Choose Your Own Project

Harvard PH125.9x Data Science: Capstone

Frankie Inguanez

2022-09-28

Contents

List of Figures

1 Introduction

This report documents the research undertaken for the MovieLens project submission in part fulfillment of the Harvard PH125.9x Data Science: Capstone module by the author, Frankie Inguanez. The data set chosen is the Dry Beans data set archived on the UCI Machine Learning Repository and researched by Koklu and Ozkan (2020);. The objective of this data set is to correctly classify seven types of dry beans using 16 features which were extracted using computer vision. These features describe 12 dimension and 4 shape characteristics of each observation.

The authors of the published research considered a number of classifers with overall accuracies ranging from 87.92% to 93.13%. The classifiers considered were Multilayer perceptron, Support Vector Machine, k_Nearest Neighbors and Decision Trees. The authors of the published research also made use of 10-fold cross validation.

For the purpose of this project k-Nearest Neighbour, Random Forest and XGBoost classifiers were considered using 5-fold cross validation with hyper parameter tuning. The data set was split into training, validation and testing. This research obtained a best overall accuracy using random forests of 87.75% using XGBoost on the final hold-out (test) data set.

This document proceeds with an overview of the data analysis undertaken (exploration, cleaning, visualisation and modelling). All findings are presented in the Results section, with final remarks found in Conclusion section.

2 Analysis

The analysis process is made up of the following stages: Exploratory Analysis; Data cleaning; Data Visualisation; Data Modelling.

After downloading the data set it has been split with 20% for testing as the final hold-out data set. The remaining was split again with 20% for validation.

2.1 Exploratory Analysis

The data set has 17 features and after removing the hold-out data set a total of 10,885 observations were left. No missing values were found. The target variable is named Class, whilst all other variables are predictors. All predictors are continuous numeric variables, whilst the target should be a factor with 7 levels.

2.2 Data Cleaning

Two tasks were needed for data cleaning: first data type conversion of predictor variables to numeric and the target variable to a factor; secondly the normalization of values. The predictor variables had different range of values and thus needed to be normalized such that each had the same scale from 0 to 1.

2.3 Data Visualisation

Data visualisation can be generally viewed in two steps. As seen in Figure 1 the distribution of the target classes is not balanced. For all predictor variables box plots were generated to determine whether any distinguishing features were observed. As seen in Figure 2 the Bombay type of bean can be easily distinguished from features such as area, perimeter, axis length, convex area and equivalent diameter. Other features such as Extent and Solidity, shown in Figures 3 and 4, appear to be irrelevant since they do not offer any distinguishing information.

2.4 Data Modelling

For data modelling the data set was split to retain 20% for validation and 80% used to train the models. 5-fold cross validation repeated for 3 times was used on three classifiers each with hyperparameter tuning. The first model is K-Nearest Neighbour with tuning of k from 1 to 50 in increments of 2. The second model is Random Forest with a tune length of 15. The last model is XGBoost with 100 and 200 rounds and different tree depth in increments of 5.

3 Results

The performance of the KNN model is shown in Figure 5. The best configuration has a k of 13 with an overall accuracy of 0.91609 and a Kappa value of 0.89847.

The performance of the Random Forest is shown in Figure 6. The best configuration has a mtry value of 5 with an overall accuracy of 0.91976 and a Kappa value of 0.90292.

The performance of the XGBoost is shown in Figure 7. The best configuration has 100 rounds, max depth of 25 with an overall accuracy of 0.92251 and a Kappa value of 0.90625.

