Georges Essoh

Final Project: Predict Customer loyalty

Date: 11/17/2021

Executive summary

The retail industry accounts for almost $ 4 trillion in sales each year. So, it is no surprise that we see companies like Walmart and Amazon, among the biggest retailers in the United States, regularly use predictive analysis to understand and learn as much as possible from their customers. For example, in 2004, during hurricane Frances, using predictive analysis, Walmart was able to predict that beyond regular products such as torches or water bottles whose demand is increasing, sales of Strawberry Pop Tarts grow more than seven times the normal during events like this. They were, therefore, able to fill their shelves in time to meet the demand. Another concern of retail businesses besides increasing sales is customer retention. Indeed, with the increasing competition from competitors, it is crucial to retain the already acquired customers. Predictive analysis can, in this case, identify the programs or actions that have helped retain customers or not. For example, this is the case for Amazon, which created personalized product recommendations based on buying patterns. Amazon intends to take the process further by setting up a program called anticipatory shipping, consisting of "shipping products to customers before they even buy them based on their behaviour on Amazon's platform." The objective of this course project is to use predictive analytics to predict the behaviour of a customer: if the customer will continue to shop at a specific location.

Throughout this project, emphasis has been placed on developing an algorithmic model capable of predicting whether a client will remain loyal and this with the greatest possible accuracy (above 85%). Prior to constructing the model, it was essential to extract the hidden patterns that comprise the data set and finally from all the variables (32) to choose the ones that have the most significant influence on a customer's loyalty. Several models (vector machine support, AdaBoost, Random tree classifier, Bagging and logistic regression) and a combination of several models (ensemble learning) were therefore tested to obtain an accuracy of more than 85% fixed as the threshold for validating a model. The result is a model capable of predicting customer loyalty with an accuracy of 90%.

Technical Report

Introduction

With ever-increasing competition, retail companies must redouble their strategies to retain their customers and attract new customers. The challenge is often to determine which strategies bear the most fruit and invest more money. How to improve a strategy for maximum return? These are the problems companies face every day to maximize their profits. This is where predictive analytics can shed some light on understanding customer's behaviour.

- Which factor most affects loyalty?
- Which location is the most visited?
- What is our best-selling product?
- How much time does a customer spend on average?
- Will the customer continue buying at Starbucks?

Methods

We have identified three main steps necessary before generating our final predictive model, namely Exploration Data Analysis, Feature / variable selection and finally the consolidation of the two previous steps.

• **Step I:** Here we will be conducting an exploratory data analysis (EDA) on our dataset. The dataset contains 32 different variables that could contribute to our predictive model. At this level, gaining insights into the data is necessary. At this level, it is crucial to identify and deal with missing values and outliers. Visualization will be handy insofar as it will allow us to understand how each variable is distributed. The type of distribution predetermines the type of model. Since most models assume that the data distribution is normal if otherwise, the distributions should be normalized. At this level, it would also be necessary to proceed to the encoding of certain categorical variables for the sake of the next step, which is the selection of variables, because most algorithms

take continuous values as input. Renaming the columns for the sake of simplicity is another thing that will be done.

• **Step II:** At this stage, the focus will be on feature/variable selection from the 32 variables within the dataset. Several methods such as Logistic regression, extra trees, random forest classifier, a recursive feature elimination, a chi-squared analysis, and a Lasso regression will be tested. The advantage of some of these methods is that they generate top feature data frames in addition to a scoring system. The overall feature score was then determined that provided the final feature rankings.

• **Step III:** At this stage, the results of Step I and Step II are combined. The final features that will be used in our initial predictive models build. The models will then be run, the summary of results generated, and discussions will follow.

Results

Step1: Exploratory Data Analysis (EDA)

- The summary of the data below was carried out after a deep cleaning and an encoding of the categorical values. There were no missing values in the initial data set but just NA values that could not be imputed and, therefore, removed.

