Predicting the USA's Honey Production: Multiple Linear Regression Against Random Forest Regression (210025601)

Keywords: multiple linear regression, MLR, random forest regression, RFR, machine learning, data, error, cross validation, regression.

01 Data collection

Description and Motivation Of the Problem

- This study will critically evaluate two supervised machine learning models.
- Linear regression will be implemented against random forest regression as the two proposed regression algorithms.
- Data reported from The USDA's National Agricultural Statistics Service (NASS) will be used to predict the USA's honey production.

Data Set Analysis & Basic Statistics

- 'Honey Production in the USA' includes time-series data from 1998-2012 which was obtained from Kaggle consisting of 626 rows (excluding variable names) and 8 columns; 1 column contains categorical data (from 44 of the 50 US states), 1 column of a series of dates (year), and the remaining 6 observations are numeric.
- Given the nature of the study, 'yieldpercol' will act as each model's target variable.
- The data was set to be cleaned, however, the data contained zero missing values, is without inconsistencies, and does not hold duplicate observations.
- Table 01 presents basic statistics in the form of mean, median, minimum, maximum, standard deviation, and skew values each of the 6 numeric variables in the data.
- Pearson correlation matrix (figure 01) presents multiple strong linear relationships among multiple of the data's explanatory variables. It can be assumed that these features will play a significant role in predicting US honey production.
- The data set's correlation coefficients will later be explained and further explored using Multiple Linear Regression.
- Boxplots (figure 02) assists in identifying variable distribution and outliers. While many outliers exist within the dataset, these will not be removed as they may later prove to be informative and provide greater statistical representation of the data.

Figure 01. Correlation matrix

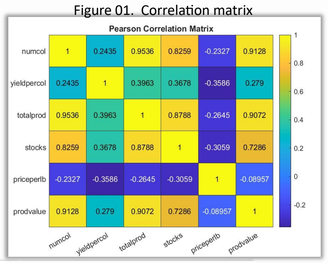
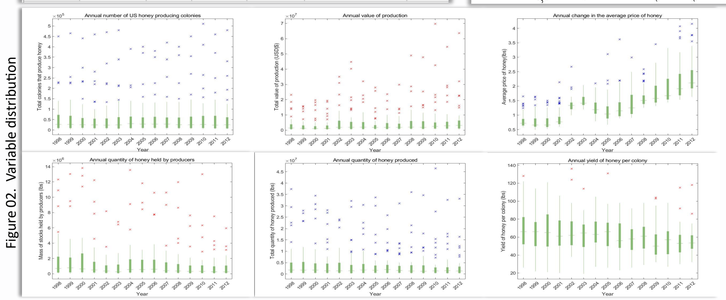


Table 01. Central tendency and variability

	Mean value	Median	Minimum	Maximum	Standard deviation	Skew
numcol	6.0284e+04	26000	2000	510000	9.1077e+04	2.7556
yieldpercol	62.0100	60	19	136	19.4600	0.6920
totalprod	4.1691e+06	1533000	84000	46410000	6.8838e+06	2.8900
stocks	1.3189e+06	439500	8000	13800000	2.2730e+06	3.2200
priceperlb	1.4100	1.3600	0.5000	4.1500	0.6400	1.1900
prodvalue	4.7157e+06	1841500	162000	69615000	7.9761e+06	3.8320

Figure 02. Variable distribution



02 Model specification

Multiple Linear Regression (MLR)

Summary

- Multiple linear regression attempts to explain the linear relationship amongst two or more independent variables in relation to the model's single metric dependent variable.
- Through regression, we aim to determine which variables greatly explain the model's variation (otherwise referred to as the r-squared value).

Advantages

- Statistical models containing multiple linear regressors can easily be modified and simplified depending on the extent to which "predictors substantially influence the dependent variable" [2].
- Relatively straight forward to implement and is an efficient method to understand the parameters of a model.
- An overfitting issue can be diminished through regularization.

Disadvantages

- MLR holds 5 main assumptions: minimal multicollinearity, normal distribution of residuals, no autocorrelation, linearity and homoscedasticity.
- Excess explanatory variables can cause an overfitting problem [2].

Random Forest Regression (RFR)

Summary

- A supervised machine learning regressor technique whereby the final outcome is dictated by a random forest.
- The process construes of a "multitude of decision trees—known as a 'forest' which have been trained with the bagging method" (Cooper, 2021).

Advantages

- Is flexible in its applications and thus, has proved successful within numerous fields [4].
- Provides great accuracy which is applicable to both linear and non-linear models [4].
- More efficient than a single decision tree classifier due to utilizing groups of training set data.

Disadvantages

- The number of trees requires manual imputation.
- While RFR aims to provide accurate results, its tendency to overfit the final model against its training dataset, thus, impacting the performance of the model's outcome [3].

03 Hypothesis statements

Multiple Linear Regression

- H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$
- H_a: $\beta_j \neq 0$ (for at least one value of β)
- Given that one of linear regression's utilizations are for model prediction, it is expected that MLR will perform significantly well, given that the model's holds true to the parameter's assumptions [3].

Random Forest Regression

- Through random forest's method of bagging, this increases the accuracy of the model [3].
- Having explored various studies comparing multiple linear regression against random forest regression, we can expect models deriving from RFR will outperform those of MLR. However, multiple linear

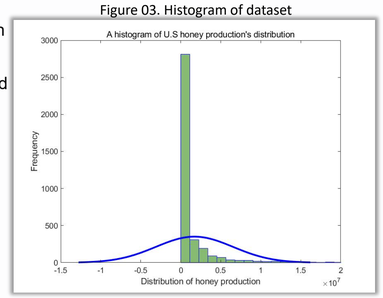
04 Training and evaluation methodology

- Considering that the dataset is relatively small comprising of 627 observations, the original csv file will be split as 75:25 for training and testing purposes—470 for training and 156 for testing.
- Given that the data is not of a Gaussian distribution, is in fact reverse J-shaped. Normalized feature scaling will be utilized.
- The selected models will exhibit the lowest root mean square error, mean squared error (MSE/training error), mean average error (MAE), the highest accuracy and r-squared value (R^2).

Calculations of k-fold loss, will be utilized to compute training accuracy for the regression models.

- Multiple linear regression model will be fit using the least-squares method.
- A random forest "bootstrap-aggregated" otherwise referred to as bagging will be implemented as decision trees [7].
- Standard CART algorithm will be used to build 30 learning cycle with a minimum leaf size of 8.
- Figure 01 exhibits a multicollinearity problem between *prodvalue* and *totalprod*, *prodvalue* and *numcol, totalprod* and *numcol*—each containing strong correlations exceeding 0.9.
- Train two different models due to meet the assumptions of MLR.
- Model 1 excludes *numcol* and *totalprod*. Model 2 excludes *numcol* and *prodvalue*.
- Due to the aforementioned size of the data, both models will initially utilize 10-fold cross validation.

Figure 03. Histogram of dataset



05 Parameters and experimental results

Multiple Linear Regression

Choice of parameters

- Model 02 —which excludes *numcol* and *prodvalue*—will be utilized to maximize performance results.
- 10 k-fold cross validation.
- Normalised feature selection.
- Least-squares linear model.
- Robust options off.

Experimental results

- Robust linear regression and stepwise linear regression only marginally increased the r-square value
- Linear regression with robust options as a hyperparameter to some degree reduced MAE and RMSE, while simultaneously increasing the r-squared value.
- PCA increased all errors, thus reducing training accuracy and r-squared values.

Experimental results

- Robust linear regression and stepwise linear regression only marginally increased the r-square value
- Linear regression with robust options as a hyperparameter to some degree reduced MAE and RMSE, while simultaneously increasing the r-squared value.
- PCA increased all errors, thus reducing training accuracy and r-squared values.

Random Forest Regression

Parameters

- Model 02 exhibited greater regression results.
- Resubstitution cross validation worked favourably across both models as opposed to 10 k-fold cross validation.
- The random forest regression will be fit using bagged trees.
- Hyperparameters—minimum leaf size: 5. Number of learning cycles: 25.

Experimental results

- PCA too created growth for errors and decreased r-squared values within the random forest model.
- Boosted trees had the same affects as PCA.

