

Prediction of Online Shoppers' Purchasing Intention Using Ensemble (Bagging, Boosting, Stacking) Learning Methods

David Akinmade
CS 831
200440384

I. INTRODUCTION

In most cases, the fate of an online business boils down to its ability to improve the turnover rates of the shoppers that browse through their website into shoppers that end up making an eventual purchase. Key to improving this metric is the ability to predict which sessions are more likely to end up in a purchase and that is what this project aims to achieve. A couple of other techniques including traditional machine learning algorithms and neural networks (Sakar et al)¹ have been applied to predict shoppers' intention. This project however explores solving this binary classification problem using the three prevalent Ensemble learning methods namely: model bagging, model boosting and model stacking.

Ensemble Learning also known as Multiple Classifier Systems have been used extensively in the world of data science ever since Professor Leo Breiman of Berkeley popularized the term 'Bagging' in his 1994 paper². The core idea of Ensemble Learning is to aggregate the results of several classifier models into a single model that performs better than any of the individual models that constitute it. Model Bagging achieves this by training several homogeneous weak learners in parallel and then averaging their results, Model Boosting trains several homogeneous weak learners sequentially before combining their results and Model Stacking combines together several heterogeneous models and makes use of a meta-model to select the best predictions of each base model for each instance. This report details the individual steps carried out in this project which was carried out entirely in Python 3 (using Jupyter Notebook) from the Objectives to the Methodology where the data pre-processing and transformation is discussed. After that, how each modelling method and algorithm implemented used is also shown and finally, a comparative analysis is conducted on the results of each of the three Ensemble methods used in order to demonstrate the limits of each approach. In its conclusion, other related work and possible areas of model accuracy improvement are discussed.

II. PROJECT OBJECTIVES

The goal of this project is two fold. First is to conduct binary classification of online shoppers sessions, using the three main Ensemble Learning methods (of bagging, boosting and stacking) into 2 categories. Secondly, is to carry out a comparative analysis of which Ensemble method performs better by producing the best level of Accuracy, F1 score , Precision and Recall on the Online shoppers' dataset. Expected outcome example:

Input:

BounceRates	ExitRates	PageValues	SpecialDay	Month	OperatingSystems	Browser	Region	TrafficType	VisitorType	Weekend	Revenue
0.200000	0.200000	0.0	0.0	Feb	1	1	1	1	Returning_Visitor	False	False
0.000000	0.100000	0.0	0.0	Feb	2	2	1	2	Returning_Visitor	False	False
0.200000	0.200000	0.0	0.0	Feb	4	1	9	3	Returning_Visitor	False	False
0.050000	0.140000	0.0	0.0	Feb	3	2	2	4	Returning_Visitor	False	False

Output:

ID	Revenue
0	0
1	0
2	1
3	0

III. METHODOLOGY

The data used in this project was the Online shoppers' intention data by UCI's Machine Learning Library obtained on Kaggle³. The dataset is comprised of feature vectors that was obtained over the course of 12,330 sessions, where each session belongs to a different shopper in a 1 year period. Of the 12,330 sessions, 10,422 (84.5%) resulted in no purchase while 1,908 (15.5%) resulted in a purchase.

A. Exploratory Data Analysis

TABLE 1: Numerical features used in the Online Shoppers' Intention analysis model⁴

Feature name	Feature description	Min. value	Max. value	SD
Administrative	Number of pages visited by the visitor about account management	0	27	3.32
Administrative duration	Total amount of time (in seconds) spent by the visitor on account management related pages	0	3398	176.70
Informational	Number of pages visited by the visitor about Web site, communication and address information of the shopping site	0	24	1.26
Informational duration	Total amount of time (in seconds) spent by the visitor on informational pages	0	2549	140.64
Product related	Number of pages visited by visitor about product related pages	0	705	44.45
Product related duration	Total amount of time (in seconds) spent by the visitor on product related pages	0	63973	1912.25
Bounce rate	Average bounce rate value of the pages visited by the visitor	0	0.2	0.04
Exit rate	Average exit rate value of the pages visited by the visitor	0	0.2	0.05
Page value	Average page value of the pages visited by the visitor	0	361	18.55
Special day	Closeness of the site visiting time to a special day	0	1.0	0.19

The Table 1 above gives the statistical overview of all the numerical features present in the dataset, while Table 2 below gives the statistical overview of all the categorical features present in the dataset. Since the dataset was gathered by the Machine Learning Library of UCI, some preliminary data cleaning had already been carried out on it such that there were no missing or null values in the data points or columns.

The tables also give a brief explanation as to what summary of what each column feature means with respect to the online shopping sessions.

TABLE 2: Categorical features used in the Online Shoppers' Intention analysis model⁴

Feature name	Feature description	Number of categorical values
OperatingSystems	Operating system of the visitor	8
Browser	Browser of the visitor	13
Region	Geographic region from which the session has been started by the visitor	9
TrafficType	Traffic source by which the visitor has arrived at the Web site (e.g., banner, SMS, direct)	20
VisitorType	Visitor type as "New Visitor," "Returning Visitor," and "Other"	3
Weekend	Boolean value indicating whether the date of the visit is weekend	2
Month	Month value of the visit date	12
Revenue	Class label indicating whether the visit has been finalized with a transaction	2

B. Data Transformation and Standardization

The next step in the data processing is the encoding of all the categorical features listed in Table 2 into numerical features. This is because none of the models we are going to be working with would be able to process data that is not entirely in numerical form. To achieve this, LabelEncoder was imported from Scikit-learn and used to one-hot encode each categorical variable.

