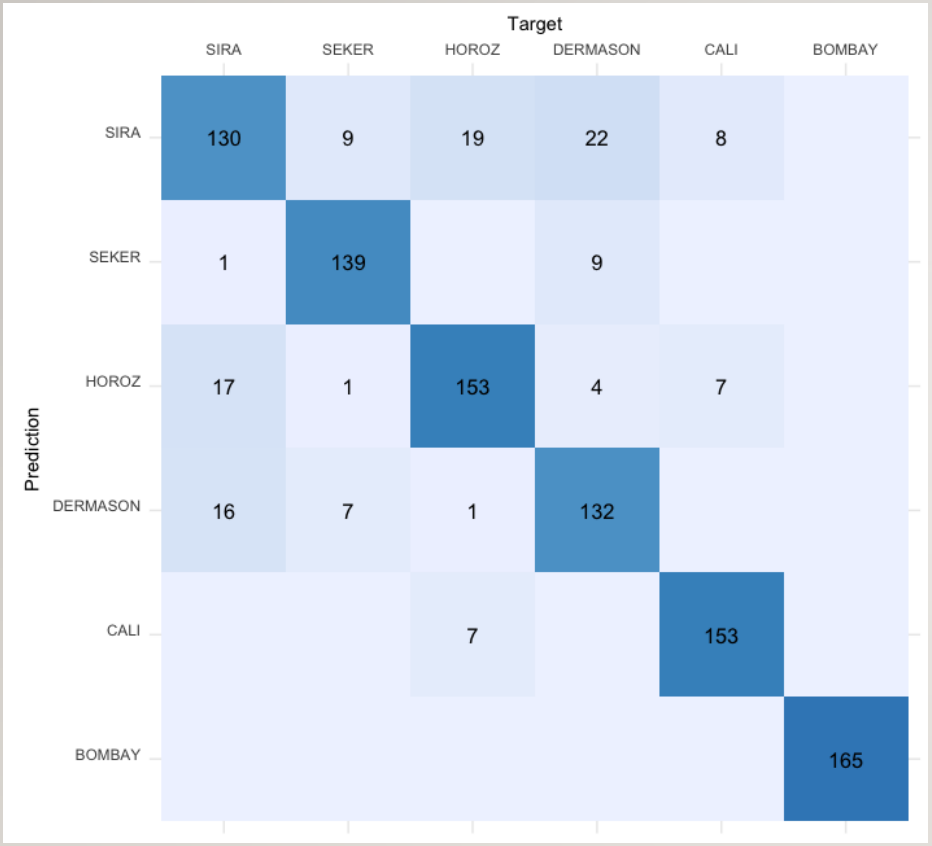
CLASSIFICATION ALGORITHMS

- **Random Forest Regression i**s used to do supervised learning. It uses an ensemble learning method to do regression**.**
  - In the ensemble learning method, predictions from several machine learning algorithms are combined in order to make more accurate predictions than a single model could make on its own. *[3]*

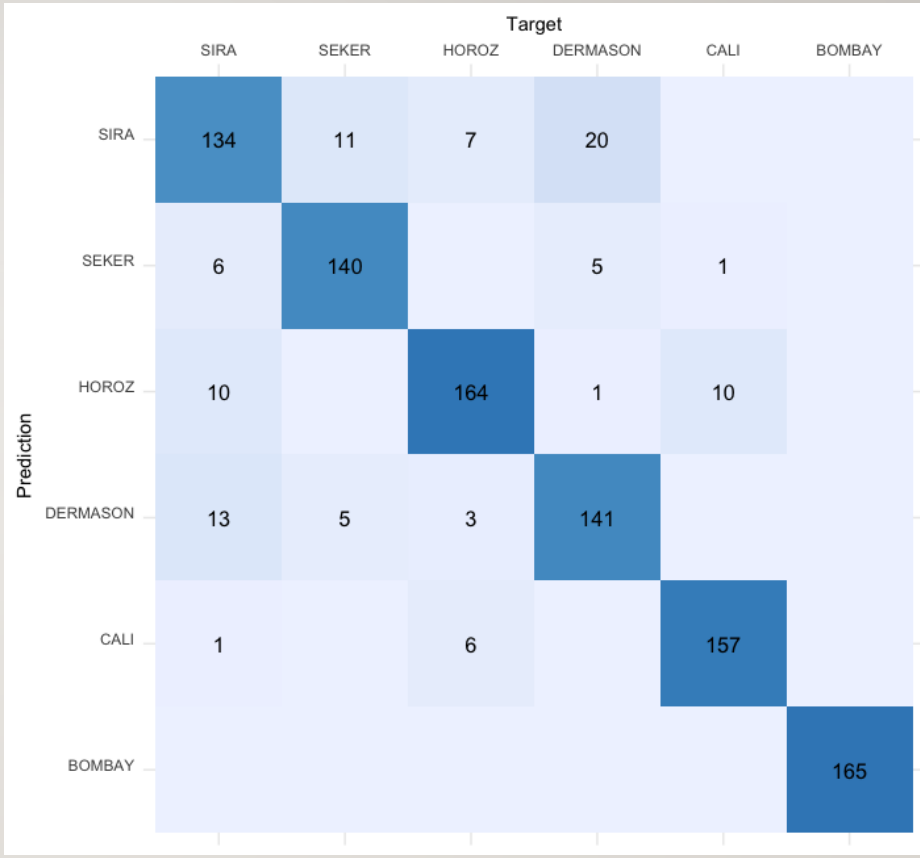
- **Generalized additive models:** Adding more variables to a standard linear model can be done with generalized additive models (GAMs). GAMs allow non-linear functions for each of the variables but keep the model's additivity.
  - GAMs can be used with both quantitative and qualitative responses, just like linear models can be used with both. *[1: Section 7.7]*