	gender	age	status	income	visitNo	method	timeSpend	location	membershipCard	itemPurchaseCoffee	...	chooseRate
count	113.000000	113.000000	113.000000	113.000000	113.000000	113.000000	113.000000	113.000000	113.000000	113.0	...	113.000000
mean	0.522124	1.185841	1.221239	0.761062	2.557522	1.070796	0.610619	1.274336	0.469027	1.0	...	3.539823
std	0.501735	0.675445	0.932877	1.087874	0.718854	0.979402	0.849723	0.804538	0.501263	0.0	...	1.026744
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.0	...	1.000000
25%	0.000000	1.000000	0.000000	0.000000	2.000000	0.000000	0.000000	1.000000	0.000000	1.0	...	3.000000
50%	1.000000	1.000000	2.000000	0.000000	3.000000	1.000000	0.000000	1.000000	0.000000	1.0	...	4.000000
75%	1.000000	1.000000	2.000000	1.000000	3.000000	2.000000	1.000000	2.000000	1.000000	1.0	...	4.000000
max	1.000000	3.000000	3.000000	4.000000	3.000000	5.000000	4.000000	2.000000	1.000000	1.0	...	5.000000

8 rows × 32 columns

Figure 1. Sample of summary statistics for dataset.

- **Histograms:**

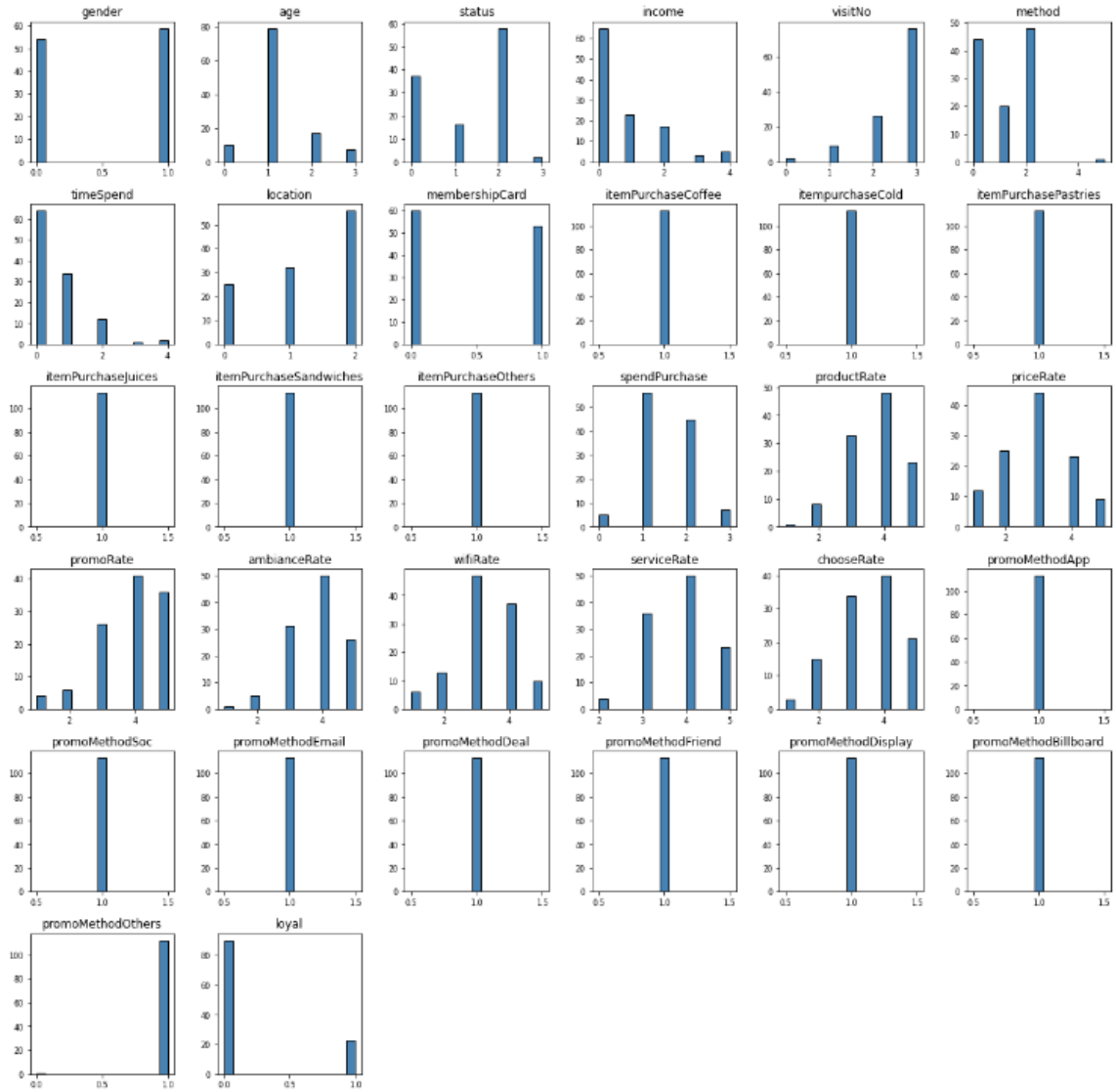


Figure 2. Histogram plots of all dataset variables.