TABLE 3: Raw features before label encoding

Related_Duration	BounceRates	ExitRates	PageValues	SpecialDay	Month	OperatingSystems	Browser	Region	TrafficType	VisitorType	Weekend	Revenue
0.000000	0.200000	0.200000	0.0	0.0	Feb	1	1	1	1	Returning_Visitor	False	False
64.000000	0.000000	0.100000	0.0	0.0	Feb	2	2	1	2	Returning_Visitor	False	False
0.000000	0.200000	0.200000	0.0	0.0	Feb	4	1	9	3	Returning_Visitor	False	False
2.666667	0.050000	0.140000	0.0	0.0	Feb	3	2	2	4	Returning_Visitor	False	False
627.500000	0.020000	0.050000	0.0	0.0	Feb	3	3	1	4	Returning_Visitor	True	False

TABLE 4: Features after label encoding

Related_Duration	BounceRates	ExitRates	PageValues	SpecialDay	Month	OperatingSystems	Browser	Region	TrafficType	VisitorType	Weekend	Revenue
0.000000	0.20	0.20	0.0	0	2	0	0	0	0	2	0	0
64.000000	0.00	0.10	0.0	0	2	1	1	0	1	2	0	0
0.000000	0.20	0.20	0.0	0	2	3	0	8	2	2	0	0
2.666667	0.05	0.14	0.0	0	2	2	1	1	3	2	0	0
627.500000	0.02	0.05	0.0	0	2	2	2	0	3	2	1	0

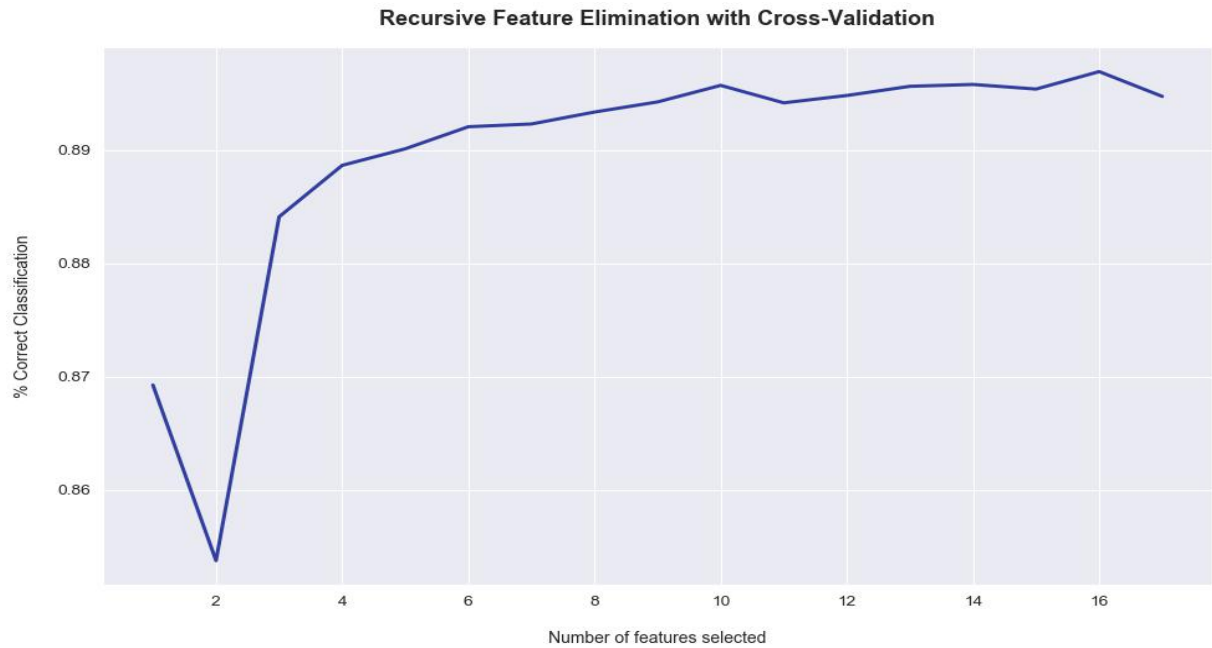
After the encoding, the target feature 'Revenue' was then removed from the dataset and the remaining features were transformed by using StandardScaler from Scikit-learn to scale them such that their mean becomes zero and the standard deviation 1. This is necessary so as to prevent data points from features with high numerical values from unfairly skewing the predictions in any particular way, therefore leading to more accurate predictions by the Ensemble models.

TABLE 5: Features after standardization (excluding Revenue feature)

2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
-0.396478	-0.244931	0.051008	0.078914	-0.351917	-0.611814	0.343874	3.713114	0.352731	0.961270	0.374389	-0.061364	-0.514182	0.407786	-0.550552
-0.396478	-0.244931	-0.623548	-0.542198	0.367291	1.171473	-0.317178	-0.308821	1.196576	-1.233426	3.286094	-0.061364	-0.762629	0.407786	1.816360
-0.396478	-0.244931	-0.623548	-0.602399	-0.457683	0.142551	-0.317178	-0.308821	-1.756881	-0.136078	4.450776	-0.894178	-0.514182	0.407786	-0.550552
-0.396478	-0.244931	-0.331240	0.005655	-0.199879	-0.114679	0.718829	3.713114	0.352731	0.961270	-0.207952	-0.477771	-0.514182	0.407786	-0.550552
-0.396478	-0.244931	0.006038	-0.191838	-0.457683	-0.874301	2.627527	-0.308821	-0.069191	-0.136078	-0.207952	1.604266	-0.514182	0.407786	-0.550552