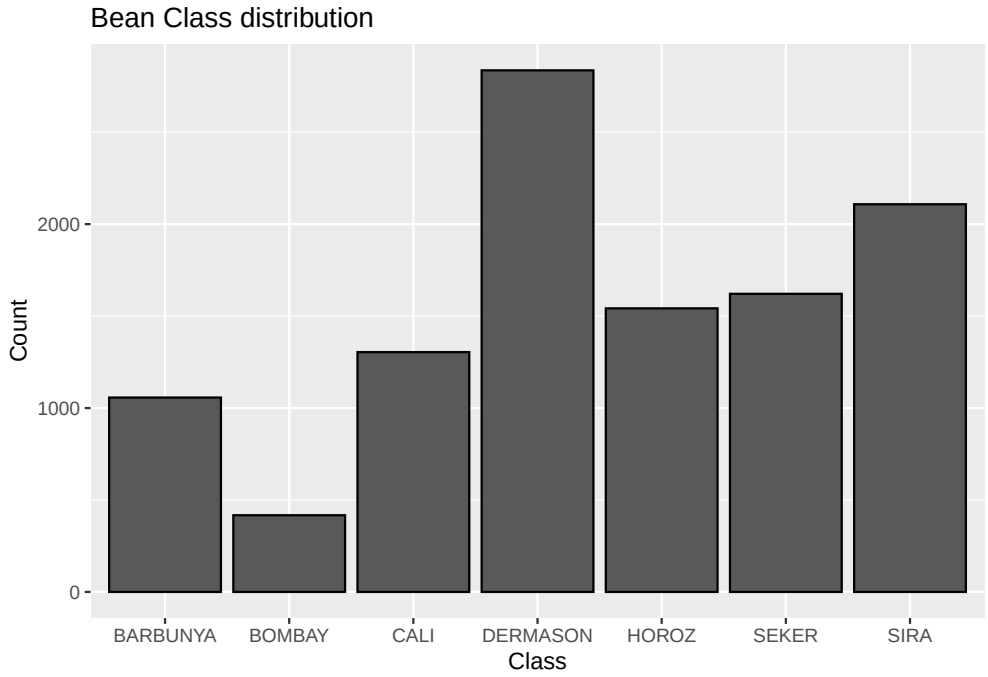

Figure 1: Target Class Distribution

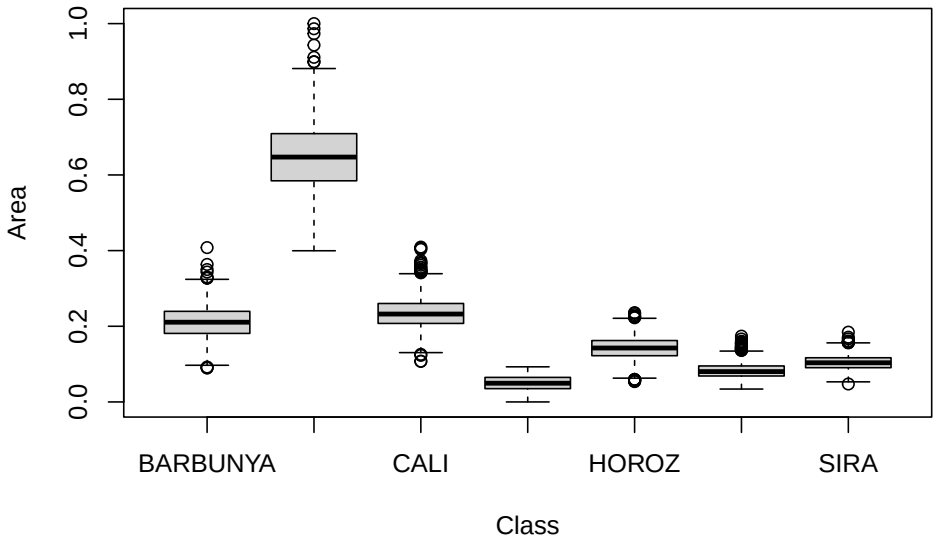

Figure 2: Box plot of Area per target class

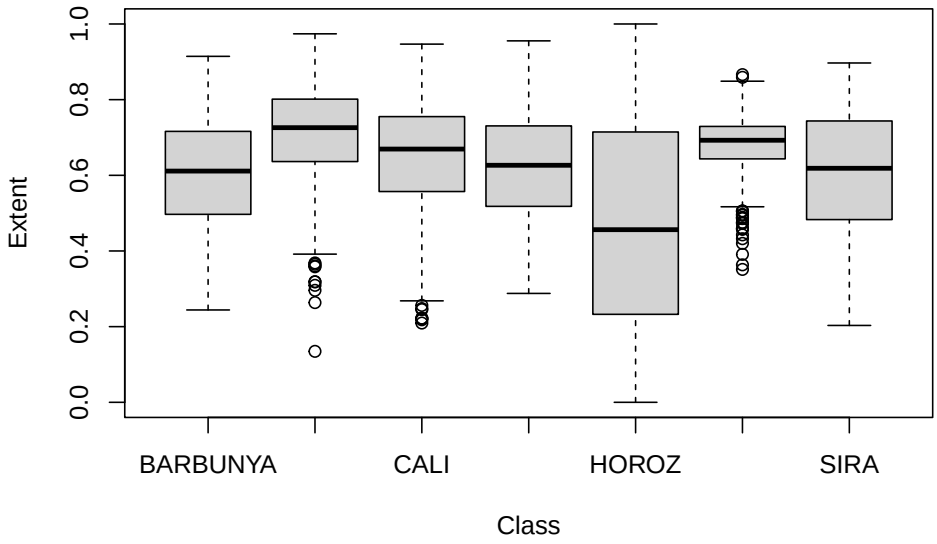

Figure 3: Box plot of Extent per target class

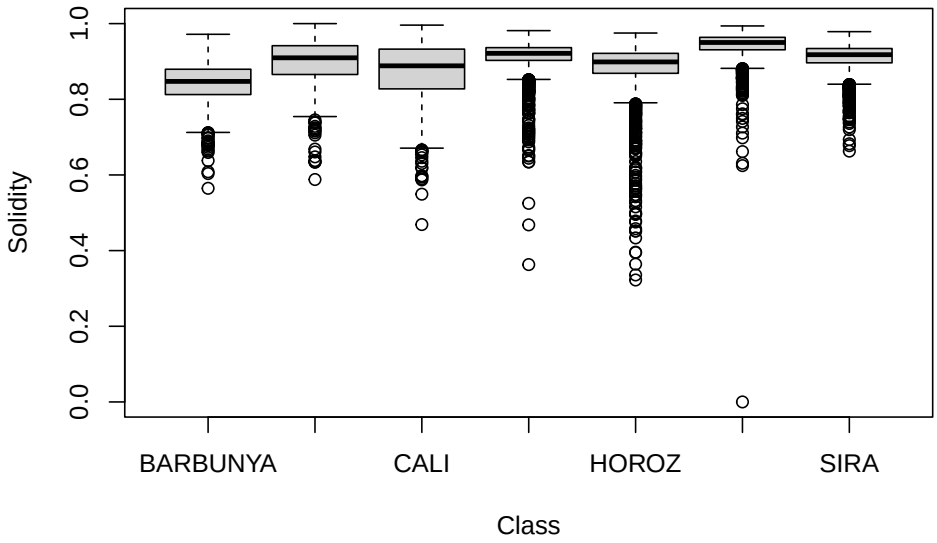

Figure 4: Box plot of Solidity per target class

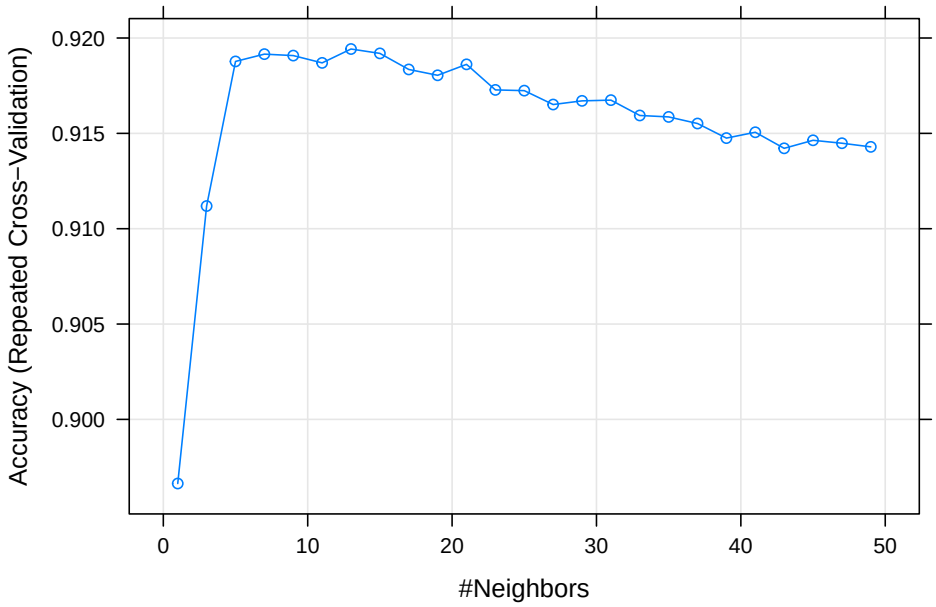

Figure 5: Training Performance of KNN via Cross Validation

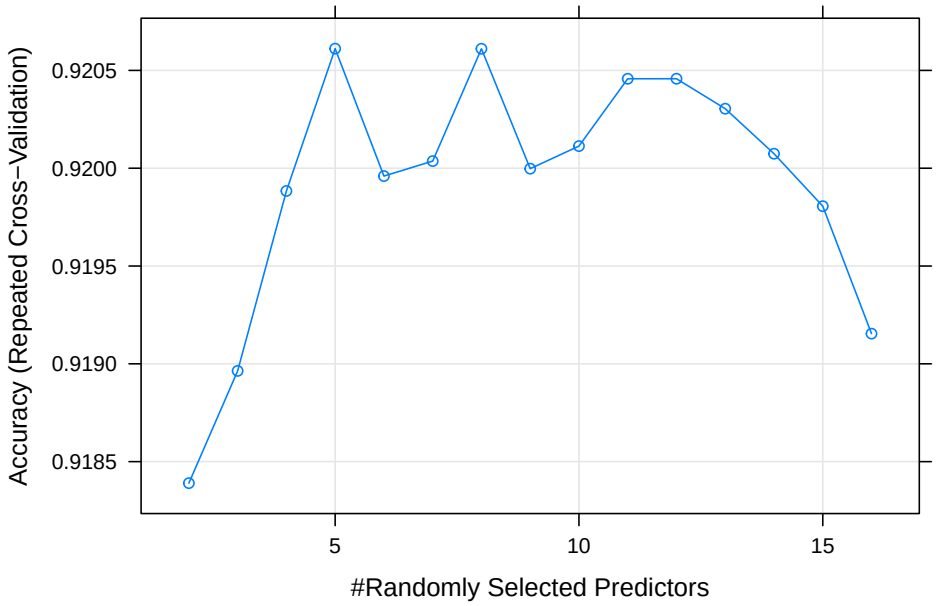

Figure 6: Training Performance of Random Forest via Cross Validation

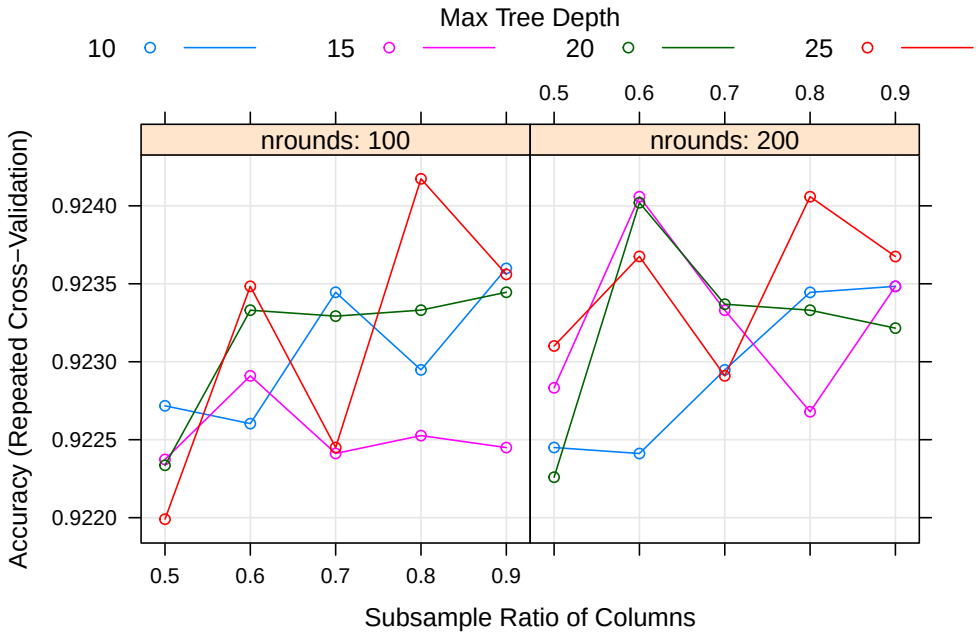

Figure 7: Training Performance of XGBoost via Cross Validation

Based on these results XGBoost was considered as the best model. Prior to applying this model on the final hold-out data set, the same cleaning processes applied on the training and validation data sets were applied to the test data