[CAPSTONE Project Notes –2]

[Supply Chain Management

(2022-2023)

TEXAS McCombs
The University of Texas at Austin
McCombs School of Business

]

Course Name

Post Graduate Program in
Data Science and Business Analytics

Batch Id

(PGP-DSBA-June22C)

Submitted by

Jayant Singh

Email Id: jayant101169@gmail.com

Literature Review

Inventory is the stock of goods held for doing business. It can be any company's most valuable asset. It is critical for each company to manage its inventory effectively with no excess stock stored and meeting the client requirements. Inventory management is primarily specifying the placement of stocked goods. The problem we are discussing here is about the instant noodles company which is experiencing inventory cost loss due to the inadequate supply management The management intends to maximise the amount of supply in every warehouse across the entire nation. The goal of this project is to create a model utilising past data that will establish the ideal weight of the product to be delivered to the warehouse on each occasion.

Visualization is the window to the data. It is essential in understanding the given records, attributes and the relationships. The Python programming language will be used to conduct this investigation using univariate and bivariate analysis. Dropping records containing missing values can lead to loss of vital information. Thus, they have been imputed using suitable measures. There are many possibilities for data analysis, making it challenging to choose which approach and machine learning model to employ because the model's effectiveness depends on the parameters included in the data. Data modelling is done using different regression models and tuning techniques.

The models are evaluated used accuracy score and RMSE values and chose the best model for further analysis. Important features that affect the optimum weight of the product to be shipped are Warehouse that are established atleast 5 years ago and its importance increases with the age of warehouse, Warehouse breakdown, refilling & transport issue. If business focus on these features and minimize the accidents happened and transport issues and increases the refilling time for older warehouses, based on this optimum weight to be shipped can be determined.

Table of Contents

1.INTRODUCTION

Problem Statement

A FMCG company has entered into the instant noodles business two years back. Their higher management has notices that there is a miss match in the demand and supply. Where the demand is high, supply is pretty low and where the demand is low, supply is pretty high. In both the ways it is an inventory cost loss to the company; hence, the higher management wants to optimize the supply quantity in each and every warehouse in entire country. The objective of this exercise is to build a model, using historical data that will determine an optimum weight of the product to be shipped each time to the warehouse. (Dataset & Data Dictionary Refer Appendix)

Purpose of Study

The company is facing inventory cost loss due to the poor supply management of the instant noodles. The management wants to optimize the supply quantity in each and every warehouse in entire country. The objective of this project is to build a model, using historical data that will determine an optimum weight of the product to be shipped each time to the warehouse. In this problem PORDUCT_WG_TON is the target variable. With a lot of options available to analyse data, it is very difficult to decide which method and machine learning model to use since the performance of the model vary on the parameters available in the data.

This project aims to compare different popular machine learning classifiers, and measure their performance to find out which machine learning model performs better. This exploration will be carried out using the Python programming language through univariate and bivariate analysis. This exploration will be carried out using the Python programming language through univariate and bivariate analysis. The presence of outliers interferes with the correct modelling and interpretation of the data. Thus, the outliers (extreme values) are identified and replaced. Since the dataset used is related to supply chain important parameters are identified and the machine learning models are trained with the dataset for detection of the optimum weight. The study helps to analyse/optimize the supply quantity at each warehouse in the country & thereby determining the advertising strategies & campaigns for specific pockets

.

2. MODEL BUILDING

Feature Selection

Feature Selection is the method of reducing the input variable to your model by using only relevant data and getting rid of noise in data. Feature selection can be done in multiple ways but there are broadly 3 categories:

1. Filter Method
2. Wrapper Method
3. Embedded Method

1. Filter Method

In this method you filter and take only the subset of the relevant features. The model is built after selecting the features. The filtering here is done using correlation matrix and it is most commonly done using Pearson correlation and VIF

Variance Inflation Factor (VIF)

The Variance Inflation Factor (VIF) measures the severity of multicollinearity in regression analysis. Collinearity is the state where two variables are highly correlated and contain similar information about the variance within a given dataset.

Here we perform the VIF and will remove the variables one by one which are highly correlated and proceeding with the other variable for modelling. We set the threshold to 10, as we wish to remove the variable for which the remaining variables explain more than 90% of the variation. One can choose the threshold other than 10. (it depends on the business requirements). Initial VIF values for all variables (Refer Appendix 5) After few iterations the VIF, **below are the variable selected to build the model**.