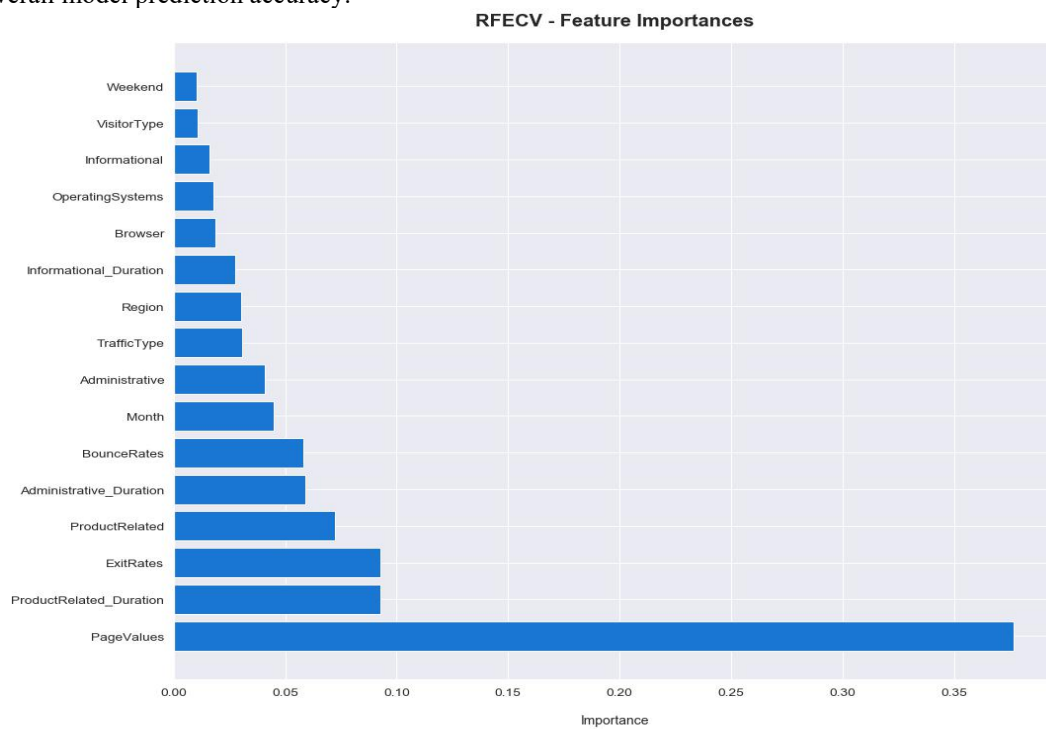
C. Feature Importance and Selection

In order to obtain a clearer view of the contribution of each feature to the overall prediction and in determining which features should be dropped, a Cross Validated Recursive Feature Elimination model (RFECV) from python feature_selection module is fitted to the entire dataset (excluding the target feature 'Revenue') using a Stratified K-fold in order to check which number of features produce the best level of accuracy for modelling.



Graph 1: Potential model performance for each number of features using RFECV

The results showed that 16 features would give the higher number of model prediction accuracy. Below is chart showing the percentage of importance of each of the selected 16 features contributes to overall model prediction accuracy.



Graph 2: Feature Importance in ascending order

IV. MODELLING

A. Data Splitting and Balancing

After the data was successfully pre-processed and transformed, we feed it into our models. But before doing that two key steps were carried out:

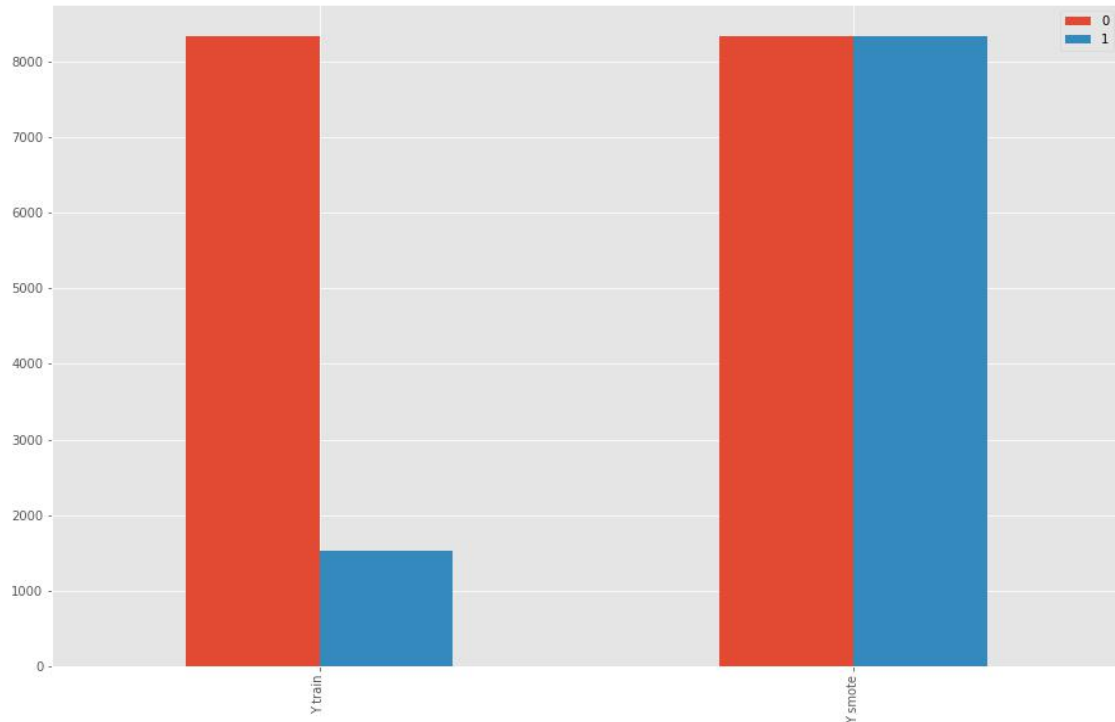
1. Data splitting

The data is split 80% train data and 20% using the Train-Test-Split function. The stratify parameter is also set on the target feature to ensure that the number of number of data points from each

class in the target feature is proportionally split into the train and test sets. 80% of the data was for training used so as to maximize the prediction accuracy of the models especially on the data points that end in a purchase (i.e where Revenue = 1) as the data is imbalanced having more 70% more data points that do not end in a purchase (I.e where Revenue = 0)

2. Data balancing

To finally address the issue of data imbalance, the data points in the training set where Revenue = 1 are over-sampled using the SMOTE function from the Imbalanced-Learn python library to match the number of data points where Revenue = 0. This raises the total number of data points from 9864 to 16676, where the new dataset has 8338 data samples with Revenue class of 1 and also 8338 data samples with Revenue class of 0, therefore balancing the total number of positive and negative target class labels.



Graph 3: Data samples of the Target variable before and after SMOTE balancing

B. Base Model

To set the tone for the entire modelling exercise, a base level of performance has to be set in order to judge if there is a progression in the prediction capability of each successive model. For this project, the base model that was used was a K Nearest Neighbors (KNN) Classifier⁵. This was chosen due to its simplicity as KNN models work by aggregating the predictions of the data points nearest to the query data point. K was set to 8 and fit to the training data. A lot of emphasis would be placed on F1-scores because F1-score is the parameter that shows how well our precision and recall scores balance each other. Below is the performance of the base model on the test data

TABLE 6: Base Model Performance

Target Class Value	Precision	Recall	F1-score	Support
0	0.93	0.80	0.87	2084
1	0.40	0.66	0.50	382
Macro Avg	0.66	0.74	0.68	2466
Accuracy				0.79

From the performance of KNN, we obtained an average F1- score of 0.68 and an overall model accuracy of 0.79 which is okay for a benchmark performance. However, the F1-score for positive data points in the target class is low at 0.50 (which is no better than a random coin toss). therefore we expect future models to perform better at predicting the positive target class by having a much better F1-score.

C. Ensemble Method Models

1. Model Bagging

The first Ensemble method that was applied is Bagging. The core idea of Bagging (Breiman, 1996c) is to combine several homogeneous weak learner models, in this case several decision trees, which are learned independently of each other and averaged together into a single prediction. The specific type of Bagging model used in this project is the Random Forests model which is the most efficient as far as Bagging models go. The model is imported from Scikit-Learn and its parameters are tuned using a cross validation process. GridSearchCV function is used for this tuning, and after the tuning is completed the model is then fit on the training data.

