|  |  |
| --- | --- |
| A picture of a winding road and trees  Social Media Advertising customer prediction | Abstract  Analysis on the Social Media Advertising  Arpita B |

Contents

[Building of Models: 2](#_Toc83468055)

[Testing of Predictive Models against test data: 3](#_Toc83468056)

[Interpretation of the Models: 4](#_Toc83468057)

[Ensemble modelling: 5](#_Toc83468058)

[Model tuning measures: 5](#_Toc83468059)

[Interpretation of the most optimum model: 6](#_Toc83468060)

# Building of Models:

The data considered here is taken from Laptop on the users’ social media profile. As part of data cleaning process all the null values have been replaced. Using Label encoding, categorical values replaced with numerical values.

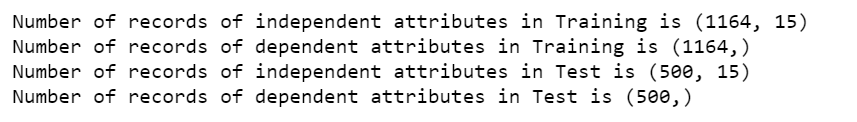
On further analysis it has been found the target variable “Taken\_Product” here is provided which indicates whether the users’ are buying the offers the site is sharing. The value of target variable is either “Yes” or “No”. Since the target variable can be considered for two values the current data set can be a classification type data and all models that support classification can be used. There is a high difference found in the target variable between records of “Taken\_Product” yes or no. Considering the data to be imbalance in that respect, SMOTE has been used.

The entire data set is separated into two sets Test and Train Data into 70:30 ratio.

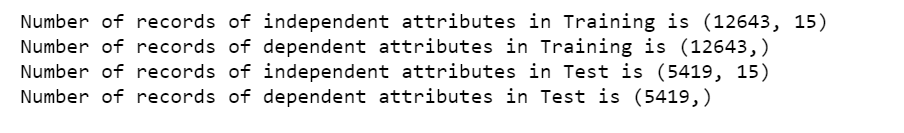
Further the attributes are separated into two different variable X\_train/X\_test for independent attributes and y\_train/y\_test for dependent attributes.

Device column has been dropped as the data is taken from a single device only.

Following is the distribution of test and train data for Laptop:



Following is the distribution of test and train data for **Laptop**:



The following models have been used for the analysis of the correct prediction of the classification of target variable for data collected from Laptop:

* Logistic Regression
* Liner discriminant Analysis
* Decision Tree
* Random Forest
* AdaBoosting
* Bagging

The following models have been used for the analysis of the correct prediction of the classification of target variable for data collected from Mobile:

* KNN
* Decision Tree
* Random Forest
* AdaBoosting
* Bagging

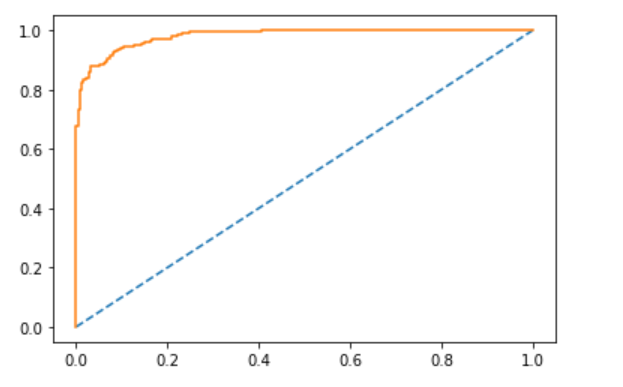
Since in case of Mobile collected data KNN model has been used which is a distance based model additional step of scaling the data to bring the attributes in a common range has been performed.