Histograms are essential because they give an idea of the distribution of the variables. The histograms confirm that our variables are categorical and with 0 and 1 of the most common values that our variables take.

With a better understanding of the type and distribution of the variables obtained thanks to the histograms and statistical summary, we can now move on to the feature selection step.

Step II: Feature selection

• **Random Forest classifier:** Each of our variables was subjected to this algorithm and then ranked on the importance to the model. The output is shown below.

	index	RF
14	priceRate	0.193635
15	productRate	0.099997
2	chooseRate	0.090620
26	spendPurchase	0.066371
28	timeSpend	0.058821
1	ambianceRate	0.056515
12	membershipCard	0.050999
24	promoRate	0.049160
27	status	0.047227
30	wifiRate	0.045708
13	method	0.044307
25	serviceRate	0.040742
29	visitNo	0.034238
3	gender	0.031760
4	income	0.029978
11	location	0.027893
0	age	0.024679
22	promoMethodOthers	0.007351
5	itemPurchaseCoffee	0.000000
6	itemPurchaseJuices	0.000000
23	promoMethodSoc	0.000000
10	itempurchaseCold	0.000000
21	promoMethodFriend	0.000000
19	promoMethodDisplay	0.000000
18	promoMethodDeal	0.000000
17	promoMethodBillboard	0.000000
16	promoMethodApp	0.000000
7	itemPurchaseOthers	0.000000
8	itemPurchasePastries	0.000000
9	itemPurchaseSandwiches	0.000000
20	promoMethodEmail	0.000000

- **Recursive Feature Elimination:** This technique begins by building a model on the entire set of predictors and computing an importance score for each predictor. The least important predictor(s) are then removed, the model is re-built, and importance scores are computed again, hence the recursive nature of the process.

	index	RFE
0	age	True
1	ambianceRate	True
2	chooseRate	True
3	gender	True
4	income	True
6	itemPurchaseJuices	True
7	itemPurchaseOthers	True
11	location	True
12	membershipCard	True
13	method	True
14	priceRate	True
15	productRate	True
22	promoMethodOthers	True
24	promoRate	True
25	serviceRate	True
26	spendPurchase	True
27	status	True
28	timeSpend	True
29	visitNo	True
30	wifiRate	True

- **Extra Tree classifier:** The Extra Trees Classifier is very similar to a Random Forest Classifier and only differs from it in the manner of construction of the decision trees in the forest. Each Decision Tree in the Extra Trees Forest is constructed

from the original training sample. Then, at each test node, each tree is provided with a random sample of k features from the feature-set. From this, each decision tree must select the best feature to split the data. This random sample of features leads to the creation of multiple de-correlated decision tree.

	index	Extratrees
14	priceRate	0.131866
15	productRate	0.084581
26	spendPurchase	0.082330
2	chooseRate	0.077282
28	timeSpend	0.065862
12	membershipCard	0.063300
1	ambianceRate	0.058600
24	promoRate	0.056891
25	serviceRate	0.055018
30	wifiRate	0.051164
13	method	0.049961
27	status	0.040661
29	visitNo	0.039031
3	gender	0.035300
11	location	0.035262
4	income	0.031887
0	age	0.026484
22	promoMethodOthers	0.014519
5	itemPurchaseCoffee	0.000000
6	itemPurchaseJuices	0.000000
23	promoMethodSoc	0.000000
10	itempurchaseCold	0.000000
21	promoMethodFriend	0.000000
19	promoMethodDisplay	0.000000
18	promoMethodDeal	0.000000
17	promoMethodBillboard	0.000000
16	promoMethodApp	0.000000
7	itemPurchaseOthers	0.000000
8	itemPurchasePastries	0.000000
9	itemPurchaseSandwiches	0.000000
20	promoMethodEmail	0.000000