[1] "An Introduction to Statistical Learning with Applications in R Second Edition" by Gareth Jones, Daniela Witten, Trevor Hastie, Robert Tibshirani
[3] https://levelup.gitconnected.com/random-forest-regression-209c0f354c84

CONFUSION MATRIX (LDA and QDA Model)

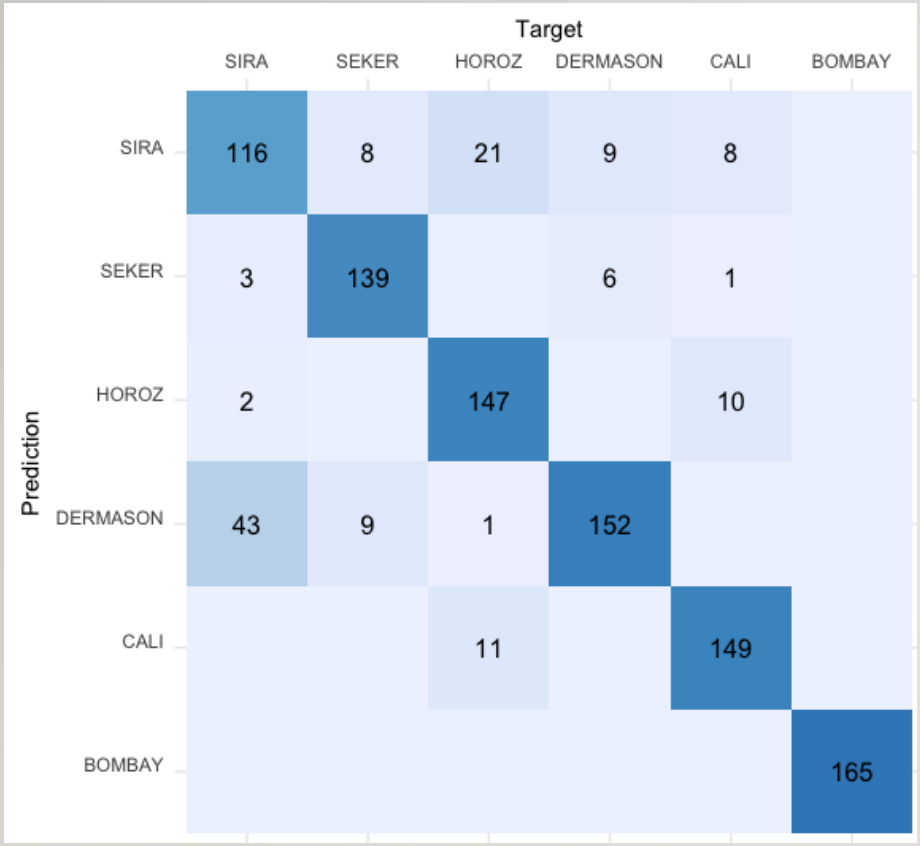


LDA Model Accuracy on Test Data:
87.20%

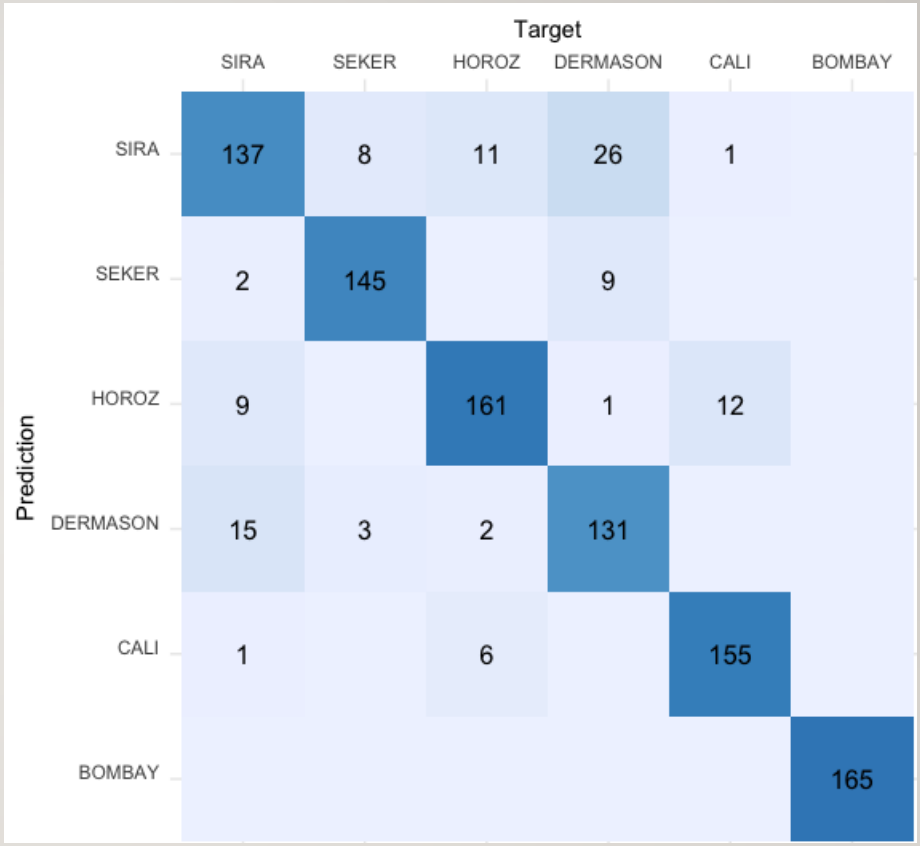


QDA Model Accuracy on Test Data:
90.10%

CONFUSION MATRIX (KNN and Random Forest)