set. The final overall accuracy obtained was of 0.88628 and a Kappa value of 0.86337. The metrics for each class based on the final model are shown in Table 1.

Table 1: Final Model Class Evaluation

	Sensitivity	Specificity	Precision	Recall	F1	Prevalence
Class: BARBUNYA	0.97358	0.96099	0.72881	0.97358	0.83360	0.09721
Class: BOMBAY	1.00000	1.00000	1.00000	1.00000	1.00000	0.03852
Class: CALI	0.82515	0.99125	0.92759	0.82515	0.87338	0.11959
Class: DERMASON	0.80986	0.98958	0.96477	0.80986	0.88055	0.26045
Class: HOROZ	0.91451	0.99359	0.95924	0.91451	0.93634	0.14160
Class: SEKER	0.97783	0.98319	0.91055	0.97783	0.94299	0.14894
Class: SIRA	0.86932	0.94631	0.79549	0.86932	0.83077	0.19369

4 Conclusion

This research considered a published data set by Koklu and Ozkan (2020); and archived on the UCI Machine Learning Repository. Three different classifier models were considered using 5-fold cross validation and hyper parameter tuning. A final overall accuracy of 0.88628 was achieved within range of the results obtained by the authors of the original research.

In order to improve on the obtained results a cross validation configuration similar to the original authors of 10-folds can be considered. This was not due to the computational resource limitations encountered in this research. More complex classifiers such as Neural Networks and Support Vector Machines could yield better results, yet requiring more computational resources. Final recommendation is to consider dimension reduction and matrix factorization.

References

Koklu, Murat, and Ilker Ali Ozkan. 2020. "Multiclass Classification of Dry Beans Using Computer Vision and Machine Learning Techniques." *Computers and Electronics in Agriculture* 174: 105507. https://doi.org/https://doi.org/10.1016/j.compag.2020.105507.