- **Chi-square:** The Chi Square Test is used in statistics to test the independence of two events. In feature selection, the two events are occurrence of the feature

and occurrence of the class. The idea here is to test whether the occurrence of a specific feature and the occurrence of a specific class are independent. If two events are independent, the observed count is close to the expected count, thus a small chi square score. A high chi square value is an indication that the hypothesis of independence is incorrect. In other words, the higher value of the chi square score, the more likelihood the feature is correlated with the dependent variable, thus it should be selected for the model.

	index	Chi_Square
14	priceRate	14.16
12	membershipCard	6.05
2	chooseRate	5.79
26	spendPurchase	4.46
15	productRate	3.95
27	status	2.25
29	visitNo	1.80
1	ambianceRate	1.68
28	timeSpend	1.40
11	location	0.94
30	wifiRate	0.73
25	serviceRate	0.67
13	method	0.54
0	age	0.49
4	income	0.45
3	gender	0.10
22	promoMethodOthers	0.03
24	promoRate	0.00
5	itemPurchaseCoffee	nan
6	itemPurchaseJuices	nan
7	itemPurchaseOthers	nan
8	itemPurchasePastries	nan
9	itemPurchaseSandwiches	nan
10	itempurchaseCold	nan
16	promoMethodApp	nan
17	promoMethodBillboard	nan
18	promoMethodDeal	nan
19	promoMethodDisplay	nan
20	promoMethodEmail	nan
21	promoMethodFriend	nan
23	promoMethodSoc	nan

- **Lasso Ridge:** This technique consists in adding a penalty to the different parameters of the machine learning model to reduce the freedom of the model and in other words to avoid overfitting. In linear model regularization, the penalty is applied over the coefficients that multiply each of the predictors. L1 has the property that can shrink some of the coefficients to zero. Therefore, that feature can be removed from the model. Lasso regression identified the three top features influencing the most customer loyalty, namely: The price, the product purchased and the choice of Starbucks for future business.

	index	L1
2	chooseRate	True
14	priceRate	True
15	productRate	True

- **Feature finalization:** Each of the variables is assigned a score of 1 if it appears in one of the five (05) feature selection steps and 0 if not. The scores are therefore added and compiled in the table below.

	index	RF	Extratrees	Chi_Square	RFE	L1	final_score
15	productRate	1	1	1	1	1	5
2	chooseRate	1	1	1	1	1	5
14	priceRate	1	1	1	1	1	5
26	spendPurchase	1	1	1	1	0	4
28	timeSpend	1	1	0	1	0	3
12	membershipCard	0	0	1	1	0	2
13	method	0	0	0	1	0	1
29	visitNo	0	0	0	1	0	1
27	status	0	0	0	1	0	1
25	serviceRate	0	0	0	1	0	1
24	promoRate	0	0	0	1	0	1
22	promoMethodOthers	0	0	0	1	0	1
1	ambianceRate	0	0	0	1	0	1
0	age	0	0	0	1	0	1
11	location	0	0	0	1	0	1
7	itemPurchaseOthers	0	0	0	1	0	1
3	gender	0	0	0	1	0	1
4	income	0	0	0	1	0	1
6	itemPurchaseJuices	0	0	0	1	0	1
30	wifiRate	0	0	0	1	0	1
20	promoMethodEmail	0	0	0	0	0	0
21	promoMethodFriend	0	0	0	0	0	0
10	itempurchaseCold	0	0	0	0	0	0
23	promoMethodSoc	0	0	0	0	0	0
8	itemPurchasePastries	0	0	0	0	0	0
19	promoMethodDisplay	0	0	0	0	0	0
18	promoMethodDeal	0	0	0	0	0	0
5	itemPurchaseCoffee	0	0	0	0	0	0
17	promoMethodBillboard	0	0	0	0	0	0
16	promoMethodApp	0	0	0	0	0	0
9	itemPurchaseSandwiches	0	0	0	0	0	0

The variables with at least a score of 2 or more are retained for the next phase, model selection. Later and depending on the result of the models, more features could be added, or useless features removed. Therefore, these variables are productRate, ChooseRate, priceRate, spendPurchase, membershipCard, and timespend**.**

Step III: Models

In a classification problem, the performance of a model is determined by three primary parameters: Accuracy, Recall and F1 score. Accuracy is how good your model is at guessing the correct labels or ground truths. A recall is a ratio of what the model predicted correctly to what the actual labels are. The higher they are,

the better the model is. F1-score is a more robust metric in the framework because it combines Precision and Recall metrics.

