Table of Contents

[MODEL BUILDING 2](#_Toc98087774)

[Data Split 2](#_Toc98087775)

[Using default parameters 2](#_Toc98087776)

[Decision Tree 2](#_Toc98087777)

[Parameters 2](#_Toc98087778)

[Results 2](#_Toc98087779)

[Random Forest 2](#_Toc98087780)

[Parameters 2](#_Toc98087781)

[Results 3](#_Toc98087782)

[Artificial Neural Network 3](#_Toc98087783)

[Parameters 3](#_Toc98087784)

[Results 3](#_Toc98087785)

[Ensemble Modelling 3](#_Toc98087786)

[Default 3](#_Toc98087787)

[Logistic Regression 3](#_Toc98087788)

[LDA 4](#_Toc98087789)

[SMOTE 4](#_Toc98087790)

[Logistic Regression 4](#_Toc98087791)

[LDA 4](#_Toc98087792)

[ADABoosting Classifier 5](#_Toc98087793)

[Logistic Regression 5](#_Toc98087794)

[LDA 5](#_Toc98087795)

[Gradient Boosting 6](#_Toc98087796)

[Logistic Regression 6](#_Toc98087797)

[LDA 6](#_Toc98087798)

# MODEL BUILDING

We have built various models and measured their accuracy in this step.

## Data Split

The data is split into train and text data with 70% data for train and rest for test.

X are all the input variables and y is the output variable (target variable)

In our case y is the Taken product column.

The data after split has

X Train = 7318 rows and 15 columns

Y train = 7318 rows and 1 column

X test = 3137 rows and 15 columns

Y test = 3137 rows and 1 column

## Using default parameters

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Train RMSE | Test RMSE | Training Score | Test Score |
| Decision Tree | 0.000000 | 0.149380 | 1.000000 | 0.840834 |
| Random Forest | 0.045210 | 0.114939 | 0.984719 | 0.905768 |
| ANN | 1.153860 | 1.203903 | 8.953617 | 9.338296 |

## Decision Tree

### Parameters

Using various model parameters, we have optimized the model at

Max depth=15

Min samples split=15

Min samples leaf=3

### Results

Train RMSE = 0.129928

Test RMSE = 0.200836

Train Score = 0.873794

Test Score = 0.712295

## Random Forest

### Parameters

Using various model parameters, we have optimized the model at

max depth=10

max features=6

min samples leaf= 3

min samples split= 30

n estimators= 500

### Results

Train RMSE = 0.215941

Test RMSE = 0.246587

Train Score = 0.651386

Test Score = 0.566284

## Artificial Neural Network

### Parameters

Using various model parameters, we have optimized the model at

Activation = ‘tanh',

hidden layer sizes = (500)

solver = 'sgd'

max iter = 10000

### Results

Train RMSE = 0.372087

Test RMSE = 0.378887

Train Score = -0.035055

Test Score = -0.023964

## Ensemble Modelling

We have used various ensemble methods to make the models more accurate.

The models built are Logistic Regression and Linear Discriminant Analysis (LDA).

## Default

### Logistic Regression

Model Score = 0.8313675486133248

Confusion matrix =

|  |  |
| --- | --- |
| 2608 | 0 |
| 529 | 0 |

Classification report =

precision recall f1-score support

0 0.83 1.00 0.91 2608

1 0.00 0.00 0.00 529

accuracy 0.83 3137

macro avg 0.42 0.50 0.45 3137

weighted avg 0.69 0.83 0.75 3137

### LDA

Model Score = 0.8498565508447561

Confusion matrix =

|  |  |
| --- | --- |
| 2527 | 81 |
| 390 | 139 |

Classification report =

precision recall f1-score support

0 0.87 0.97 0.91 2608

1 0.63 0.26 0.37 529

accuracy 0.85 3137

macro avg 0.75 0.62 0.64 3137

weighted avg 0.83 0.85 0.82 3137

## SMOTE

### Logistic Regression

Model Score = 0.6168313675486133

Confusion matrix =

|  |  |
| --- | --- |
| 1583 | 1025 |
| 177 | 352 |

Classification report =

precision recall f1-score support

0 0.90 0.61 0.72 2608

1 0.26 0.67 0.37 529

accuracy 0.62 3137

macro avg 0.58 0.64 0.55 3137

weighted avg 0.79 0.62 0.66 3137

### LDA

Model Score = 0.716289448517692

Confusion matrix =

|  |  |
| --- | --- |
| 1840 | 767 |
| 123 | 406 |

Classification report =

precision recall f1-score support

0 0.94 0.71 0.81 2608

1 0.35 0.77 0.48 529

accuracy 0.72 3137

macro avg 0.64 0.74 0.64 3137

weighted avg 0.84 0.72 0.75 3137

## ADABoosting Classifier

### Logistic Regression

Model Score = 0.8380618425247052

Confusion matrix =

|  |  |
| --- | --- |
| 2628 | 1 |
| 506 | 1 |

Classification report =

precision recall f1-score support

0 0.84 1.00 0.91 2629

1 0.50 0.00 0.00 508

accuracy 0.84 3137

macro avg 0.67 0.50 0.46 3137

weighted avg 0.78 0.84 0.76 3137

### LDA

Model Score = 0.8559132929550526

Confusion matrix =

|  |  |
| --- | --- |
| 2552 | 77 |
| 375 | 133 |

Classification report =

precision recall f1-score support

0 0.87 0.97 0.92 2629

1 0.63 0.26 0.37 508

accuracy 0.86 3137

macro avg 0.75 0.62 0.64 3137

weighted avg 0.83 0.86 0.83 3137

## Gradient Boosting

### Logistic Regression

Model Score = 0.8380618425247052

Confusion matrix =

|  |  |
| --- | --- |
| 2628 | 1 |
| 507 | 1 |

Classification report =

precision recall f1-score support

0 0.84 1.00 0.91 2629

1 0.50 0.00 0.00 508

accuracy 0.84 3137

macro avg 0.67 0.50 0.46 3137

weighted avg 0.78 0.84 0.76 3137

### LDA

Model Score = 0.8559132929550526

Confusion matrix =

|  |  |
| --- | --- |
| 2552 | 77 |
| 375 | 133 |

Classification report =

precision recall f1-score support

0 0.87 0.97 0.92 2629

1 0.63 0.26 0.37 508

accuracy 0.86 3137

macro avg 0.75 0.62 0.64 3137

weighted avg 0.83 0.86 0.83 3137