2. Model Boosting

The second Ensemble model tested on the training data was the Boosting Ensemble. But here instead of using the regular adaptive boosting model or any of its variants, the Extreme Gradient Boosting (XGB) algorithm was used. Extreme Gradient Boosting like other Boosting⁶ models (Freund & Schapire, 1996; Schapire, 1990) similarly uses a combination of several weak learners (decision trees), which are learned sequentially such that each tree builds upon the results of the previous one, and then it combines the prediction of each tree by using a number of hyper-parameters to ensure that no individual tree has an inordinate amount of influence on the final prediction. Here also, GridSearchCV is used to tune all the hyper-parameters to find the best combination for our provided dataset. Also very important to note here is that the parameters for maximum tree depth and number of estimators for the XGB model were set to 14 and 600 respectively to further strengthen its performance.

3. Model Stacking

The third and final Ensemble model that was used to make predictions on the data was a stacked generalization. Unlike Bagging and Boosting, Stacking⁷ (Wolpert, 1992) can combine several heterogeneous or dissimilar weak learner models together in order to make a prediction. It also differs from them such that it does not simply average the predictions of the weak models in order to make a final prediction, rather it fits a 'meta-model' on top of them, and the job of this meta-model is to learn when it is appropriate to use each of the meta-models to make predictions on the test data. This thereby results in a combination that uses the strengths of each model and an overall model that is better than any of its individual weak models.

For this project, the stacking model was kept simple by using only two models: The cross validated XGB model from the Bagging test, and a Logistic Regression model that has been tuned using GridSearchCV. Then a simple un-tuned Logistic Regression model was fitted as the meta-model to make the final predictions. The simple Logistic Regression was chosen as a meta-model because it is usually better to have a simpler model rather than a complex one to reduce the complications of choosing a final prediction.

V. RESULTS

A. Performance Metrics

The following are the performance metrics used in evaluating and comparing the performances of each model:

1. Accuracy: total number of correct predictions / total number of predictions
2. Precision: true positive/ (true positive + false positive)
3. Recall: true positive/ (true positive + false negative)
4. Support: This refers to the number of data samples used from each class
5. F1-score: This refers to the ratio of the product of Precision and Recall to their sum multiplied by two. This can be expressed mathematically as:

$$F_1 = \frac{2(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})}$$

The F1-score is very important in this project, especially the F1-score on positive class data samples since it is very important for online companies to know how to predict sessions that end up in Revenue more than they need to predict sessions that don't.

5.2 Results

TABLE 7: Bagging (Random Forest) Model Performance on Test Data

Target Class Value	Precision	Recall	F1-score	Support
0	0.94	0.92	0.93	2084
1	0.62	0.69	0.65	382
Macro Avg	0.78	0.81	0.79	2466
Accuracy				0.89

TABLE 8: Boosting (Extreme Gradient Boosting) Model Performance on Test Data

Target Class Value	Precision	Recall	F1-score	Support
0	0.94	0.94	0.94	2084
1	0.68	0.67	0.67	382
Macro Avg	0.81	0.81	0.81	2466
Accuracy				0.90

TABLE 9: Stacking (XGB + Logistic Regression) Model Performance on Test Data

Target Class Value	Precision	Recall	F1-score	Support
0	0.95	0.92	0.93	2084
1	0.61	0.73	0.67	382
Macro Avg	0.78	0.82	0.80	2466
Accuracy				0.89

TABLE 10: Stacking (XGB + Random Forest) Model Performance on Test Data

Target Class Value	Precision	Recall	F1-score	Support
0	0.94	0.93	0.93	2084
1	0.63	0.68	0.65	382
Macro Avg	0.78	0.80	0.79	2466
Accuracy				0.89

TABLE 11: Overall Model Performance Comparison on Test Data

Model	Avg Precision	Avg Recall	Avg F1-score	Accuracy	Rank
KNN	0.66	0.74	0.68	0.79	Base
Random Forest	0.78	0.81	0.79	0.89	3rd
XGB	0.81	0.81	0.81	0.90	1st
XGB + Logistic Regression stack	0.78	0.82	0.80	0.89	2nd
XGB + Random Forest stack	0.78	0.80	0.79	0.89	4th

The results at the end of modelling show the following:

1. All ensemble models used generally performed much better than the KNN model (which was used as the benchmark for performance) on all performance metrics.
2. The stacking ensemble of XGB and Random Forest comes in at fourth place, performing worse as an ensemble than either of its individual base models when used alone. It had an average precision of 0.78, an average Recall of 0.80 and an overall accuracy of 0.89
3. In third place was the bagging ensemble method, Random Forest. It had a decent level of overall accuracy at 0.89, but ranked lowest of the three ensemble models because it had the lowest F1-score in predicting the positive Revenue label (at 0.65) and also the lowest average F1-score of the three at 0.79
4. In terms of accuracy and F1-score (on positive class where Revenue = 1), the stacking ensemble of XGB and Logistic Regression performed better than Random Forest but less than the boosting ensemble model, putting it comfortably in second place. It had an overall accuracy score of 0.89 and an average F1-score of 0.80, but performed better at predicting positive Revenue classes with an F1-score of 0.67.

5. Finally the Boosting model emerged as the champion, but by a slight margin as it ties with the Stacked model with an F1-score of 0.67 in predicting positive Revenue classes. Where it does shine better than the other models is on its overall accuracy which is 0.90 and its average F1 score which is 0.81.

VI. RELATED WORK

In a study by Suchacka and Chodak⁸, they categorized shoppers online behaviour using web server log data and then used association rule mining on the resulting log data to obtain potential knowledge about different customer profiles. Another study used Support Vector Machines (SVM) to classify online bookstore user data into browser and buyer sessions⁹. K-nearest neighbors classifier, which is one of the models that was used in this project, has also been used on the Online shoppers' dataset to classify the user sessions into those that generate Revenue and those that do not¹⁰. Finally, Sakar et al also work on this same online shoppers' dataset but instead classify each user session by implementing multi-layer and Long Short Term Memory (LSTM) recurrent neural networks

VII. LIMITS and EXTENSIONS

The scope of this work covers only the application of Ensemble learning methods to this binary classification problem of shopping data. As was shown, this method is limited first by the fact that not enough positive data samples for Revenue were available. Also is the fact that even though GridSearchCV was used to tune each model, a lot more time could be dedicated to tuning the hyper-parameters using dedicated python libraries like: Ray-Tune, Optuna, Hyperopt, Mlmachine, Polyaxon, BayesianOptimization, Talos, SHERPA.