Figure 04. Predictions of model 02

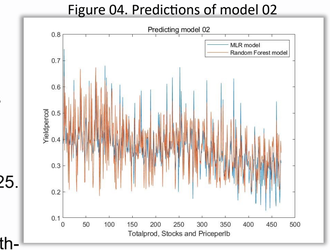


Table 02. Honey production training results

	Linear model 01	Linear model 02	RF model 01: 10-Fold	RF model 01: Resubstitution	RF model 02: 10-Fold	RF model 02: Resubstitution
Rsquared	0.1800	0.2200	0.3400	0.5700	0.3300	0.6000
RMSE	0.1484	0.1450	0.1338	0.1081	0.1344	0.1041
TrainError	0.0228	0.0220	0.0183	0.0114	0.0174	0.0112
TrainAccuracy	0.9773	0.9780	0.9817	0.9886	0.9826	0.9888

06 Analysis and critical evaluation of results

Multiple linear regression

- Multiple linear regression final model equation: Y = 0.40523 + 0.34848 + 0.024384 - 0.24739. Thus, holding other variables constant yieldpercol will increase on average by 0.34848lbs for each 1lb increase in total production. Holding other variables constant yieldpercol will increase on average by 0.024384lbs for each 1lb increase in stocks. Holding other variables constant yieldpercol will decrease on average by 0.24739lbs for each 1 US dollar increase in priceperlb.
- The final multiple linear regression model appears to display a slight under fitting problem. The training error of the model lies close to the test error which explains why under fitting is present. To increase the fitting of the model the model could undergo further complexity, consist of improved features and noise could be reduced from the data [8].

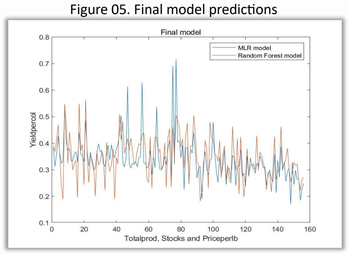
Table 03. Final model results

	Linear final model	RF final model
Rsquared	0.2616	0.3200
RMSE	0.1331	0.1401
TestError	0.0191	0.0104
TestAccuracy	0.9809	0.9896

Random Forest Regression

- While is was predicted that multiple linear regression would outperform the results of the random forest regression, this is not exhibited in the final model's results. Random forest presents a r-squared value of 0.32, thus explaining an additional 5.84%
- The final random forest regression model exhibits a fitting with again a slight underfitting problem. The training error of the random forest model, as presented in table 02 equates to 0.0112 while the computer test error is equivalent to 0.0104. Although the difference between the two errors is not substantial, nor is their value, this to an extent displays that the model is too simplistic and similar to MLR, additional features should be considered.

Figure 05. Final model predictions



07 Lessons learned and future work

- Only one feature scaling technique was used throughout all stages of the training and testing process, due to the fact that the dataset had displayed a J-shaped skew. Manipulating data with other methods of feature scaling such as standardized feature scaling could have perhaps presented unlimited results.
- Gradient descent algorithm could have been applied to the linear regression model to optimize results.
- Future projects will ensure larger datasets are used and models are optimized by adding additional features and have greater complexity.
- Considering that each model contained 3 predictors, and given the results of both the multiple linear regression and random forest models, other explanatory factors should be explored to improve the variation of yieldpercol.

08 References

[1] Karadas, K. and Kadirhanogullari, I. (2017) Predicting Honey Production using Data Mining and Artificial Neural Network Algorithms in Apiculture. *Pakistan Journal of Zoology*, 49(5), 1611-1619. doi:10.17582/journal.pjz/2017.49.5.1611.1619

[2] Hazra, A. and Gogtay, N. (2017) Biostatistics Series Module 10: Brief Overview of Multivariate Methods. *Indian Journal of Dermatology,* 62(4), 258-366. doi:10.4103/ijd.IJD_296_17

[3] Cooper, K. (2021) 14 Essential Machine Learning Algorithms, Springboard. Available at: https://www.springboard.com/blog/ai-machine-learning/14-essential-machine-learning-algorithms/ (Accessed 30th November 2021)

[4] Agrawal, R. et al (2021) Machine Learning: Theoretical Foundations and Practical Applications [E-Book], Singapore: Springer, pg 70-72.

[5] Newbold, P., Carlson, W., Thorne, B., and Aitken, J. (2019) Statistics for Business and Economics [E-book], Harlow: Pearson Education limited. Accessed 2nd December 2021)

[6] Kumar, A. (2021) K-Fold Cross Validation—Python Example, *Vital Flux.* Available at: https://vitalflux.com/k-fold-cross-validation-python-example/amp/ (Accessed 30th November 2021)

[7] MathWorks. (2021) TreeBagger, *MathWorks.* Available at: https://uk.mathworks.com/help/stats/treebagger.html (Accessed 5th December 2021)

[8] Kaggle. (2020) Underfitting vs. Just right vs. overfitting in machine learning, *Kaggle*. Available at: https://www.kaggle.com/getting-started/166897 (Accessed 10th December 2021)