- **Logistic Regression:** Logistic regression is a linear model for binary classification predictive modeling. The parameters of the model can be estimated by maximizing a likelihood function that predicts the mean of a Bernoulli distribution for each example.

Model	Accuracy	Recall	F1
Logit	**85.0%**	**0.5**	**0.5**

Table 1. Logistic Regression: Results

- **Random Forest Classifier**: It's more accurate than the decision tree algorithm. It can produce a reasonable prediction without hyper-parameter tuning. It solves the issue of overfitting in decision trees**.**

Model	Accuracy	Recall	F1
RFE	**87.5%**	**0.5**	**0.55**

Table 2. Random Forest Classifier Results

- **Bagging:**

  Bagging, also known as bootstrap aggregation, is the ensemble learning method that is commonly used to reduce variance within a noisy dataset. In bagging, a random sample of data in a training set is selected with replacement—meaning that the individual data points can be chosen more than once.

Model	Accuracy	Recall	F1
Bagging	**85.0%**	**0.16**	**0.25**

Table 3. Bagging Classifier Results

- **AdaBoost Classifier:** With Adaptive Boosting, the weights are re-assigned to each instance, with higher weights assigned to incorrectly classified instances.

Model	Accuracy	Recall	F1
Adaboost	**85.0%**	**0.16**	**0.25**

Table 4. AdaBoost Classifier Results

- **Support Vector Machine (SVM):** Chosen because of their relative simplicity and flexibility for addressing a range of classification problems, SVMs distinctively afford balanced predictive performance, even in studies where sample sizes may be limited.

Model	Accuracy	Recall	F1
SVM	**90.0%**	**0.67**	**0.67**

Table 5. Support Vector Machine Results

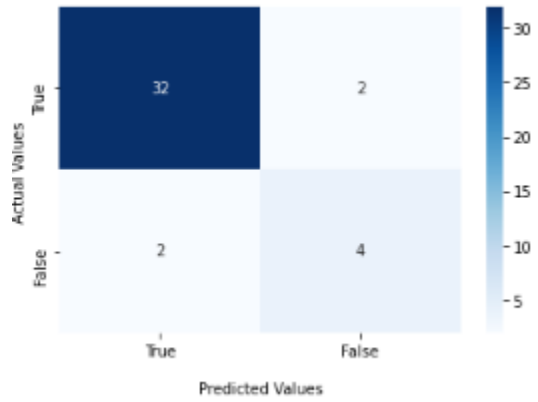
Conclusion

To begin, exploratory data analysis was performed on the variables to help identify which ones might not be good predictors of salary. Plotting the variables as histograms helped to visualize each variables point and confirmed that most of them are categorical. After understanding the type of data that our data set contains, the next step is the choice of variables for modeling. Several feature selection techniques were

used to help with this decision, including random forest classifier, RFE, ExtraTreesClassifier, Chi Square, and Lasso regression. These techniques were helpful in feature selection and several variables were removed from the analysis if their correlation value was relatively small when compared to other variables. At the end of the selection process, from 32 features at the start, we now have six (06) features for modelling.

Given the kind of problem (classification) we have, our initial model was logistic regression. The results were decent, with an accuracy of 85%, a recall of 0.5 and an F1-score of 0.55. Of the five models we considered, the Support Vector Machine (SVM) proved to me the most accurate. We obtained an accuracy value of 90% for our test subset of data which is 5% above the initial threshold needed for model validation Additionally, we built a confusion matrix and a roc curve (receiver operating characteristic curve) to visualize better the model's performance on new data on the test subset.

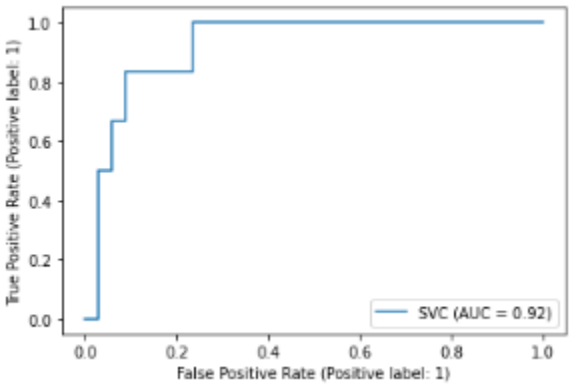
- **Confusion matrix:**



Interpretation

32 customers were classified as loyal that were loyal, 2 customers were classified as not loyal that were loyal, 2 customers were classified as loyal that were not loyal and 4 customers were classified as not loyal that were not loyal.

