Performance Prediction of the Food Processing Companies in India

Name:	**Ardhana M Prabhash**
Registration No./Roll No.:	2220902
Institute/University Name:	IISER Bhopal
Program/Stream:	Economics
Problem Release date:	August 17, 2023
Date of Submission:	November 19, 2023

1 Introduction

This study mainly focuses on developing a supervised machine learning framework for the performance prediction and missing value imputation of the data associated with the Food Processing Companies in India (in terms of sales of products in millions) using the CMIE ProwessIQ dataset (starting from 2011 to 2021). The number of training instances of this data set is 10254 and features are 8. Therefore, the shape of the training data set is (10254,10) including the company name and year. And also test instances are 1140, so the shape is (1140,10). However, there are some missing values in the data. So, here I am trying to impute the missing values in the data set by using mean, median, and interpolation methods [1]. The major features of this company data are Total assets, Raw materials, stores & spares, Packaging and packing expenses, Power, fuel & water charges, Rent & lease rent, Repairs & maintenance, Total capital, and Miscellaneous expenditures. Based on this we want to predict the sales of the given companies.

2 Methods

The training dataset is split into train data and test data in 70:30 ratio.The missing values are imputed by using mean,median and interpolation methods.Then applied various regressors like Decision Tree Regressor, Random forest,Ridge Regressor, Linear Regressor, Lasso Regressor,and Adaptive Boosting in each imputed train data.The data had undergone feature selection and feature extraction. After that we will apply machine learning regression models for the training dataset to train the model. This train model will be then tested on test dataset and validation dataset for checking the accuracy of the model[2].

- 2.1 Hyperparameter Tuning

  Hyperparameters are specified parameters that can be used to tune the behaviour of a machine learning algorithm. These are initialized before the training and supplied to the model. To perform better and improve on the evaluation metric, hyper parameters are tuned by selecting the ideal values.To tune models, all feasible permutations of the hyper parameters for a specific model are used, and the best-performing ones are chosen for the test data.

- 2.2 Regression Methods

  Since this is a regression problem, here I used some existing regression methods. All the below models are tested and trained for their best performance using various cross validation techniques and fine tuning. The value of cv was fixed to 10 for cross-validation over each model. And also to evaluate the model Root Mean Squared(RMSE),Mean Square Error(MSE) and R^2 score are used. The regression methods used are:

- Decision Tree Regressor:

  Hyper-parameters used are:

  (random state=40), rgr parameters:rgr criterion:(squared error,friedman mse,poisson), rgr max features:(None,auto, sqrt, log2), rgr max depth:(10,40,45,60), rgr ccp alpha:(0.009,0.01,0.05,0.1) Values for criterion, max feature, and max depth are found through grid search, and keeping these values constant, the MSE,RMSE and R^2 is measured for different values of ccp alpha from 0 to 0.1.

- Random Forest Regressors: Hyperparameters used are: rgr criterion:(squared error,friedman mse,poisson), rgr n estimators:(30,50,100), 'rgr max depth':(10,20,30), 'rgr max features':(None, 'auto', 'sqrt', 'log2'), Values for criterion, max feature, and max depth are found through grid search. The values for these parameters, which gave the maximum value of R^2 measure, is chosen.

- Adaptive Booster: Hyperparameters used are: $rgr = AdaBoostRegressor(nestimators = 100)rgrparametersrgrbaseestimator : (be1, be2), rgrrandomstate : (0, 10)$, Values are found through grid search. The values for these parameters, which gave the maximum value of R^2 measure are chosen.

- Ridge regression: Hyperparameters used are: $rgr = Ridge(alpha = 1.0, positive = True)$ $rgrparametersrgrsolver : (auto, lbfgs)$, The values for these parameters, which gave the maximum value of R^2 score-measure, are chosen.

- Linear Regressor : Here MSE,RMSE and R^2 measure is found for different values of n and which gave the maximum value of R^2 score,is chosen. $rgrparameters\ rgrpositive : (True, False)$,

2.1 Github link

Performance Prediction of the Food Processing Companies in India

3 Experimental Setup

Based on the data , we have to find a suitable model or regression model from the algorithms that we had run using the trained model, like Linear Regression, Decision Tree, Ridge Regession, Random Forest etc using grid search cv and also by analyzing the evaluation matrices. Based on the scores of the different evaluation criteria we are selecting the best model and apply the model to the given test data.The evaluation criteria are:
The Mean Square Error function computes the mean squared error, which is a risk function corresponding to the expected value of the squared error loss or quadratic loss.