# Testing of Predictive Models against test data:

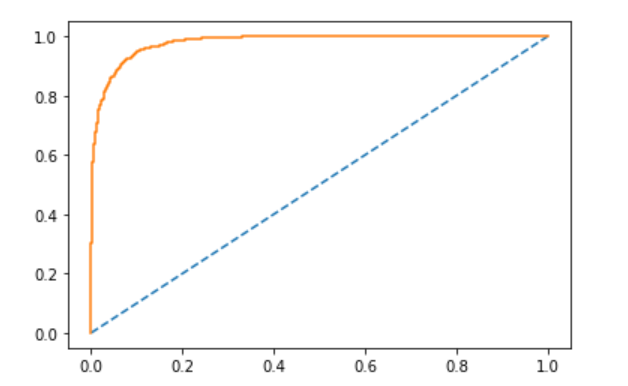
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Device: Laptop** |  |  |  |  |
| **Model Names** | **Overall Accuracy** | **Recall %** | **F1 Score %** | **AUC Score %** |
| Logistic Regression | 65 | 66 | 65 | 69.7 |
| Liner discriminant Analysis | 72 | 70 | 72 | 78.5 |
| Decision Tree | 78 | 74 | 77 | 85.8 |
| Random Forest | 89 | 85 | 89 | 97.8 |
| AdaBoosting | 84 | 81 | 83 | 91.5 |
| Bagging | 91 | 89 | 91 | 98.1 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Device: Mobile** |  |  |  |  |
| **Model Names** | **Overall Accuracy** | **Recall %** | **F1 Score %** | **AUC Score %** |
| KNN | 78 | 93 | 81 | 88.4 |
| Decision Tree | 85 | 86 | 85 | 92.4 |
| Random Forest | 87 | 89 | 88 | 95.5 |
| AdaBoosting | 80 | 80 | 80 | 87.8 |
| Bagging | 92 | 95 | 93 | 98 |

The above chart displays the model’s performance against the test set data for individual devices Laptop and Mobile:



ROC AUC curve of the model with highest accuracy is shown for device Laptop.



ROC AUC curve of the model with highest accuracy is shown for device Mobile.

**Model**: Bagging

While overall accuracy is a strong indicator on how well the Model is performing with test data, here in this case the Recall and F1 Score plays an equal level of importance. High Recall rate shows how many of the customers identified to have taken the product has taken in actual. So, high rate shows correct prediction of the same and that offers less risk to the business and help determine the company the correct set of customers.

# Interpretation of the Models:

**Recall**(Sensitivity) - Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes.

The question recall answers here is: Of all the customers that truly opted for the product, how many could we label?

**Precision** on the other hand is the ratio of correctly predicted positive observations to the total predicted positive observations.

The question that this metric answer is of all customers identified to be “Taken\_Product” as yes, how many has taken the product? High Precision shows a low false positive rate.

**F1 score** - F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Accuracy works best if false positives and false negatives have similar cost. However, in this data set the cost of false positives and false negatives are very different. Every false positive case will cost the company a lot compare to the expenditure associated with the advertisement. Hence in this case it’s better to look at both Precision and Recall. So the F1 score is also referred here.

Comparing all the different values of Recall, F1 and precision, it has been found Bagging model is returning high values in all the 3 metrics.

# Ensemble modelling:

Here I have used 2 ensemble techniques:

* Bagging
* AdaBoosting

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Device: Laptop** |  |  |  |  |
| **Model Names** | **Overall Accuracy** | **Recall %** | **F1 Score %** | **AUC Score %** |
| AdaBoosting | 84 | 81 | 83 | 91.5 |
| Bagging | 91 | 89 | 91 | 98.1 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Device: Mobile** |  |  |  |  |
| **Model Names** | **Overall Accuracy** | **Recall %** | **F1 Score %** | **AUC Score %** |
| AdaBoosting | 80 | 80 | 80 | 87.8 |
| Bagging | 92 | 95 | 93 | 98 |

When models are created using classification models like Random Forest/Decision Tree, only one tree out of the given data is created and accuracy is based on that single outcome. However, in case of ensemble techniques, here Bagging with decision tress, outcome of multiple such decision trees are learned and an average of the different outcomes considered thereby giving a better model performance and accuracy compare to the rest.

This helps to correctly identify the segment of the customers who are most likely going to accept the product. So, company can decide to advertise the product accordingly to these target customers with higher success rates.

## Model tuning measures:

Apart from Bagging with Decision tree that has resulted to a high accuracy, Recall and F1 score , I have used different other models as following:

* Logistic Regression
* Liner discriminant Analysis
* Decision Tree
* Random Forest

Using best\_grid features of Decision tree, Random Forest the best parameters w.r.t max\_depth of the tree, number of features etc. were identified and considered to get a better outcome from the individual models.

# Interpretation of the most optimum model:

The most accurate model as identified with this data set is Bagging with Decision Tree.

The highest accuracy along with best recall and F1 score considered it to be the best of all models. This model helped to identify the customers who is largely going to accept the products if advertised at their social site. This model could correctly predict with higher percentage of those customers.

Hence company could mainly focus on those customers and advertisement cost can be planned accordingly with a higher chance of return.