- **Roc curve:** a graph showing the performance of a classification model at all classification thresholds.



Interpretation

Understanding what AUC is crucial in interpreting a Roc curve. AUC provides an aggregate measure of performance across all possible classification thresholds. One way of interpreting AUC is as the probability that the model ranks a random positive example more highly than a random negative example. AUC ranges in value from 0 to 1. A model whose predictions are 100% wrong has an AUC of 0.0; one whose predictions are 100% correct has an AUC of 1.0. With AUC = 0.92, it means that our model does quite a good job on new values.

References

Pickell, D. (2019, May 14). 8 Examples of Industries Using Predictive Analytics Today. g2. Retrieved September 12, 2021, from https://www.g2.com/articles/predictive-analytics-examples#retail.

Provost, F., & Fawcett, T. (2013). Data Science for Business: What you need to know about data mining and data-analytic thinking. O'Reilly.

Hamzah, M. (2020, May 16). Starbucks Customer Survey. Kaggle. Retrieved September 12, 2021, from https://www.kaggle.com/mahirahmzh/starbucks-customer-retention-malaysia-survey.

Abbott, D. (2014). Applied predictive analytics: Principles and techniques for the professional data analyst. Wiley.

Choueiry, G. (n.d.). George Choueiry. Quantifying Health. Retrieved September 12, 2021, from https://quantifyinghealth.com/variables-to-include-in-regression/.

Krishnan, S. (2019, December 20). Variable Selection using Python - Vote based approach. Retrieved from https://medium.com/@sundarstyles89/variable-selection-using-python-vote-based-approach-faa42da960f0.

[Dubey, A. (2018, December 15). Feature Selection Using Random forest. Retrieved from https://towardsdatascience.com/feature-selection-using-random-forest-26d7b747597f.

Rade, D. (2019, September 2). Feature Selection in Python - Recursive Feature Elimination. Retrieved from https://towardsdatascience.com/feature-selection-in-python-recursive-feature-elimination-19f1c39b8d15.

[Gupta, A. (2019, July 26). ML: Extra Tree Classifier for Feature Selection. Retrieved from https://www.geeksforgeeks.org/ml-extra-tree-classifier-for-feature-selection/.

[Chi Square test for feature selection. (2018, July 5). Retrieved from http://www.learn4master.com/machine-learning/chi-square-test-for-feature-selection.

Dubey, A. (2019, February 4). Feature Selection Using Regularisation. Retrieved from https://towardsdatascience.com/feature-selection-using-regularisation-a3678b71e499.

Kanstrén, T. (2021, May 19). *A look at precision, recall, and F1-score*. Medium. Retrieved October 31, 2021, from https://towardsdatascience.com/a-look-at-precision-recall-and-f1-score-36b5fd0dd3ec#:~:text=F1

GetYourGuide. (2021, May 4). *What is a good F1 score?* Inside GetYourGuide. Retrieved October 31, 2021, from https://inside.getyourguide.com/blog/2020/9/30/what-makes-a-good-f1-score#:~:text=Clearly%2C%20the%20higher%20the%20F1%20score%20the%20better%2C,answer%20depends%20on%20the%20specific%20prediction%20problem%20itself. Pykes, K. (2021, January 24). *Cohen's kappa*. Medium. Retrieved October 31, 2021, from https://towardsdatascience.com/cohens-kappa-9786ceceab58.

Kumar, A. (2021, July 30). *AdaBoost algorithm: Boosting Algorithm in machine learning*. GreatLearning Blog: Free Resources what Matters to shape your Career! Retrieved November 17, 2021, from https://www.mygreatlearning.com/blog/adaboost-algorithm/.

What is bagging? IBM. (n.d.). Retrieved November 17, 2021, from https://www.ibm.com/cloud/learn/bagging.

Google. (n.d.). *Classification: Roc curve and AUC | machine learning crash course*. Google. Retrieved November 17, 2021, from https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc.