Predicting Bean Type Based on Metamorphic Measurements

**Charts mentioned can be found in the PowerPoint slides with the corresponding heading and can be reproduced via the annotated code with the corresponding heading*

Table of Contents

Intro and Objective

For this project, algorithms covered in this course will be tested in their ability to make predictions on beans using features consisting of their measurements. The project relies on a dataset obtained from a study that took the morphometric measurements of several types of beans found in Turkey in order to separate white beans from the others. (KARA, SAYINCI, ELKOCA, ÖZTÜRK, & ÖZMEN, 2013) Furthermore, the factor of cost for each type of bean will be applied to this analysis, as incorrect predictions could financially harm a customer purchasing beans or a business selling the beans. Thus, not only accuracy of the predictive models will be tested in their application to the dataset but also the average cost of error created by the models to determine the best model or models. The objective of this project is to find a model that obtains an accuracy measure of at least 80%, maximizes accuracy, and minimizes the weight and cost differential between predictions and actual values, in order to give a recommendation on which type of algorithm has the most potential for accurately sorting the beans. No major assumptions about the dataset are being made other than that it consists of 3000 observations, 500 observations of each bean type, 7 independent variables, 6 dependent variable classes, and the info shown in tables 1 and 2 about the beans and variables.

The type of analysis required by this problem is predictive analytics utilizing multiclass classification as there are 6 classes to fit the bean data in with 7 continuous independent variables to utilize in order to classify them. In this type of analysis, the values of selected independent variables for each observation are used in conjunction with statistical models in order to predict the dependent variable, bean type. The results of the predictions are then compared with the actual results in order to test for stats such as accuracy, weight differential, and price differential which can then be used to check how useful the models are.

The dependent variables of the dataset used in this analysis include the bean classes which are as follows: Seker, Bombay, Cali, Horoz, Sira, and Dermason. The independent variables of this dataset include the area, perimeter, major axis length, minor axis length, eccentricity, convex area, and extent measurements of the beans. These variables will be analyzed for each model on their ability to predict the bean classes, and then utilized for the predictions based on the log odds that they attribute towards the classes of beans.

Finally, in order to measure the cost of error associated with inaccurate predictions, additional data for the beans was provided for this project. This data includes the price per pound of each bean and the weight in grams per each seed. This information will be used to obtain approximate values of these stats for each bean and therefore the average cost associated with the error rates of the predictive models to decide on the optimal model(s).

Methods

For this project, the dataset will be randomly split 80% for training/validation and 20% for testing of the final chosen model in order to maximize the amount of data used for creating the model and minimize bias as well as potential for overfitting. This should still allow for enough data in the test dataset to test the final model for accuracy and error rates. For the training/validation method, the leave one out cross validation approach was chosen over the validation set approach. This is due to advantages it has over the validation set approach such as being able to maximize the dataset used in the training process and its tendency to not overestimate the validation error rate. Also, its main disadvantage compared to the validation set approach and other cross validation approaches which is that it is computationally expensive is minimized in this project due to the dataset not being immense and only having 3000 observations.

The leave one out cross validation approach involves repeatedly running the model on the training/validation dataset while changing what's used for the training/validation split with each iteration. (James, Witten, Hastie, & Tibshirani, 2021) On each iteration, all but one of the training/validation dataset observations are used to train the model, and the unused observation is used to validate the model. This is repeated until each observation is used once to validate the model so that eventually every observation in the training/validation dataset is used in the training as well as validation processes.

For normalizing the data, a z-score methodology was chosen due to it being common and applicable to this dataset. Also it is known for handling outliers well by giving them slightly more weight than the rest of the data. (Bobbitt, 2019) Regarding the calculation for Z-Score, the standard calculation is used by first calculating the mean and standard deviation of each feature, and then calculating the z score for each individual feature by subtracting the value by the mean and dividing this new value by the standard deviation. (Bobbitt, 2019) The mean and standard deviation of the training dataset are saved for use in future test and production datasets in order to maximize effectiveness of the trained statistical models.

For feature selection, best subset selection via the leaps library in R was chosen for this project. This method and library test all possible combinations of models for the independent variables and the dependent variable via multinomial logistic regression, and chooses the best ones based on residual sum of squares. (Clark Science Center, 2016) It then reveals the best variable combinations and statistics for each possible number of variables. In this case, there were seven possible independent variables and thus this method revealed 7 possible combinations of variables, and statistics for them. Next, a variable combination was chosen based on adjusted R squared.

Statistical models chosen for this project include multinomial logistic regression, linear discriminant analysis, KNN classifier, quadratic discriminant analysis, and Naive Bayes classifier.

Multinomial logistic regression is a model where the log odds of the dependent variable are predicted based on a linear combination of independent variables. (UCLA: Statistical Consulting Group, n.d.)

$$ln(\frac{P(BeanClass=1)}{P(BeanClass=2)})=b_1(Variable_1)+b_2(Variable_2)+b_3(Variable_3)\ldots$$

$$ln(\frac{P(BeanClass=3)}{P(BeanClass=4)})=b_1(Variable_1)+b_2(Variable_2)+b_3(Variable_3)\ldots$$

(UCLA: Statistical Consulting Group, n.d.)