If $\hat{y}_i$ is the predicted value of the $i-th$ sample and y_i is the corresponding true value, then the Mean Squared Error (MSE) estimated over n samples and the RMSE is taking the square root of the mean squared error term. Both are defined as

$\text{MSE}(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

$RMSE = sqrt[(\sum(Pi–Oi)^2)/n]$

This calculation serves as a measure of the differences between values predicted by a model and the values observed in reality. The R^2 score function computes the Coefficient of Determination. It provides a measure of how well future samples are likely to be predicted by the model.

If $\hat{y}_i$ is the predicted value of the $i-th$ sample and y_i is the corresponding true value, then the score R^2 estimated over $n$samples is defined as

$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ The value of R^2 is in the range of 0 to 1. The closer the value toward 1, the more accuracy of the model.

Evaluation Matrices of Different Regressors

Measures such as MSE, RMSE, R^2 scores can be used for calculating the accuracy of the model. Here are the three different tables to show the experimental results. The table1 shows the performance of different regression methods using the Median imputation, the table2 gives the performance of different regression methods using the Mean imputation, and table3 shows the performance of different regression methods using the Interpolation imputation.[3]

Table 1: Performance Of Different Regressors Using Median imputation

Regressor	MSE	RMSE	R^2 Score
Adaptive Boosting	114867607.9	10717.6307	0.709445725
Decision Tree	68462384.53	8274.199933	0.826826389
Ridge Regression	72437233.31	8511.006598	0.816772124
Linear Regression	69888894.62	8359.957812	0.823218073
Random Forest	42557126	6523.582298	0.892352987

From this table, we can observe the relative performance of different regression algorithms.In this, Random Forest has the lowest MSE, RMSE, and highest R^2 Score(0.8923), suggesting it performs well on the model with Median imputation, followed by Decision Tree regression and Linear Regression. That is 89.23% of the data fits the regression model.

Table 2: Performance Of Different Regressors Using Mean imputation

Regressor	MSE	RMSE	R^2 Score
Adaptive Boosting	138167922.9	11754.48522	0.704172380023737
Decision Tree	95762324.7672109	9785.822641	0.794965864545915
Ridge Regression	101894700.951427	10094.29051	0.781835999
Linear Regression	100078023.878325	10003.9004332473	0.785725637
Random Forest	69816909.0271944	8355.65132273926	0.850516895686635

From this table, we can observe the relative performance of different regression model with Mean imputation. In this, Random Forest has the lowest MSE, RMSE, and highest R^2 Score of 0.8505. That is, 85.05% of the data fit the Random forest model.

Table 3: Performance Of Different Regressors Using Interpolation imputation

Regressor	MSE	RMSE	R^2 Score
Adaptive Boosting	152038066.5	12330.37171	0.674475389
Decision Tree	138737193.2	11778.67536	0.702953531
Ridge Regression	111186720.5	10544.51139	0.761941106
Linear Regression	111191612.7	10544.74337	0.761930632
Random Forest	63156257.4	7947.091128	0.864777838

From this table, we can observe the relative performance of different regression models with Interpolation imputation. In this also, Random Forest has the lowest MSE and RMSE values with 0.8647 as the R^2 Scores, which is comparatively higher than all other models.

4 Results and Discussion

The best set of parameters for the test and train data are Random Forest Regressor(max depth=10, max features=None, n estimators=30))]) with test size as 0.3 with median imputed data, which is

giving high R^2 score and minimum of the MSE and RMSE values. When we are applying the same algorithm to the interpolated data,median imputed data and mean imputed data, the median imputed data gives the highest R^2 score with the Random Forest algorithm. So the Random Forest algorithm works well and it fits the data (the goodness of fit).So I choose the the Random forest regression model for predicting the labels of the given test data , by imputing the missing values using the Median imputation in the given test data.

5 Conclusion

From this study, we can analyze the performance of the food processing companies in India, and also by using the supervised machine learning algorithm we can predict the performance of various companies and impute the missing values in the data. By considering different algorithms, the Random forest model gives better predictions of the target variables.

5.1 Scope

For further development of the study, we can apply the Support Vector Machine (SVM) Regression for the prediction. Here I do not use this SVM Regression because of it's computational time.

References

[1] Wei-Chao Lin and Chih-Fong Tsai. Missing value imputation: a review and analysis of the literature (2006–2017). *Artificial Intelligence Review*, 53:1487–1509, 2020.

[2] Luise Patricia Lago. Imputation of household survey data using mixed models. 2015.

[3] Alireza Farhangfar, Lukasz A Kurgan, and Witold Pedrycz. A novel framework for imputation of missing values in databases. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 37(5):692–709, 2007.