	VIF_Factor	Features
0	9.58	Competitor_in_mkt
1	7.27	distributor_num
2	7.15	dist_from_hub
3	6.23	wh_breakdown_l3m
4	5.99	govt_check_l3m
5	4.82	WH_regional_zone_Zone_6
6	3.97	WH_regional_zone_Zone_5
7	3.84	WH_regional_zone_Zone_4
8	3.69	num_refill_req_l3m
9	3.07	electric_supply
10	3.01	WH_regional_zone_Zone_2
11	2.96	WH_regional_zone_Zone_3
12	2.26	WH_capacity_size_Small
13	2.25	AgeGroup_10to15
14	2.17	approved_wh_govt_certificate_Aplus
15	1.99	AgeGroup_15to20
16	1.98	AgeGroup_20to25
17	1.98	zone_West
18	1.98	temp_reg_mach
19	1.96	approved_wh_govt_certificate_C
20	1.96	AgeGroup_5to10
21	1.92	wh_owner_type_Rented
22	1.88	approved_wh_govt_certificate_Bplus
23	1.86	approved_wh_govt_certificate_B
24	1.81	zone_South
25	1.62	transport_issue_l1y
26	1.16	flood_impacted
27	1.10	Location_type_Urban
28	1.08	flood_proof

Table 3 VIF for selected variables

Train – Test Split

After selecting the variable for model building, we performed the train test split.

- X= Copy all the predictor variables & y= target into the y dataframe().
- Splitting the X and y into training and test set in 70:30 ratio with random_state=1

```
The dimension of X_train is (17500, 29)
The dimension of X_test is (7500, 29)
```

Scaling

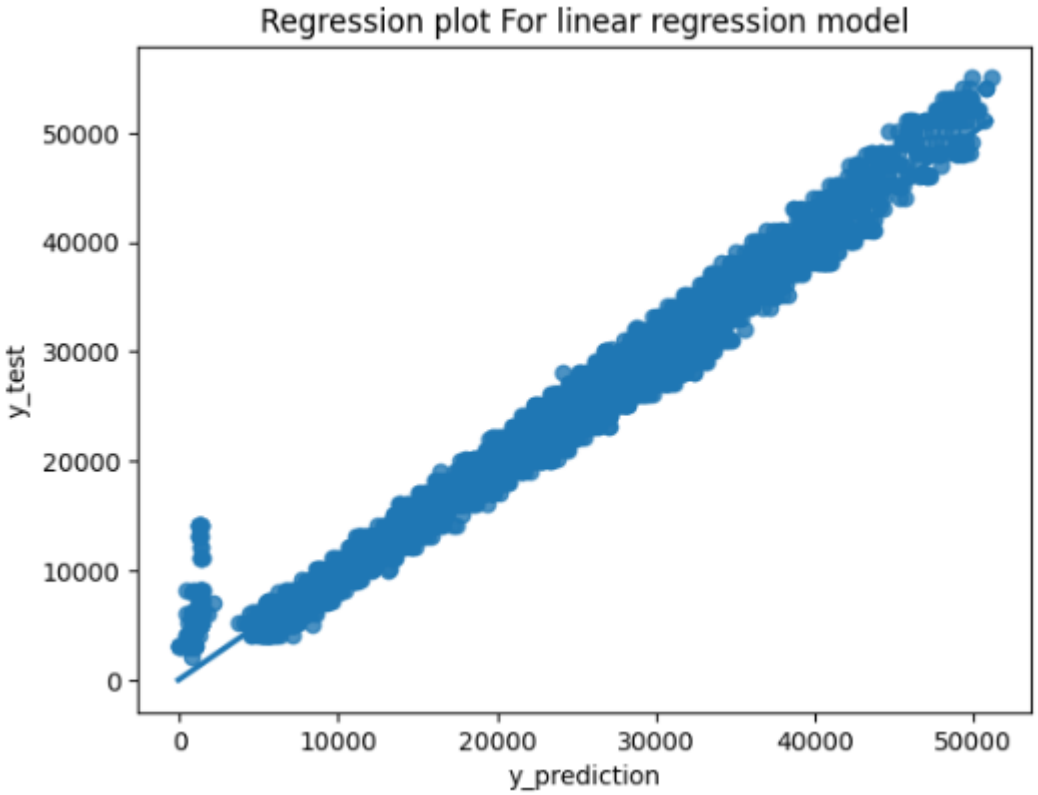
Data standardization is the process where using which we bring all the data under the same scale. Here, we are building a model, to predict optimum weight of the product to be shipped each time to the warehouse. In this case we are expected to build model using LinearRegression, LDA, Ridge, Lasso, ANN etc. So, we are scaling the data (x_train_scaled,x_test_scaled) and will use this scaled data to perform the models were scaling is necessary.