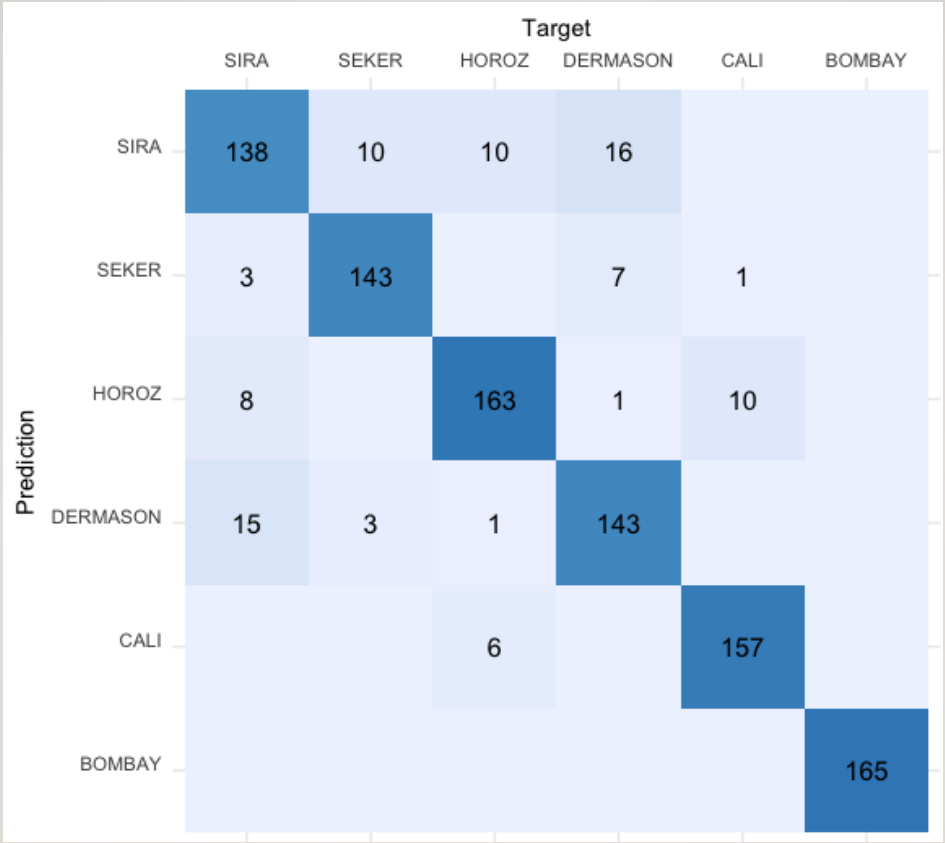


KNN Model Accuracy on Test Data:
86.80%



Random Forest Accuracy on Test Data:
89.40%

GENRALIZED ADDITIVE MODEL (GAM)



GAM Accuracy on Test Data: **90.90%**

GENRALIZED ADDITIVE MODEL (GAM)

ANOVA for GAM with and Without Splines

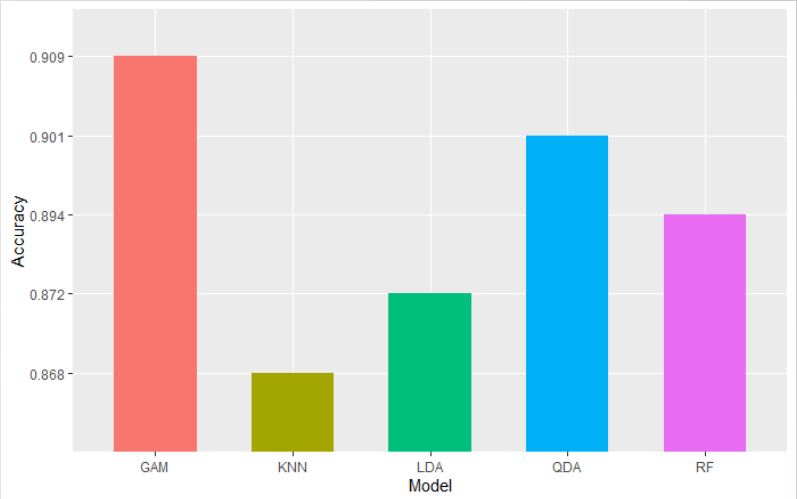
Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
9980	1162.493	NA	NA	NA
9970	1140.049	10	22.44384	0.0129965

- Feature engineering was used in the development of the GAM model where a natural spline with 4 degrees of freedom was used on the log-transform of Area.