Furthermore, more improvements in accuracy could be obtained by applying Neural networks and Multi-layer Perceptrons to model the data like Sakar et al¹¹ did in their paper

VIII. CONCLUSION

In conclusion, we have been able to successfully classify the test data of online shoppers' user intention into the sessions that generate revenue and the ones that do not, with an overall accuracy of 0.90. This was achieved using the ensemble boosting model called Extreme Gradient Boosting (XGB), which emerged the winner out of a collection of four different ensemble learning models. In second place was the stacking ensemble of XGB and Logistic regression which produced an overall accuracy of 0.89 and coming in close at third place was the bagging ensemble model of Random forests which also had an overall accuracy of 0.89, but lower average precision, recall and F1-score values.

Although this makes it to appear as that model boosting was the best ensemble learning method of them all, it is very important to note that an ensemble learning method's performance is highly dependent on the type of data being used. On a different data type like text data or image data, it might get outperformed by other types of ensemble models. Also, model accuracy can be improved if more positive Revenue class data samples can be obtained to balance the data instead of having to create artificial data samples using SMOTE or other oversampling techniques. Additionally, other modelling approaches like Neural networks and Deep learning have generally shown themselves to be superior to ensemble learning methods and might be a better option for real-world use case scenarios.

IX. REFERENCES

- [1] Sakar, C.O., Polat, S.O., Katircioglu, M. et al (2019). Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks. *Neural Computing & Application* 31, 6893–6908.
- [2] Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24 (2), 123–140.
- [3] <https://doi.org/10.1007/s00521-018-3523-0>
- [4] Sakar, C.O., Polat, S.O., Katircioglu, M. et al (2019). Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks. *Neural Computing & Application* 31, 6896.
- [5] D. Hand, H. Mannila, P. Smyth (2001). *Principles of Data Mining*. The MIT Press.
- [6] Freund, Y., & Schapire, R. (1996). Experiments with a new boosting algorithm. In *Proceedings of the Thirteenth International Conference on Machine Learning*, pp. 148–156
Bari, Italy
- [7] Wolpert, D. (1992). Stacked generalization. *Neural Networks*, 5, 241–259.
- [8] Suchacka G, Chodak G (2017) Using association rules to assess purchase probability in online stores. *IseB* 15(3):751–780

- [9] Suchacka G, Skolimowska-Kulig M, Potempa A (2015) Classification of e-customer sessions based on support vector machine. ECMS 15:594–600
- [10] Suchacka G, Skolimowska-Kulig M, Potempa A (2015) A k-nearest neighbors method for classifying user sessions in e-commerce scenario. J Telecommun Inf Technol 3:64

APPENDIX

Code and Experimental Results


```
In [2]: import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import sklearn.metrics as metrics
from numpy import mean
from numpy import std
from pandas import read_csv
from sklearn.model_selection import cross_val_score
from sklearn.metrics import accuracy_score
from sklearn.model_selection import RepeatedStratifiedKFold
from sklearn.neighbors import KNeighborsClassifier
from sklearn.linear_model import LogisticRegression
from xgboost import XGBClassifier
from sklearn.ensemble import StackingClassifier
from sklearn.svm import SVC
from sklearn.model_selection import RandomizedSearchCV
from sklearn.preprocessing import StandardScaler
from sklearn.pipeline import Pipeline
from sklearn.model_selection import train_test_split, GridSearchCV, StratifiedFold
from imblearn.over_sampling import RandomOverSampler, SMOTE
from sklearn.model_selection import GridSearchCV
from sklearn.datasets import make_classification
from sklearn.ensemble import RandomForestClassifier
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
```

Data Pre-processing, EDA and Cleaning

```
In [3]: # importing the online shopping data file
df = pd.read_csv('online_shoppers_intention.csv')
df.head(10)
```

```
Out [3]:
```