Models

Since this a supervised regression problem will be performing some of the regression models below. Two metrics that statisticians often use to quantify how well a model fits a dataset are the root mean squared error (RMSE) and the R-squared (R^2).

Base models used for model building

Linear Regression: Linear Regression is the supervised Machine Learning model in which the model finds the best fit linear line between the independent and dependent variable. It is mostly used for finding out the relationship between variables and forecasting.



As we can see from the above regression plot that most of the data points lie on our regressor line hence our Linear regressor model is performing really good.

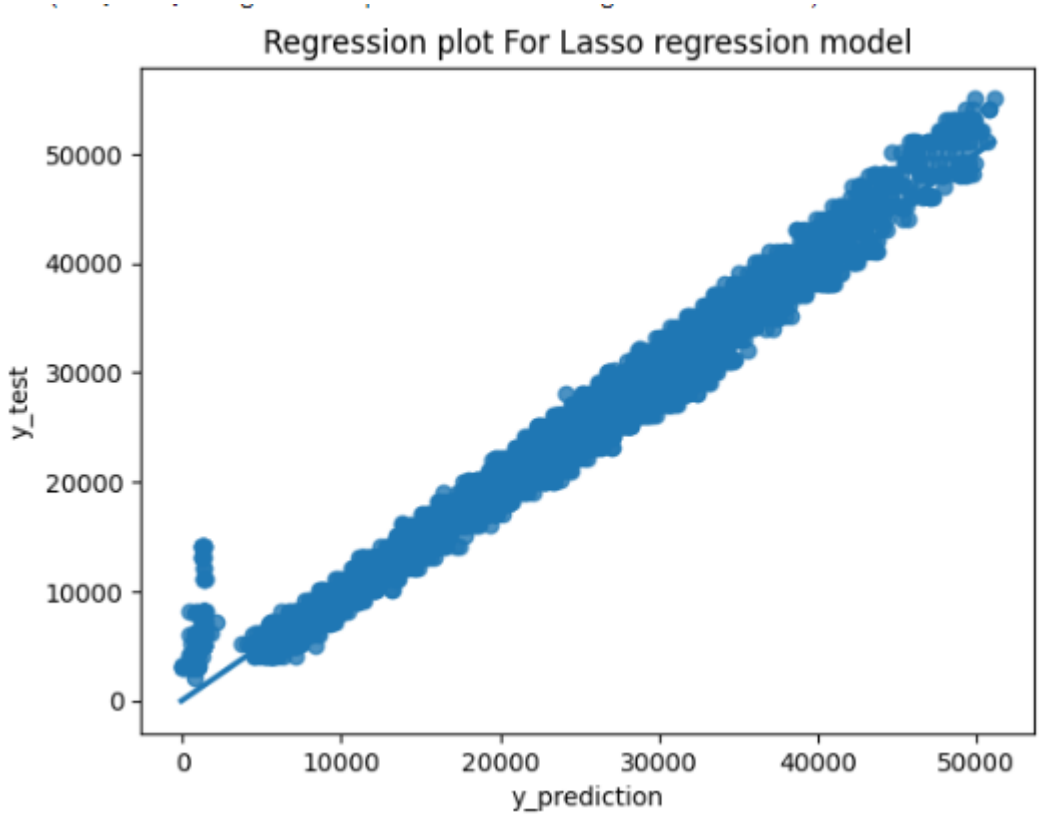
Our Linear Regressor Model is Giving out R2 score of 0.97 i.e 97% that is good as the mse and rmse is also low

Test results:

	RMSE	Accuracy Score
Train	6309.61	0.97
Test	6116.42	0.997

Coefficients of Linear Regression model (refer Appendix 6)

Lasso Regression: Lasso regression is like linear regression; Linear regression gives you regression coefficients as observed in the dataset. The lasso regression allows you to shrink or regularize these coefficients to avoid overfitting and make them work better on different datasets



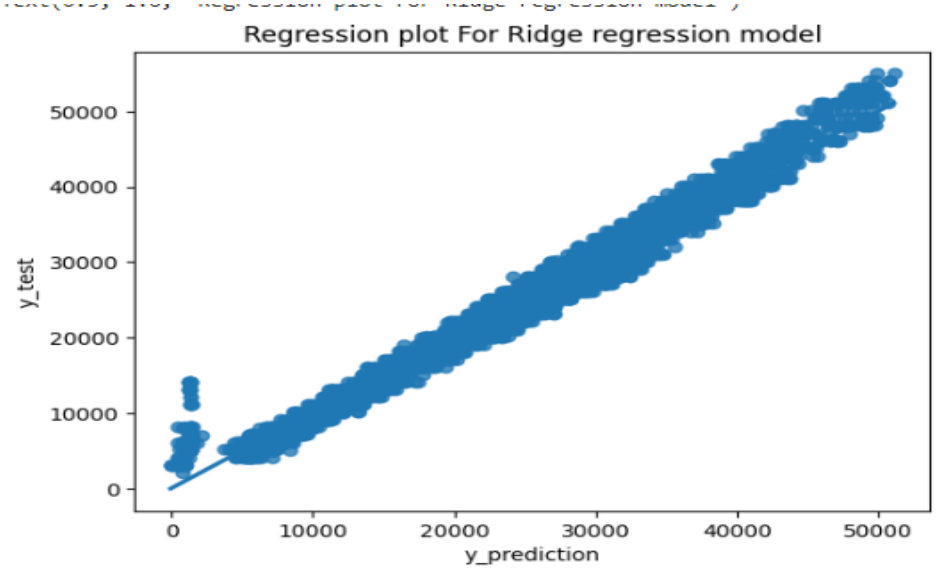
As we can see from the above regression plot that most of the data points lie on our regressor line hence our Lasso regressor model is performing really good.

Test results:

	RMSE	Accuracy Score
Train	6309.61	0.97
Test	6116.42	0.97

Coefficients of Lasso Regression model (refer Appendix 7)

Ridge Regression: Ridge regressor is basically a regularized version of a Linear Regressor. The regularized term has the parameter 'alpha' which controls the regularization of the model.



As we can see from the above regression plot that most of the data points lie on our regressor line hence our Ridge regressor model is performing good

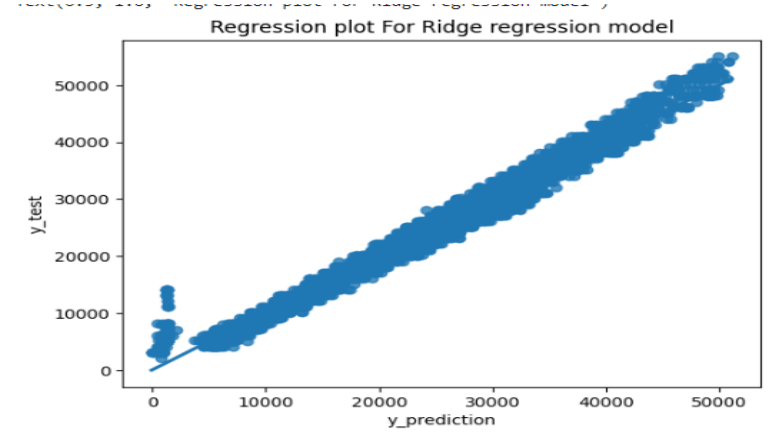
Our Ridge Regressor Model is Giving out R2 score of 0.97 i.e 97% that is good as the mse and rmse is also low as compared to lasso regression.

. Test Results:

	RMSE	Accuracy Score
Train	6309.61	0.97
Test	6116.42	0.97

Coefficients of Ridge Regression model (refer Appendix 8)

Decision Tree Regressor: Decision tree builds regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed.



Test Results:

	RMSE	Accuracy Score
Train	0.00	1.00
Test	8279.36	0.99

Important features of Decision tree Regressor model (refer Appendix 9)

Insights:

- We see the results are almost similar for linear, lasso and ridge regression. Since the features are selected using VIF method, lasso and ridge are performing same as linear regression
- Decision tree and Random Forest's nonlinear nature gives better results than linear regression. Decision tree's accuracy shows that it is overfitting, so does random forest's results show
- Linear regression and other methods can understand only linear relationships, to understand non-linear relationships ANN works better. Looking at the result ANN performs better than Linear and regularization methods. Real life data is supposed to have complex non-linear relationships, that's why ANN is giving better results than linear model
- From the models we could see that warehouse established year, number of refills, warehouse breakdown, distribution from hub etc are few of the importance features effecting the optimum weight shipment
- We can use grid search to tackle this problemLater, we will try to tune the models and will see whether the model performance improves

Based on R2 scores from our Linear, Ridge, Lasso and Decision Tree Regressor Model we can conclude that Decision Tree Regressor model outperforms the other models as it produces an accuracy of 100% percent on our dataset that is really remarkable for a model to achieve it. We have tuned various parameters like random state & max_depth of our Decision tree model to achieve better results by gradiently decreasing both paramater

5. MODEL VALIDATION

Two metrics there are use to validate how well a model fits a dataset are the root mean squared error (RMSE) and the R-squared (R^2), which are calculated as follows:

- **RMSE**: A metric that tells us how far apart the predicted values are from the observed values in a dataset, on average. The lower the RMSE, the better a model fits a dataset
- **R^2**: A metric that tells us the proportion of the variance in the response variable of a regression model that can be explained by the predictor variables. This value ranges from 0 to 1. The higher the R^2 value, the better a model fits a dataset.

Final Model Comparison

Models	Train RMSE	Test RMSE	Training Score	Test Score
Ridge Regression	6309.62	6113.95	0.71	0.72
Lasso Regression	6309.63	6113.68	0.71	0.72
Linear Regression	6309.61	6113.80	0.71	0.72
Decision Tree Regressor	0.00	8160.90	1.00	0.50
Random Forest Regressor	2257.56	5941.59	0.96	0.73
ANN Regressor	5946.95	5838.63	0.74	0.74
AdaBoostRegressor	6847.90	6724.97	0.65	0.66
GradientBoostingRegressor	5974.91	5830.55	0.74	0.74
BaggingRegressor	2661.88	6164.40	0.95	0.71
XGBRegressor	4364.92	6001.27	0.86	0.73
Ridge Regression with GridSreachCV	6522.99	6366.76	0.69	0.69
Lasso Regression with GridSreachCV	6310.90	6115.72	0.71	0.72
Linear Regression with GridSreachCV	6309.61	6116.42	0.71	0.72
Decision Tree Regressor with GridSreachCV	5529.03	5944.87	0.77	0.73
Random Forest Regressor with GridSreachCV	4998.51	5753.39	0.82	0.75
ANN Regressor with GridSreachCV	6270.82	6080.12	0.71	0.72
AdaBoostRegressor with GridSreachCV	6702.77	6521.56	0.67	0.68
GradientBoostingRegressor with GridSreachCV	5536.09	5733.53	0.77	0.75
BaggingRegressor with GridSreachCV	6289.33	6810.36	0.71	0.65
XGBRegressor with GridSreachCV	5364.44	5724.93	0.79	0.75

Table 7 Model Comparison

Based on R2 scores from our Linear, Ridge, Lasso and Decision Tree Regressor Model we can conclude that Decision Tree Regressor model outperforms the other models as it produces an accuracy of 100% percent on our dataset that is really remarkable for a model to achieve it. We have tuned various parameters like random state & max_depth of our Decision tree model to achieve better results by gradiently decreasing both paramater