PERFORMANCE MATRIX OF ALL APPLIED MODELS

Performance Measures by Model Type

	KNN	LDA	QDA	Random Forest	GAM
Accuracy (%)	86.80	87.20	90.10	89.40	90.90
Precision (%)	87.44	87.79	90.24	89.70	91.05
Recall (%)	86.87	87.25	90.07	89.44	90.91



$$\text{Accuracy} = \frac{tp+tn}{tp+fp+tn+fn}$$

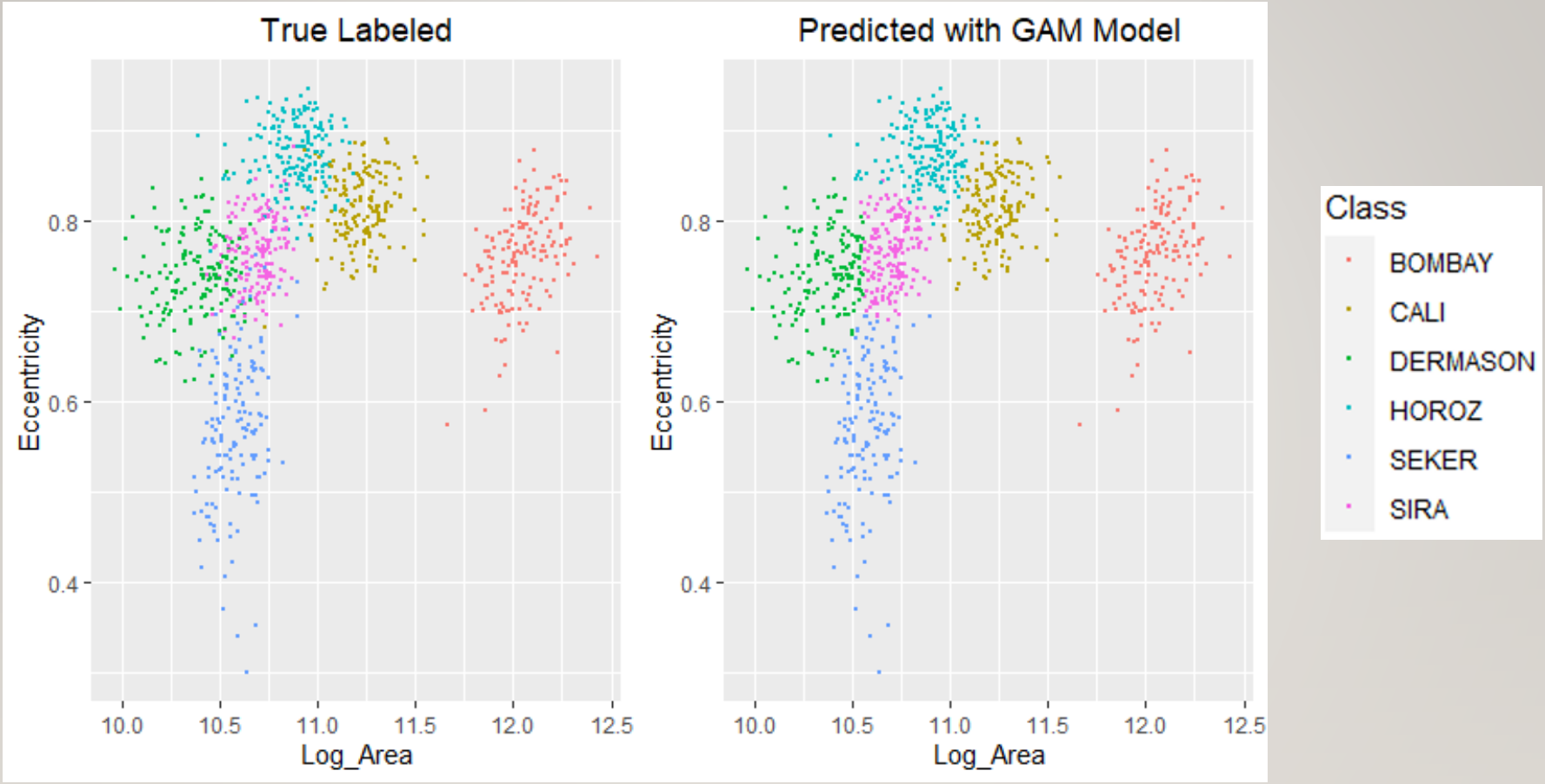
$$\text{Precision} = \frac{tp}{tp+fp}$$

$$\text{Recall} = \frac{tp}{tp+tn}$$

where tp = true positive, tn = true negative, fp = false positive, fn = false negative

https://www.rdocumentation.org/packages/caret/versions/3.45/topics/confusionMatrix

TRUE VS GAM COMPARISON



Bean Market Value Reference Table

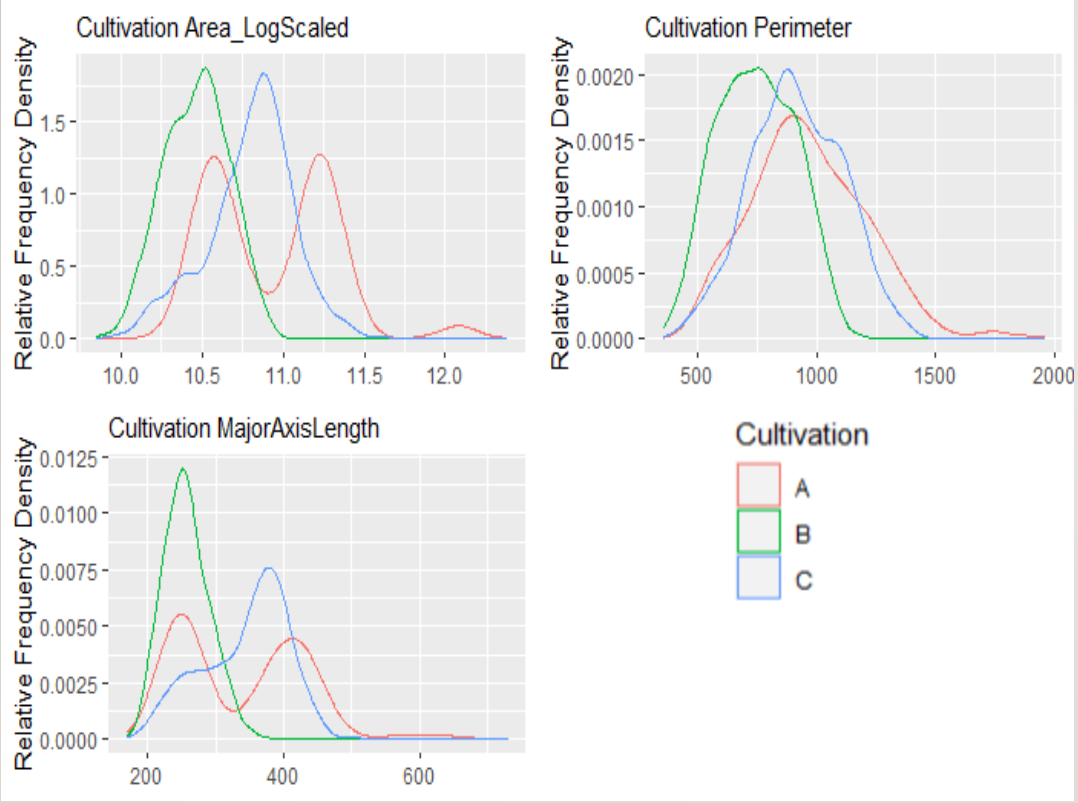
Local Market Price and Grams/Seed Reference by Bean Variety

Class	GramsPerSeed	DollarPerPound
BOMBAY	5.56	1.92
CALI	6.02	0.61
DERMASON	1.98	0.28
HOROZ	2.43	0.52
SEKER	2.72	0.49
SIRA	5.40	0.38

[a] ~453.592 grams per pound

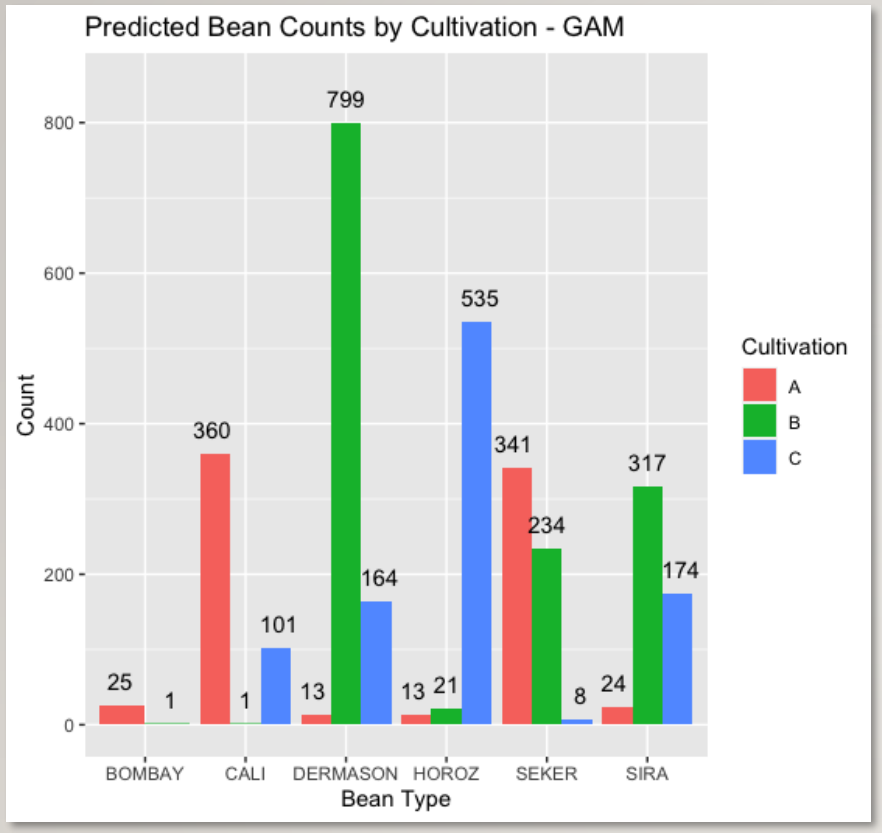
- Predicted Market Value = GramsPerSeed x DollarPerPound/453.592 grams per pound

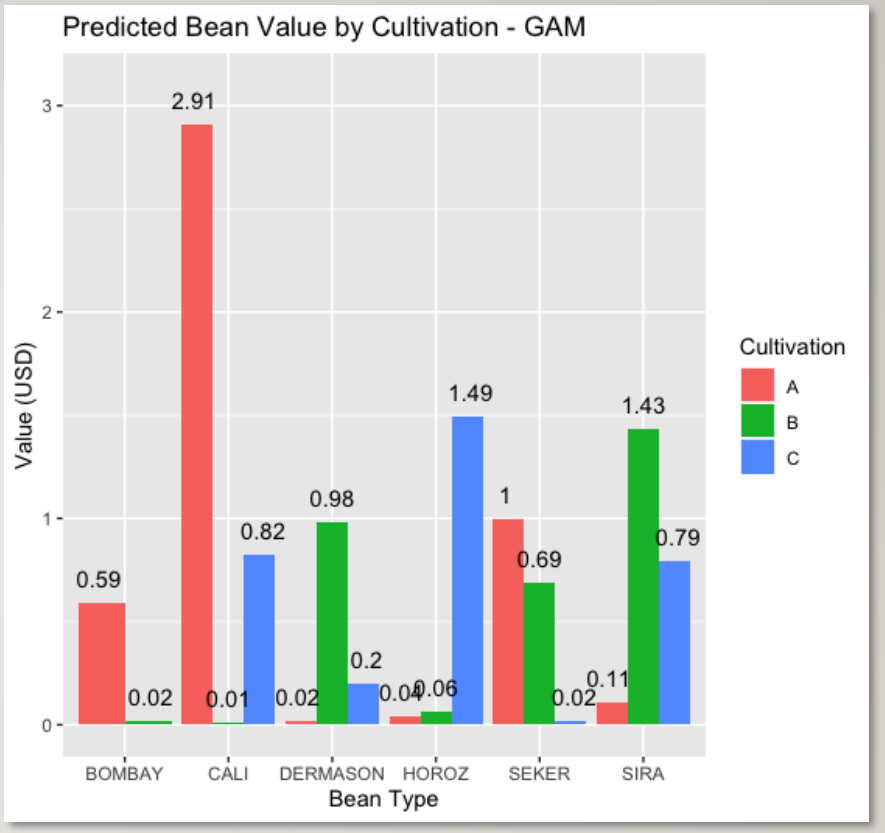
DATA EXPLORATION OF CULTIVATIONS



- The density plots from the unlabeled data reveal that the variables exhibit multimodal behavior.
- This indicates that at the cultivations are likely to have a significant difference in their makeup of bean classes.

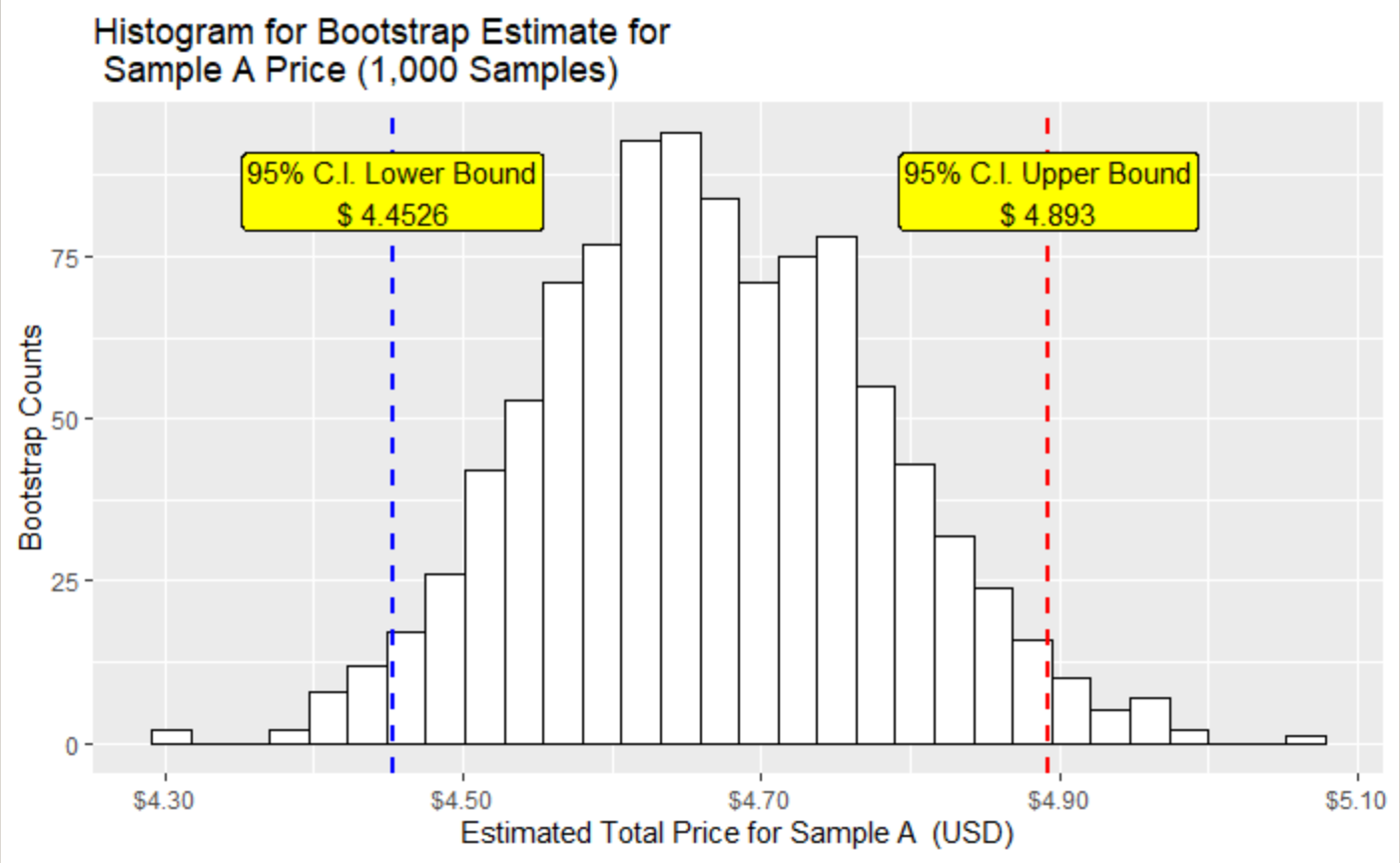
PREDICTED BEAN COUNTS AND VALUE BY CULTIVATION



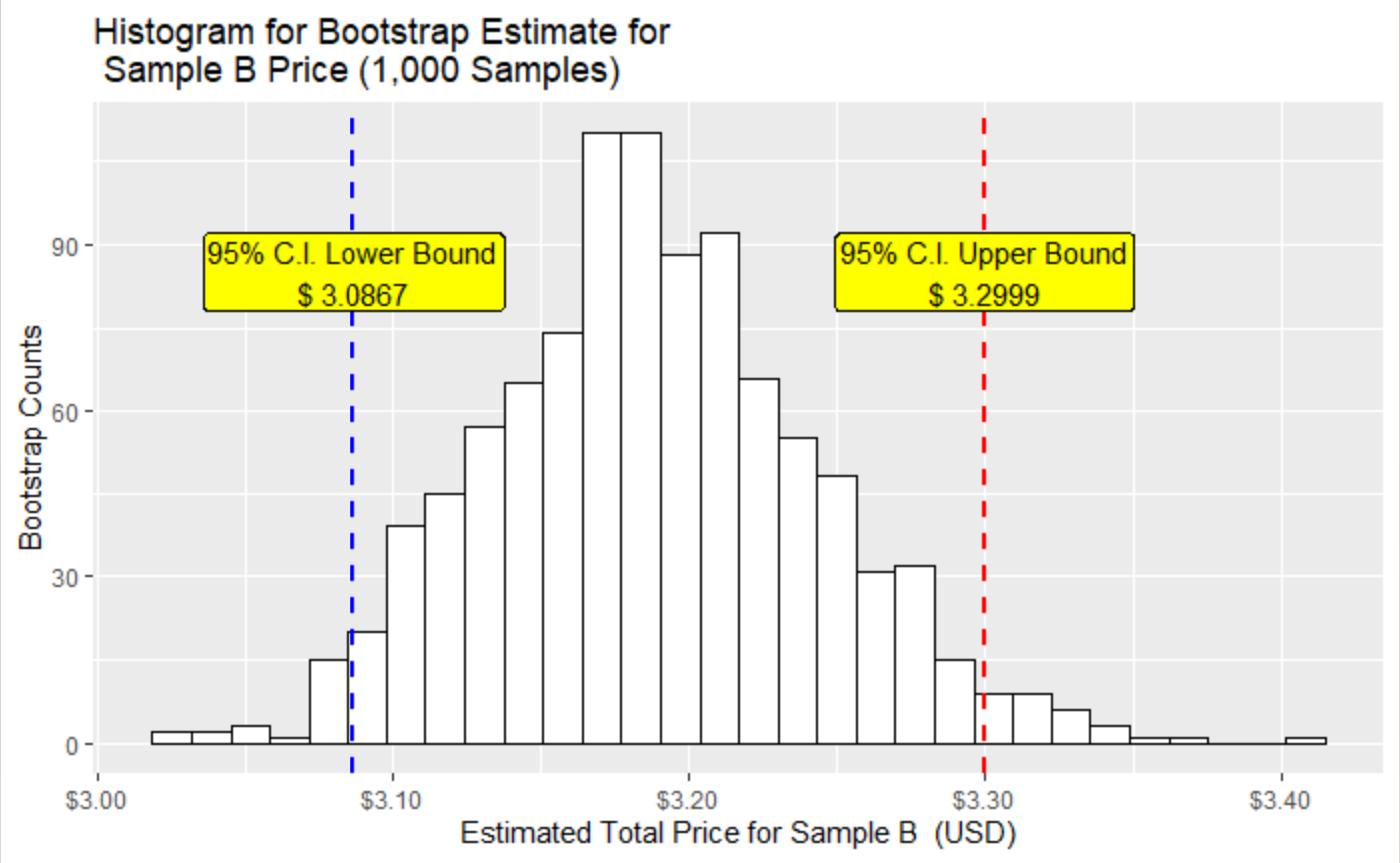


- Bean Counts were calculated by predicting bean class for each cultivation using the GAM model.
- Bean Values were calculated by taking the grams per seed for the predicted bean class, divided by grams per pound and multiplied by dollar per pound for the predicted bean class.