	Administrative	Administrative_Duration	Informational	Informational_Duration	ProductRelated	ProductRelated_Duration	BounceRates
0	0	0.0	0	0.0	1	0.000000	0.200000
1	0	0.0	0	0.0	2	64.000000	0.000000
2	0	0.0	0	0.0	1	0.000000	0.200000
3	0	0.0	0	0.0	2	2.666667	0.000000
4	0	0.0	0	0.0	10	627.500000	0.200000
5	0	0.0	0	0.0	19	154.216667	0.015789
6	0	0.0	0	0.0	1	0.000000	0.200000
7	1	0.0	0	0.0	0	0.000000	0.200000
8	0	0.0	0	0.0	2	37.000000	0.000000
9	0	0.0	0	0.0	3	738.000000	0.000000
10	0	0.0	0	0.0	3	395.000000	0.000000
11	0	0.0	0	0.0	16	407.750000	0.018750
12	0	0.0	0	0.0	7	280.500000	0.000000
13	0	0.0	0	0.0	6	68.000000	0.000000
14	0	0.0	0	0.0	2	98.000000	0.000000
15	2	63.0	0	0.0	23	1668.285119	0.008333
16	0	0.0	0	0.0	1	0.000000	0.200000
17	0	0.0	0	0.0	13	334.966667	0.000000
18	0	0.0	0	0.0	2	32.000000	0.000000
19	0	0.0	0	0.0	20	2981.166667	0.000000

```
In [4]: df.shape
Out [4]: (12330, 18)
```

```
In [5]: # defining a function to carry out exploratory data analysis on the null, unique, duplicate and missing values in the data
def explore_data(data, fr_columns):
    print(col, 'value_counts: ', data_fr[col].value_counts())
    print(col, 'total values: ', len(data_fr[col]))
    print(col, 'null values: ', data_fr[col].isna().sum())
    print(col, 'non-null values: ', data_fr[col].notnull().sum())
    print('is unique: ', data_fr[col].nunique()/len(data_fr[col]))
    print(col, 'duplicate values: ', (len(data_fr[col]) - data_fr[col].nunique()))
    print('\n')
```

```
In [7]: info(df)
```

```
Administrative value_counts: 0      5768
1      1154
2      1314
3      915
4      765
5      575
6      792
7      338
8      287
9      225
10     153
11     105
12      86
13     144
14      44
15      38
16      24
17      16
18      12
19      16
24      4
22      4
23      1
20      1
21      2
26      1
27      1
Name: Administrative, dtype: int64
Administrative total values: 12330
Administrative null values: 0
Administrative non-null values: 12330
is unique: 0.0021897810218978104
Administrative duplicate values: 12303

Administrative_Duration value_counts: 0.000000      5903
0.000000      56
5.000000      53
7.000000      45
11.000000      42
294.070513      1
90.875000      1
97.333333      1
53.166667      1
247.083333      1
Name: Administrative_Duration, Length: 3335, dtype: int64
Administrative_Duration total values: 12330
Administrative_Duration null values: 0
Administrative_Duration non-null values: 12330
is unique: 0.2704785070478507
Administrative_Duration duplicate values: 8995

Informational value_counts: 0      9699
1      1041
2      1728
3      380
4      222
5      99
6      15
7      36
8      14
9      7
10      7
12      5
14      2
15      1
13      1
24      1
16      1
Name: Informational, dtype: int64
Informational total values: 12330
Informational null values: 0
Informational non-null values: 12330
is unique: 0.00137875101378751
Informational duplicate values: 12313

Informational_Duration value_counts: 0.0      9925
0.000000      13
6.0      26
10.0      26
7.0      26
291.5      1
43.2      1
338.4      1
86.6      1
145.6      1
Name: Informational_Duration, Length: 1238, dtype: int64
Informational_Duration total values: 12330
Informational_Duration null values: 0
Informational_Duration non-null values: 12330
is unique: 0.2704785070478507
Informational_Duration duplicate values: 11072

ProductRelated value_counts: 1      622
2      465
3      458
4      404
6      396
377      1
385      1
292      1
409      1
339      1
Name: ProductRelated, Length: 311, dtype: int64
ProductRelated total values: 12330
ProductRelated null values: 0
ProductRelated non-null values: 12330
is unique: 0.02522303252230333
ProductRelated duplicate values: 12019

ProductRelated_Duration value_counts: 0.000000      755
17.000000      21
8.000000      17
11.000000      17
0.000000      1
6560.007540      1
821.893333      1
2004.500000      1
266.500000      1
1919.550000      1
Name: ProductRelated_Duration, Length: 9551, dtype: int64
ProductRelated_Duration total values: 12330
ProductRelated_Duration null values: 0
ProductRelated_Duration non-null values: 12330
is unique: 0.7746147607461477
ProductRelated_Duration duplicate values: 2779

BounceRates value_counts: 0.000000      5518
0.200000      700
0.666667      134
0.028571      115
0.050000      113
0.023457      1
0.003901      1
0.005074      1
0.016735      1
0.007356      1
Name: BounceRates, Length: 1872, dtype: int64
BounceRates total values: 12330
BounceRates null values: 0
BounceRates non-null values: 12330
is unique: 0.151388
BounceRates duplicate values: 10458

ExitRates value_counts: 0.200000      710
0.100000      338
0.050000      329
0.033333      291
0.066667      267
0.025325      1
0.020586      1
0.084444      1
0.055882      1
0.010710      1
Name: ExitRates, Length: 4777, dtype: int64
ExitRates total values: 12330
ExitRates null values: 0
ExitRates non-null values: 12330
is unique: 0.38742903487429037
ExitRates duplicate values: 7533

PageValues value_counts: 0.000000      9600
53.988000      6
42.293068      3
40.278152      2
12.558859      2
1.625051      1
20.157102      1
8.181818      1
12.587222      1
30.203577      1
Name: PageValues, Length: 2704, dtype: int64
PageValues total values: 12330
PageValues null values: 0
PageValues non-null values: 12330
is unique: 0.21930251419302516
PageValues duplicate values: 9626

SpecialDay value_counts: 0.0      11079
0.6      351
0.8      325
0.4      243
0.2      178
1.0      154
Name: SpecialDay, dtype: int64
SpecialDay total values: 12330
SpecialDay null values: 0
SpecialDay non-null values: 12330
is unique: 0.000486180048618007
SpecialDay duplicate values: 12324

Month value_counts: May      3364
Nov      2998
Mar      1907
Dec      1727
Oct      549
Sep      448
Aug      433
Jul      432
June      288
Feb      184
Name: Month, dtype: int64
Month total values: 12330
Month null values: 0
Month non-null values: 12330
is unique: 0.000811030081103001
Month duplicate values: 12320

OperatingSystems value_counts: 2      6601
1      2585
3      2555
4      478
8      79
6      19
7      7
5      7
Name: OperatingSystems, dtype: int64
OperatingSystems total values: 12330
OperatingSystems null values: 0
OperatingSystems non-null values: 12330
is unique: 0.0006488240064882401
OperatingSystems duplicate values: 12322

Browser value_counts: 2      7961
1      2463
4      736
5      467
6      174
163      135
3      105
13      61
7      49
12      10
11      6
9      1
Name: Browser, dtype: int64
Browser total values: 12330
Browser null values: 0
Browser non-null values: 12330
is unique: 0.00105433901054339
Browser duplicate values: 12317

Region value_counts: 1      4780
3      2403
4      1182
2      1136
6      805
7      761
9      511
8      434
5      318
Name: Region, dtype: int64
Region total values: 12330
Region null values: 0
Region non-null values: 12330
is unique: 0.00072992700729927
Region duplicate values: 12321

TrafficType value_counts: 2      3913
1      2451
3      2052
4      1069
13      738
10      450
6      444
8      343
5      260
11      42
20      198
9      47
7      40
14      13
18      10
17      3
12      1
17      1
Name: TrafficType, dtype: int64
TrafficType total values: 12330
TrafficType null values: 0
TrafficType non-null values: 12330
is unique: 0.0016220600162206002
TrafficType duplicate values: 12310

VisitorType value_counts: Returning_Visitor      10551
New_Visitor      85
Other      85
Name: VisitorType, dtype: int64
VisitorType total values: 12330
VisitorType null values: 0
VisitorType non-null values: 12330
is unique: 0.0002433900243309004
VisitorType duplicate values: 12327

Weekend value_counts: False      9462
True      2868
Name: Weekend, dtype: int64
Weekend total values: 12330
Weekend null values: 0
Weekend non-null values: 12330
is unique: 0.00016220600162206002
Weekend duplicate values: 12328

Revenue value_counts: False      10422
True      1908
Name: Revenue, dtype: int64
Revenue total values: 12330
Revenue null values: 0
Revenue non-null values: 12330
is unique: 0.00016220600162206002
Revenue duplicate values: 12328
```

```
In [6]: #dropping any row with missing values
df = df.dropna()
df.shape
df.columns
```

```
Out [6]: Index(['Administrative', 'Administrative_Duration', 'Informational',
'Informational_Duration', 'ProductRelated', 'ProductRelated_Duration',
'BounceRates', 'ExitRates', 'PageValues', 'SpecialDay', 'Month',
'OperatingSystems', 'Browser', 'Region', 'TrafficType', 'VisitorType',
'Weekend', 'Revenue',
dtype='object'])
```

```
In [7]: #using label encoder to encode all categorical input feature
from sklearn.preprocessing import LabelEncoder
label_enc = LabelEncoder()
df['Revenue'] = label_enc.fit_transform(df['Revenue'])
df['feature'] = label_enc.fit_transform(df['feature'])
df.head(5)
```

```
Out [7]:
```

	Administrative	Administrative_Duration	Informational	Informational_Duration	ProductRelated	ProductRelated_Duration	BounceRates	ExitRates
0	0	0.0	0.0	0	0.0	1	0.000000	0.20
1	0	0.0	0.0	0	0.0	2	64.000000	0.00
2	0	0.0	0.0	0	0.0	1	0.000000	0.20
3	0	0.0	0.0	0	0.0	2	2.666667	0.05
4	0	0.0	0.0	0	0.0	10	627.500000	0.02

```
In [8]: #make every column a float, but keeping revenue as a bool
X_train, X_test, y_train, y_test = train_test_split(df, y, test_size=0.2, random_state=42, stratify=y)
df.dtypes
df.head(5)
```

```
Out [8]:
```

	Administrative	Administrative_Duration	Informational	Informational_Duration	ProductRelated	ProductRelated_Duration	BounceRates	ExitRates
0	0.0	0.0	0.0	0.0	0.0	1.0	0.000000	0.20
1	0.0	0.0	0.0	0.0	0.0	2.0	64.000000	0.00
2	0.0	0.0	0.0	0.0	0.0	1.0	0.000000	0.20
3	0.0	0.0	0.0	0.0	0.0	2.0	2.666667	0.05
4	0.0	0.0	0.0	0.0	0.0	10.0	627.500000	0.02

Feature Scaling

```
In [113]: #In general it is a good idea to scale the data
scaler = StandardScaler()
scaler.fit(X)
X = scaler.transform(X)
```

```
In [9]: #train test split
X, y = df.drop(['Revenue'], axis=1), df['Revenue']
#In general it is a good idea to scale the data
scaler = StandardScaler()
scaler.fit(X)
X = scaler.transform(X)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42, stratify=y)
X_train.shape
```

```
Out [9]: (9864, 17)
```

Feature Selection

```
In [18]: # plotting X_train into a dataframe for rfecv evaluation
X_df = pd.DataFrame(X_train)
X_df.head()
```

```
Out [18]:
```

	0	1	2	3	4	5	6	7	8	9	10	11	
0	0.1711449	3.688152	-0.396478	-0.244031	0.051008	0.078914	-0.351917	-0.611814	0.343874	3.713114	0.352731	0.961270	0.37
1	-0.069993	0.457191	-0.396478	-0.244031	-0.623548	-0.542198	0.367291	1.171473	-0.317178	-0.308821	1.196576	-1.233426	0.28
2	-0.069993	0.457191	-0.396478	-0.244031	-0.623548	-0.602399	-0.457683	0.142551	-0.317178	-0.308821	-1.756881	-0.130780	0.45
3	-0.094862	1.454877	-0.396478	-0.244031	0.331240	0.005655	-0.199879	-0.114879	0.718829	3.713114	0.352731	0.961270	-4.20
4	-0.069993	0.457191	-0.396478	-0.244031	0.060338	-0.191838	-0.457683	-0.873401	2.627527	-0.308821	-0.069191	-0.136078	-0.20

```
In [160]: from sklearn.model_selection import StratifiedFold
from sklearn.feature_selection import RFECV
rfc = RandomForestClassifier(random_state=101)
rfecv = RFECV(estimator=rfc, step=1, cv=StratifiedKFold(10), scoring='accuracy')
rfecv.fit(X_train, y_train)
```

```
Out [160]: RFECV(cv=StratifiedKFold(n_splits=10, random_state=None, shuffle=False),
estimator=RandomForestClassifier(random_state=101), scoring='accuracy')
```

```
In [161]: # plotting the rfecv chart for feature selection
plt.figure(figsize=(16, 9))
plt.plot(rfc.feature_importances_, 'Month', 'OperatingSystems', 'Browser', 'Region', 'TrafficType', 'VisitorType',
'TrafficType', 'Weekend', 'Revenue'])
df['feature'] = label_enc.fit_transform(df['feature'])
df.head(5)
```

```
Out [161]:
```

	Administrative	Administrative_Duration	Informational	Informational_Duration	ProductRelated	ProductRelated_Duration	BounceRates	ExitRates
0	0.0	0.0	0.0	0.0	0.0	1.0	0.000000	0.20
1	0.0	0.0	0.0	0.0	0.0	2.0	64.000000	0.00
2	0.0	0.0	0.0	0.0	0.0	1.0	0.000000	0.20
3	0.0	0.0	0.0	0.0	0.0	2.0	2.666667	0.05
4	0.0	0.0	0.0	0.0	0.0	10.0	627.500000	0.02

Modeling

Base Model - KNN

```
In [178]: # creating a KNN model
knn = KNeighborsClassifier(n_neighbors = 8)
knn.fit(X_train, y_train)
```

```
Out [178]: KNeighborsClassifier(n_neighbors=8)
```

```
In [179]: # checking the accuracy, precision, recall and f1-score of the KNN model
y_pred_knn = knn.predict(X_test)
print(accuracy_score(y_test, y_pred_knn))
print(metrics.classification_report(y_test, y_pred_knn))
```

```
Out [179]: 0.7923763179237631
precision    recall  f1-score   support
0          0.93          0.82          0.87        2084
1          0.40          0.66          0.50         382
accuracy          0.66          0.74          0.68        2466
weighted avg          0.83          0.79          0.81        2466
```

Model Bagging - Random Forest

```
In [135]: #Random Forest model
rfc = RandomForestClassifier(n_jobs=-1, max_features='sqrt', n_estimators=50, oob_score = True)
param_grid_rf = [
    {'n_estimators': [200, 700],
     'max_features': ['auto', 'sqrt', 'log2']}
]
```

```
Out [135]: GridSearchCV(cv=5,
estimator=RandomForestClassifier(max_features='sqrt',
n_estimators=50, n_jobs=-1,
oob_score=True),
param_grid=[{'max_features': ['auto', 'sqrt', 'log2'],
'n_estimators': [200, 700]}])
```

```
In [136]: # checking the accuracy, precision, recall and f1-score of the Random forest model and
y_pred = rf_stack.predict(X_test)
print(accuracy_score(y_test, y_pred))
print(metrics.classification_report(y_test, y_pred))
```

```
Out [136]: 0.855447680564477
precision    recall  f1-score   support
0          0.94          0.92          0.93        2084
1          0.62          0.69          0.65         382
accuracy          0.89          0.81          0.89        2466
weighted avg          0.78          0.89          0.89        2466
```

Model Boosting - XGB

```
In [137]: # splitting the dataset before applying smote oversampling
X_train, X_test, y_train, y_test = train_test_split(df, y, test_size=0.2, random_state=42, stratify=y)
X_train.shape
```

```
Out [137]: (9864, 16)
```

```
In [101]: #applying smote oversampling to address data imbalance
smote = SMOTE(random_state=1, k_neighbors=6)
X_train_smote, y_smote = smote.fit_resample(X_train, y_train)
print(X_train_smote.shape)
```

```
Out [101]: (16676, 17)
```

```
In [24]: #converting y_train to a dataframe for plotting
y_df = pd.DataFrame(y_train)
print(y_df['Revenue'].value_counts())
```

```
Out [24]: 0      8338
1      1526
Name: Revenue, dtype: int64
```

```
In [23]: #converting y_smote to a dataframe for plotting
y_df = pd.DataFrame(y_smote)
print(y_df['Revenue'].value_counts())
```

```
Out [23]: 1      8338
0      8338
Name: Revenue, dtype: int64
```

```
In [31]: #plotting the dataset before and after smote oversampling
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.datasets import load_iris
y_comb = pd.DataFrame(data=y_dic)
print(y_comb)
```

```
Out [31]: # plotting the xgb values of each feature on the train and test data side by side as a multiple bargrap
h
style.use('ggplot')
y_comb = y_comb.plot(kind='bar', figsize=(15,10))
```

```
Out [31]: <matplotlib.axes._subplots.AxesSubplot at 0x1ae6abd4708>
```


Modelling

Base Model - KNN

```
In [178]: # creating a KNN model
knn = KNeighborsClassifier(n_neighbors = 8)
knn.fit(X_train, y_train)
```

```
Out [178]: KNeighborsClassifier(n_neighbors=8)
```

```
In [179]: # checking the accuracy, precision, recall and f1-score of the KNN model
y_pred_knn = knn.predict(X_test)
print(accuracy_score(y_test, y_pred_knn))
print(metrics.classification_report(y_test, y_pred_knn))
```

```
Out [179]: 0.7923763179237631
precision    recall  f1-score   support
0          0.93          0.82          0.87        2084
1          0.40          0.66          0.50         382
accuracy          0.66          0.74          0.68        2466
weighted avg          0.83          0.79          0.81        2466
```

Model Bagging - Random Forest

```
In [135]: #Random Forest model
rfc = RandomForestClassifier(n_jobs=-1, max_features='sqrt', n_estimators=50, oob_score = True)
param_grid_rf = [
    {'n_estimators': [200, 700],
     'max_features': ['auto', 'sqrt', 'log2']}
]
```

```
Out [135]: GridSearchCV(cv=5,
estimator=RandomForestClassifier(max_features='sqrt',
n_estimators=50, n_jobs=-1,
oob_score=True),
param_grid=[{'max_features': ['auto', 'sqrt', 'log2'],
'n_estimators': [200, 700]}])
```

```
In [136]: # checking the accuracy, precision, recall and f1-score of the Random forest model and
y_pred = rf_stack.predict(X_test)
print(accuracy_score(y_test, y_pred))
print(metrics.classification_report(y_test, y_pred))
```

```
Out [136]: 0.855447680564477
precision    recall  f1-score   support
0          0.94          0.92          0.93        2084
1          0.62          0.69          0.65         382
accuracy          0.89          0.81          0.89        2466
weighted avg          0.78          0.89          0.89        2466
```

Model Boosting - XGB

[illegible]

```

loss = -(transformed_labels * np.log(y_pred)).sum(axis=1)
C:\Users\akimade\AppData\Local\site-packages\xgboost\_metrics\_classification.py:2240: RuntimeWarning:
g: invalid value encountered in multiply
    loss = -(transformed_labels * np.log(y_pred)).sum(axis=1)
[Parallel(n_jobs=1)]: Done 245 out of 245 | elapsed: 12.6min finished

Out[12]: GridSearchCV(estimator=XGBClassifier(base_score=None, booster=None,
      colsample_bylevel=None,
      colsample_bynode=None,
      colsample_bytree=None, gamma=None,
      gpu_id=None, importance_type='gain',
      interaction_constraints=None,
      learning_rate=None, max_delta_step=None,
      max_depth=None, min_child_weight=None,
      missing=None, monotone_constraints=None,
      n_estimators=100, n_jobs=None,
      num_parallel_tree=None, random_state=None,
      reg_alpha=None, reg_lambda=None,
      scale_pos_weight=None, subsample=None,
      tree_method=None, validate_parameters=None,
      verbosity=None),
      param_grid={'max_depth': [2, 4, 6, 8, 10, 12, 14],
                  'n_estimators': [50, 100, 200, 300, 400, 500, 600]},
      scoring='neg_log_loss', verbose=1)

```

```

In [139]: # checking the accuracy, precision, recall and f1-score of the xgb_stack model tuned up to 14 md and 6
00 ne
y_pred_xgb = xgb_stack.predict(X_test)
print (accuracy_score(y_test, y_pred_xgb))
print (metrics.classification_report(y_test, y_pred_xgb))

0.8990267639902676

```

	precision	recall	f1-score	support
0	0.94	0.94	0.94	2084
1	0.68	0.67	0.67	382
accuracy			0.90	2466
macro avg	0.81	0.81	0.81	2466
weighted avg	0.90	0.90	0.90	2466

Model Stacking

```

In [14]: # defining the function to create a stacked model ensemble
def get_stacking(y):
    # define the base models
    level0 = list()
    # level0.append('RandomForest', rf_stack))
    # level0.append(('xgbboost', xgb_stack))
    # level0.append(('lr', lr_stack))
    level0.append(('lr_b', lr))
    # define meta learner model
    level1 = LogisticRegression()
    # define the stacking ensemble
    model = StackingClassifier(estimators=level0, final_estimator=level1, cv=5)
    return model

```