Feature Importance

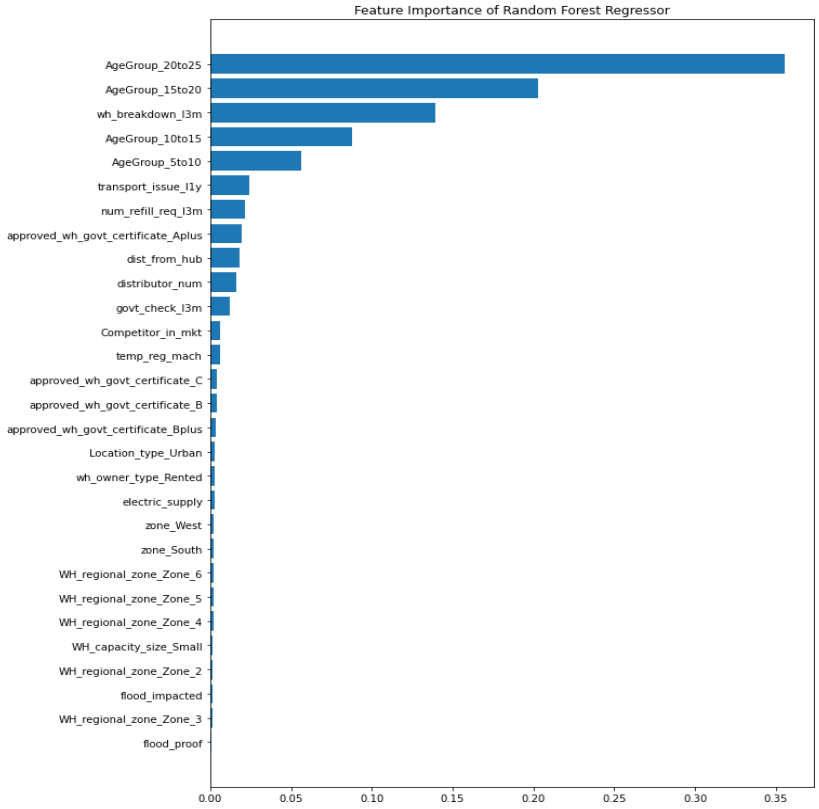


Figure 12 Feature Importance of Random Forest

These features have maximum effect on optimum weight (product_wg_ton):

- Warehouse Breakdown in last 3 months
- Transport issue
- Refilling time in last 3 months
- Warehouse established year

6. FINAL INTERPRETATION / RECOMMENDATION

- Since this is regression problem, we have tried out different regression models to confirm which performs well and gives the best accuracy
- To handle overfitting, we performed hyperparameter tuning using GridsearchCV.
- Comparing all models, we have obtained the best results for Random Forest regressor → Random Forest regressor is giving 82% accuracy In Train & 75% in Test with RMSE value of Train as 4998.51 & Train as 5753.39

The warehouse breakdowns due to both internal & external factors results on itsinventory management & leading manufactures

- Accident or Product stolen shall be another factor which can lead to optimum weight mismatch at the time of delivery, resulting in supply constraints
- Delay in stock refilling hampers reduced stock during high demand times
- Features that affect product_wg_ton which specifies the optimum weight of the product to be shipped are Warehouse that are established at least 5 years ago and its importance increases with the age of warehouse.

Recommendations

- Set up a governing council that offers a clear strategy for functionality and efficiency, thus reducing Warehouse breakdown factors. The council's aim is to give directions and align the supply chain strategy with the company's core goals. The council helps in removing barriers within the organization.
- Review policies and procedures to ensure efficiency and compliance. It also helps avoid bottlenecks in the supply chain, streamline operations and mitigate the risks of theft and fraud. Regular reviews help in identifying different risk elements and estimating their financial impact.
- Include demand planning and forecasting to improve refilling. Incorporate warehouse operations that are proficient in managing inventory and accurate inventory records. This helps to know whenever there is a refill required immediately and won't affect supply.
- Need to perform frequent audits upon warehouse operation standards. Use technology to improve the supply chain. Review all the existing processes that are affecting the inventory management. Determine the areas where implementing technology could improve the processes. The right strategies have the potential to transform your supply chain and increase revenue

East zone has fewer warehouses, retail outlets and distributors. It has much higher number of competitors yet has the same product demand as other zones. This might bea result of the popularity of the product in this region, encouraging the marketing department to pay greater attention to this region

1. HyperTuning Parameters

Hypertune Parameters	Description
alpha	Constant that multiplies the L2 term, controlling regularization strength. alpha must be a non-negative float
Tol	Precision of the solution
cv	the cross-validation generator or an iterable, in this case, there is a 10-fold cross validation
Solver	This parameter represents which algorithm to use in the optimization problem
Max_depth	The maximum depth of the tree. If None, then nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples.The max_depth value should be adjusted according no to high/low to avoid overfitting/underfitting
min_samples_split	int, float, optional (default=2). min_samples_split is used to control over-fitting. depending on the level of underfitting or overfitting
min_samples_leaf	int, float, optional (default=1). The minimum number of samples required to be at a leaf node. min_samples_leaf is also used to control over-fitting by defining that each leaf has more than one element.