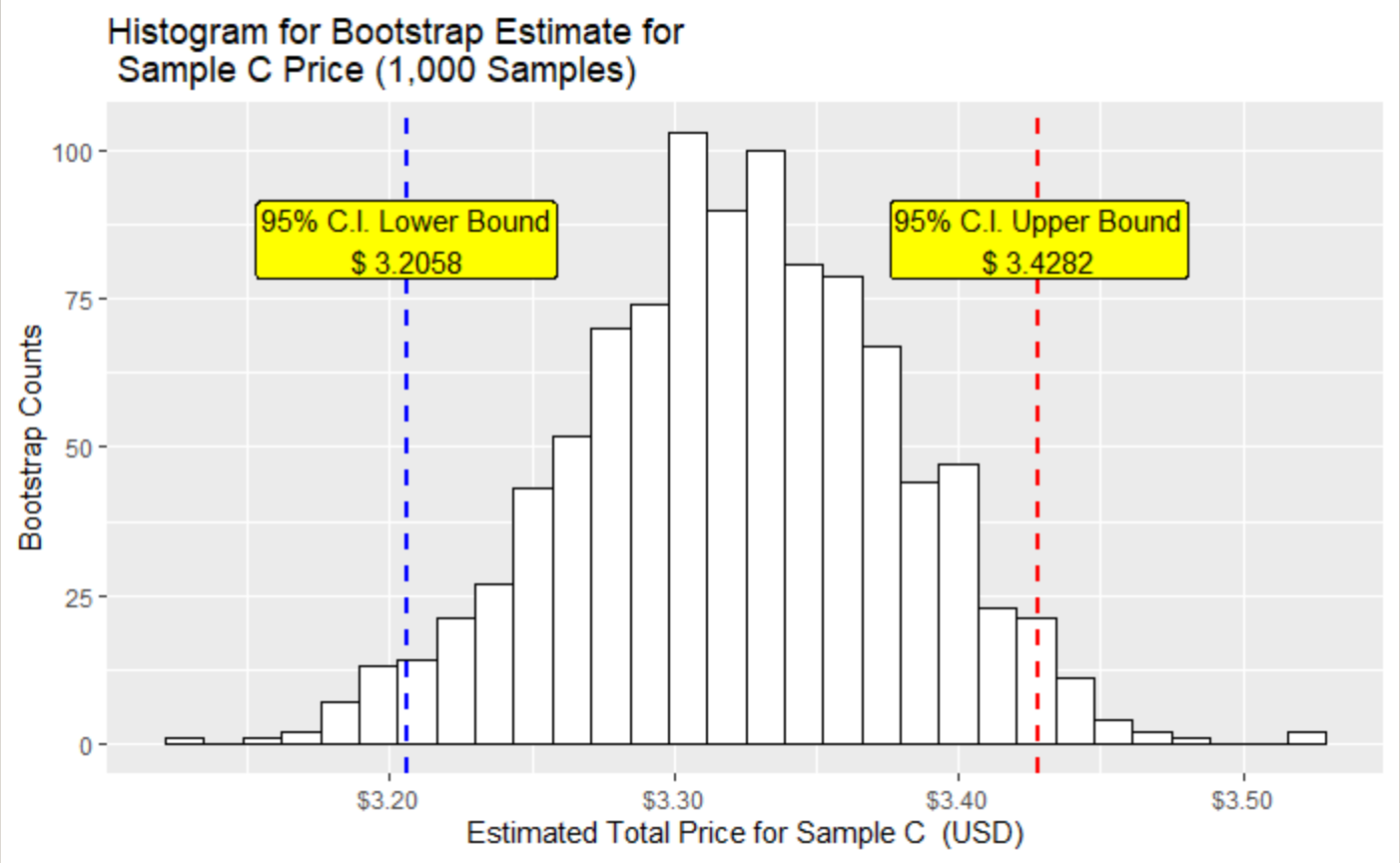
PLOT OF BOOTSTRAP 95% CONFIDENCE INTERVAL FOR A CULTIVATION [SAMPLE A]



PLOT OF BOOTSTRAP 95% CONFIDENCE INTERVAL FOR A CULTIVATION [SAMPLE B]



PLOT OF BOOTSTRAP 95% CONFIDENCE INTERVAL FOR A CULTIVATION [SAMPLE C]



CONCLUSION

Value (USD) of One Pound Bean Cultivation (Calculated from Bootstrap of 1,000 One Pound Bean Samples)

	Mean	Standard Error	Lower Confidence Interval	Upper Confidence Interval
A	4.6680	0.1122	4.4511	4.8792
B	3.1894	0.0566	3.0788	3.3010
C	3.3189	0.0581	3.2022	3.4342

- The GAM algorithm is recommended as the best model based on producing the highest accuracy in bean classification.
- The selected GAM model was used to predict the bean type as well as the market value associated with each bean type for each cultivation sample.
- Predicted values for each cultivation are based on 1,000 bean samples randomly selected with replacement from the unlabeled dataset.
- The measure of the accuracy of the predicted value of a pound of bean for each sample is given by the lower and upper 95% confidence interval in the chart above.

REFERENCES

Koklu, M., Ozkan, I. A. (2020). Multiclass classification of dry beans using computer vision and machine learning techniques. Computers and Electronics in Agriculture, 174.

James, G., Witten, D., Hastie, T., Tibshirani, R. (n.d.). 6. Linear Model Selection and Regularization. In An Introduction to Statistical Learning with Applications in R.

Hossin, Mohammad & M.N, Sulaiman. (2015). A Review on Evaluation Metrics for Data Classification Evaluations. International Journal of Data Mining & Knowledge Management Process. 5. 01-11. 10.5121/ijdkp.2015.5201.