Linear discriminant analysis attempts to map observations of the dependent variable to a data space and utilize a discriminant rule to divide the data space into regions based on the independent variables, utilizing maximum likelihood or Bayesian rules. (Xiaozhou, 2020) It assumes that the function of the class is a density function for a multivariate normal random variable with a class specific mean and a shared covariance matrix.

$$f_k(x)=\frac{1}{\sqrt{2\pi}\sigma_k}exp(-\frac{1}{2\pi\sigma_k^2}(x-\mu_k)^2)$$

(Xiaozhou, 2020)

K-nearest neighbor classifier is a model that, given a positive integer, K, and observation, attempts to classify the observation based on the K most similar observations in the training dataset. It improves over time as observations are added to its training dataset.

$$Pr(T=j|X=x_0)=\frac{1}{K}\sum_{i=\epsilon N_0}I(y_i=j)$$

(James, Witten, Hastie, & Tibshirani, 2021)

Quadratic discriminant analysis is similar to linear discriminant analysis although it assumes that each class has its own covariance matrix and utilizes a quadratic function. (James, Witten, Hastie, & Tibshirani, 2021) It assumes that the function of the class is a density function for a multivariate normal random variable with class-specific mean and covariance matrix.

$$\delta_k(x) = -\frac{1}{2}(x-\mu_k)^T \sum_k^{-1}(x-\mu_k) - \frac{1}{2}log|\sum_k| + log\pi_k$$

The Naive bayes model makes class predictions by first assuming that within the kth class, the p predictors are independent, and then utilizes this assumption in a modified Bayes Theorem function.

$$Pr(Y=k|X=x) = \frac{\pi_x \times f_{k1}(x_1) \times f_{k2}(x_2)\times\ldots\times f_{kp}(x_p)}{\sum_{l=1}^{K} \pi_l \times f_{l1}(x_1) \times f_{l2}(x_2)\times\ldots\times f_{lp}(x_p)}$$

(James, Witten, Hastie, & Tibshirani, 2021)

Exploratory Analysis

For the exploratory analysis, the data was grouped by bean type and then box plots and histograms of each variable were created in order to check the distributions and outliers of the data. Most of the data seems to be normally distributed with varying levels of kurtosis while the rest have varying levels of skewness, especially within the extent variable. No assumptions have been made that the data has been previously cleaned of any error or abnormalities, and thus the outliers will be removed based on anything that is above quartile 3 plus 1.5 times the interquartile range or below quartile 1 minus 1.5 times the interquartile range.

After removing the outliers, the dataset has been reduced from 3000 observations to 2731 observations with 422 being the lowest amount of observations in a class. These outliers are visibly gone from the upper and lower limits of the box plots as well as the tails of the histograms.

Predictive Analysis

The predictive analysis included several steps to first prepare the data, create models, and then run the models. It involved first converting the dependent variable into a numeric format between 1 and 6 based on the BeanKey.df dataframe. Next it involved normalizing the data via converting them into Z-scores and variable selection utilizing the Best Subset Selection method in conjunction with the adjusted r^2 measure. The variable selection process determined that the best combination of variables to use for this analysis were "Area", “Perimeter”, "Eccentricity", ConvexArea, and "Extent". Afterwards the dataset was randomly split, 80% for training/validation and 20% for testing. The models were created with the formula “ClassID = Area + Perimeter + Eccentricity + ConvexArea + Extent”, set up in Leave-One-Out-Cross-Validation, and modified via some minor hyperparameter tuning such as optimization of K Nearest Neighbors K value. Finally, the models were run and the best model was chosen based on accuracy, price comparison between predictions and actual results, and weight comparison between predictions and actual results.

Results Analysis

The results of testing the models on the validation dataset via leave one out cross validation were that the most optimal models were multinomial logistic regression (MLR) and linear discriminant analysis (LDA) as MLR had an accuracy rate over 93% and LDA had the lowest cost differential compared to the other models while keeping an accuracy rate over 87% when tested on the validation dataset. The model with the lowest weight differential between predictions and actual values was K Nearest Neighbors with 30 neighbors, but this was not chosen as an optimal model due to having an accuracy race below 80% and due to accuracy rate and cost differential being prioritized. Thus, the multinomial logistic regression model as well as the linear discriminant analysis model were both chosen to be run on the test dataset to get final prediction values as they are both superior in different measures.

The results on the test dataset indicated strong results on all three measures by both models with each maintaining similar accuracy rates and cost differentials.

Conclusions

The project resulted in two models being chosen for optimally predicting bean data, multinomial logistic regression, and linear discriminant analysis. This is due to them both having the highest accuracy rate and lowest cost differential respectively when compared to the other tested models. Furthermore, these two models were consistent in their results when tested on a holdout test dataset, showing that they can be relied on to have similar results on similar datasets.

References

Bobbitt, Z. (2019, April 20). *How to Normalize Data in R*. Retrieved from Statology: https://www.statology.org/how-to-normalize-data-in-r/

Clark Science Center. (2016). *Lab 8 - R*. Retrieved from Clark Science Center @ Smith College: https://www.science.smith.edu/~jcrouser/SDS293/labs/lab8-r.html

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning with Applications in R Second Edition.* Springer.

KARA, M., SAYINCI, B., ELKOCA, E., ÖZTÜRK, İ., & ÖZMEN, T. B. (2013). Seed Size and Shape Analysis of Registered Common Bean (Phaseolusvulgaris L.) Cultivars in Turkey Using Digital Photography. *Journal of Agricultural Sciences*, 220-221.

UCLA: Statistical Consulting Group. (n.d.). *Multinomial Logistic Regression | R Data Analysis Examples*. Retrieved from UCLA Advanced Research Computing: https://stats.oarc.ucla.edu/r/dae/multinomial-logistic-regression/

Xiaozhou, Y. (2020, May 9). *Linear Discriminant Analysis, Explained*. Retrieved from Towards Data Science: https://towardsdatascience.com/linear-discriminant-analysis-explained-f88be6c1e00b